PROCEEDINGS OF THE 25TH INTERNATIONAL CONFERENCE ON AUDITORY DISPLAY (ICAD 2019)

Northumbria University
Newcastle upon Tyne

June 23–27, 2019

ICAD 2019
Sonification for Everyday Life

NSF HYUNDAI Northumbria University NEWCASTLE

Imprint

Proceedings of the 25th International Conference on Auditory Display (ICAD 2019)

Editors: Paul Vickers, Matti Gröhn, and Tony Stockman
Publisher: Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom.
Date: June, 2019

ISBN: 0-9670904-6-6
DOI: 10.21785/icad2019.000

Chairs' Welcome

It is with great pleasure that we welcome everyone to ICAD 2019, the 25th International Conference on Auditory Display in the Department of Computer and Information Sciences at Northumbria University in the great northern-English city of Newcastle upon Tyne. The theme of this year's conference is "Digital Living: Sonification for Everyday Life". Digital technology and artificial intelligence are becoming embedded in the objects all around us, from consumer products to the built environment. Everyday life happens where People, Technology, and Place intersect. Our activities and movements are increasingly sensed, digitised and tracked. Of course, the data generated by modern life is a hugely important resource not just for companies who use it for commercial purposes, but it can also be harnessed for the benefit of the individuals it concerns. Sonification research that has hit the news headlines in recent times has often been related to big science done at large publicly funded labs with little impact on the day-to-day lives of people. At ICAD 2019 are exploring how auditory display technologies and techniques may be used to enhance our everyday lives. From giving people access to what's going on inside their own bodies, to the human concerns of living in a modern networked and technological city, the range of opportunities for auditory display is wide. The diversity of international sonification research is highlighted at ICAD 2019, as reflected in the titles of this year's paper sessions:

- Assisting with Everyday Life
- Widening Participation I
- Widening Participation II
- Sonification Theory, Philosophy, and Ethics
- Perception
- Sonification Techniques and Models I
- Sonification Techniques and Models II
- The Hyundai Motors Design Challenge

The sessions on everyday life and widening participation are bolstered by two keynote talks by Alexandra Supper (Maastricht University) and Jude Brereton (University of York), two researchers who have approached sonification from the point of view of understanding how sonification is perceived as a discipline and how its use can be broadened out from laboratory settings to the wider population. The conference theme is further explored in the Hyundai Motors Design Challenge session in which we look at how sonification might impact one of the most prosaic activities of modern life, driving a car. We are very grateful to Dong Chul Park and Taekun Yun from Hyundai Motors Sound Design Research Lab for again agreeing generously to provide sponsorship to ICAD. Our thanks also go to Myounghoon (Philart) Jeon, this year's Sponsorship Chair for nurturing ICAD's cooperative relationship with Hyundai. Of course, none of these sessions would have happened without the heroic efforts of our papers chair this year, Tony Stockman (Queen Mary, University of London) who did a sterling job of managing the submissions, recruiting the reviewers, managing the reviewing process, and author notifications.

As usual, this year's conference begins with the ICAD Student Think Tank. Chaired by Areti Andreopoulou, the Think Tank brings together 14 doctoral and masters students from a diverse range of countries and backgrounds and, supported by a panel of six experienced sonification researchers, helps those students to develop their research ideas. Thanks go to the US National Science Foundation for providing funds to support the running of this year's think tank.

Four workshops are being offered to delegates this year covering a broad range of interests, and our thanks go to Derek Brock (US Naval Research Laboratory) for pulling this vital part of the conference programme together.

As has become custom, this year's ICAD reprises the sonification concert event which is hosted across the road at the neighbouring Newcastle University, ably chaired and curated by Bennett Hogg and Tim Shaw. This year's concert features eight diverse pieces, descriptions of which can be found in this volume. Complementing the concert are two installation pieces which straddle the divide between formal programmed aesthetic experience and oral presentation of research findings.

New to this year's conference is the ICAD 2019 Algorave, chaired by Shelly Knotts from the University of Durham. Algorave musicians use algorithms to produce music and shape it in response to way an audience dances to it. At the ICAD Algorave the performers will use data sonification practices to generate the music and will rewrite the algorithms at performance time.

With the banquet at the amazing Wylam Brewery venue in the Exhibition Park taking up the third evening of the conference, this year's ICAD offers a packed and, we hope, stimulating programme.

No conference happens without the significant efforts of a team of volunteers. Katie Wolf, the Webmaster and Accessibility Chair has been invaluable in her work to ensure a quality and accessible web presence for the conference and also for working with the local organisers to ensure the conference is as accessible as possible. New this year is the Audio Map which Katie developed with Brandon Biggs to further enhance the accessibility of the venue. David Verweij, Vice-Local Chair for Media (Northumbria University) has done a fabulous job working on the graphic design elements of the conference, most notably the conference booklet. Ben Levien and Alan Fuesdale have provided the vital technical support on which the conference depends. Barry Nicholson and his team in Campus Services are warmly thanked for

handling the registration and booking side of the conference.

Finally, the unsung hero of this year's conference is our Local Chair, Selina Sutton. A huge proportion of everything that happened that enabled the conference to take place was a result either of Selina organising it, or reminding me to do it. From scouting potential venues for the banquet, to putting together the goody bag, organising poster boards, booking rooms, buying velcro corners for the posters, organising the army of volunteers, and ordering the catering everyday, etc. etc, Selina has been an invaluable asset. Please make sure to thank Selina when you see her!

Whether you are an authors of a paper or extended abstract, a performer of a sonification concert or Algorave piece, a presenter of an installation, a Think Tank participants, or simply a delegate who has come to absorb it all, we bid you a warm welcome. Without you there is no ICAD and we hope that you will have a canny good time at ICAD 2019 in Newcastle upon Tyne and that you're not too paggered by the end of it!

Paul Vickers and Matti Gröhn

ICAD 2019 General Chairs

The ICAD 2019 Team

Conference Chairs	Paul Vickers NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK Matti Gröhn GLUE, HELSINKI, FINLAND
Papers Chair	Tony Stockman QUEEN MARY, UNIVERSITY OF LONDON, UK
Workshops & Installations Chair	Derek Brock UNITED STATES NAVAL RESEARCH LABORATORY, WASHINGTON, DC, USA
Concert Chairs	Bennett Hogg NEWCASTLE UNIVERSITY, NEWCASTLE UPON TYNE, UK Tim Shaw NEWCASTLE UNIVERSITY, NEWCASTLE UPON TYNE, UK
Think Tank Chair	Areti Andreopoulou NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS, ATHENS, GREECE
Local Chair	Selina Sutton NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK
Vice-Local Chair (Media)	David Verweij NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK
Algorave Chair	Shelly Knotts DURHAM UNIVERSITY, DURHAM, UK
Sponsorship Chair	Myounghoon (Philart) Jeon VIRGINIA TECH, BLACKSBURG, VA, USA
Accessibility Chair & Webmaster	Katie Wolf MEASURINGU, DENVER, CO, USA
Volunteers	Simran Chopra NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK Megan Doherty NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK Justin Kuhn RAYLEIGH, NC, USA Masooma Masooma NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK Kelly Snook CONCORDIA David Verweij NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK Jamie Webster NORTHUMBRIA UNIVERSITY, NEWCASTLE UPON TYNE, UK

Reviewers

And big thanks go especially to all our reviewers!

Areti Andreopoulou, Jack Armitage, Massimo Avantaggiato, Nida Aziz, Mariam Bahameish, Mark Ballora, Amit Barde, Stephen Barrass, Dan Bennett, Braxton Boren, Megen Brittell, Derek Brock, Julius Bucsis, Ivica Bukvic, Elliot Canfield-Dafilou, Eugene Cherny, Teresa Connors, Allan Coop, Antonio D'Amato, Luke Dahl, Feng Feng, Jason Fick, Adrián García Riber, Matti Gröhn, Thomas Hermann, Robert Höldrich, Bennett Hogg, Joseph Hyde, Robert Jack, Myounghoon Jeon, Alex Krasnoskulov, Steven Landry, Sara Lenzi, Doon Macdonald, Justyna Maculewicz, Angela Mcarthur, Oussama Metatla, Thomas Mitchell, Yota Morimoto, Michael Nees, Matthias Rauterberg, Mike Richardson, Davide Rocchesso, Alessio Rossato, Niklas Rönnberg, Disha Sardana, Holger Schultheis, Stefania Serafin, Seth Shafer, Tim Shaw, Kelly Snook, Tony Stockman, Anna Terzaroli, David Verweij, Paul Vickers, Katieanna Wolf, Tim Ziemer.

Sponsors

We extend our warm gratitude to our generous sponsors:

- Hyundai Motors for general financial support, the prizes, and the special session;
- The US National Science Foundation for providing funds to support the think tank under grant #1924796 (PI: Bruce N. Walker, Georgia Tech);
- Northumbria University for hosting the conference and providing the facilities and support staff.

Contents

Keynotes

Papers

Extended Abstracts

Concert Pieces

Installations

Index

Keynotes

Alexandra Supper

The Everyday Life of Sonification
Monday 24th June.

Alexandra Supper is an assistant professor in the Department of Society Studies at Maastricht University. Trained in sociology and science & technology studies, she is interested in studying the dynamics of (disciplinary and interdisciplinary) academic communities. For her PhD research on the sonification of scientific data, she conducted ethnographic research at ICAD conferences, paying particular attention to how the sonification community seeks to establish the scientific legitimacy of listening to scientific data, and in doing so, negotiates the meanings of 'objectivity' and the relationship between science and art. Her work has been published, among others, in the Oxford Handbook of Sound Studies and various peer-reviewed journals, including Social Studies of Science, Science as Culture and Sound Studies.

Jude Brereton

Listener and Performer Perception: Opportunities for Engagement with Auditory Display
Thursday 27th June.

Jude Brereton is a Senior Lecturer (T&S) in Audio and Music Technology in the Department of Electronic Engineering at the University of York. She teaches a number of modules in the areas of acoustics, psyschoacoustics, virtual acoustics and auralization, music performance analysis, voice analysis and synthesis on postgraduate and undergraduate programmes. She designed the department's popular MSc in Audio and Music Technology and was programme leader until 2017. Her research interests include: the performance and perception in virtual acoustic environments; the use of spatial sound to enhance performer and listener experience and interaction; the analysis, perception and evaluation of musical performance; the analysis and synthesis of the human voice. In 2008 Jude was winner of the British Voice Association Van Lawrence Prize for Voice Research.

https://icad2019.icad.org/keynotes/

Papers

CONGRUENT AUDIO-VISUAL ALARMS FOR SUPERVISION TASKS

Eliott Audry

Omnicontact-SafetyData,
2 allée Santos Dumont,
92150 Suresnes
eliott.audry@enac.fr

Jérémie Garcia

ENAC-Université de Toulouse,
7 Avenue Edouard Belin,
31400 Toulouse
jeremie.garcia@enac.fr

ABSTRACT

Operators in surveillance activities face cognitive overload due to the fragmentation of information on several screens, the dynamic nature of the task and the multiple visual or audible alarms. This paper presents our ongoing efforts to design efficient audio-visual alarms for surveillance activities such as traffic management or air traffic control. We motivate the use of congruent cross-modal animations to design alarms and describe audio-visual mappings based on this paradigm. We ran a preference experiments with 24 participants to assess our designs and found that specific polarities between visual and audio parameters were preferred. We conclude with future research directions to validate the efficiency of our alarms with different cognitive load levels.

1. INTRODUCTION

Maritime or aeronautical surveillance systems allow the recovery and fusion of information from ships and aircraft (type, position, speed, etc.) for traffic monitoring purposes via a display device. In both areas, the priority for operators is to guarantee safety through the prevention and resolution of potential conflicts (risk of collision, breakdowns, etc.). In addition, the detection of abnormal behavior and the early identification of associated threats (disaster, illegal or criminal activity, pollution, terrorist act, etc.) are major challenges for all surveillance operators.

To carry out their monitoring tasks, operators rely on complex systems, mainly graphical, to represent all traffic on a map and perform operations such as filtering certain information or selecting an element to obtain detailed information [17]. The systems also include visual or audible notifications and alarms when one or more algorithms integrated into the systems triggers an event [1,17,22].

As with most surveillance activities, a major problem concerns the cognitive overload and underload of operators [15,26]. This cognitive load problem is mainly due to the fragmentation of information on several screens but also to the dynamic nature of the task, visual and auditory distractions as well as interruptions. This overload can lead to blindness or unintentional deafness [4], [20] that prevents the perception of a visual notification or audible alarm when the user is overly solicited by the visual search for an element on the interface, for example. On the other hand, the phenomenon of cognitive underload, when traffic is calm, causes vigilance and attention maintenance problems that also have a negative impact on the quality of surveillance since operators can miss alarms.

Our goal is to rethink the design of audible alarms for surveillance by focusing on redundant modalities: instead of conceiving visual information and audible alarms as separate entities from monitoring systems, our approach consists in integrating several modalities in congruence with the sound to strengthen its perception and more effectively inform the monitoring operator even in cognitively complex situations.

2. BACKGROUND AND MOTIVATION

To support users reacting to dangerous or unpredicted events detected by algorithms, surveillance systems rely on audio or visual alarms. On one hand, visual animations are often used for helping users perceiving changes [24] or to shift their attentions [13], [18]. On the other hand, audible signals transmit important information or alert users through an item requiring immediate attention regardless of where users' current visual focus is.

The work by Gaver et al. highlights the ability of sound to provide useful information on processes and problems [10]. Several guides and experiments have been developed to guide sound interface designers to draw attention to and communicate the urgency of notification [14], [30], facilitate situational awareness of other operators [12] or for use in aircraft systems [22] or rail systems [23]. Texeira et al. [27] propose a gradual design of audible alarms allowing operators to distinguish the criticity level of alarms. The results of the implementation of such alarms suggest that more intelligible information reduces stress and the time spent verifying ambiguous cases or false alarms.

While sound interfaces offer potential benefits for monitoring activities, they are generally considered in isolation of visual components in current systems. Existing design guidelines rarely deal with their explicit combination, which would, among other advantages, improve situational awareness during change [24]. Our perception of the world takes advantage of all our senses and we constantly combine the different ways we understand and interact with our environment. One of the mechanisms we use to merge the inputs of these different channels is frequently defined as cross-modal interaction [25]. One of the main characteristics of a cross-modal interface is the transmission of information through two or more modalities, for example when oral comprehension is facilitated if the speaker's lip movements are visible.

Research on multi-sensory experience often uses the term congruence or cross-modal correspondence to refer to non-arbitrary associations between different modalities and their consequences on the processing of human information. For example, studies have revealed cross-modal associations between high-pitched sounds and bright, small objects at upper spatial locations, and between low-pitched sounds and dark rounded objects at lower locations [21]. This cross-modal congruence was identified as relevant for interface design [8], [25], and exploited in particular by Hoggan et al. [16] who showed that the perceived quality of the buttons on a touch screen was correlated with the congruence between the visual

https://doi.org/10.21785/icad2019.002

and audio/tactile feedback used to represent them. Other studies suggest that bimodal feedback can increase performance and reduce perceived mental workload [28].

To address the challenges faced by operators with notifications and audio or visual alarms, designing cross-modal signals seems like a promising way to improve both the quality and the quantity of information transmitted to the users even with cognitive load issues. In the context of air or maritime fleet control, operators are required to pass multiple medical checks, including vision and auditory tests, to be fit for the position. Thus, we do not consider issues related to color blindness or deafness in this paper.

3. CROSS-MODAL ANIMATIONS DESIGNS

Before designing new systems for surveillance activities, we first wanted to explore congruent audio-visual mappings for simple animations. We define an animation as a temporal evolution of one or more audio and visual parameters of a multimodal stimuli. The temporal evolution is driven by a modulation signal that will be mapped to one or many audio-visual parameters.

The stimulus is made of a circular colored shape and a sound produced with frequency modulation synthesis [5]. This stimulus is meant to be overlaid on any item that raises an alarm in a surveillance system. For instance, such an alarm can be triggered when an aircraft altitude is too low, or when two ships are not respecting the minimal distance between them.

3.1. An ecological approach

We follow an ecological approach to the design of the stimulus, i.e. relationships that exist in the world such as a bigger object produces lower resonances or the closer the louder when an object moves. This approach is inspired by the sonic finder [9], work in designing audio alarms for medical contexts [7] or background monitoring [6].

In our application domain, the congruence of visual and spatial position seems appropriate. Indeed, the spatial position of the items on the map is already used to represent their GPS coordinates so we can match the position of the sound source to the location on screen with sound spatialization techniques.

3.2. Criticity levels

Complex surveillance systems are likely to produce a multitude of alerts, with the possibility of them happening simultaneously. To resolve the conflicts, the operator needs to perceive the level of criticity and assign them the proper amount of cognitive charge to be efficient. Here we designed two criticity levels, low and high.

Sound and visual parameters offer several possibilities for creating appropriate warning scales. For instance fundamental frequencies, harmonic series, envelope shape and modulation speed can influence the perceived urgency of sounds [7], [11]. These results have several implications for our two criticity levels. First, high criticity sound use a higher inharmonicity ratio in the synthesis to produce more inharmonic spectrum. Second, we add distortion to produce higher frequencies that also enhance the perceived emergency [11]. Finally, the animation, i.e. the temporal changes, should be different so that the induced changes are slow and "round" for low criticity and fast and sharp for high criticity. We use two different modulation envelopes : a sine function for low criticity and a sawtooth function with doubled speed for high criticity. Figure 1 illustrates these two modulation settings.

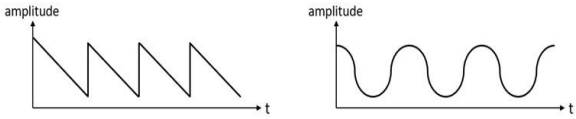


Figure 1: Modulation functions of animation parameters. Left: sawtooth. Right: sine

Regarding the visual parameters, we decided to mimic several existing systems by using the color hue to encode the criticity levels. We use yellow for low criticity and red for high criticity. This choice is intended for the lab experiment setup as a common design guideline but should not be interpreted as fixed rule. We are aware that for an end-user environment experiment, the designers will have to compose with the limitations of the panel of colors available to them.

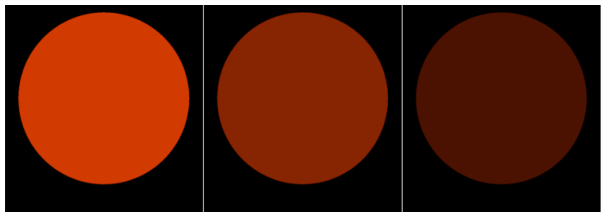


Figure 2: animating the size of the shape

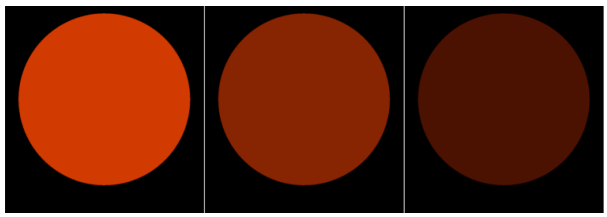


Figure 3: animating the brightness of the shape

The remaining visual, non-positional parameters that seem suitable to be animated are the size of the shape and its brightness as illustrated in Figure 2 and Figure 3. The size of the shape creates a motion that can guide the users' attention [13], [18]. The brightness has the advantage of preserving the shape which can be useful when the shape communicates the type of ships or other relevant information. Regarding audio parameters, we decided to animate the amplitude of the sound source, the pitch, and the dry/wet reverberation ratio and the lowpass filter cutoff frequency.

3.3. Congruent mappings

Based on the available parameters and our ecological approach, we propose four mappings between audio and visual parameters:

- M1 uses size as visual parameter and amplitude as sound parameter. It mimics a moving object going back and forth.
- M2 uses size as visual parameter and pitch as sound parameter. It mimics an object increasing or reducing its size which should respectively produce lower or higher sounds.
- M3 uses brightness as visual parameter and the dry/wet reverb ratio as sound parameter. It creates a diffuse sound stimulus similar to a temporal blur.
- M4 uses brightness as visual parameter and lowpass filter cutoff frequency as sound parameter. It mimics a fog that has a dampening effect on higher pitches [29].

4. PREFERENCE STUDY

Before evaluating the impact of cross-modal congruent alarms on surveillances tasks, we first need to validate our design approach. We conducted a preference study to better characterize subjective preferences on audio-visual mappings.

4.1. Hypothesis

We hypothesize that the ecological mappings should be preferred over non-ecological ones on both the associations between parameters and the polarity, i.e. whether an increase in the sound parameter should indicate an increase or decrease in the visual dimension [29]. For instance, size with pitch should be perceived as a better association than size with reverberation amount. Conversely, brightness with low pass filter cutoff frequency should be perceived as a better association than brightness with amplitude. Regarding polarity, we expect that the polarity suggested in M1, M2, M3 and M4 mappings to be preferred over opposites polarities.

4.2. Method

We ran an online preference test with 24 participants, 15 men and 8 women (M: 36 years; SD: 10,7 years) recruited with various research diffusion lists. One of them indicated being a professional in surveillance systems.
The first part of the online experiment introduces the tasks and indicates guidelines such as being in a quiet environment or wearing headphones before starting the experiment. The second part contains the tasks and the last part gathers information on the participants such as their age or their experience with surveillance systems and sound synthesis. The results were collected and anonymized before performing statistical analyses.

4.3. Task design

For each task, there is an animated visual (brightness or size) and a sound playing. The participant must rate the degree of harmony of the matching between the sound and the video.
We followed a [2×4×4×2] within-subject design with 4 primary factors: VISU ∈ [SIZE, BRIGH], AUDIO ∈ [AMP, PIT, REV, LPF], POLAR ∈ [NO, VR, AR, 2R], CON ∈ [CONT, DISC], as detailed below.
We tested two visual parameters (VISU): the size of the shape (SIZE) and its brightness (BRIGH). We tested four different audio parameters (AUDIO): the amplitude (AMP); the pitch (PIT); the reverberation ratio (REV); and the lowpass filter frequency (LPF).
Polarity (POLAR) is represented by the way one variable vary in association with another. There are two possible polarities: positive where both variables vary in the same direction, negative where variables vary in opposite directions. Based on those, we defined four orders of playing our audiovisual items: the visual variable is played forward and the audio variable is also played forward (NO), the visual is played forward and the sound in reverse (SR), the visual is played in reverse and the sound forward (VR), and both are played in reverse (2R).
We created two different conditions (CON) to challenge the robustness of the participants' preferences. A condition will consist in one of the two modulation curves, i.e. the function controlling the animation. The modulation curve is either a sawtooth function (SAW) or a sine function (SIN) as presented in Figure 1.

These conditions create a set of 64 possible mappings. To avoid fatigue and concentration biases we created two sets of 32 items. The different parameters are fairly divided between the 2 blocks, and each participant will be randomly affected to one of them. Participants were presented all items in a randomized order and had to rate each of them as illustrated in Figure 4.

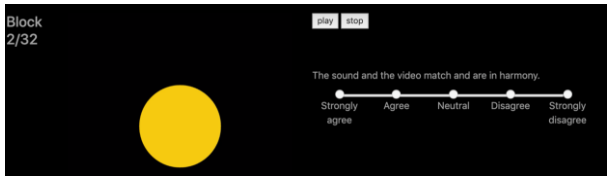


Figure 4: Example of an animated audio-visual item to be rated by the subjects

The rating of the harmonicity of the association between the audio and the visual is done on a Likert scale, from 1 to 5: The lowest rating corresponding to "Strongly disagree", then "Disagree", "Neutral", "Agree", and the highest rating "Strongly agree".

4.4. Results

We proceed with a statistical analysis of the results, first for the global audio-visual mappings, then on more detailed variables with polarities or modulations and compared the results with our assumptions.

We first ran test preferences between each audio-visual possible combination with repeated measures ANOVA. The aggregated results of the possible audio-visual association without considering the modulation nor the polarity resulted in a neutral score for each mapping and none is standing out as statistically significant.

	Amplitude	Pitch	Reverberation	LPF
Size	**m+ = 3,83** (M1) m- = 2,73 p < 0,001	**m+ = 4,00** m- = 2,94 (M2) p < 0,001	m+ = 2,85 **m- = 3,81** p < 0,001	**m+ = 3,94** m- = 2,88 p < 0,001
Brightness	**m+ = 3,81** m- = 2,83 p < 0,001	**m+ = 3,38** m- = 2,46 p < 0,001	m+ = 2,52 **m- = 3,83** (M3) p < 0,001	**m+ = 3,73** (M4) m- = 2,9 p = 0,001

Table 1: Significance of the effect of polarities for each mapping (bold = favorite polarity), and their mean score value.

We then studied preferences between polarities with dependent Student's t-test. The results indicate a significant effect of polarities on the users' preferences. Table 1 shows the mean score of each polarity within each mapping, and the result of the dependent samples t-test between them. Every single test returned a significative result ($p < 0.05$) on the effect of polarity on the preference score.

The distribution of the score only between the preferred polarity of each mapping is presented in Figure 5. No visual variables are significantly preferred in association with Amplitude, Reverberation, and Low-pass filer. However, the pitch is significantly preferred ($p < 0.001$) when associated with size Size (M = 4.0) than with Brightness (M=3.38).

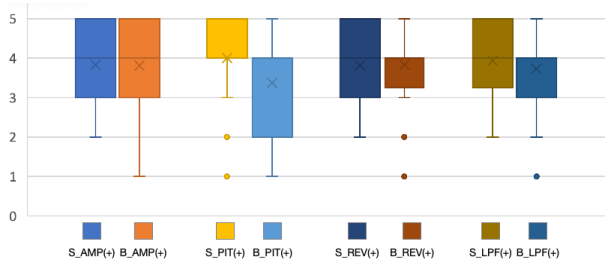


Figure 5: Box plots of score distribution for the preferred polarity of each association

We investigated the effect of the continuous or discontinuous modulation type with dependent Student's t-test. In six out of height cases the mapping preferences were robust to the variation of modulation. However, the preference for polarities did vary depending on the modulation function for two of the mappings, both in the positive polarity setup. The Size and AMP association is preferred with the discontinuous modulation over the continuous one ($M = 3.1$ vs $M = 2.3$, $p = 0.015$). The Size and LPF association is also preferred with the discontinuous modulation over the continuous one ($M = 3.3$ vs $M = 2.5$, $p = 0.015$).

5. DISCUSSION

Our results show that for each possible audio-visual association, there is a preferred polarity. These preferences are consistent with our ecological mappings M1, M3 and M4 but not for M2, that associates a size increase with a pitch decrease. In fact, the opposite preference was observed. We believe that this might be due to the fact that size is a physical invariant in everyday life, thus making it unlikely to change dynamically. Cases involving size change might imply transformations such as stretching the object which might produce a higher pitch.

We assumed that size would be preferred with amplitude (M1) or pitch (M2) and brightness with the reverberation ratio (M3) or lowpass filter frequency (M4). While our ecological approach seems appropriate, the results of the study does not show preferences for specific associations between visual parameters and audio parameters expect for the pitch. Even if the polarity is not the one we hypothesized, users seem to favor an association of pitch with size rather than brightness.

Regarding the other associations, it is possible that the brightness and the amplitude can be related via an intensity metaphor. Similarly, the dry/wet reverberation ratio can also be perceived with an object moving further away in a reverberating room and because there is also a attenuation of the high frequencies with the distance.

Regarding the effect of modulation on polarities, we observed an effect of the discontinuous over the continuous one with the negative polarity for Size and AMP and Size and LPF. This might be due to the fact that synchronization perception is facilitated with a discontinuity.

6. CONCLUSION AND PERSPECTIVES

Our goal is to design efficient audio-visual alarms to support fleet surveillance activities. We motivated the use of cross-modal congruent parameters interactions between sound alarms and visual animations, to improve operators' reaction time, ease of use and localization of alarms. We proposed audiovisual congruence interactions based on an ecological approach and conducted an experiment to assess user preferences of the possible associations.

While the results do not suggest preferences for specific associations, we found that for each possible association, a polarity is significantly preferred. These particular polarities can be used by designers to combine audio and visual stimuli.

To better characterize our design, we also need to validate our criticity level guidelines and to investigate the effect of congruency on attention-related tasks in a surveillance context. We are currently setting up another study to assess the effect of these new interactions on operators' reaction time and error rate against the existing alarm designs.

In our study, we only tested a subset of correlations between visual and sound that seemed the most relevant in our application domain but we are not excluding other cross-modal correlations to be promising and will further investigate these in future work. We are also concerned by the difference between an abstract warning signal designed in a lab, and an alarm signal in a professional environment associated with a strong mental representation [12]. For this reason, future work will focus on conducting field studies with maritime fleet centers and air traffic controllers.

7. ACKNOWLEDGEMENTS

We would like to thank Stephane Conversy and Jean-Luc Marini for their help and support in this project. This project has received funding from ANRT.

8. REFERENCES

[1] Sylvie Athènes, Stéphane Chatty, and Alexandre Bustico. 2000. Human factors in ATC alarms and notifications design: an experimental evaluation. *Proceedings of the USA/Europe Air Traffic Management R&D Seminar*.A. Bee, C. Player, and X. Lastname, "A correct citiation," in *Proc. of the 1st Int. Conf. (IC)*, London, UK, 2001, pp. 1119-1134.

[2] Michel Beaudouin-Lafon and William W. Gaver. 1994. ENO: Synthesizing Structured Sound Spaces. *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, ACM, 49–57.

[3] Tifanie Bouchara, Christian Jacquemin, and Brian F. G. Katz. 2013. Cueing Multimedia Search with Audiovisual Blur. *ACM Trans. Appl. Percept.* 10, 2: 7:1–7:21.

[4] Mickaël Causse, Jean-Paul Imbert, Louise Giraudet, Christophe Jouffrais, and Sébastien Tremblay. 2016. The role of cognitive and perceptual loads in inattentional deafness. *Frontiers in human neuroscience* 10: 344.

[5] John M. Chowning. 1973. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the audio engineering society* 21, 7: 526–534.

[6] Stephane Conversy. 1998. Ad-hoc synthesis of auditory icons. Georgia Institute of Technology.

[7] Judy Edworthy, Sarah Loxley, and Ian Dennis. 1991. Improving Auditory Warning Design: Relationship

between Warning Sound Parameters and Perceived Urgency. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 33, 2: 205–231.

[8] Thomas K. Ferris and Nadine B. Sarter. 2008. Cross-modal links among vision, audition, and touch in complex environments. *Human Factors* 50, 1: 17–26.

[9] William W. Gaver. 1989. The SonicFinder: An interface that uses auditory icons. *Human–Computer Interaction* 4, 1: 67–94.

[10] William W. Gaver, Randall B. Smith, and Tim O'Shea. 1991. Effective sounds in complex systems: The ARKola simulation. *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, ACM, 85–90.

[11] A. Guillaume, C. Drake, M. Rivenez, L. Pellieux, and V. Chastres. 2002. Perception of urgency and alarm design. Georgia Institute of Technology.

[12] Carl Gutwin, Oliver Schneider, Robert Xiao, and Stephen Brewster. 2011. Chalk sounds: the effects of dynamic synthesized audio on workspace awareness in distributed groupware. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, ACM, 85–94.

[13] Johanna Haider, Margit Pohl, and Peter Frohlich. 2013. Defining Visual User Interface Design Recommendations for Highway Traffic Management Centres. *2013 17th International Conference on Information Visualisation*, IEEE, 204–209.

[14] Elizabeth J. Hellier, Judy Edworthy, and I. A. N. Dennis. 1993. Improving auditory warning design: Quantifying and predicting the effects of different warning parameters on perceived urgency. *Human factors* 35, 4: 693–706.

[15] Helen M. Hodgetts, François Vachon, Cindy Chamberland, and Sébastien Tremblay. 2017. See no evil: Cognitive challenges of security surveillance and monitoring. *Journal of applied research in memory and cognition* 6, 3: 230–243.

[16] Eve Hoggan, Topi Kaaresoja, Pauli Laitinen, and Stephen Brewster. 2008. Crossmodal congruence: the look, feel and sound of touchscreen widgets. *Proceedings of the 10th international conference on Multimodal interfaces*, ACM, 157–164.

[17] Anne R. Isaac and Bert Ruitenberg. 2017. *Air traffic control: human performance factors*. Routledge.

[18] Björn B. de Koning, Huib K. Tabbers, Remy M. J. P. Rikers, and Fred Paas. 2009. Towards a Framework for Attention Cueing in Instructional Animations: Guidelines for Research and Design. *Educational Psychology Review* 21, 2: 113–140.

[19] Lester F. Ludwig, Natalio Pincever, and Michael Cohen. 1990. Extending the notion of a window system to audio. *Computer* 23, 8: 66–72.

[20] Arien Mack and Irvin Rock. 1998. *Inattentional blindness*. MIT press Cambridge, MA.

[21] Geoffrey R. Patching and Philip T. Quinlan. 2002. Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance* 28, 4: 755.

[22] Roy D. Patterson. 1982. *Guidelines for auditory warning systems on civil aircraft*. Civil Aviation Authority.

[23] UK Rail Safety. *Standards Board. Alarms and alerts guidance and evaluation tool*.

[24] Céline Schlienger, Stéphane Conversy, Stéphane Chatty, Magali Anquetil, and Christophe Mertz. 2007. Improving Users' Comprehension of Changes with Animation and Sound: An Empirical Assessment. In C. Baranauskas, P. Palanque, J. Abascal, and S.D.J. Barbosa, eds., *Human-Computer Interaction – INTERACT 2007*. Springer Berlin Heidelberg, Berlin, Heidelberg, 207–220.

[25] Charles Spence and Jon Driver. 1997. Cross-modal links in attention between audition, vision, and touch: Implications for interface design. *International Journal of Cognitive Ergonomics*.

[26] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Elsevier, 37–76.

[27] Bruno Teixeira De Sousa, Alessandro Donati, Elif Özcan, et al. 2016. Designing and deploying meaningful audio alarms for control systems. *14th International Conference on Space Operations*, 2616.

[28] Holly S. Vitense, Julie A. Jacko, and V. Kathlene Emery. 2003. Multimodal feedback: an assessment of performance and mental workload. *Ergonomics* 46, 1–3: 68–87.

[29] Bruce N. Walker and Gregory Kramer. 2004. Ecological Psychoacoustics and Auditory Displays: Hearing, Grouping, and Meaning Making. *Ecological psychoacoustics*: 150–175.

[30] Marcus O. Watson and Penelope M. Sanderson. 2007. Designing for attention with sound: challenges and extensions to ecological interface design. *Human Factors* 49, 2: 331–346.

AN INVESTIGATION INTO CUSTOMISABLE AUTOMATICALLY GENERATED AUDITORY ROUTE OVERVIEWS FOR PRE-NAVIGATION

Nida Aziz, Dr. Tony Stockman

Queen Mary University of London
London, UK
n.aziz@qmul.ac.uk

Dr Rebecca Stewart

Imperial College London
London, UK

ABSTRACT

While travelling to new places, maps are often used to determine the specifics of the route to follow. This helps prepare for the journey by forming a cognitive model of the route in our minds. However, the process is predominantly visual and thus inaccessible to people who are either blind or visually impaired (BVI) or doing an activity where their eyes are otherwise engaged. This work explores effective methods of generating route overviews, which can create a similar cognitive model as visual routes, using audio. The overviews thus generated can help users plan their journey according to their preferences and prepare for it in advance. This paper explores usefulness and usability of auditory routes overviews for the BVI and draws design implications for such a system following a 2-stage study with audio and sound designers and users. The findings underline that auditory route overviews are an important tool that can assist BVI users to make more informed travel choices. A properly designed auditory display might contain an integration of different sonification methods and interaction and customisation capabilities. Findings also show that such a system would benefit from the application of a participatory design approach.

1. INTRODUCTION

Maps have always been central in supporting travellers to plan their journeys. As technology is becoming ubiquitous in the world, mapping has evolved and many different ways of using and interacting with maps have been developed. They can be found in mobile phones, sat-nav systems and on computers in general. With easy access to these tools, travelling to unknown or far-off places has become easier for the general population. Often, sighted people plan their journeys before undertaking travel to reduce the overhead of wrongly taken turns leading to wasted time and effort, as well as to gain confidence in where they are going. However, since maps are typically a visual representation of the geographical world, they are only readily available to sighted people. The blind and visually impaired (BVI) community, which, according to WHO [1], is almost 3% of the total population in the UK, does not have easy access to these tools. Even though many applications have been designed to aid navigation, none of them provides a means to plan the journey ahead of time. This puts the BVI community at a distinct disadvantage, as they can not familiarise themselves with the journey ahead. Having means to peruse a map and query it to customise a journey according to their own specific requirements and comfort could potentially improve this experience for them.

This project seeks to design a system to provide route information, such as shape of the route and the passing landscapes, as an audio based overview. The intent is to provide the user with the ability to "experience" the route by creating an image or a mental map through the audio. Furthermore, it would provide the user with the chance to utilise functional route information, such as distance and duration, to make informed decisions regarding the route. As a result, the users will be able to familiarise themselves with the journey or decide whether they want to take that route or not. Other information that users can deduce from this system may be how complicated or busy a route is, whether it is safe or not, is an alternate route better suited, etc. These questions distinguish this work from other works in the field as they allow the user to make informed choices according to their own navigation preferences before undertaking the journey.

For this project, audio is the preferred medium of output because of its ubiquitousness, little to no cost and almost universal availability. In addition, according to [1], many BVI travellers highly value environmental acoustic information and rely on it for navigation purposes. This means that having an auditory display (AD) or soundscape of the route beforehand can improve the experience for them.

The design of studies is based on participatory design concepts including needs assessment and requirement gathering through surveys and informal discussions as described in [2, 3]. The relevance and importance of designing an interactive system with user involvement has been discussed by Spinuzzi in [3]. They provide guidelines for conducting participatory design based research, stating it as an iterative method for developing a system which simultaneously constitutes the expertise and knowledge of both the designer(s) and the users of the design [2, 3].

Some of the research questions this paper seeks to examine are:

- Are apriori auditory overview of routes valued by the BVI community?
- What are the best methods to obtain requirements?
- What types of information are important to represent?
- Which sonification techniques should be included? What should be the balance between speech and non-speech sounds? What should be the balance between functionality

[1] World Health Organisation

https://doi.org/10.21785/icad2019.029

and aesthetics?

- How should different route elements be represented?
- How can such a system best be evaluated?

We do not expect to fully answer to all of these questions within this study, but we strive to provide evidence-based findings towards determining a subset.

1.1. Auditory route overviews

An auditory route overview is an audio based representation of the summary of a route, where hearing is the primary interface channel for communicating information, as opposed to visual. ADs often use a technique called sonification for converting data and interactions into sound, as for example in [4, 5, 6]. Sonification renders sound from data from different sources and produces a time-ordered sequential audio data stream [7]. Auditory displays generally employ three distinctive techniques to convey information such as identities, status, notifications, and events: auditory icons, earcons and speech: either synthetic or pre-recorded. Auditory icons, which were developed in 1989 by Gaver [8], can be described as "caricatures of naturally occurring sounds such as bumps, scrapes, or even files hitting mailboxes". Earcons were developed by Blatner, who defines them as musical motives that can grouped as a family to represent similar meaning sounds [4].

An auditory route overview provides route information, such as direction, distances, landmarks, etc. as audio, utilising the various techniques of sonification. The purpose is to provide the user with information that would help them learn more about the route before embarking on the journey. Some of the scenarios it could help in could be

- Deciding whether to make a journey or not;
- Length and complexity of the journey;
- Choosing between different possible routes, etc.

This paper is organised as follows: first, we review the related work and the motivation for developing auditory route overviews. This includes analysis of a survey to show that there is in fact a demand of such a system by the BVI community. Then we describe our first study in which we conducted a workshop with expert audio designers to gather design ideas, followed by the design implications drawn from the findings. This is followed by a feedback session on the aforementioned designs by BVI users. Here we explore user requirements as they listen to the AD designed by the experts and answer some related questions. This is followed by a discussion of the gaps between the design implications and user requirements and how this gap can be filled. Last we conclude with a brief conclusion and our future plans.

2. THE NEED FOR AUDITORY ROUTE OVERVIEWS

2.1. Related work

The past two decades have seen a significant increase in accessible technology research, including using audio for providing navigational maps. This information, which is mostly geographical, is transformed to be represented as an AD. Depending on the techniques used to transform and represent this information, the user can then learn to interpret this audio input to extract meaningful geographical information. Hennig et al. [9] provide design guidelines and recommendations for developing accessible maps including verbal description of the map content, amount of information suitable, order in which this information should be presented and the best ways of representing functional information like distance and direction. Auditory mapping and navigation is a useful application area for both BVI and sighted users. It promotes and facilitates independent mobility for the BVI users while providing eye-free interaction to sighted users. Auditory maps can either be used passively to explore unknown areas or actively, as a navigation aid to guide en route. There are a number of mechanical and electronic navigation aids such as [10, 11, 5], to name a few.

Pielot et al.[12] designed auditory maps using a movable marker on a table top, encouraging users to explore a virtual space passively. The system determined orientation through rotation of the tangible device, showing improved usability of an auditory map. Similarly, Heuten et al. [13] used a torch metaphor whereby users could only hear objects within a particular radius. Both of these encouraged an exploration of surrounding environment. Auditory route overviews, on the other hand, provide an overview of the route from one location to another, rather than an exploration of an entire area. In his seminal paper, Shneiderman [14] advocates the use of overviews as the first step in any information seeking task: overview first, zoom and filter, then details-on-demand. He argues that the overview allows users to conceptualise the information being presented. Auditory overviews are not currently a major part of auditory interfaces except for a few niche areas like information seeking, graphs, web browsing, etc [15]. Zhao et al. [16] introduced information seeking tasks in auditory overviews by presenting geo-referenced data as audio. They presented various types of information, such as total population, population by age range, etc in menus, sub-menus and multiple pages. The auditory overview swept through the locations horizontally and a spatialised tone played the value range of each state. This could be further extended for subsections of the map to create sub-overviews. They found that the auditory overview successfully displayed patterns in the data and encouraged exploration of areas of interest.

Researchers have used different sonification schemes for representing their data in audio. A. Brown et al. [17] represented line/node graphs as audio and designed guidelines for their mapping as well. These guidelines included mapping y-values to pitch, choosing appropriate MIDI note ranges, exploring different methods to represent multi line graphs, etc. L. Brown et al. also showed that a quick representation, like an overview, can provide the gist of the graph without the need for lengthy explorations. Other researchers suggest using musical timbres as well as panning and stereo output to make interpretation easier [18]. Kildal [19, 20] provided an auditory overview of complex numerical data tables by mapping data values to pitches and then playing them quasi-concurrently to give sense of the magnitude of the data. They also designed an interactive and customisable interface to encourage users to explore the data row or column-wise to reveal patterns and trends.

Guerreiro et al. [5] created auditory overviews of routes to provide navigation information, such as turn instructions and distances, as an overview from one point of interest (POI). They evaluated the merit of apriori mental maps through experiments using route reconstruction as well as real world exposure and showed that users were able to recreate both the sequential structure of the route as well as the approximate locations of the POIs, thereby substantiating the importance of using overviews of route for creating an apriori mental map. However, their application was navi-

gation and not information-seeking.

According to our best knowledge, no significant work has been done on developing playback auditory overviews of routes for information-seeking, planning or preparation of a journey.

2.2. Motivation

To gauge the demand and acceptance of auditory overviews, a online survey was conducted with BVI individuals having at least some experience with assistive technology. The outcomes of the survey helped give a better understanding of which applications could benefit from having overviews as required by people with visual impairments. 15 BVI individuals participated in the survey (5 female and 10 male; aged between 17 and 72) and were recruited from target groups on social media.

As an example, the participants were presented with two non-speech based auditory weather forecast overviews to give an idea of what auditory overviews can sound like. They were then posed 5 questions relating to the importance of overviews, target applications, length of the overviews and type of information presented. The survey showed that BVI individuals favoured applications that would facilitate travelling, as shown in Fig. 1

The first question in the survey established usefulness and importance of auditory overviews, with 10 out of 14 respondents considering them useful. None of the participants thought they were useless.

The subsequent question was regarding the applications for auditory overviews. The participants were asked to rate 6 potential applications that could benefit from having them. The rating was on a scale of 1 to 5, with 5 being most useful. The graph in Fig. 1 shows that overview of route before travel was considered the most useful. Overview of a document, web-page and website were also considered quite useful. This response corroborated the belief that there was in fact a real demand for route overviews to facilitate travel for BVI individuals. The next question in the survey was re-

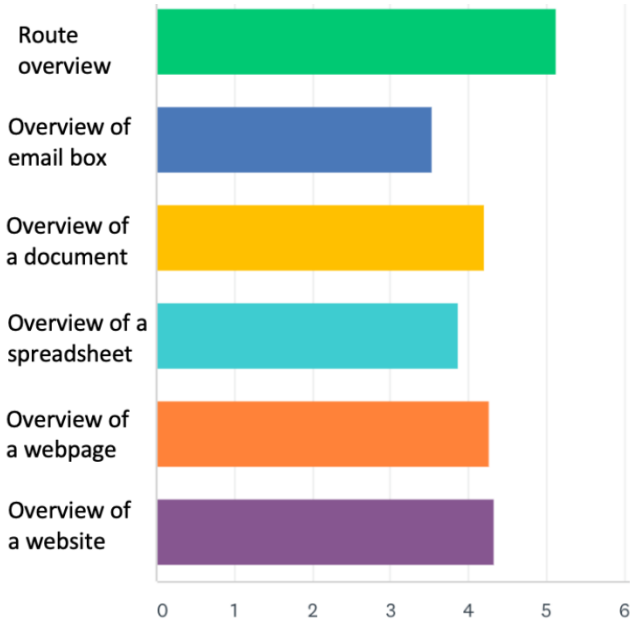


Figure 1: Graph showing responses for preference of different applications of auditory overviews

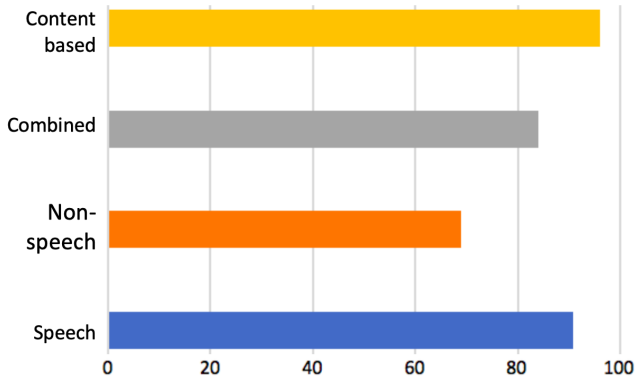


Figure 2: Graph showing preference between having speech, non-speech, a combination of speech and non-speech, and deciding based on application/ scenario in the auditory display

garding the length of the AD. According to literature [16], an AD that is too long will have a larger cognitive load. In their paper, Zhao et al. suggested that an AD of up to 10s would be optimal for their application. However, it was felt that for route-based applications this duration might be too short to provide any substantial information. Thus, this question was posed to the users to get a preliminary feel of their preference. Seven out of 15 participants disagreed with Zhao's duration, while 4 agreed. The remaining were neutral. One of the participants P1 commented that,

> P1: *"They should be as long as the information they need to convey. If a user finds them too long or gets used to them the user can always silence the description as we do in many circumstances. You could also consider the ability to change the verbosity level, i.e., perhaps an expert mode for those who are used to the information and only want to hear certain details."*

Similarly, participant P15 stated:

> P15: *"I believe they should be customised for the subject. One might only require a few seconds while an other could require a lot more."*

Thus, P1 and P15, in addition to giving opinions about duration, also established a need for customisation in the output AD.

The final question was regarding the constituent content of the auditory display. The participants were asked whether they preferred the overviews to be comprised of synthetic speech sounds only, non-speech sounds only or a combination of both. The participants preferred having content chosen on the basis of application and context, as shown in Fig. 2, however showed partiality towards speech-based information too.

In summary, the BVI participants preferred having route overviews with customisable duration and a speech-based audio content with or without a non-speech matter. This survey helped establish a basic framework of requirements. Henceforth, further studies were done to gather more insight into the design and usage of route based auditory overviews.

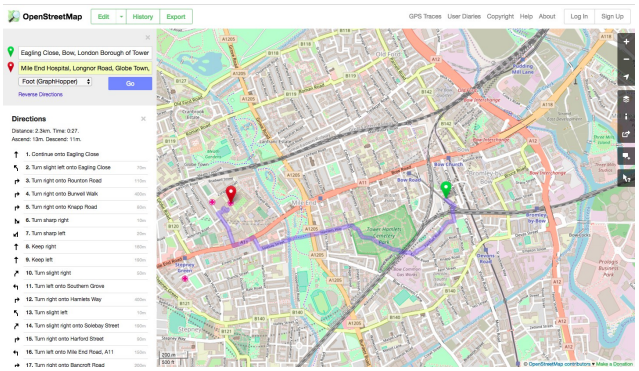
3. DESIGN APPROACH

After analysing the results of the survey and gathering preliminary design requirements, a two-part study was arranged to gather de-

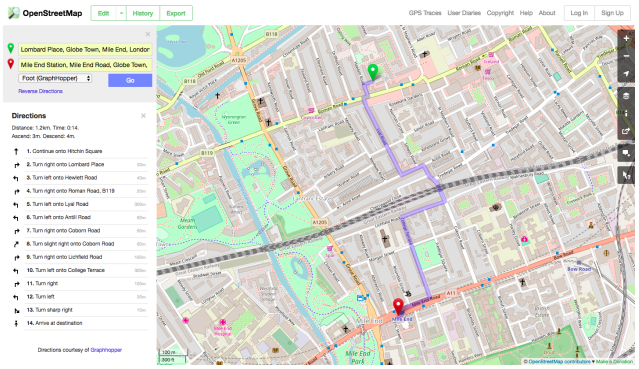
sign perspectives. In the first part, Study 1, a workshop was held with audio and/or design experts. They were given sample routes (Fig. 3) and asked to design auditory route displays manually. Some of the questions that the study was looking to address are given below:

1. What are some of the considerations that the designers have while designing auditory route overviews?
2. Do the designers have the same considerations regarding duration and content as the BVI users while designing auditory route overviews?
3. What are some of the techniques that the designers employ while designing auditory route overviews?

Eight participants took part in the workshop. They were MSc or PhD students of either audio engineering or sound design and were chosen for their domain knowledge. Participants were encouraged to ask questions and think-aloud while designing if they felt comfortable doing so. It was fully explained that they were part of a targeted user-group and that, as a result, their opinion was important to the iterative design process.



(a) Route 1



(b) Route 2

Figure 3: Routes for study 1

3.1. Task

The study took place over multiple sessions, with a pair of participants in each session. The participants were each provided with a route on Open Street Maps[2]. They were requested to examine their route and design an auditory display for it, individually. Afterwards, they were asked to discuss their design strategies as well as evaluate their partner's design. The evaluation process included listening to the AD and drawing the route on paper. They were also asked to rate the intuitiveness and aesthetic of the AD on a scale of 1 to 5.

[2] http://www.openstreetmaps.org

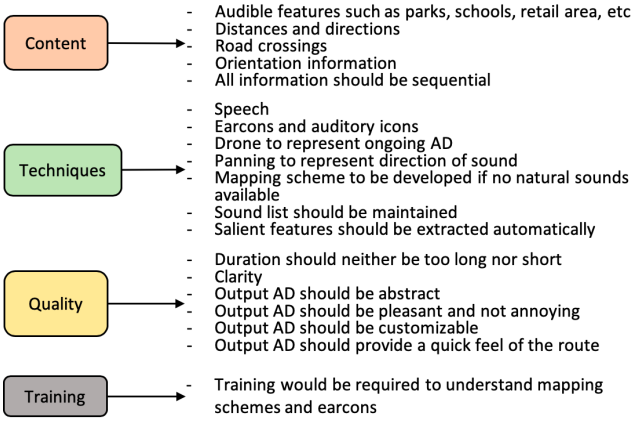


Figure 4: Themes that emerged from applying a thematic analysis to study 1

3.2. Analysis

The data gathered from this study was subjected to thematic analysis, as explained by Braun & Clarke[21]. Ideas which appeared to have some potential but did not recur were also considered. Data was investigated semantically based on a theoretical approach to determine design requirements specific to auditory overview of routes only. The initial themes were identified by analysing the data in a number of ways, including frequently occurring responses or ideas, acknowledging stress words, colour coding similar topics and comparing responses between participants. If an answer mentioned several topics, all of them were considered individually. Themes were extracted from this analysis and then analysed from a broader perspective to draw design implications from the designer's perspective. The themes that emerged from this study are shown in Fig. 4.

These design implications drawn from the designers perspectives give a good sense of how an auditory display for route overviews should be designed, what it should encompass and what qualities it should possess.

4. USER FEEDBACK AND REQUIREMENTS

In the next phase, Study 2, some of the designs developed by the designers of the first study were presented to the BVI users to obtain their feedback. For this purpose, a call for participation was made on groups for BVI on Facebook. Some of the groups which garnered responses are Blind & visually impaired travels –emotioneyes, Blind advocates Texas, Blind and vision impaired community, Blind chat, tips, views and information sharing community, Blind help project, Blind penpals, Blind veterans UK and Wales, Blindwebbers, British blind community, Hadley Institute for the Blind and Visually Impaired support group, iPhone and iPad apps for the blind and visually impaired, RNIB Con-

Table 1: Survey questions to obtain feedback on auditory route overviews

1. Kindly listen to Design 1 and in a few words explain what you understood.
2. The beeping sounds in Design 1 are actually representative of the turn numbers (re-affirmed by the speech sound), for example the third turn on left is shown by three beeping sounds in the left ear
   a. Does having this information change your perspective about this design at all?
   b. Do these two overlain sounds increase or decrease the information provided?
   c. What are the good or useful features in this design?
   d. What are the bad or useless features in this design?
3. Design 1 provides direction information while design 2 provides landscape information. Which one do you prefer and why? If you have no preference, please feel free to make any other comment you wish.
4. Design 2 provides only landscape information, while Design 3 provides landscape information overlaid with direction information. Which one do you prefer and why? If you have no preference, please feel free to make any other comment you wish.
5. Is there any scenario (or application) where you would prefer Design 1 over the other two designs?
6. Is there any scenario (or application) where you would prefer Design 2 over the other two designs?
7. Is there any scenario (or application) where you would prefer Design 3 over the other two designs?
8. Kindly rate the difficulty level of the designs, in terms of understanding and retaining the information provided. The scale is 0 to 10, where, 0 is very easy and 10 is very hard.
9. Kindly rate the length of the designs, in terms of understanding and retaining the information provided. The scale is 0 to 10, where, 0 is too short and 10 is too long.
10. If you are not satisfied with the length of the display, how long do you think it should have been?
11. Do you have any other suggestions for improving this design?

nect–London, RP fighting blindness London, SiteAppsClub, Solent Active Visually Impaired Group and Visually impaired/ blind adults.

Forty three people responded showing interest, however only 8 individuals (4 M, 4 F) from these actually participated in the study. They were situated in different parts of the world, including UK, USA, Canada and Indonesia, and participated virtually via email. Even though the data set of users seems smaller than average, this is common practice when working with a niche population, as seen in many studies [11, 22, 23, 24].

4.1. Task

The study took place on-line. The participants were provided with 3 different auditory displays designed by the participants in Study 1. These files were named Design 1, Design 2 and Design 3. They were also provided with the survey form shown in Table. 1 to evaluate the different aspects in the design files.

4.2. Analysis

The participant responses were collated and analysed to uncover recurrent ideas and underlying themes to guide the design process. Additionally, any ideas that were potentially useful but did not recur were also considered. Similar to the first study, the data was generally investigated semantically. However, in certain instances, implied mentions in participants' answers were also considered to ensure that no important pattern in the data was missed. A theoretical approach to analysis was adopted to determine design requirements specific auditory overview of routes from the user's perspective. Themes were extracted from this analysis and analysed from a broader perspective to draw the design implications, as shown in Fig. 5.

The figure shows different ideas and User feedback from the study. It was mostly regarding the content of the auditory displays presented to them. They considered speech to be of utmost importance in providing functional route information, and wanted it to be clear and coherent. They thought that the auditory icons used to present the information of audible features and points of interest in the landscape were useful to give a feel of the area, as well as aid orientation and localisation. However, by themselves, without any other AD component like speech, they were extremely hard to understand and retain. Thus, auditory icons for landscape features must be accompanied by some other functional information, such as speech for directions. Moreover, they felt that if any transitory audible feature, such as traffic, was not present on the actual route while navigating, it could potentially confuse the user, as they would be expecting to hear those sounds. This showed that the users did not fully grasp the idea that this tool was an informatory tool which would augment their navigation aids during journey.

An integrated AD consists of all or a combination of speech, auditory icons and earcons, presenting different kinds of information via different modalities. All of the designs presented to the BVI users consisted of integrated displays. The users believed that having an integrated display increased their confidence to travel to new places. At the same time, the overlaying audio in such displays obscured information. Thus care must be taken while combining different methods to ensure clarity and effectiveness. The users also stressed on the importance of making the overviews interactive and customisable in order to suit every individual's personal requirements.

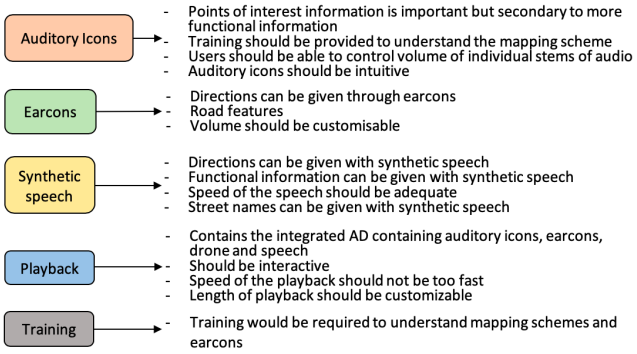


Figure 5: Themes that emerged from applying a thematic analysis to study 2

5. DISCUSSION

Since all users have different preferences and expertise, there is a wide variety of opinions as can be seen from the analysis. As a result, a high level of flexibility is very desirable. This is in line with [24], which states that the success of an assistive system depends on providing means to consider individual preferences. Some of the aspects which highlight the similarities and differences in user and designer perspectives have been discussed in detail below.

According to the BVI participants, direction is the most important route information for them. Moreover, they prefer to have it as speech as it is more explicit and requires no learning. Alternatively, however, most of the designers believe that speech takes up more time and is less flexible in terms of the things that it can represent. They focused their energy on designing soundscapes containing more abstract representations, using auditory icons, earcons and/or panning. This discrepancy is representative of the gap between expert designs and user requirements and shows why products which are designed without user participation are not readily accepted by the concerned community [25]. A BVI user in an informal discussion on social media stated that:

> *"I have noticed with most apps and their flaws [sic.] simply because the developers fail to include VI individuals in their R & D which leads to an app that becomes very frustrating when used by someone who is losing their sight."*

Auditory icons can be used to represent landscape and landmark features, including permanent features like parks or train stations as well as temporary ones like traffic. Some BVI users showed concern that absence of these sounds during the actual journey might cause confusion. For example, if the user is expecting to hear chirping birds depicting a park, but can not hear those sounds due to any reason, such as time of the day, weather, etc. during the journey, they will lose their orientation and feel lost. This problem can be handled by teaching the users to consider this tool as a guide to give a feel of the route, which is designed to augment navigational aids rather than replacing them.

Participating users also stated that landscape/landmarks information alone was useless for guiding travel and must be augmented with a more concrete direction information. They felt that the design in which the components were integrated , i.e. having all or a combination of auditory icons, earcons and speech was most effective.

Both the designers and the users agreed that the 10s time duration, as suggested by [16], was insufficient to provide rich, sequential information. They suggested that the duration should be proportional to the amount of information being imparted.

Representing distance information was one of the major challenges faced by the designers of the study. Some of them proposed to use time delay to depict distance between two instances on the route, such as changing directions or between points of interest, etc. However, none of them was able to effectively include sound representative of distances in their AD as it was difficult in the 10s output time-frame that they had; and hence the users couldn't evaluate or comment on it. Some of the designers and even the users suggested that earcons may be a good choice.

> User_2: *"If the beeps are spaced proportionally to the distance between the streets, I would pay attention to them. For example, if the first crossing is 200 meters away and the second is 100 meters after that, the time interval before the first beep could be twice as long as the time interval between the first and second beeps. That would be useful information to have."*

Using earcons can make the distance information more tangible and easy to follow, as mentioned by User_2. However, whenever earcons are used, a certain level of training is required to understand their meaning. This can be arranged by adding audioclips explaining the earcons and providing an option for training before using the system.

The designers predominantly thought that the AD should be more abstract. They believed that overviews are supposed to be more metaphorical rather than literal. Furthermore, an abstract auditory display might give quicker representation while a literal description using only speech would require more time. This is again in contrast to the users who considered functional information to be more important. However, it comes down to what is considered an overview and what is useful for the users. This is one of the areas that requires more in depth analysis and comparison of designer vs user perspectives. Related to this is the the issue of how to represent objects that stretch over significant geographical areas, such as a school, university or park. If this information is represented literally, this would mean a repetition of the same auditory icon for the length of the geographical area, which can be tedious and annoying. A way of handling this could be to add a fast-moving footsteps sound to represent an unchanging scene briskly. Another way could be to add the sound of opening a door to represent the starting and closing door to show ending of an area that is stretched over a significant distance, making the whole thing more metaphoric.

There was agreement among the designers and users of the two studies conducted, that the display should be interactive and customisable. Both believed that there should be some provision to choose the amount and type of information available at the output. Moreover, the users wanted an interactive display where they can control the level of volume and speed of the different components of the AD, i.e., the earcons, auditory icons and speech, separately as well. This can be solved by setting up a query system, as suggested by [1], where the users can initiate a request to increase or reduce volume or mute some aspect of the AD altogether. This will provide the user with the flexibility to issue requests when they want. Fig. 6 shows the final design implications drawn from both the studies.

These design implications, along with recommendations from literature [9], are being used to inform the design of the AD in the prototype system. Sample auditory overviews can be found at this link. [3]

6. CONCLUSION

In this paper we propose to develop a system which provide auditory route overviews to assist in independent travel for the blind and visually impaired individuals. For this purpose, we explored several research questions in order to develop an informed design in line with user requirements.

Initially, we tried to determine whether a system for providing auditory overview of routes prior to walking was of any value to the BVI community. We posed this question as a survey and determined that there is in fact a high demand for auditory overviews

[3] https://soundcloud.com/anon-5210/sets/audio-routes

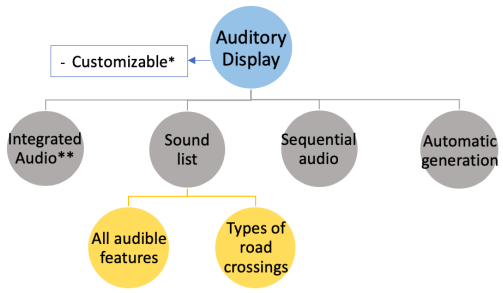
in general and particularly for a system that helps in independent travel.

The next step was to determine user requirements, so we presented some examples to the BVI communities on social media and obtained their feedback. This was an informative exercise and gave insight into their requirements. However, gathering feedback through questionnaires provides limited information. A better way would be to have open-ended discussions/ semi–structured interviews to gain more in-depth knowledge.

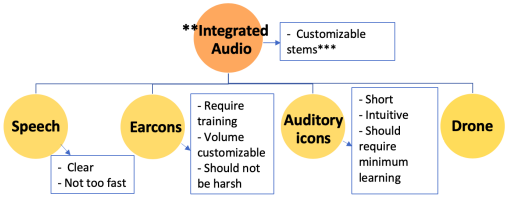
With regards to the techniques of developing auditory displays, it was found that the participants predominantly preferred the use of synthetic speech for obtaining functional information. They also believed that functional information was the most important part of a route map. However, they also felt that having landscape information gave them a feel of the route and landmarks helped them localise and orientate themselves. Hence, we tried to determine how to choose the right ratio of each component in the AD, so as to maximise information while preventing sensory overload. This is an open-ended question in other domains of research with ADs as well, as shown in [4]. The participants suggested that providing the AD with means to control the amount, duration, and amplitude of individual components would allow the users to access each part of the information according to their own needs. However, further user tests are required to determine the effectiveness of this method. This also leads to requirement of exploring different interactive controls and determining which of them are effective.

Similar to selection of components is the question of mapping them to different route elements. Even though the designers in study 1 tried to assign elements of the audio to different road features, the BVI users had a hard time understanding this mapping, whether it was auditory icons or earcons. This showed the need for some training before using the AD or perhaps choosing better representations. This is one the major areas that requires more intense research, and user testing.

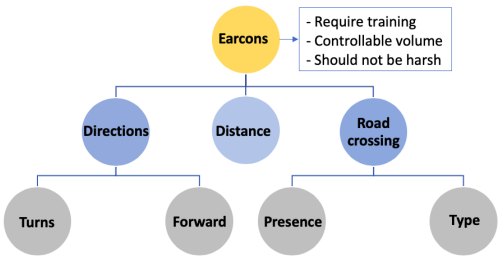
An interesting disparity noticed between designers and users was the balance between functionality and aesthetics in the auditory displays, where designers considered aesthetics to be more important but the users completely disagreed. Since the users will be using the system to aid travel, functionality would definitely have to be given higher priority. However, determining the extent of importance of aesthetics still needs further exploration.



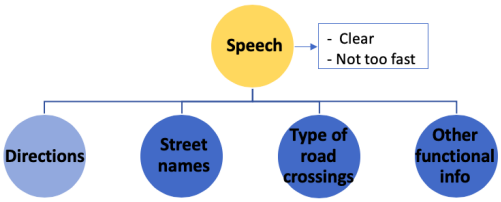
(a) Components of the auditory display



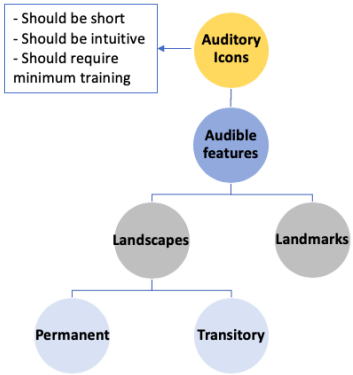
(b) Components of the integrated audio



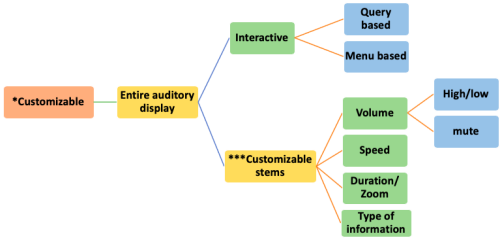
(c) Design implications for earcons



(c) Design implications for speech



(d) Design implications for auditory icons



(e) Design implications for customisable automatic AROs

Figure 6: Design implications for customisable auditory route overviews

7. REFERENCES

[1] A. Arditi and Y. Tian, "User interface preferences in the design of a camera-based navigation and wayfinding aid," *Journal of Visual Impairment and Blindness*, vol. 107, pp. 118–129, 03 2013.

[2] M. J. Muller, D. M. Wildman, and E. A. White, "Taxonomy of PD Practices: A Brief Practitioner's Guide," *Commun. ACM*, vol. 36, pp. 26–28, 1993.

[3] C. Spinuzzi, "The Methodology of Participatory Design," *Technical Communication*, pp. 163–174, 2005.

[4] E. Loeliger and T. Stockman, "Wayfinding without Visual Cues: Evaluation of an Interactive Audio Map System," *Interacting with Computers*, vol. 26, no. 5, pp. 403–416, Sept. 2014. [Online]. Available: https://academic.oup.com/iwc/article-lookup/doi/10.1093/iwc/iwt042

[5] J. Guerreiro, D. Ahmetovic, K. M. Kitani, and C. Asakawa, "Virtual Navigation for Blind People: Building Sequential Representations of the Real-World," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '17*. Baltimore, Maryland, USA: ACM Press, 2017, pp. 280–289. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3132525.3132545

[6] P. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, Feb. 1992. [Online]. Available: http://ieeexplore.ieee.org/document/121642/

[7] T. Hermann, A. Hunt, and J. Neuhoff, *The Sonification Handbook*. Logos Publishing House, Germany, 01 2011.

[8] W. W. Gaver, "Auditory Icons: Using Sound in Computer Interfaces," *HumanComputer Interaction*, vol. 2, no. 2, pp. 167–177, June 1986. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1207/s15327051hci0202_3

[9] S. Hennig, F. Zobl, and W. W. Wasserburger, "Accessible Web Maps for Visually Impaired Users: Recommendations and Example Solutions," *Cartographic Perspectives*, vol. 0, no. 88, pp. 6–27, Nov. 2017. [Online]. Available: http://cartographicperspectives.org/index.php/journal/article/view/1391

[10] J. Wilson, B. N. Walker, J. Lindsay, C. Cambias, and F. Dellaert, "SWAN: System for Wearable Audio Navigation," in *2007 11th IEEE International Symposium on Wearable Computers*. Boston, MA, USA: IEEE, Oct. 2007, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/4373786/

[11] L. Dunai, G. P. Fajarnes, V. S. Praderas, B. D. Garcia, and I. L. Lengua, "Real-time assistance prototype a new navigation aid for blind people," in *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, Nov 2010, pp. 1173–1178.

[12] M. Pielot, N. Henze, W. Heuten, and S. Boll, "Tangible User Interface for the Exploration of Auditory City Maps," in *Haptic and Audio Interaction Design*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4813, pp. 86–97. [Online]. Available: http://link.springer.com/10.1007/978-3-540-76702-2_10

[13] W. Heuten, N. Henze, and S. Boll, "Interactive exploration of city maps with auditory torches," in *CHI '07 extended abstracts on Human factors in computing systems - CHI '07*. San Jose, CA, USA: ACM Press, 2007, p. 1959. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1240866.1240932

[14] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *IN IEEE SYMPOSIUM ON VISUAL LANGUAGES*, 1996, pp. 336–343.

[15] L. V. Nickerson, "Overviews and their effect on interaction in the auditory interface," p. 221, 2012. [Online]. Available: https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/8687/Nickerson_L_PhD_final.pdf?sequence=1&isAllowed=y

[16] H. Zhao, C. Plaisant, B. Shneiderman, and J. Lazar, "Data Sonification for Users with Visual Impairment: A Case Study with Georeferenced Data," *ACM Transactions on Computer-Human Interaction*, vol. 15, no. 1, pp. 1–28, May 2008. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1352782.1352786

[17] A. Brown, R. Stevens, and S. Pettifer, "Audio representation of graphs: A quick look," in *Proceedings of the International Conference on Auditory Displays - ICAD '06*, 2006, p. 8.

[18] L. Brown, S. Brewster, S. Ramloll, R. Burton, and B. Riedel, "Design guidelines for audio presentation of graphs and tables," in *Proceedings of the International Conference on Auditory Displays - ICAD '03*, 07 2003.

[19] J. Kildal and S. A. Brewster, "Exploratory strategies and procedures to obtain non-visual overviews using TableVis," *International Journal on Disability and Human Development*, vol. 5, no. 3, Jan. 2006. [Online]. Available: https://www.degruyter.com/view/j/ijdhd.2006.5.3/ijdhd.2006.5.3.285/ijdhd.2006.5.3.285.xml

[20] J. Kildal and S. Brewster, "Non-visual overviews of complex data sets," in *CHI '06 Conference on Human Factors in Computing Systems*. Montral, Qubec, Canada: ACM Press, 04 2006, pp. 947–952.

[21] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, pp. 77–101, 01 2006.

[22] M. Dascalu, A. Moldoveanu, O. Balan, R. G. Lupu, F. Ungureanu, and S. Caraiman, "Usability assessment of assistive technology for blind and visually impaired," in *2017 E-Health and Bioengineering Conference (EHB)*, June 2017.

[23] E. Striem-Amit, L. Cohen, S. Dehaene, and A. Amedi, "Reading with Sounds: Sensory Substitution Selectively Activates the Visual Word Form Area in the Blind," *Neuron*, vol. 76, no. 3, pp. 640–652, Nov. 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0896627312007635

[24] K. Yamamoto, K. Suganuma, D. Sugimori, M. Murotani, T. Iwamoto, and M. Matsumoto, "Walking support system with robust image matching for users with visual impairment," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*. Anchorage, AK, USA: IEEE, Oct. 2011, pp. 1100–1105. [Online]. Available: http://ieeexplore.ieee.org/document/6083821/

[25] N. Sachdeva and R. Suomi, "Assistive technology for totally blind barriers to adoption," p. 16, 2013.

DESIGN AND EVALUATION OF AN AUDIO GAME-INSPIRED AUDITORY MAP INTERFACE

Brandon Biggs
OCAD University,
Toronto, ON, M5T 1W1, Canada
3164451@student.ocadu.ca

James M. Coughlan
Smith-Kettlewell Inst.,
San Francisco, CA,
94115, United States
coughlan@ski.org

Peter Coppin
OCAD University,
Toronto, ON, M5T 1W1, Canada
pcoppin@faculty.ocadu.ca

ABSTRACT

This study evaluated a web-based auditory map prototype built utilizing conventions found in audio games and presents findings from a set of tasks participants performed with the prototype. The prototype allowed participants to use their own computer and screen reader, contrary to most studies, which restrict use to a single platform and a self-voicing feature (providing a voice that talks by default). There were three major findings from the tasks: the interface was extremely easy to learn and navigate, participants all had unique navigational styles and preferred using their own screen reader, and participants needed user interface features that made it easier to understand and answer questions about spatial properties and relationships. Participants gave an average task load score of 39 from the NASA Task Load Index and gave a confidence level of 46/100 for actually using the prototype to physically navigate.

1. INTRODUCTION

Visual maps have been a part of civilization for many years, but it has only been in the last couple of decades that these visual maps have been turned into digital audio [1], [2]. Despite a number of digital auditory interfaces being presented in the academic literature [1], [3], [4], [5], governments and large mapping companies still do not offer effective nonvisual digital maps commercially, and the Google Maps and ESRI interfaces do not follow auditory display conventions described in the literature [6], [7], [8], [9]. It is difficult to pinpoint why the digital auditory interfaces from the academic literature have not made it into commercial mapping products thus far, but some possible reasons include the need to train users to use an unfamiliar paradigm, an inability to customize the few auditory interfaces that exist, and a limited number of published interface evaluations.

[10] describes a "natural laboratory" in the form of audio games, games that can be played completely using audio, a domain in which extensive iteration in a commercial market has created a set of effective conventions for auditory digital maps that are already familiar to a community of nonvisual users. The present study examines what happens when experienced Audio Gamers interact with a complex digital map that utilizes familiar Audio Game interface conventions identified in [10]. The hypothesis here is that participants would leverage their implicit knowledge of conventions from audio games and find the proposed interface faster and easier to use than the alternatives introduced thus far in the existing auditory display research literature. The findings of the study did not offer a valid comparison in many cases with other studies due to missing data in other studies or due to the data set used in this study not being the dataset used in other studies. This study did highlight that several audio game conventions, such as a scan function, allowing the use of a personal screen reader, having multiple interface types, and combining speech with audio, should be employed in future auditory map designs. Audio game interfaces often undergo rigorous beta testing, and users find the interfaces easy and fun enough to use. The evidence of this is their willingness to pay for the game [11], [12], [13], [2]. [10] outlined a set of interface conventions present in audio games utilized by the prototype in this study, similar to the audio game A Hero's Call [11]. The objective of this study was to evaluate reactions and performance of blind participants on a map utilizing audio game conventions.

1.1. Definition of digital map

For the purposes of this paper, a digital map is conceptualized as a dynamic representation of items configured in spatial and topological relations with each other, represented in a virtual sensory format. This excludes much of the research on interactive maps that use a combination of digital and non-dynamic and non-refreshable physical displays, such as raised-line paper maps over the top of a touch screen and other examples that can be found in Brock and Jouffrais [14].

2. AUDIO GAME CONVENTIONS

The three types of audio game interfaces utilized in this prototype were grid-based, first-person, and tree-based. [10] presents these interfaces: "Grid-based maps are based on a set of coordinates representing squares placed together in a column-row relationship" that are navigated through using the arrow keys. When a user enters a cell, a spearcon (a short speech message [15]) along with a short auditory icon (an iconic sound of an object [16], [17]) play, followed by the cell's coordinates [11], [18]. Grid interfaces are best for getting an overview of a map such as in strategy games [18]. First-person interfaces utilize 3D audio to position objects around the player through looping auditory icons of an object. The use of footstep sounds tell the user what type of terrain they are walking on and how fast they are going. First-person is used to give the player a realistic connection to the real world because the cues presented bear an ecological resemblance to an experience in a real physical environment [19]. Tree interfaces are composed of hierarchical parent-child relationships showing in a hierarchy such as a menu. Games often use tree interfaces for complex menus [20]. Most games, such as [18], [19], and [11] use tree interfaces

https://doi.org/10.21785/icad2019.051

to list locations or options users can select, often with child menus with further options.

3. BACKGROUND

Several promising studies report on auditory digital maps that utilize multiple interfaces such as first-person and grid, but the influence of audio game conventions remains limited.

The map presented in [5] and [21] is the most promising, given that it is a downloadable Windows application and follows many audio game conventions. [5] utilizes a first-person interface and a tree interface, along with a "scan function" to "scan" through points of interest around the player. In the first-person view, looping auditory icons convey the spatial location of points of interest, like the clinking of dishes for restaurants and a fast-moving stream for rivers, that are placed using 3D audio and that change as the user moves around the map. The menus representing different locations one can go to is in a tree interface.

[21] utilized an automatic orientation adjustment to keep participants on a path. In contrast, most first-person interfaces in audio games do not have an automatic orientation adjustment because users can get extremely disoriented, and this is what the study found. The choice to use earcons rather than footstep sounds also could have contributed to the difficulties they had with distance estimation.

Other studies, such as [1], [3], and [22], attempted to utilize a first-person interface, but their systems were often considered complex by participants, even though these studies also found that utilizing auditory icons through 3D audio allowed participants to develop a mental map of a location.

[23] and [4] presented iSonic, a grid-based interface that allowed users to observe trends in data across different geographical regions by listening to speech and musical sounds while the participant arrowed around a grid of the U.S. The most significant feature they found was that participants loved the ability to switch between viewing a table of regional data and switching to the current region on the map, allowing multiple modes for navigation. Their interface, however, differed significantly from that used in audio games [18]. For example, the participant did not jump a fixed distance when moving around the map; instead they jumped region by region. When a participant pressed the up arrow while on Washington state, they went to Alaska; but when they pressed the down arrow to go back to Washington, they landed in Hawaii instead. Their interface also had a training time of 1.82 hours, which is much longer than the 2.5 minutes it takes to read (with a screen reader) the three-page user guide for the audio game Tactical Battle with a grid interface and/or get used to the interface in the tutorial levels [24].

It is difficult to quantify the effectiveness of many of these interfaces, such as [1], [3], and [5], because these papers contain limited results that can be used to compare across studies. Customizability for navigation modes, platform preferences, and synthesizer choice remain extremely limited in all the above prototypes.

4. MATERIAL

4.1. Platform

One of the major objectives of the prototype design was to allow participants to use their own computer and screen reader. This was a deliberate choice that was contrary to most studies, which restrict use to a self-voicing feature (provides a voice that talks by default) and single platform [14], [15], [21], [5]. The reason for this choice was to allow participants to focus completely on the interface, rather than being required to split their attention by learning an unfamiliar synthesizer, although self-voicing was provided by default. The prototype presented in this study was programmed in Javascript and React [25] to be used in the web browser. Audio was played using the Web Audio API and text to speech was obtained either through triggering the participant's screen reader through using ARIA live regions, or used the Web Speech API. The prototype only allowed for keyboard access.

4.2. Map data

The map data was compiled from a combination of measuring shapes from Google Earth and manual measurements taken at the Magical Bridge Playground in Palo Alto, California [26]. The playground map was based off a rectangle that encompassed an area 76 meters wide by 62 meters long.

4.3. Interface design

The auditory interface prototype utilized three modes of navigation: a first-person view, a grid view, and a tree view. The grid view and first-person view utilized the same position and step size settings, so there was no disorientation when alternating between views. It was expected that participants would utilize the tree interface to quickly move between objects, the grid interface to get shape information and spatial relationships between objects, and first-person to walk routes between objects. Each interface had a particular specialty and it was expected participants would utilize the most effective interface for each task. It was not possible to complete the tasks with the tree interface, because there was no information on route information, object shapes, or distance. Allowing these tasks to be completed with the tree interface will be work for future iterations of this project. All modes used the same data from the array of objects. The first-person and grid interfaces used data from the participant's current location to construct their experience.

The first-person interface had a locked orientation with the participant facing the top of the playground. When the participant pressed the arrow keys, the character used footsteps to walk a specified distance every 0.3 seconds. When the participant entered a polygon (i.e., a 2D polygonal region defining an object on the playground), a recorded label would play saying the name of the object. (The polygon shapes are shown in Fig. 1.) Several of the objects, such as the long ramp, had a material attribute set, such as "wood". Footsteps of that material would play when the participant walked over the objects.

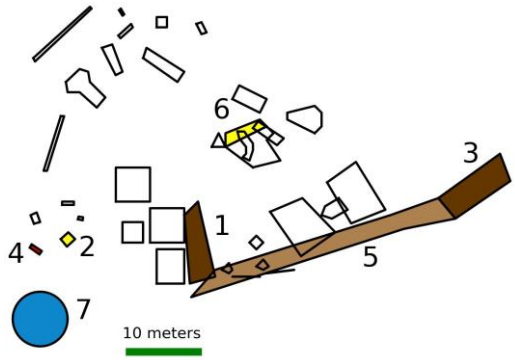


Figure 1. Polygon shapes shown on playground map. Each polygon is drawn with a black outline; polygons that were addressed in the participant tasks are filled in color, with a number label from 1 to 7 printed nearby. The number labels correspond to the following structures: 1 = Ava's Bridge, 2 = Climbing giraffe, 3 = creek bridge, 4 = KinderBells, 5 = long ramp, 6 = roller slide, 7 = stepping sounds. The green bar near the bottom indicates the scale of the map.

The grid interface had more speech and auditory feedback. Every time a participant moved to a new square in the grid interface, a spearcon (a short speech message [15]) would say the name attribute of the polygon followed by the coordinates. The default spearcon was called "Playground Walkway". Several of the objects had short, less than 0.7 second, auditory icons that would play when the participant entered the square with the polygon. The auditory icons were unique identifying clips from the recordings of the object being used. The spearcon and auditory icon would play together. The default sound was an unobtrusive scuff sound.

The tree interface listed the items all together in the object menu, where the name attribute of the object was read out as a spearcon as the participant moved through the menu [15], [18]. The object menu was effectively the map key. Pressing Enter on each object brought up a submenu with the options:

- Go: take the player to the center of the object polygon.
- Listen: hear the sound associated with the object in isolation from the other sounds.
- Description: Hear the textual description of the object, if any.
- Directions: Say where the object was in relationship to the participant's current position and the nearest point. The key "d" would then be set to quickly replay updated directions relative to the player's current location.

The main menu brought up a list of most commands that could be done in the game along with their key shortcut. For example, "Toggle Sounds, t" was the first item. Both the menus were closed by pressing Escape.

5. METHOD

5.1. Structure

The qualitative study comprised two phases: the first was an interview asking participants about their experience with maps and technology, and the second was to show participants a prototype and evaluate their usage and comments on the prototype. The whole study was estimated to take approximately one hour. The studies were all conducted remotely over Skype. Skype was a deliberate choice as it is widely used by the blindness community and allows users to share system audio on Windows. Participants were asked to make sure they had Skype, an updated browser, and headphones.

5.2. Study

All the participants were asked to complete eight tasks (listed below), then rate their performance on the NASA Task Load Index [27], [28]. The NASA Task Load Index is an established method of obtaining a subjective assessment for human-computer interactions and provides a simple numeric score for comparison across multiple tests and interfaces. The eight tasks were chosen to explore the aspects of navigation identified in [29] and [30] such as getting an abstract overview of a map, getting an overview of what is around a location, getting routes between locations, and the exact placement of specific locations. Most of the tasks revolved around participants developing and demonstrating route, landmark, and survey knowledge of the map [30]. Tasks 6 and 7 were used to evaluate if this type of map could be used for scatterplots, heat maps, or other types of representations that require the identification of trends such as those in [4].

Each task was timed starting from when the participant began to complete the task and finished when they completed the task or when they verbally indicated they were done with the task. All the participants were able to ask for the task instructions to be repeated. The headings in the results section were the text that the interviewer said. If the participant asked for clarification a short description or reiteration of the task was given. For example, "Locate the climbing giraffe" could be described as: "Go to the climbing giraffe in any way you wish". The clarification was mostly used by the four participants for whom English was a second language. Participants were not given the definition of each object before starting the task. The eight tasks participants were asked to complete are as follows and are described further in the results section: 1. Locate the climbing giraffe. 2. Describe the route from the stepping sounds to the roller slide. 3. Describe the shape of the KinderBells. 4. What are the objects on both ends of the long ramp? 5. Describe the shape of the long ramp. 6. What is the smallest item on the map? 7. Where is the highest density of items? And 8. Describe the overall layout of the map.

5.3. Participants

Ten congenitally blind male participants were recruited from a forum post on audiogames.net. The study was approved through the institutional review board from OCAD University and no compensation was given for the study. The participants ranged from 16 to 43 years old. The participants were from many different countries including India, South Africa, Romania, Canada, United States, and Iran. All the participants had audio game experience and all of them had used a screen reader for at least five years. All but one user used Nonvisual Desktop Access (NVDA) [31], and one participant used JAWS for Windows [32]. Six participants used Firefox and four used Chrome. None of the participants were familiar with the Magical Bridge playground in Palo Alto. Seven of the participants had no vision, one participant had light perception, and two participants were considered

very low-vision, to the point where they used a screen reader to read large print (one participant said their vision was 20/800 and the other did not know). The analysis of results showed no difference in the performance of the different participants, so they were all aggregated together in the results section.

6. RESULTS

6.1. Exploration Phase: Please explore the map and let me know when you feel comfortable with the interface.

During the exploration time, the researcher gave hints of buttons to press to insure every participant explored the entire interface. The main hints were to press t to toggle the sounds, backslash to toggle between text to speech and the screen reader, escape to bring up the main menu, dash and equals to zoom in and out, and to make sure each participant explored grid view and the objects menu. When the participant finished exploring each part of the interface, the researcher prompted: "Let me know when you feel comfortable using this interface, then we can move on to the tasks." There are three methods that have been explored in the literature for map exploration: [21] and [30] gave a time limit of 15 and 10 minutes respectively to explore the interface before starting the tasks. [4] had a tutorial that took 1.82 hours on average to complete. The approach in this study was similar to [29] that took between 5-10 minutes where they let participants say when they felt comfortable with the interface.

On average, the participants in this study spent 9.87 minutes (SD 6.07) exploring with the fastest being 2.6 minutes and the longest being 19.5 minutes. Five of the participants took less than eight minutes to explore the interface and the other five took more than eleven minutes. It's important to note that the participant who took the longest to explore the interface went to all 43 objects on the map before saying they were comfortable. The fastest participant quickly moved through all the features. There was no major difference between the performance of the slower explorers and the faster explorers. The Faster explorers accomplished 7/8 of the tasks 3 minutes faster on average than the slower explorers. Finding the climbing giraffe took the faster explorers 1.2 minutes and the slower explorers 0.9 minutes. Future studies should compare the performance of slow explorers when timed on a tutorial vs allowing them to feel comfortable with the interface. This exploration method seems faster than the other methods of exploration. There were 43 objects on this map, 8 objects in [29], and 50 objects in [4] and the other studies did not indicate the number of objects on their maps.

6.2. Task 1: Locate the climbing giraffe.

The climbing giraffe is a giraffe leaning over with its neck horizontally curved covered in handholds and toys for kids to play with. The climbing Giraffe was randomly selected from the list of 16 objects that contained sounds and that was not the "Stepping Sounds" which is the first object participants encounter on the map. Participants were asked this question after they felt comfortable using the interface and had explored all the interface features. This task was to evaluate how a participant would find a specific location/landmark on the map. The expected use case for this map included the user knowing the name of an object and wanting to find that object. This is similar to a participant knowing an address and needing to find the address. This task was also going to be repeated for tasks 2 through 5, so it was critical participants knew how to quickly locate items on the map.

There were three methods participants could have used to complete this task: 1. First, they could have moved around in either grid or first-person view and found the object by hearing the sound or hearing the label announced while exploring the map. One of the 10 participants accomplished the task in first-person view doing this method. It took 2.32 minutes. 2. They could have used the Object Menu to get "directions" and walked to the object using the directions. Six of the 10 participants used this method with their times in minutes being: 1.43, 1.18, 6.83, 0.83, 1.5, and 0.97. The participant who took 6.83 minutes tried finding the object first through exploring, then gave up and used the object menu to get directions. 3. They could have used the "go" option to jump to the object. Three of the 10 participants used this method with their times in minutes being: 0.65, 0.47, and 0.38.

The results of this task were not necessarily predictive of future behavior. Nine of the 10 participants used both the "go" and "directions" option at least once during the study with the sole exception being the participant who only moved in first-person during the study. The average time to find the object was 1.66 minutes (SD 1.91).

6.3. Task 2: Describe the route from the stepping sounds to the roller slide.

Stepping sounds are an art installation with a speaker that plays different footstep sounds as users walk in front of a motion sensor. The roller slide is a slide made out of long rotating dowels that spin under the person sliding. This task assessed the ability of users to find a route between two objects. Many map studies use a task to travel between objects as one of the major factors in assessing the effectiveness of a map [21], [29], [30], [5]. [21] describes "decision points" participants encountered during the exploration which were basically intersections or turns. This map had no barriers, so intersections were not applicable. Participants did need to choose the method for travel between objects and identify the objects between the start and end of the route. These two objects were chosen because they both had a sound, and they were relatively far apart (from the nearest point they were 39 squares diagonally apart) with most of the objects between. [5] had success with blind participants describing routes using "free text". The theory was that verbal descriptions and free text would yield similar results, but verbal would be faster and give more detail as participants did not need to type every obstacle and turn they made.

There were three methods participants used to find the route between the two objects: 1. Seven of the 10 participants used the "go" option in the menu to get to one of the objects, then used the "directions" option in the menu to get to the other object. The times in minutes it took to complete the task were: 5.8, 5.32, 4.23, 3.07, 2.65, 3.68, and 6.28. 2. Two of the 10 participants used "go" to get to an object and relied on both the scan function and their memory to locate the second object. The times in minutes it took were: 9.78 and 4.6. 3. One of the 10 participants used first-person to navigate between the objects from memory. It took 3.75

minutes for them to walk to the stepping sounds and find the roller slide.

On average it took all the participants 4.92 minutes (SD 5.93) to navigate and describe the route. In [21] it took participants 16 minutes on average to navigate their route, although there was no number of squares given between the start and end points, so a comparison is difficult to make. They also indicate interruption time separate from navigation time. In this study, participants gave feedback while navigating, so it was not possible to separate navigation from interruption times. [21] also stated their participants had five types of keyboard error: Orientation errors, Omitting error, Unintentional pressing, Incorrect keystrokes while self-orienting, and Miss-keying. None of these errors occurred with the participants in this study. Three of the 10 participants did get lost during the study, but they were able to complete the task with minimal prompting: One of the three participants was prompted "You can use the menu to navigate" when they verbally expressed they were lost and they were able to "go" to the object and make their way to the other object without further prompting (this was the participant that took 9.78 minutes to complete the task). One of the other participants suggested they thought in routes rather than a map, so this task was very easy.

All of the participants managed to navigate between the objects, but all of the routes were slightly different from one another. Each participant was able to articulate the objects they passed and the route they took. For example (starting from the stepping sounds): "Go up, past the mini slide, go a few steps up (maybe 5 or 6), then go right. You pass the disk swings and keep going right, you pass a slide, then you're there." (This participant took 4.23 minutes and used "directions" eight times.) This description is very similar to the text descriptions given in [5]: "Leave Shakespeare's Globe Theatre and turn right along the river. Walk on until you reach your destination, Pizza Express". Future studies should evaluate how participants physically navigate between the objects. Three of the 10 participants expressed their route was not realistic because of needing to cross over the ramp which could not be crossed in real life. This interface should also evaluate the same route in [21], although there is no mention of the start and end points they evaluated on.

6.4. Task 3: Describe the shape of the KinderBells.

KinderBells are a set of bells children can bang with a ball to ring them. It is not clear how important shape recognition is in digital maps. [29] and [3] attempted shape recognition in a 3D auditory landscape, but the "shape of the drawn objects often differs clearly from the real shapes". This description is also valid for the findings in this study. More focused auditory shape recognition has been investigated in several studies such as [33], [34], and [35], and several applications for auditory shape recognition and creation have been developed such as [36], [37], [38], and [39]. For this task, participants were asked to verbally describe the shape of an irregular symmetrical shape. Most studies ask participants to draw shapes or ask participants to describe recognizable shapes such as stars or squares [29], [35]. Physically drawing on swell paper was not possible through the remote medium this study employed and utilizing an application such as [38] would have defeated the cross-platform ability of the study.

The grid medium in this modality meant that the descriptions were all tile based. A slant or curve would look like "steps". The KinderBells are small, so participants were required to zoom in to the highest level to view the shape. The below "tiles" are at the highest zoom level. The exact description of the KinderBells set by the researcher was: "A symmetrical 4-step object with 2 tiles on the top and 2 tiles on the bottom with a single tile nob on either end on the second level. Starting from the top, the horizontal tile width of the levels are 2, 5, 4, 5, 2. The tile length of each level from the top, going to the right is: 2, 2, 1, 2, and the top level has a single square step going to the left." None of the participants gave this level of a description. Five of the 10 participants expressed they did not know how to describe the shape. Two of the 10 participants did not want to switch to the grid view which, in this version, was the only way to get the 2D shape. Three of the 10 participants were able to describe a basic shape: "It's like a sideways rectangle with points on each end. The points are 1 wide… They are off-set… They are at an angle… It's like a crescent with a thicker end and a thinner end. It curves to the bottom of the map."

What should improve the result is the addition of optional borders to object polygons, so that users are able to stay in a polygon if they wish, rather than needing to exit and reenter the polygon every time they move past the edge. Future work needs to incorporate a better shape description system, either using something like [38], or having participants list the points of the polygon.

6.5. Task 4: What are the objects on both ends of the long ramp?

The long ramp is a 44 square long ramp that outlines the bottom right edge of the play area and slants up to the right 13 squares. It has 11 steps and ranges from one to four squares wide. This task tested the ability of participants to follow a path and getting an overview of what is around a location. [21] had participants follow a route, but it was not a single path. [3] has "following paths" as future work that needs to be done.

Seven out of ten participants were able to identify both objects on either end of the long ramp. One participant suggested that along with borders along the edge of the path, earcons of beeps and buzzes representing openings, doors, and objects should be used, similar to those in [11]. There were three methods that participants used to accomplish this task: 1. Four out of seven participants followed the ramp landings until they went out of the object, then they checked if the ramp went up or down from their current location until they reached the end of the ramp. They all started by using the "go" option to get to the center of the ramp. 2. One out of seven participants read the description of the long ramp to answer the question. 3. Two of the seven participants remembered objects from past exploration.

6.6. Task 5: Describe the shape of the long ramp.

Seven out of 10 participants were able to follow the ramp from start to finish and described the ramp as "steps going up to the right". The other three out of 10 participants followed the ramp at least 13 squares to the right and five squares up (four out of 11 "steps").

6.7. Task 6: What is the smallest item on the map?

This question was to evaluate the effectiveness of this map in dealing with something like a scatter plot such as in [4]. Only

one out of 10 participants was able to answer this question correctly. This is because he systematically used the "go" option in the Objects Menu on the highest zoom setting and explored the size of objects in grid view. Once he reached the first object that was one square, he stopped and said that object was the smallest. It took him 6.97 minutes. Seven out of 10 participants started doing this task correctly, but gave up around the 13th (out of 43) object. It would have been much more efficient to have a sound mapped to the area of each object and play that sound as participants arrowed through the Object Menu, or had a sorting option for the Object Menu, similar to [4]. There was no task completion time given in [4], and participants were not identifying the size of objects, so it is difficult to compare the two studies, but the above methods would reduce the amount of steps currently required to review size.

6.8. Task 7: Where is the highest density of items?

This question was to test how effective the map is at conveying clusters of data points. Nine out of 10 participants found one of the two areas with the highest density of items (average minutes = 1.51, SD = 1.13). Three of those nine participants employed scan to count the number of items that were nearby (Average minutes = 2.46, SD = 0.96), five of the nine participants mentioned that they listened for the highest number of sounds clustered together (average minutes = 1.53, SD = 0.95), and one participant used their past knowledge of the map to identify the highest density of items in 0.02 minutes. Seven of the nine participants expressed uncertainty with their choice "I wouldn't say if it is the most clustered, but there is a lot going on".

6.9. Task 8: Describe the overall layout of the map.

This is the first task sighted users do when viewing a map and it is one of the most important uses of a map [29]. Both [29] and [3] evaluate sketches participants drew after hearing their auditory map. The sketches in [29] showed all eight objects properly identified and spatially placed correctly. The sketch method was not possible in this study, so a free verbal description was asked for.

One problem that made itself apparent very quickly was that the participants did not have the vocabulary or chunking skills to systematically describe the map. A common sentiment was: "I don't know how to put all that into words, how things are located." Or "I wouldn't be able to tell you exactly where something is". This response meant that the participants needed a framework to put their responses into. The researcher broke the playground into nine squares: Top right, top middle, top left, middle right, center, middle left, bottom right, bottom middle, and bottom right. The researcher then asked the participant to describe generally what was in each area one section at a time using chunking [40]. It was not practical for participants to remember all 43 objects, especially if the chunks were not extremely clear. This meant that accuracy was evaluated on the percentage of objects correct in each chunk. Five out of 10 participants were able to give a 100% accurate overview with all correct objects in each chunk, four of the 10 participants were able to give a pretty accurate overview with only one or two items incorrect, and one participant was unable to describe any overview. When participants were exploring the interface to get an overview, seven participants switched to grid view and held down the keys so they only heard the auditory icons in each tile. When they heard a sound they didn't know, they would stop, investigate the items, then continue moving as fast as possible to the edge. They performed this action in a grid pattern so they could get what was in each tile. Several comments were that there needed to be sounds for each object to maximize the effectiveness of this strategy. One participant even turned off his screen reader completely and just used the sounds to get an overview of the playground. The average time in minutes for getting an overview was 6.12 (SD 3.19).

This method of evaluation was not ideal as it was difficult to quantify. Future work needs to explore better methods of getting an overview of large-scale landscapes.

6.10. Other Results

- Participants were asked to rate their comfort level physically navigating between two objects that were on either ends of the map. The mean score was 46 (SD = 30.89) with the min score of 0 and a max score of 90, a median of 35 and a mode of 30. 0 was not at all confident and 100 was very confident. The participant with the highest score admitted that he would need his mobility equipment which included his white cane and Sunu band, a wrist band that uses haptic feedback to alert users of obstacles to their upper body [41].
- Eight of the participants used all three interface types to accomplish the tasks and two participants never used the grid interface past the initial exploration stage despite it being the best interface for getting the shape of an object. All the participants also expressed a preference for either grid or first-person for the majority of their navigation. This means that users have a preference for a mode and some will stick with their preference, even if it may not give the information they need. This means it's important that each interface convey the same level of information, such as object shape, spatial relations, and texture.
- All the participants elected to use their own screen reader to accomplish the study. It took less than a minute for all the participants to get the prototype running on their machine. Prior testing showed the prototype working perfectly with Macintosh and Windows platforms, both with self-voicing and screen readers. [1], [5], [3], and [4] all require participants to use the self-voicing feature, rather than use their own screen reader. These results suggest participants prefer the ability to use their own screen reader, like they can do in games such as [11] and [19].
- Nine out of 10 participants repeatedly used the Object Menu to either "go" to an object or get "directions" to an object. [4] presented a function they called a "spreadsheet" interface that listed objects in a list that could be navigated using up and down arrow keys and navigated focus to the selected object when focus was given to the map. Participants were very enthusiastic about this feature in [4], and most participants really liked the feature in this interface.
- All participants made extensive use of the "scan" function. The suggestions were to make

instructions more accurate, so rather than saying "far off, behind and to the left", it would say something similar to "4 meters behind and 10 meters to the left". Also, participants really wanted to adjust the distance of the scan function rather than having it locked at 10 meters.

- The "directions" need to give more constant and accurate feedback. Although directions were extensively used by nine of the 10 participants, the usage pattern was quite excessive. Participants pressed the d key every three seconds when looking for an object. Using beacons similar to [19] and [11] would give a more steady source of the participant's current location relative to the target.

6.11. Task Load Index ratings

The overall workload score in all categories for the NASA TLX was an average of 39 (SD = 10.58). The NASA Task Load Index is a method of obtaining a subjective score for mental load when completing a task. Scores can be used as a baseline when evaluating future work on the same or similar projects [42], [43]. Participants were asked to rate their experience in six subscales on a scale of 0-100, where 0 was as little as possible and 100 was as much as possible. The subscales and their mean scores are: mental demand: 55.1 (SD = 20.58), Physical demand: 5.5 (SD = 7.52), Temporal demand: 38.5 (SD = 19.59), Performance: 58.1 (SD = 21.39), Effort: 50 (SD = 31.62), and Frustration level: 27.5 (SD = 22.88). Other auditory map interfaces have not been evaluated for mental task load.

6.12. Feedback on the prototype

Participants were asked their general thoughts on the prototype. Three participants said they "really liked it" and five said they liked it or thought it was cool because of the familiar interface, ability to get a detailed overview, and sounds. The users who were more moderate in their feedback said it was interesting, but of limited use, and they didn't think they could do anything with it. In general, participants said they found the controls intuitive and very easy because of their resemblance to audio games. All the participants liked the idea of allowing the user to dictate their mode of navigation, either through grid view or first-person, similar to [11]. Each participant was asked why they used each mode of navigation: tree, grid, or first-person. Their responses are summarized as follows:

1. Tree was used for quick navigation through the map.
2. Grid view was used to quickly navigate and get an overview of the map.
3. First-person allowed users to "relate" to the space.

The final question asked users for any final thoughts they had about the prototype. Six of the participants reiterated that they wanted to see a map like this made for more locations: "It was quite fun. If this was released, I would be so happy and use it on a daily basis." Another participant wanted first-person to match the exact navigation system (with ability to change orientation and earcons for surrounding items) as [11].

7. CONCLUSION

The prototype in this study evaluated the use of common audio game conventions to display topological objects on a map. There were several major findings from the tasks: the interface was extremely easy to learn and navigate, participants all had unique navigational styles and preferred using their own screen reader, and participants needed user interface features that made it easier to understand and answer questions about spatial properties and relationships. Future studies need to figure out a more effective way of evaluating the shapes blind users recognize and create a better method for giving a general overview of the map.

8. ACKNOWLEDGMENTS

Dr. Coughlan would like to acknowledge funding support from NIH grant 1R01EY029033 and NIDILRR grant 90RE5024-01-00.

9. REFERENCES

[1] P. Parente and G. Bishop, "BATS: The blind audio tactile mapping system," in Proceedings of the ACM Southeast Regional Conference, 2003, pp. 132–137 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.6189&rep=rep1&type=pdf

[2] J. Friberg and D. Gärdenfors, "Audio games: New perspectives on game audio," in Proceedings of the 2004 acm sigchi international conference on advances in computer entertainment technology, 2004, pp. 148–154.

[3] W. Heuten, N. Henze, and S. Boll, "Interactive exploration of city maps with auditory torches," in CHI'07 extended abstracts on human factors in computing systems, 2007, pp. 1959–1964.

[4] H. Zhao, C. Plaisant, B. Shneiderman, and J. Lazar, "Data sonification for users with visual impairment: A case study with georeferenced data," ACM Transactions on Computer-Human Interaction (TOCHI), vol. 15, no. 1, pp. 1–28, 2008.

[5] E. Loeliger and T. Stockman, "Wayfinding without visual cues: Evaluation of an interactive audio map system," Interacting with Computers, vol. 26, no. 5, pp. 403–416, 2014.

[6] "Cal fire." State of California, 2019 [Online]. Available: http://www.fire.ca.gov/general/firemaps

[7] T. Logan, "Accessible maps on the web," 2018 [Online]. Available: https://equalentry.com/accessible-maps-on-the-web/

[8] Google, "Accessibility in Google Maps," 2019 [Online]. Available: https://support.google.com/maps/answer/6396990?co=GENIE.Platform%3DDesktop&hl=en

[9] ESRI, "A11y-map," 2018 [Online]. Available: https://github.com/Esri/a11y-map

[10] B. Biggs, L. Yusim, and P. Coppin, "The audio game laboratory: Building maps from games," 2018 [Online]. Available: http://icad2018.icad.org/wp-content/uploads/2018/06/ICAD2018_paper_51.pdf

[11] Out of Sight Games, "A hero's call." 2019 [Online]. Available: https://outofsightgames.com/a-heros-call/

[12] audiogames.net, "AudioGames, your resource for audiogames, games for the blind, games for the visually

impaired!" 2018 [Online]. Available: http://audiogames.net/

[13] E. Adams and A. Rollings, Fundamentals of game design. Prentice-Hall, 2006.

[14] A. M. Brock and C. Jouffrais, "Interactive audio-tactile maps for visually impaired people," ACM SIGACCESS Accessibility and Computing, no. 113, pp. 3–12, 2015.

[15] B. N. Walker et al., "Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus," Human Factors: The Journal of Human Factors and Ergonomics Society, vol. 55, no. 1, pp. 157–182, 2013.

[16] E. Brazil and M. Fernstrom, "Chapter 13 auditory icons," in The sonification handbook, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin, Germany: Logos Publishing House, 2011 [Online]. Available: http://sonification.de/handbook/download/TheSonificationHandbook-chapter13.pdf

[17] W. Gaver, "Auditory icons: Using sound in computer interfaces," Human-Computer Interaction, vol. 2, no. 2, pp. 167–177, 1986.

[18] I. Reed, "Tactical battle." 2013 [Online]. Available: https://blindgamers.com/Home/IanReedsGames

[19] J. Kaldobsky, "Swamp," 2011 [Online]. Available: http://www.kaldobsky.com/audiogames/

[20] MetalPop LLC, "Crafting kingdom." 2018 [Online]. Available: http://metalpopgames.com/forum/index.php?/forum/11-crafting-kingdom/

[21] F. Feng, T. Stockman, N. Bryan-Kinns, and D. Al-Thani, "An investigation into the comprehension of map information presented in audio," in Proceedings of the XVI International Conference on Human Computer Interaction, 2015, p. 29.

[22] A. P. Milne, A. N. Antle, and B. E. Riecke, "Tangible and body-based interaction with auditory maps," in CHI'11 extended abstracts on human factors in computing systems, 2011, pp. 2329–2334.

[23] H. Zhao, C. Plaisant, and B. Shneiderman, "ISonic: Interactive sonification for non-visual data exploration," in Proceedings of the 7th international acm sigaccess conference on computers and accessibility, 2005, pp. 194–195.

[24] I. Reed, "User Guide for Tactical Battle," 2013 [Online]. Available: https://blindgamers.com/Home/IanReedsGames

[25] "React: A JavaScript library for building user interfaces." Facebook Inc., 2018 [Online]. Available: https://reactjs.org/

[26] N. Ofiesh and L. Poller, "A playground for the entire community: The design of magical bridge playground," Magical Bridge Foundation, 2018 [Online]. Available: https://slc.stanford.edu/sites/default/files/a_playground_for_the_entire_community_final_submitted_to_mb_and_pw_0.pdf

[27] NASA, "TLX," 2018 [Online]. Available: https://humansystems.arc.nasa.gov/groups/TLX/

[28] Human Performance Research Group (NASA Ames Research Center), "Task load index (nasa-tlx) v 1.0" [Online]. Available: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000021488.pdf

[29] W. Heuten, D. Wichmann, and S. Boll, "Interactive 3D sonification for the exploration of city maps," in Proceedings of the 4th Nordic conference on Human-Computer Interaction: Changing roles, 2006, pp. 155–164.

[30] A. M. Brock, P. Truillet, B. Oriola, D. Picard, and C. Jouffrais, "Interactivity improves usability of geographic maps for visually impaired people," Human–Computer Interaction, vol. 30, no. 2, pp. 156–194, 2015.

[31] NV Access, "NVDA 2017.4 user guide," 2017 [Online]. Available: https://www.nvaccess.org/files/nvda/documentation/userGuide.html

[32] Freedom Scientific, "JAWS, job access with speech" [Online]. Available: http://www.freedomscientific.com/products/software/jaws/

[33] S. Uno, Y. Suzuki, T. Watanabe, M. Matsumoto, and Y. Wang, "Sound-based image and positon recognition system: SIPReS," 2018.

[34] F. Bermejo, E. A. Di Paolo, M. X. Hüg, and C. Arias, "Sensorimotor strategies for recognizing geometrical shapes: A comparative study with different sensory substitution devices," Frontiers in psychology, vol. 6, p. 679, 2015 [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00679

[35] M. Rice, R. D. Jacobson, R. G. Golledge, and D. Jones, "Design considerations for haptic and auditory map interfaces," Cartography and Geographic Information Science, vol. 32, no. 4, pp. 381–391, 2005 [Online]. Available: https://www.researchgate.net/publication/228948257_Design_Considerations_for_Haptic_and_Auditory_Map_Interfaces

[36] D. Greve, "Blind paint." 2009 [Online]. Available: http://www.tichnut.de/blindpaint/

[37] Sysop-Delco, "AudiMesh3D." 2017 [Online]. Available: https://sysop-delco.itch.io/audimesh3d

[38] Sysop-Delco, "BrushTone." 2015 [Online]. Available: https://sysop-delco.itch.io/brushtone

[39] S. Balanced, "New smarter balanced desmos calculators are free and accessible to students," Smarter Balanced, 2017 [Online]. Available: http://www.smarterbalanced.org/new-smarter-balanced-desmos-calculators-free-fully-accessible/

[40] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Psychological Review, vol. 63, no. 2, p. 81, 1956 [Online]. Available: http://www.psych.utoronto.ca/users/peterson/psy430s2001/Miller%20GA%20Magical%20Seven%20Psych%20Review%201955.pdf

[41] "Sunu band." Sunu, Inc., 2019 [Online]. Available: https://www.sunu.io/en/index.html

[42] N. Meshkati, P. A. Hancock, M. Rahimi, and S. M. Dawes, "Techniques in mental workload assessment." 1995 [Online]. Available: https://www.researchgate.net/profile/Peter_Hancock2/publication/232471724_Techniques_in_mental_workload_assessment/links/0c960532a1f2b630b0000000/Techniques-in-mental-workload-assessment.pdf

[43] Agency for Healthcare Research and Quality, "NASA task load index," 2019 [Online]. Available: https://healthit.ahrq.gov/health-it-tools-and-resources/evaluation-resources/workflow-assessment-health-it-toolkit/all-workflow-tools/nasa-task-load-index

STUDIES IN SPATIAL AURAL PERCEPTION: ESTABLISHING FOUNDATIONS FOR IMMERSIVE SONIFICATION

Ivica Ico Bukvic
Virginia Tech
SOPA, ICAT, C+I
Blacksburg, VA, USA
ico@vt.edu

Gregory Earle
Virginia Tech
ECE
Blacksburg, VA, USA
earle@vt.edu

Disha Sardana
Virginia Tech
HCD
Blacksburg, VA, USA
dishas9@vt.edu

Woohun Joo
Virginia Tech
HCD
Blacksburg, VA, USA
joowh@vt.edu

ABSTRACT

The Spatial Audio Data Immersive Experience (SADIE) project aims to identify new foundational relationships pertaining to human spatial aural perception, and to validate existing relationships. Our infrastructure consists of an intuitive interaction interface, an immersive exocentric sonification environment, and a layer-based amplitude-panning algorithm. Here we highlight the system's unique capabilities and provide findings from an initial externally funded study that focuses on the assessment of human aural spatial perception capacity. When compared to the existing body of literature focusing on egocentric spatial perception, our data show that an immersive exocentric environment enhances spatial perception, and that the physical implementation using high density loudspeaker arrays enables significantly improved spatial perception accuracy relative to the egocentric and virtual binaural approaches. The preliminary observations suggest that human spatial aural perception capacity in real-world-like immersive exocentric environments that allow for head and body movement is significantly greater than in egocentric scenarios where head and body movement is restricted. Therefore, in the design of immersive auditory displays, the use of immersive exocentric environments is advised. Further, our data identify a significant gap between physical and virtual human spatial aural perception accuracy, which suggests that further development of virtual aural immersion may be necessary before such an approach may be seen as a viable alternative.

1. INTRODUCTION

Human interfaces to the natural world are inherently multisensory [1]. In simulated environments we often mimic our interaction with the natural world by combining sensory mechanisms to broaden our cognitive bandwidth [2], and to reinforce comprehension [3] and learning [4], [5]. A 1997 report to the National Science Foundation [6] defines sonification as "the use of nonspeech audio to convey information". Simplistic examples of sonification include warning "beeps" that sound when a piece of heavy machinery backs up, and the click-frequency associated with Geiger counters [7], but the full potential of sonification and the multidimensionality of sound is only starting to be explored, particularly in the context of multidimensional datasets. This conceit echoes studies spanning the past two decades: in 1999 Hermann and Ritter suggested that sonification is an "underused perceptual channel for man-machine interaction" [8], and in 2007 Nasir and Roberts stated that "researchers have not fully utilized the maximum potential of spatial sound" [9]. More recently, a 2014 paper by Thomas Hermann suggests that sonification is still in its infancy [10], while in a 2018 publication Paul Vickers notes that "our knowledge of sonification design and theory is still fairly primitive" [11].

Unlike data visualization, which has a long history and a clear set of foundational guidelines [12], sonification is a nascent field [11] that has not yet produced a counterpart to Tufte's seminal work on data visualization [12]. The lack of such knowledge may be one of the major obstacles to the broader adoption of sonification. Sound is inherently multidimensional–each sound has multiple properties that can be assigned to independent variables, or combined to reinforce the perception of a single variable. Such dimensions include timbre, pitch, amplitude, psychoacoustic meaning, source location, and movement. This content richness, when coupled with the innate human ability to simultaneously detect and discriminate between multiple sound sources, supports the contention that sonification affords tremendous promise for analysis of large, complex, multidimensional datasets. Research into sonification may lead to new ways to understand and interact with data, and may significantly enhance and extend traditional data analysis techniques.

1.1. Immersive Exocentric Sonification

In the field of user interfaces the term exocentric environment refers to a virtual reality or other immersive environment that completely encompasses the user [13], [14]. In a previous publication we extended this definition into the aural domain to make a case for an environment that offers all the affordances of the way we interact with the real world [15]. Thus a live concert is experienced in an exocentric environment, but this element is lost when music is heard through headphones that attempt to mimic an exocentric environment but fail to account for user's change in location and orientation. A key difference is that head-motions, echoes, and phase and amplitude differences based on proximity, orientation, and environmental characteristics are fully experienced in an exocentric environment. This specific meaning of the term "exocentric environment" is used throughout this paper. Central to the exocentric environment is its focus on producing sounds whose qualities remain stable throughout the space. Our exocentric environment renders sound sources only around the space perimeter, with no at-

https://doi.org/10.21785/icad2019.017

tempt to emulate virtual sources inside or outside the physical volume. This environment is in contrast to an egocentric environment that prevents changes in sound due to head and body movement, rotation, and orientation. These disparate environments may lead to different conclusions pertaining to human spatial aural perception acuity. This is in part because the egocentric approach does not reflect our real-world capacity for processing sound and interacting with natural aural environments. It is worth noting that there are real-world scenarios in which the egocentric approach may be necessary, because significant head and body movement, rotation, and/or orientation is either not possible or is discouraged. As such, the exploration of both approaches continues to be relevant to the field of sonification.

2. THE SADIE PROJECT

The Spatial Audio Data Immersive Experience (SADIE) is a project whose goal is to increase our capability to create, manage, and understand data and information, with an emphasis on immersive exocentric sonification of three-dimensional multivariate coupled systems. SADIE aims to study spatially distributed data by creating a natural aural environment that leverages intuitive affordances of the immersive exocentric sonification environment, including:

- Utilization of a physical space—recognition and utilization of acoustics, reverberance, and reflections;
- Location-based perception—sound amplitude that is dependent on the user's location within the acoustic field, creating an environment that builds upon natural human perception capabilities;
- Individual variance—aural perception is unique to each individual, and this limits the effectiveness of Head Related Transfer Functions (HRTF) that tacitly employ a one-size-fits-all approach.

2.1. Project Goals

The SADIE project has two primary goals. The first is to develop a powerful, flexible, and reproducible set of tools and techniques with minimal idiosyncrasies, through which it is possible to explore immersive exocentric sonification. The findings of this work will guide further research, including immersive sonification studies of inherently multidimensional spatial data, which may help to quantify its utility for research, teaching, and real-world applications. To minimize idiosyncrasies the project focuses on a geospatial environment model, which has inherently spatial qualities that are directly mapped onto the spatial domain around space perimeter. The second goal of the SADIE project is to develop intuitive approaches to interactions with aural data, including both scientific and artistic scenarios. The data obtained from test subjects have identified several fundamental findings relevant to both goals.

2.2. Side-stepping Idiosyncrasies

Cross-domain-mapping maps elements from a source domain onto a target domain to add an additional level of meaning to the target domain [16]. The approach described herein leverages the human capacity for cross-domain mapping while minimizing potential idiosyncrasies. For example, when we listen to sounds we use our vantage point, location, and motion to accurately pinpoint the sound source, thereby reinforcing our perception by using cross-domain-mapping. In contrast, in the existing immersive audio research literature we observe extensive work in studying human aural perception egocentrically, or in isolation from other senses [15]. One reason for this is the lack of access to infrastructure that is capable of rendering an immersive exocentric aural environment while also tracking users as they traverse the space. Some research has attempted to develop a simulated algorithm to address this problem [5], [17], [18]. Such implementations tend to introduce compounding idiosyncrasies [15] whose impact on the study data may be underestimated. For instance, consider the front-back confusion idiosyncrasy that is inherent to binaural virtual systems; it cannot be addressed without introducing a head-tracking system, but doing so creates latency issues that further compound the problem by introducing new idiosyncrasies.

Another area of concern that may hamper the ability to work with empirical data is the artificial way in which users interact with the system. Interaction that is complex and unnatural may yield biased findings. We posit that systems designed to identify foundational relationships need to be as natural and intuitive as possible. In the SADIE project we focus on allowing subjects to interact naturally with their surroundings, including the ability to freely navigate the space, and to manipulate the properties of spatial aural sources using simple and intuitive hand gestures. We achieve this by using a glove-based gesture interface that is tracked by a motion capture system.

2.3. Infrastructure

A key aspect of SADIE infrastructure is a unique Virginia Tech facility known as the Cube, an immersive cuboid audio facility that measures 50x40x32 feet [19]. The features of the facility that are most relevant to this study are the motion capture capability and the loudspeaker array. The latter includes 124 homogeneous speakers distributed across 5 layers within the facility (3 catwalks and 2 ceiling layers), 4 subwoofers in quad configuration covering frequencies down to 50 Hz, and two 17-inch subwoofers responsible for frequencies below 50 Hz. This configuration enables rendering of a cuboid hemisphere with listeners able to freely traverse the equatorial cross-section of sonified, inherently spatial data. The facility is conducive to all current spatialization algorithms, including both physical and virtual, and thereby allows for testing foundational assumptions and identifying the underexplored potential of immersive sonification.

3. IMPLEMENTATION

Central to SADIEs implementation are three components that constitute the Locus system [20]: the glove-based interaction interface coupled with a motion tracking system, Unity [21] middleware designed to translate captured data into easily interpreted and manipulated Open Sound Control (OSC)-like [22] network packets, and a MaxMSP [23] patch that renders spatial sound and responds to user interaction based on the Unity data stream. We discuss each component in greater detail below.

3.1. Interaction Interface

3.1.1. Prior Work

Within the ICAD community, Beilharz [24] proposed a gestural interaction interface designed to affect sonified data and en-

hance interactions with sound. There were different approaches to a gesture-control interface for use in sound mixing [25] and sound position adjustment of multi-track audio [26]. Sheridan et al. [27]. introduced hand gesture-based software called Soundstudio4D, which allows users to synthesize, spatialize and edit sound. Sterkenburg et al. [28] conducted research on how hand movement can be productively used in connection with auditory displays.

3.1.2. LOCUS

As part of Locus we developed a wearable device to facilitate natural user interaction with spatial aural content in the immersive exocentric environment. It uses two off-the-shelf gloves fitted with retroreflective markers, in conjunction with the 24-camera Qualisys Oqus 500+ motion capture system to allow users to point towards a perceived direction of a sound. To facilitate accurate tracking of both hands, including varying finger positions we used the AIM (Automatic Identification of Markers) model offered by the Qualisys QTM software. Once properly trained, the AIM model is capable of identifying the trained object regardless of hand size, finger or hand position, or orientation. Simple hand gestures are extracted from the Qualisys' spatial marker data using the Unity gaming engine-based toolkit. In this study we focus on the finger pointing gesture that offers a proven natural interaction [29] with minimal impact on the user performance [30]. It is coupled by a thumb trigger gesture consisting of thumb touching the side of the index finger that users can employ to mark the perceived location of the source. We use Unity's vector and raytracing processing capabilities to accurately detect the user's pointing location on the periphery of the domain with submillimeter accuracy, while simultaneously monitoring and responding to other gestures, e.g. thumb trigger. A visualization framework designed to accelerate system setup and troubleshooting illustrates these features on a computer screen.

Unity toolkit's OSC-like output formatting allows it to interface with a wide variety of network-enabled digital signal processing software. Once the motion-capture data are processed they are sent to MaxMSP that responds to the captured data and user's gestures. Doing so allows rapid prototyping by leveraging the functionality of the D4 audio spatialization library that was designed specifically for use with HDLAs in low-latency interactive scenarios with focus on sonification of multidimensional scalar arrays [31]. The resulting infrastructure allows us to distribute sound across the 124.6 HDLA with a high degree of control and interactivity. As described above, interactions employed in this study include pointing towards the perceived location of a spatialized sound source and marking such a location using a thumb trigger motion.

The system enables the support of both egocentric and exocentric environments for a wide array of creative scenarios. Its implementation in the facility further allows for a comparison of various spatialization techniques. Conversely, it allows validation of known sonification ground truths, as well as identification of entirely new ones. These unique affordances have inspired the following research questions.

4. RESEARCH QUESTIONS

The infrastructure described above allows us to study a number of key questions, with the highest priority being given to:

1. What techniques are best-suited to sonifying scalar arrays in an immersive space, and how can they be utilized to facilitate pattern perception in multivariate scenarios?
2. What is a normal user's spatial aural perception capacity, and for what idiosyncrasies must we compensate in interpreting our data?
3. How does the human ability to pinpoint sources and perceive patterns in an immersive exocentric sonification compare to that of more commonly studied egocentric and egocentric-like scenarios?
4. How does sonification in our immersive space compare to that of virtual systems, such as headphone-based binaural systems?

Observing our perception capacity limitations while addressing these topics may better inform the design process and the subsequent implementation of auditory displays. Consequently, the ensuing ground truths may help to create a foundation for a Tufte-like treatise in the audio domain.

5. EXPERIMENTAL APPROACH

Our initial case study focuses on system validation, and on testing the boundaries of human perceptive capabilities in the immersive environment described above. We seek to design tests from which we can infer the limitations of human interactions with the system. A simple example relates to the ability of users to locate sounds that move, are emitted from different locations, or are dispersed over a range of positions. We perform these studies via a sequence of game-like scenarios in which users are asked to identify the sound source location, while we create an anonymous database of user responses. Doing so allows us to determine the normal limits of human perception, and to assess the statistical significance of various tests. As evidenced by the existing body of research, casting our initial studies in game-form was expected to decrease stress among users, while providing a playful environment that may lead to improved retention for sequential studies [32].

5.1. Sonifying Data

Sonification studies can use synthesized and/or sampled sounds. Synthesis offers flexibility in how various parameters may be mapped to the sound generating properties, including simple data audification at a human-audible rate. This can result in widely varying and unnatural sounds. On the other hand, sampled sounds offer a sense of familiarity, and in some instances their psychoacoustic meanings can aid data interpretation. Furthermore, natural or familiar sounds may minimize fatigue and/or annoyance. A notable subset of the aforesaid two approaches are earcons and auditory icons [33] that have a proven role and value in notifying users. Of particular interest is faster response time associated with the auditory icons that, under the right conditions, can be seen as a form of sonification using sampled sounds. Consequently, in our study we opted for a sound that has the following qualities:

- Familiarity;
- Minimal fatigue and/or annoyance factor;
- Broad spectrum that enables greater spatial localization potential, allows for various processing/filtering techniques, and minimizes chances of the sound being masked by other sounds, and
- Consistent amplitude to enable detectable amplitude modulation.

A pre-recorded sound loop of cicadas meets all of these goals. This sound has been used throughout all studies conducted so far in conjunction with a low-frequency (4Hz) exponential inverted sawtooth waveform that modulates the sound's amplitude. Its use resulted in an impulse-like presence of a sound resembling pink noise with a short decaying envelope, followed by a near-silent moment that highlights dissipating reflections.

The study was further complemented by earcons that provided user feedback, which helps to promote a game-like experience for the user. Given the combination of sampled material, its manipulation through amplitude and pulse frequency modulation, and the use of earcons, our system is a hybrid of the aforesaid approaches.

	Stationary	Moving
Physical	Users stand in the middle of the room and are allowed to rotate head and body to locate the source, but are not allowed to move.	Users are encouraged to rotate their head and body, and to walk within the space to help locate the source. A boundary is maintained via a warning sound if the user leaves the motion tacked area.
Virtual	Users locate the sound sources using binaural rendition via motion tracked headphones. They are allowed to rotate their body and head but are not allowed to move.	Users wear headphones while moving around the room to locate the source. The motion tracking system monitors their position and orientation, modifying the sound accordingly.

Table 1: A 2x2 matrix of test scenarios and their variants.

5.2. Study Scenarios

Variables controlled during testing included room lighting, system calibration settings, speaker positions, and randomness of sound source locations and presentation order. The study consists of two scenarios. The first focuses on physical perception of point sound sources in an immersive exocentric environment. A virtual counterpart to using a headphone-based binaural implementation is the second focus, and both studies make use of the system's motion capture capability to account for changes in the user's orientation, head rotation, and position. In the binaural study we mount a rigid body onto the headphones and compensate for the difference between the location of the rigid body and the user's ears (15 cm downward offset against the local Y axis). The motion capture system records the user's head position and rotation and adjusts the output accordingly.

Two scenarios further explore two variants of exocentric environments, resulting in a 2x2 matrix shown in table 1. The first leverages the full potential of an immersive exocentric sonification environment in which users can move and orient freely. They can further invoke head rotation and motion though the space to improve their ability to locate the sound, thus mimicking the way we interact with real-world sound cues. In the second variant the user's location is fixed, but head and body rotation/orientation are allowed. The latter case is a hybrid that has elements of both exocentric and egocentric environments. Because the two scenarios were a part of a larger study, their order was kept consistent for the sake of minimizing the time overhead in reconfiguring the system, while the two scenario variants were presented in random order.

Each scenario consists of 10 trials per user. For each test question a sound is played through the speakers from a random location on the space perimeter, including the ceiling. In human hearing, spatial accuracy decreases with elevation of the sound source. To prevent potential data bias that may ensue from a batch of tests that may use a larger number of randomly generated higher elevations, the elevation choices were limited to 0-90 degrees in 10-degree increments that were consistently utilized in all scenarios, with each elevation being utilized only once per scenario. By phasing the sound sources we create both real (single speaker) and virtual sources, where in the latter case the sound appears to emanate from a region between speakers. Users are asked to find the location of the sound source under various conditions, using their dominant hand for both pointing and marking/triggering functions. With only a short practice session users became adept at interacting with the system.

5.3. Data Processing

Data processing involves calculating the miss-distance in spherical coordinates between the actual sound source and the locations to which the subjects point. The azimuthal and elevation angles of the pointing location are recorded during the study, along with the actual source angles, the time required for users to localize a perceived sound source (in ms), the relative accuracy on a scale of 1-5, and the final game score. Accuracy data are binned in 5 degree increments, so a user pointing to a location within 5 degrees of the correct azimuth and elevation receives the highest possible score for that test. Each lower level of accuracy corresponds to increasing the previous error radius by 5 degrees (10, 15, 20, and 25 or more). Each accuracy level is accompanied by a corresponding earcon. Total game scores allow us to track the best performers among our test subjects in both the stationary and moving scenarios. This competitive aspect of the game adds a degree of excitement, and encourages the test subjects to attempt to beat the all-time highest score, and/or be ranked among the top 10.

5.4. Participant Demographics

After an initial round of beta-testing to identify and remove inconsistencies and biases, a total of 20 test subjects have participated in the study. A hearing test administered prior to the games allows us to screen out persons with hearing impairments. Participants to date are all adults, largely comprised of Virginia Tech students, faculty and staff, and several individuals unaffiliated with Virginia Tech. Test subjects were 30% female and 70% male, ranging in age from 18-55, with a mean age of 25.65. All participation was voluntarily, and no financial or other rewards were given to encourage participation. 90% of the test subjects are right-handed, 75% have had previous experience with a gesture-based device and 45% had been previously exposed to some form of spatial sound environment. Only one of the subjects reported that they were not very interested in music. Others classified themselves on a scale ranging from those who sometimes listened to music, to those who were music majors. All subjects were asked to confirm that they knew how to abort the test before the study began. Qualitative data are drawn from pre- and post-session questionnaires filled out by each user and archived to allow subsequent correlative studies.

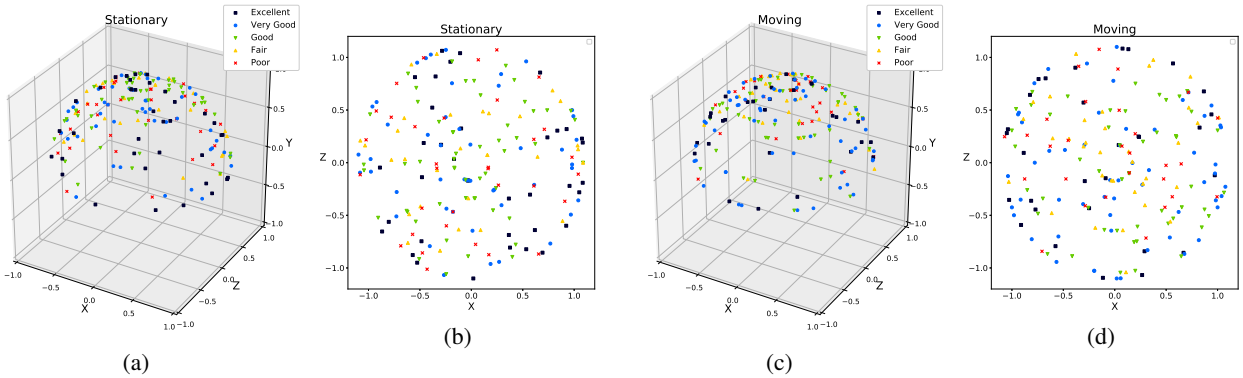


Figure 1: 3D (a and c) and top-down (b and d) color- and shape-coded projections of users' input in stationary and moving variants of the physical immersive scenario, showing consistent accuracy across the entire hemisphere.

6. HYPOTHESES

Several testable hypotheses are investigated using our infrastructure, although more users are needed to generate statistical significance. Prior to testing we formed several hypotheses that were tested using our group of 20 subjects:

1. The human aural perception in an immersive exocentric environment will produce more accurate sound source localization than the egocentric environment;
2. The physical immersive exocentric environment will produce more accurate localization of sound sources than the virtual (binaural, headphone-based) environment;
3. In both physical and virtual scenarios users will perform better in the variant that allows for movement.

7. DATA PRESENTATION AND DISCUSSION

7.1. Quantitative Data

Figure 1 shows the accuracy of our test subjects' ability to identify the source location for sounds emanating from random azimuths and elevations in both stationary and moving exocentric scenarios. Recall from section 5.3 that the color scale of the points corresponds to errors separated into 5 degree bins, ranging from 5 degrees of error to more than 25 degrees. Although the distance between stimuli and users in our study is significantly greater, a subset of this data can be compared to a subset from Figures 5 and 6 of the paper by Oldfield and Parker [29] regarding such errors in a well-controlled egocentric scenario that match the interaction technique while employing a similar sound source (white noise). Whereas both studies cover the full azimuthal plane with the cited study implicitly mirroring one side to another, the proposed comparison only makes sense within the elevation angles available in both studies. The consistent accuracy across the entire hemisphere evident in our data is likely associated with the ability of our users to orient themselves towards the source, thereby utilizing the strongest acuity of their anterior spatial aural perception while side-stepping biological limitations, such as the cone of confusion [29]. This preliminary comparison suggests that hypothesis #1 is correct. While seemingly obvious, this observation may be particularly important given the prevalent use of egocentric scenarios in auditory display research to drive the design and implementation decisions. As a result, we may need to carefully consider how the design of auditory displays can fully utilize the real-world human spatial aural perception capacity and cognitive bandwidth. The confirmation of the hypothesis #1 further warrants research into a more comprehensive exploration of egocentric immersive sonification and its comparison with the exocentric approach.

Figure 2 shows a bar chart that compares the physical and virtual scenarios. All the results are obtained in the same environment, so apart from the technology necessary to allow the virtual scenario to provide immersive exocentric capability, the physical conditions of the two are essentially identical. It is therefore noteworthy that the means, medians, and standard deviations of the errors in identifying the source of a sound are all significantly larger for the binaural data, confirming the hypothesis that human perception is enhanced in the physical immersive environment. The headphone-based tests simulate the immersive environment, but the measured performance results show that these simulations add a significant error, and could in fact lead to invalid conclusions about the utility of sonification as a data analysis tool.

The angular miss-distance (E) measured in our tests and shown on the left axis of the figure is defined as

$$E = cos^{-1}[\sin(\theta_1)\sin(\theta_2) + \cos(\theta_1)\cos(\theta_2)\cos(\phi_1 - \phi_2)],$$

where, θ_1= Perceived elevation angle, θ_2 = True elevation angle, ϕ_1= Perceived azimuth angle, ϕ_2= True azimuth angle

The data in Figure 2 confirm hypothesis #2 above. While expected, this result suggests that efforts to virtualize sonification and audio immersion may impede progress in sonification research by failing to utilize the full range of human auditory capacity, including cross-domain-mapping. This suggests that further development of the binaural approach to representing immersive aural content may be warranted before we begin relying on its economy and convenience, particularly in virtual/augmented/mixed reality scenarios that may benefit from heightened aural localization resolution.

Figure 3 demonstrates that the physical immersive environment data are indicative of better performance. Somewhat surprisingly, the immersive tests do not show marked improvement when subjects are allowed to move, as compared to the cases in which they were required to stand in the middle of the room. Even in

cases where headphones are worn, the increased volume and improved angular sensitivity that might be expected to improve performance are not evident in the data in any statistically significant way. These results are therefore mixed.

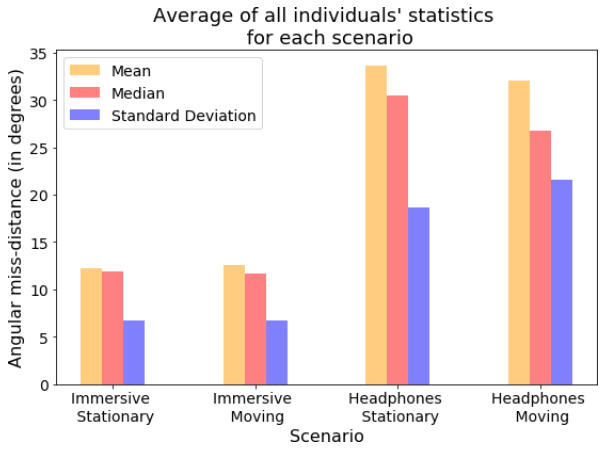
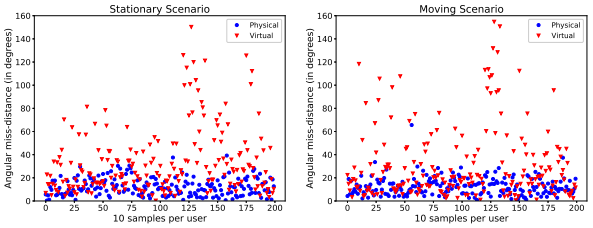


Figure 2: Comparison of the mean, median and standard deviation of the error in localizing a sound source from a group of 20 test subjects. The immersive exocentric environment consistently yields better performance than the virtual, headphone-based approach.

The preliminary data from 20-users suggests that hypothesis #3 is not true, despite the logical assumption that as a listener approaches the source their accuracy is expected to increase. While unexpected, this may be also seen as an advantage in terms of the applicability of the exocentric scenario, whereby head and body orientation may be sufficient to capitalize on the additional perception resolution afforded by the exocentric environment. We conclude that further study in this area may be warranted due to several factors:

- While accurate, human finger pointing at a distance in a space may result in deviations in the perceived location of the sound source for which the current dataset does not accurately account. We aim to address this in follow-on studies by providing more focused training of participants, which should improve their pointing accuracy. Further, the moving component may require a larger space to fully realize its impact and therefore separate the data from the two environments in a statistically significant way.
- In the exocentric scenario that allows for motion, users were confined to the central 20x20-foot space, where the motion capture worked most reliably. In the follow-on studies we intend to expand this to the edges of the space to allow for better-resolved comparisons of the two scenarios.
- Our preliminary qualitative data suggest that users who have more experience with sound and music generally perform better at localizing sound sources. We also observe that participants in general did not feel as comfortable moving around the space as they did when standing in the center, as if they were preconditioned to the stationary scenario. This interesting result warrants further study.



(a) (b)

Figure 3: The left plot shows data from 20 test subjects (with 10 points each) who were asked to locate the source of a sound from a location in the center of the Cube, and the right plot shows the result when the same subjects were allowed to move around the room during the test. Blue dots are results obtained in the physical immersive exocentric environment, and red dots in the virtual counterpart using headphones.

7.2. QUALITATIVE DATA

In addition to the quantitative data shown above, study participants were asked to answer a series of questions both before and after their experience. These data have not yet been fully analyzed, but for a few particular questions a consensus appears to emerge. These are enumerated below:

1. The overwhelming majority of test subjects experienced no discomfort or disorientation as a result of the testing process;
2. The single hearing-impaired individual among our test group became very frustrated and terminated the headphone-based test, but experienced no such effects in the physical immersive environment. Further study is warranted here, as this single data point raises intriguing questions relevant to whether immersive environments are demonstrably better for teaching hearing- and/or vision-impaired individuals who may prefer not to have their ears occluded by headphones [34].
3. The majority of the responses indicate that “Interacting with the gloves was comfortable.” (Strongly agree 14, Somewhat agree 5, Strongly disagree 1). The single person who strongly disagreed mentioned in the feedback that “gloves are too tight for large people!”
4. 17 out of 20 people reported that the headphone-based test was the most challenging part of the game. A few were more specific, observing that they faced difficulties in locating sound elevations. One of the users specifically mentioned that finding the right elevation was the most challenging part of the test.

In terms of responses pertaining to improving the system, a few users mentioned that it would have been better if they were allowed to move more, and if cameras were tracking a larger area, thereby allowing them to move farther from the center. One of the users said “sometimes the glove wouldn’t respond to gestures in certain places,“ indicating limited trackable area to move around”.

8. SEEKING PATTERNS IN GEOSPATIAL DATA

Much of what is discussed here revolves around ground truths and primitives. We consider these essential elements to use as we work toward a larger goal of sonifying geospatial data. Geospatial data from the low-Earth orbit environment is a prime example of big

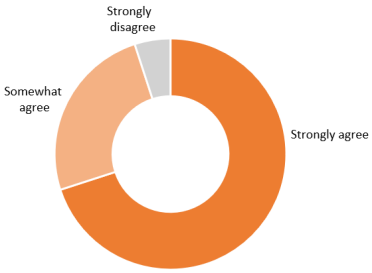


Figure 4: "Interacting with the gloves was comfortable."

data, and its inherent 3D mapping helps us side-step any potential idiosyncrasies that may be associated with arbitrary assignment of the spatial dimension in the process of sonification. As part of this project we have developed a pipeline that allows for importing such data from an empirical model and converting it into the D4 library's time-based changing spatial mask. We have extracted the model data and mapped the ensuing spatial mask to a sonification model that combines amplitude and pulse modulation of the sound source. This combination has shown greatest perception accuracy for the 20 users in our study. The geophysical data vary widely over the spatial domain, and have temporal, latitudinal, longitudinal, seasonal, and solar cycle variability. Our goal is to determine whether such large and complex data sets can be better understood using sonification techniques, and if so, to identify the sonification approaches that yield the best results. While a preliminary demonstration and a production pipeline has been implemented, a number of challenges remain in terms of appropriate sonification techniques in multilayered, multivariate scenarios in conjunction with the aforesaid spatial mask.

9. UNKNOWNS

A facet of this research that may require further attention is identifying the accuracy of the pointing technique. While clearly natural and intuitive, there is a need to further refine the interaction interface to potentially amplify the differences between static and moving scenarios. There are also other considerations, such as occlusion of the ears by long hair, and whether this may also have an effect on the observed data. The Locus system is also easily adapted to accommodate purely egocentric scenarios. Doing so will allow for a more accurate comparison of the two environments and may reveal additional ground truths. The ensuing data will serve as a foundation for a model that capitalizes on the cross-domain mapping to study human sonification capacity in the context of how we interact with the real world. Lastly, in respect to rendering point sound sources, further comparison of the algorithm utilized by D4 to other known spatialization approaches may be warranted.

10. CONCLUSIONS & FUTURE WORK

The results of our initial study confirm some of our initial hypotheses and refute others, but in almost all cases they suggest a need for additional research. We look forward to continuing these studies with significantly larger groups of test subjects in order to add statistical significance to our results. While it is too early to reach solid conclusions, our results so far suggest that the human ability to localize point-based sound sources is significantly better in physical immersive exocentric environments than in its virtual counterpart. Surprisingly, we have been unable to confirm the seemingly obvious prediction that an ability to move around in a space is helpful to the act of localizing a sound source. This suggests the potential for broader applicability of exocentric environments, including scenarios where movement is not an option. Perhaps most importantly, we show that the acuity in the immersive exocentric environment is far greater than that of the egocentric environment, which may warrant a rethinking of how we study spatial aural perception and its use in sonification and other real-world scenarios.

Future work on these and other topics is advised. In particular, a wide variety of sounds and modulation techniques should be studied to determine ground-truths that can be broadly applied to sonification of real-world data. Our work suggests that erroneous conclusions could be reached if poor choices are made in the modulation techniques applied. We have not yet tested a number of key questions, including:

1. How many distinct sounds can a user identify and/or correlate with one another in the immersive exocentric environment?
2. At what point does sonification reach the limits of human perception capabilities, and how can we recognize when this occurs?
3. How can we use the infrastructure developed for our sonification studies to enhance a user's understanding of complex multidimensional datasets?
4. Do immersive environments offer more promise than binaural techniques for teaching individuals with hearing and/or vision impairments?
5. Can we develop techniques to represent vector quantities using sonfication, and if so, how do perceptive abilities change in such cases?
6. What are the opportunities and advantages of collaborative sonification?
7. How does the conditioning in one sonification scenario translate into better performance in that specific case, and in others?
8. What role do time and stress play in localizing sources in both exocentric and egocentric scenarios?

We believe the unique infrastructure at our disposal may allow significant progress to be made on many of these topics. We look to continue this research to gain additional knowledge about the nascent field of sonification, and how it can be used to improve understanding and/or pedantic techniques. It is exciting to imagine a future in which facilities such as ours are common, and are used routinely to explore complex problems and discover new relationships in large and complex datasets. Finally we note that the infrastructure described here may be attractive to performing artists, opening doors to new means of artistic expression. There is surely much more to be learned from sonification studies, and new discoveries awaiting those with the ability to explore its potential. To facilitate this progress our goal is to make the software infrastructure and the supporting documentation publicly available to promote reproducibility and hasten progress towards the sonification of large multi-dimensional datasets and/or a Tufte-like treatise in the audio domain. This publication is an early step toward those goals.

11. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1748667.

12. REFERENCES

[1] G. L. Young, "Human ecology as an interdisciplinary concept: a critical inquiry," in *Advances in ecological research*. Elsevier, 1974, vol. 8, pp. 1–105.

[2] P. A. Kirschner, "Cognitive load theory: Implications of cognitive load theory on the design of learning," 2002.

[3] K. Hussein, E. Tilevich, I. I. Bukvic, and S. Kim, "Sonification design guidelines to enhance program comprehension," in *2009 IEEE 17th International Conference on Program Comprehension*. IEEE, 2009, pp. 120–129.

[4] P. Lennox and T. Myatt, "Concepts of perceptual significance for composition and reproduction of explorable surround sound fields." Georgia Institute of Technology, 2007.

[5] N. Mariette, "Mitigation of binaural front-back confusions by body motion in audio augmented reality." Georgia Institute of Technology, 2007.

[6] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. H. Flowers, N. Miner, and J. Neuhoff, "Sonification report: Status of the field and research agenda," 2010.

[7] J. P. Bliss and R. D. Spain, "Sonification and reliability-implications for signal design." Georgia Institute of Technology, 2007.

[8] T. Hermann and H. Ritter, "Listen to your data: Model-based sonification for data analysis," *Advances in intelligent computing and multimedia systems*, 1999.

[9] T. Nasir and J. C. Roberts, "Sonification of spatial data." Georgia Institute of Technology, 2007.

[10] T. Hermann, "Taxonomy and definitions for sonification and auditory display." International Community for Auditory Display, 2008.

[11] M. Quinton, I. McGregor, and D. Benyon, "Investigating effective methods of designing sonifications." Georgia Institute of Technology, 2018.

[12] E. R. Tufte, *The visual display of quantitative information*. Graphics press Cheshire, CT, 2001, vol. 2.

[13] "Exocentric environment," Jan 2013. [Online]. Available: https://en.wikipedia.org/wiki/Exocentric_environment

[14] A. L. Shelton and N. Yamamoto, "Visual memory, spatial representation, and navigation," *The visual world in memory*, pp. 140–177, 2009.

[15] I. I. Bukvic and G. D. Earle, "Reimagining human capacity for location-aware aural pattern recognition: A case for immersive exocentric sonification." Georgia Institute of Technology, 2018.

[16] G. Lakoff and M. Johnson, *Metaphors We Live By*, 1st ed. University of Chicago Press.

[17] G. Kramer, *Auditory display: sonification, audification and auditory interfaces*. Addison-Wesley Longman Publishing Co., Inc., 2000.

[18] M. J. Morrell and J. D. Reiss, "Inherent doppler properties of spatial audio," in *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[19] E. Lyon, T. Caulkins, D. Blount, I. Ico Bukvic, C. Nichols, M. Roan, and T. Upthegrove, "Genesis of the cube: The design and deployment of an hdla-based performance and research facility," *Computer Music Journal*, vol. 40, no. 4, pp. 62–78, 2016.

[20] D. Sardana, W. Joo, I. I. Bukvic, and G. Earle, "Introducing locus: a nime for immersive exocentric aural environments," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. Porto Alegre, Brazil: Federal University of Rio Grande do Su, June 2019.

[21] J. Halpern, "Introduction to unity," in *Developing 2D Games with Unity*. Springer, 2019, pp. 13–30.

[22] M. Wright, A. Freed, *et al.*, "Open soundcontrol: A new protocol for communicating with sound synthesizers." in *ICMC*, 1997, pp. 101–104.

[23] M. Puckette, "Max at seventeen," *Computer Music Journal*, vol. 26, no. 4, pp. 31–43, 2002.

[24] K. Beilharz, "Wireless gesture controllers to affect information sonification." Georgia Institute of Technology, 2005.

[25] M. Lech and B. Kostek, "Gesture-controlled sound mixing system with a sonified interface." Georgia Institute of Technology, 2013.

[26] M. J. Morrell, J. D. Reiss, and T. Stockman, "Auditory cues for gestural control of multi-track audio." International Community for Auditory Display, 2011.

[27] J. Sheridan, G. Sood, T. Jacob, H. J. Gardner, S. Barrass, *et al.*, "Soundstudio 4d: A vr interface for gestural composition of spatial soundscapes." in *ICAD*, 2004.

[28] J. Sterkenburg, S. Landry, and M. Jeon, "Influences of visual and auditory displays on aimed movements using air gesture controls." Georgia Institute of Technology, 2017.

[29] S. R. Oldfield and S. P. Parker, "Acuity of sound localisation: a topography of auditory space. i. normal hearing conditions," *Perception*, vol. 13, no. 5, pp. 581–600, 1984.

[30] P. Majdak, M. J. Goupell, and B. Laback, "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, Perception, & Psychophysics*, vol. 72, no. 2, pp. 454–469, Feb. 2010. [Online]. Available: https://doi.org/10.3758/APP.72.2.454

[31] I. I. Bukvic, "3d time-based aural data representation using d4 librarys layer based amplitude panning algorithm." International Community on Auditory Display, 2016.

[32] M. Krause, M. Mogalle, H. Pohl, and J. J. Williams, "A playful game changer: Fostering student retention in online education with social gamification," in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 2015, pp. 95–102.

[33] M. P. Bussemakers and A. De Haan, "When it sounds like a duck and it looks like a dog... auditory icons vs earcons in multimedia environments," in *Proceedings of the international conference on auditory display*, pp. 184–189.

[34] H. Petrie, V. Johnson, T. Strothotte, A. Raab, R. Michel, L. Reichert, and A. Schalt, "Mobic: An aid to increase the independent mobility of blind travellers," *British Journal of Visual Impairment*, vol. 15, no. 2, pp. 63–66, 1997.

SONIFICATION OF THE RIEMANN ZETA FUNCTION

Nick Collins

Department of Music
Durham University
Palace Green
Durham, DH1 3RL
nick.collins@durham.ac.uk

ABSTRACT

The Riemann zeta function is one of the great wonders of mathematics, with a deep and still not fully solved connection to the prime numbers. It is defined via an infinite sum analogous to Fourier additive synthesis, and can be calculated in various ways. It was Riemann who extended the consideration of the series to complex number arguments, and the famous Riemann hypothesis states that the non-trivial zeroes of the function all occur on the critical line $0.5 + ti$, and what is more, hold a deep correspondence with the prime numbers. For the purposes of sonification, the rich set of mathematical ideas to analyse the zeta function provide strong resources for sonic experimentation. The positions of the zeroes on the critical line can be directly sonified, as can values of the zeta function in the complex plane, approximations to the prime spectrum of prime powers and the Riemann spectrum of the zeroes rendered; more abstract ideas concerning the function also provide interesting scope.

1. INTRODUCTION

The Riemann zeta function [1, 2] is a construction in analytic number theory of great beauty and wide scope, with a central assertion in its theory, the Riemann Hypothesis (RH), that has remained unsolved since its original statement in 1859. Musical analogies have often been made in referring to the problem, with Fourier analysis a tool in the analysis of the equation, and the dual structure of the non-trivial zeroes of the function and the prime numbers analogous to the spectral and time domain viewpoints of a sound signal [3, p. 89]. There is a great deal of interesting mathematics surrounding the zeta function, commensurate with the efforts of mathematicians for centuries to gain handles on the RH that all the non-trivial zeroes of the function appear only along one 'critical line' in the complex number plane. The fuller exploitation of equations and data relating to RH for musical purposes is the subject of this present paper, and we treat direct synthesis ('audification'), as well as sonification of rhythms and pitch structures.

This paper does not present the first ever sonification of the zeta function. Multiple authors have synthesized the zeroes of the function in particular, including Jeffrey Stopple (http://web.math.ucsb.edu/~stopple/explicit.html), Robert Munafo (https://mrob.com/pub/ries/zeta.html) and Andrey Kulsha (http://empslocal.ex.ac.uk/people/staff/mrwatkin/zeta/kulsha.htm). Such sonifications tend to be based on sinusoidal resynthesis following the gradual journey along the critical line where all known zeroes have been found, incorporating the contribution of each zero as it arises. In perhaps the most developed precedent, the distinguished physicist Michael V Berry explores a number of sonifications [4], including a sum of sinusoids corresponding to the Riemann zeroes, and direct synthesis of the zeta function along the critical line based on the Riemann-Siegel formula. We differ from this prior work in considering direct synthesis based on the naive approach of summing the zeta function, on exploring rhythm and scales, and in a greater willingness to accept any 'noisy' outputs as acceptable within the wider space of sound available in computer music. We also provide SuperCollider code to accompany the paper, providing immediate sound examples and realtime interactive synthesis capability.

This work is in the spirit of composers who have integrated mathematics into the core of their music compositions, perhaps foremost of which was Iannis Xenakis, who adapted such content as hyperbolic curves, probability theory and statistical functions, group theory and game theory [5, 6]. The inter-relationship of music and mathematics is a wider topic than we have space to fully survey here [7], but composers have demonstrated a number of approaches to the incorporation of algorithms into their practice, from strict observance of algorithmic output data to taking various liberties [8, 9, 10]. Prime numbers have often appeared in composer's work, from just intonation theory using small integer ratios often favouring primes, to sonification of the sequence of prime numbers.[1] In terms of the model of Vickers and Hogg [11] the present work is more abstract, as pertaining to a Platonic space of mathematics rather than real world data, and central within the continuum between music and scientific sonification. We are interested in new sonic resources [12, 13], and do not harbour illusions that sonifications of the zeta function rather than hard mathematics will somehow resolve the RH. However, the consideration of such mathematics does widen the appreciation of beautiful ideas and human ingenuity, genuinely inspiring for new musical creation, and illuminating with respect to the audibility (or otherwise) of transplanted advanced mathematics.

[1] For example, the links at http://empslocal.ex.ac.uk/people/staff/mrwatkin//zeta/curiosities.htm provide some online projects.

https://doi.org/10.21785/icad2019.003

2. CALCULATION AND DIRECT SYNTHESIS

Whilst the zeta function for complex argument s is written

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \tag{1}$$

it is far simpler for computation to consider the related eta function:

$$\eta(s) = \sum_{n=1}^{\infty} \frac{-1^{n-1}}{n^s} \tag{2}$$

where $\eta(s) = (1 - 2^{1-s})\zeta(s)$.[2]

Since we are considering complex argument s we write $s = \alpha + ti$ and use standard identities

$$\begin{aligned} n^s = \\ n^{\alpha+it} = \\ \exp\log(n^{\alpha+it})) = \\ \exp((\alpha+it)\log(n)) = \\ \exp(\alpha\log(n))\exp(it\log(n)) = \\ n^{\alpha}(\cos(t\log n) + i\sin(t\log n)) \end{aligned} \tag{3}$$

Thus in the last step using the Euler form of a complex exponential as cosine and sine terms, and rewriting with $-s = -\alpha - it$:

$$\eta(\alpha+it) = \sum_{n=1}^{\infty} -1^{n-1} n^{-\alpha}(\cos(t\log n) - i\sin(t\log n)) \tag{4}$$

we end up with a Fourier-like representation. The complex number result can be expressed as two infinite sums, one over cosines and one over sines (which suggests immediate additive synthesis rendering):

$$\begin{aligned} &\eta(\alpha+it) = \\ &\sum_{n=1}^{\infty} -1^{n-1} n^{-\alpha}\cos(t\log n) - i\sum_{n=1}^{\infty} -1^{n-1} n^{-\alpha}\sin(t\log n) \end{aligned} \tag{5}$$

The recurring term $t\log(n)$ expresses a scaling by t of $\log(n)$, which when passed as argument to a trigonometric function of period 2π, pushes the terms more or less far in phase. The zeroes of the Riemann zeta function occur in the remarkable situation that both the sum of cosines and of sines cancel out.

For additive sound synthesis, brute force summation can be carried out for s in regions of convergence of the η function ($\alpha > 0$), though less efficiently than some series acceleration methods allow. Euler-Maclaurin summation or the Riemann-Siegel formula on the critical line [1], the Borwein method [14] or the FastZeta algorithm [15] are possible improvements, though best convergence still requires on the order of $t^{1/2}$ summands, and complexity of calculation is thus highly dependent on the height of t. In the naive sum, the $n^{-\alpha}$ coefficients in combination with the sinusoids

[2]Zeroes due to the term $1 - 2^{1-s}$ only occur for $\alpha = 1$ and $t = \frac{k2\pi}{\log 2}$, for integer k thus outside the ordinary area of interest of the function

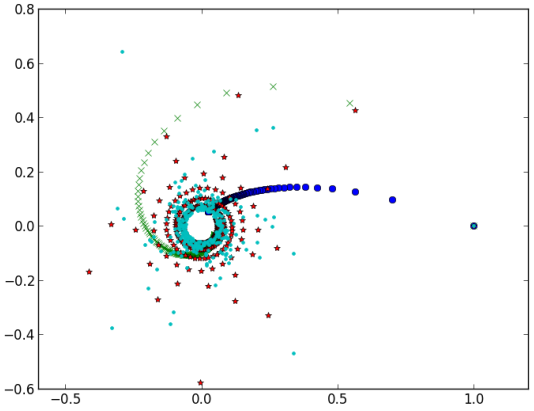


Figure 1: Example sequences of $n^{-0.5}(\cos(t\log n) + i\sin(t\log n))$ for $n = 1..100$ and $t = 0.2$, 1, 10 and 1000

cause a spiralling in of the magnitude of the complex numbers, though this is often a slow process (see Figure 1). Indeed, in perceptual terms, the $n^{-\alpha}$ prefix does not drop off at all quickly for $0 < \alpha < 1$ in the region of most interest to studies of the zeta function. A -60 dB drop requires $-dbamp(60)^{\frac{-1}{\alpha}} = 0.001^{\frac{-1}{\alpha}}$ terms, so for $\alpha = 0.5$ on the critical line, one million terms. Note though that the cosine and sine components will periodically drop to zero, and that depending on t and n, the summands can enter long runs at particular phases corresponding to positions near trigonometric zeroes, due to the slowing of $\log(n)$.

If synthesis of the whole zeta function sum is carried out for increasing t on the critical line $\alpha = 0.5$, zero amplitude of the function will be heard at the famous zeroes.

For high t on the critical line, the sinusoidal components of the zeta function sum revolve incredibly quickly until $\log(n)$ is changing slowly enough to offset large t; this will drop to under one cycle per n at $2\pi = t\log(n+1) - t\log(n) = t\log\frac{n+1}{n}$ so $\exp\frac{2\pi}{t} = 1 + \frac{1}{n}$ and therefore $n = \frac{1}{\exp\frac{2\pi}{t} - 1}$. Close movement in phase may also accompany higher multiples of 2π so this is just the point past which every update is under a cycle in difference; the formula is quickly adjusted for under ϵ in difference. Under 1 unit in difference has the approximate solution $n = t$ since a rough approximation, especially applicable for higher t as $\log(n)$ changes more and more slowly, follows from the derivative $\frac{d\log(n)}{dn} = 1/n$ so that a change of 1 corresponds approximately to the difference $1/n = \log(n+1) - \log(n)$.

If a zero s off the critical line was ever found with $0 < \alpha < 1$, the functional equation of the zeta function, and the fact that complex conjugates of zeroes are also zeroes implies that four different equations are true, namely $0 = \zeta(\alpha+ti) = \zeta(\alpha-ti) = \zeta(1-\alpha+ti) = \zeta(1-\alpha-ti)$ and four versions of the sums in cos and sin above sum to zero (there aren't eight because equating complex parts to zero, negative or positive versions of the real and imaginary sums are both zero), so the two sums already in (5) and further $\sum_{n=1}^{\infty} -1^{n-1} n^{-1+\alpha}\cos(t\log n)$ and $\sum_{n=1}^{\infty} -1^{n-1} n^{-1+\alpha}\sin(t\log n)$. The sonification of positions off the critical line could consider rendering these sums and thus illuminating their difference from each other and zero.

As an alternative sound synthesis resource, expressions of the form $\cos(t \log(n))$ are actually quite productive, including for real rather than integer n. Conversion of the sum over n to an approximating integral and error term, as in the derivation of the Euler-Maclaurin formula, is a precedent to consider continuous n.

3. SUPERCOLLIDER SYNTHESIS EXAMPLES

The domain specific audio programming language SuperCollider [16] was used for sonifications; it has the great advantage that it is designed for realtime sound synthesis with interactive coding allowing for fast prototyping [17]. As an example of coding in the language, the first code block presented here generates the correct shape on the critical line for the eta function (here, $0 \leq t \leq 30$). Sound and code examples are available along with the release of this paper.[3]

```
(0,0.1..30.0).collect{|t|
var real ,imag;
var signal = Array.fill(100,{|i|
      var n = i+1;
      var phase = t * log(n) + (pi*i);
      //+0.5pi to make cosine, pi*i is (-1)
          **(n-1) when put through cosine

      (n**(-0.5))*[sin(phase+0.5pi),sin(
          phase)];
});

signal = signal.sum;

real = signal[0];
imag = signal[1];

((real*real) + (imag*imag)).sqrt;
}.plot
```

This static generation can be turned into live synthesis. There is a limit to the number of sinusoids summed within a single SynthDef, which can be overcome by writing a new UGen specific to the synthesis capability desired. The EtaFunction UGen utilises a pre-calculated listing of the natural logs of the first 5000 positive integers, as part of strategies for sample by sample rendering of the naive sum.

```
({
var t = Phasor.ar(0,MouseX.kr(1,1000)/
    SampleRate.ir,10,30);
var n = K2A.ar(MouseY.kr(1,200));
var zeta = EtaFunction.ar(DC.ar(0.5),t,n);
var real = zeta[0]; var imag = zeta[1];
Limiter.ar(((real*real) + (imag*imag)).sqrt
     *0.5);
}.scope
)
```

Berry [4] builds up a picture of the primes from the Riemann zeta function zeroes (the equation is further discussed by Mazur and Stein [3, p.110]). In this sense, the prime numbers (due to technicalities, along with the prime powers p^n) have a spectrum defined by the zeta function zeroes ρ, and vice versa.

[3] https://composerprogrammer.com/research/ICAD2019examples.zip

$$primeapproximation_N(x) = -\sum_{n=1}^{N} \cos(\rho_n \log x) \quad (6)$$

$$riemannapproximation_N(\theta) = 2 \sum_{p^n < N} p^{-n/2} \log(p) \cos(\theta n \log p) \quad (7)$$

Taking the first of these, the approximation of a prime spectrum from zeta zeroes, in SuperCollider code this might be calculated in the language assuming ~riemannzeroes is an array of the first 1000 non-trivial zeroes:

```
var primesignal = Array.fill(1000.min(~
    riemannzeroes.size),{|i|
      cos((~riemannzeroes[i]) * log(x))
});
primesignal.sum.neg.plot;
```

The equation can also be rendered via UGens for direct synthesis:

```
var primesignal = Mix.fill(1000.min(~
    riemannzeroes.size),{|i|
      cos((~riemannzeroes[i]) * log(K2A.ar(
          MouseX.kr(2,100)).lag(0.1)))
});
```

In order to explore this further live, a UGen *PrimeSpectrum* was created, with critical input parameters x (for position along the real axis) and N (for the number of zeta zeroes utilised in the sum):

```
{
var x = SinOsc.ar(MouseX.kr(1,100)).range
    (2,100);
var n = K2A.ar(MouseY.kr(1,1000));
//limit and scale down due to larger
    outputs well outside -1 to 1
Limiter.ar(PrimeSpectrum.ar(x,n)*0.1);
}.scope
```

The live modulation of the number of zeroes made available in the approximation is an interesting effect; the clarity of the image of primes and their powers is enhanced the more zeroes are committed. As well as losing their sharp spikes at primes and prime powers, the y values fall off to the negative as more and more zeroes are missed off. The UGen has a maximum of 5000 zeroes based on its internal database of zeroes and the viability of calculation. Very high x can be calculated without sufficient zeroes in the calculation, and the resulting noisy output is still an attractive synthesis resource. This openness to noise in results differs from Berry's strait-laced and rather tonal-centric earlier conception. Similarly, a RiemannSpectrum UGen was built, through a precalculated table of prime powers and with arguments for the position along the spectrum and the number of components in the sum.

Expressions of the form $\cos(t \log(n))$ are easily deployed in SuperCollider unit generator graphs to create novel sound timbres.

The log acts to compress larger numbers more, that is, provide a nonlinear waveshaping (requiring strictly positive input; ≥ 1 is used in the examples here). The t scales the output within the phase input for a cosine function. There is a relationship here to the nonlinearities possible through frequency modulation synthesis [18]. Unlike calculating full sums, use of this formula for an individual summand is very low cost on a modern CPU, with synthesis of these patches just a matter of percentage points.

In the first SuperCollider example, a sine oscillator is used to sweep up and down in n, with the range of sweep under mouse control on the Y axis, and a further mouse control on the X axis for t:

```
{cos(MouseX.kr(1,100) * log( abs(SinOsc.ar(
    MouseY.kr(1,1000,'exponential'))*10) +
    1))}.play
```

The abs function causes full wave rectification, which has the spectral effect of creating many harmonics of the base frequency with amplitude fall off as per $\frac{1}{1-4k^2}$ for harmonic multiple k. The log function can be analysed as follows:

$$\begin{aligned} \log(10|\cos(x)|+1) = \\ \log 10 + \log(|\cos(x)| + \frac{1}{10}) = \\ \log 10 + \log(|\cos(x)| - \frac{9}{10} + 1) = \\ \log 10 + \sum_{k=1}^{\infty} \frac{(|\cos(x)| - \frac{9}{10})^k}{k} \end{aligned} \tag{8}$$

using the power series formula for $\log(1+x)$ since $||\cos(x)|-\frac{9}{10}| < 1$. The final expression illustrates the complexity of harmonics that will arise from the composition of absolute and log operators here; the powers by k of a sum of harmonics will lead to a further set of ring modulated components at sum and difference frequencies (at further multiples of the harmonics, reinforcing in a complex way the amplitude of the original absolute of sinusoid harmonics). The final expression is then the modulator for frequency modulation, leading to complicated sidebands from modulations of modulations, with amplitudes following a product of Bessel functions [19]. Since all the input sinusoids forming the modulator are harmonically related, the output of FM will be itself at harmonic frequencies, with a very complicated distribution of energy. The t scale factor will scale the indices of modulation, increasing the audibility of sidebands (harmonics) for larger t.

The second SuperCollider interactive example demonstrates a nice combination, two trigonometric expressions of the argument $t\log(n)$ fluttering against each other, with user control via mouse of t and the contrasting rate of sweep on n between the two components:

```
( {
var y = MouseY.kr(1,1000,'exponential');
var n = SinOsc.ar(4,0,y).abs +1;
var m = SinOsc.ar(7.9,0,y).abs +1;
var t = MouseX.kr(1,100);

(n.squared.reciprocal)* cos(t * log(n) ) +
(m.squared.reciprocal) * sin(t * log(m) )
}.play
)
```

The last example is a variant of the main naive sum formula for complex output, here sonified to left and right stereo positions, with 40 summands, and a phasor continually incrementing t in a sawtooth rise, the rate of progress controlled by the user via mouse X position on screen. The tanh function is used to keep the final output within bounds and provide a little distortion edge:

```
({
var alpha = 0.5;
var t = Phasor.ar(0,MouseX.kr(1,10000,'
    exponential')/SampleRate.ir,1,100);

tanh(
      Mix.fill(40,{|i|
      var n = i+1;
            var temp = t * log(n);
            (n**(alpha.neg)) * ((-1)**i) *
                [cos ( temp), sin ( temp)] ;
}));

}.play
)
```

4. SPACINGS BY ZEROES

The set of zeroes ρ of the zeta function are an interesting resource for the spacing of events.

Figure 2 presents the spacing of the first seventeen zeros of the zeta function,[4] as a rhythm.

To four decimal places, the rhythmic events are at:

14.1347, 21.022, 25.0109, 30.4249, 32.9351, 37.5862, 40.9187, 43.3271, 48.0052, 49.7738, 52.9703, 56.4462, 59.347, 60.8318, 65.1125, 67.0798

The inter-onset interval (the gaps) corresponding to these, including the rest at the start before the first event, are:

14.1347, 6.8873, 3.9888, 5.414, 2.5102, 4.6511, 3.3325, 2.4084, 4.6781, 1.7687, 3.1965, 3.4759, 2.9008, 1.4847, 4.2808, 1.9673, 2.4666

An optimal re-scaling was sought by exhaustive grid search to bring this set of event times as close as possible to uniform 24^{th} notes (0.16666 of a beat), leading to the quantised solution:

1, 0.5, 0.3333, 0.3333, 0.1667, 0.3333, 0.1667, 0.1667, 0.3333, 0.1667, 0.1667, 0.1667, 0.1667, 0.1667, 0.3333, 0.1667, 0.1667

Listening to the original zero spacing, versus this quantisation, the two are clearly distinct, though the general shape of the former is captured by the approximation.

Results on the number of zeroes up to height T tend towards:

[4] Tables of zeroes are available online, see for instance `http://www.dtc.umn.edu/~odlyzko/zeta_tables/index.html` as well as multiple associated pages across the On-Line Encyclopedia of Integer Sequences `https://oeis.org/A013629`. There are web pages that allow access to higher runs of the zeroes `http://www.lmfdb.org/zeros/zeta/?limit=100&t=100000`

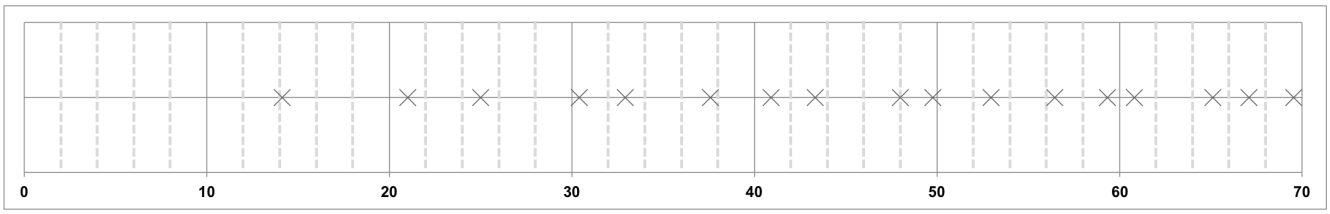


Figure 2: A Riemann zeta function rhythm: the spacing of the first seventeen zeroes

$$N(T) \approx \frac{T}{2\pi} \log \frac{T}{2\pi e} \tag{9}$$

The spacing of the zeroes is closer over increasing t, the difference being approximately $\frac{2\pi}{\log n}$,[5] asymptotically the n^{th} zero appearing at $\frac{2\pi n}{\log n}$. This is actually an inverse dual to the primes, which are approximately distributed such that the number to n are $\frac{n}{\log(n)}$ and the n^{th} prime appears around $n \log(n)$.[6] High up the critical line Odlyzko demonstrates three consecutive zeroes separated by only around 0.05 for t proximate to 10^{22} [20]. Synthesis for low $n = 1..10$ for the approximate equation (and cumulative sum) leads to the values

9.065, 14.784, 19.316, 23.22, 26.727, 29.956, 32.977, 35.837, 38.566, 41.186

We can synthesize rhythms using this approximation equation, or quantised approximations to the approximation, and scale them as needed. Due to the slowing of the log function, which can be thought of as only really increasing by 1 on the order of each additional digit in the decimal number[7], short runs in n at high t will be approximately linear. So the most interesting part of this formula for rhythms is the early part. There are studies which show that for high t the spacing of the zeta function takes on a random aspect relating to the eigenvalues of certain random matrices studied in physics [21, 22]; in a related publication, Michael Berry has sonified various categories of random matrix, including in comparison to Riemann zero data, using frequency modulation [23].

Following the dual viewpoint of rhythmic grids and tuning systems [24], the data set of the zeroes is also immediately applicable to construct pitch systems.[8] A scale based on the spacing of the first 10 zeroes, if constrained proportionally within one octave and with initial gap from 0 preserved, would be mapped to the interval [0,1] as follows to 4 d.p.:

0, 0.284, 0.4224, 0.5025, 0.6113, 0.6617, 0.7551, 0.8221, 0.8705, 0.9645, 1

Adding 1 to all values would shift these to be ratios from unison 1 to standard octave 2.

Or as cents (value*1200), to 2 d.p. given how a single cent is well under the just noticeable difference, in the absence of beating phenomena from simultaneous presentation:

0, 340.77, 506.82, 602.99, 733.51, 794.03, 906.17, 986.51, 1044.57, 1157.36, 1200

This reveals a proximity to 12TET 4th, tritone, flattened sixth, and sixth. If quantised to bare 12TET MIDI notes the mapping is non-injective:

60, 63, 65, 66, 67, 68, 69, 70, 70, 72, 72

Of course, any number of zeroes can be taken, and the squashing of zeroes together with increasing t will lead to the top part of the scale having increasingly many microtones relative to the initial steps. A scale can be constructed with respect to any enclosing ratio (in the manner of the Bohlen-Pierce 'tritave' of a ratio of 3 or arbitrary ratio r [25]). Scales can also be devised on the primes or prime powers: Though tuning systems are often built using prime powers (for example, 3-limit Pythagorean tuning constructed by rationals of powers of 2 and 3), a tuning system literally lifted from the prime number (or prime powers) spacing is a rarer beast, Roger Dean providing one counter-example by constructing scales based on prime harmonics of a fundamental [26].[9] Given the dual location equations between the primes and the zeroes at $n \log n$ and $\frac{n}{\log n}$ respectively, one attractive potential musical resource is an alternation of expanding and contracting spacings following these formulae.

The signals discussed earlier in the paper, for instance, the approximate prime and Riemann spectra or the eta function rendered over time for changing s, can themselves be the trigger for discrete materials, by the use of such techniques as peak picking and onset detection reacting to extrema in signal or derivative rather than zero crossings. Such discrete sets of values, or the original zeroes, might also be scaled and rounded off to become indices into any set of musical objects, such that aside from the spacing of events and materials for pitch systems, the sequence of positions could control arbitrary parameters in sound synthesis and algorithmic composition.

[5] $\frac{2\pi(n+1)}{\log(n+1)} - \frac{2\pi n}{\log(n)} \approx \frac{2\pi(n+1)}{\log(n)} - \frac{2\pi n}{\log(n)} = \frac{2\pi}{\log n}$

[6] Rosser's theorem states $p_n > \frac{n}{\log(n)}$; the actual prime counting function is a result of Riemann including an effect of the zeroes of the zeta function.

[7] To a scale factor of $\log(10)$, the number of decimal digits in a number is $\log_{10}(n) = \frac{\log(n)}{\log(10)}$

[8] Berry acknowledges that arbitrary enclosing intervals can be taken, but restricts himself in examples to the piano keyboard [4]

[9] A fascinating unpublished manuscript by Peter Buch available online [27] uses the Riemann zeta function as a way to find low integer steps per octave that best approximate pure just intonation ratios; the correspondence that arises with often mentioned steps per octave in the tuning literature (7, 12, 19 et al.) is impressive.

5. CONCLUSIONS

This paper has explored some further mapping possibilities to sound for equations associated with the Riemann zeta function. We have embraced some noisier possibilities and not assumed 12 note equal temperament or any other discrete system is the final aim in rendering to musical sound, though examples have included applications in rhythms and pitch scales. In a number of places we have observed some perceptual limits on the use of such mathematics, and an observer is highly unlikely to recover deep mathematical knowledge of the zeta function or primes from the sonifications. Aesthetic choices in mapping delimit the scientific result [28, 29], and we tend more to the aesthetic potential here.

There remains a huge amount of fascinating mathematics to explore for novel musical mappings. Accompanying the core Riemann Hypothesis are a host of mathematical equivalents, including statements about such mathematical objects as Farey sequences of rationals and permutation groups, of potential applicability in artistic sonification [30]. Indeed, number theory contains many more special kinds of number and number theoretic functions of potential interest to composers.

6. REFERENCES

[1] H. M. Edwards, *Riemann's zeta function*. Dover Publications, Inc., 1974.

[2] H. Iwaniec, *Lectures on the Riemann zeta function*. American Mathematical Society, 2014, vol. 62.

[3] B. Mazur and W. Stein, *Prime numbers and the Riemann hypothesis*. Cambridge University Press, 2016.

[4] M. Berry, "Hearing the music of the primes: auditory complementarity and the siren song of zeta," *Journal of Physics A: Mathematical and Theoretical*, vol. 45, no. 38, p. 382001, 2012.

[5] I. Xenakis, *Formalized Music*. Stuyvesant, NY: Pendragon Press, 1992.

[6] J. Harley, *Xenakis: His Life in Music*. New York, NY: Routledge, 2004.

[7] G. Loy, *Musimathics, Volume 1*. Cambridge, MA: MIT Press, 2007.

[8] P. Doornbusch, "Composers' views on mapping in algorithmic composition," *Organised Sound*, vol. 7, no. 2, pp. 145–156, 2002.

[9] A. McLean and R. T. Dean, *The Oxford Handbook of Algorithmic Music*. New York, NY: Oxford University Press, 2018.

[10] C. Beans, "Science and culture: Musicians join scientists to explore data through sound," *Proceedings of the National Academy of Sciences*, vol. 114, no. 18, pp. 4563–4565, 2017.

[11] P. Vickers and B. Hogg, "Sonification abstraite/sonification concrète: An 'aesthetic perspective space' for classifying auditory displays in the ars musica domain," in *Proceedings of the International Conference on Auditory Display*, London, 2006.

[12] N. Collins, "Errant sound synthesis," in *Proceedings of the International Computer Music Conference (ICMC)*, Belfast, August 2008.

[13] C. Roads, *Composing Electronic Music: A New Aesthetic*. New York, NY: Oxford University Press, 2015.

[14] P. Borwein, "An efficient algorithm for the Riemann zeta function," in *Canadian Mathematical Society Conference Proceedings*, vol. 27, 2000, pp. 29–34.

[15] K. Fischer, "The zetafast algorithm for computing zeta functions," *arXiv preprint arXiv:1703.01414*, 2017.

[16] S. Wilson, D. Cottle, and N. Collins, *The SuperCollider Book*. The MIT Press, 2011.

[17] T. Bovermann, J. Rohrhuber, and A. de Campo, "Laboratory methods for experimental sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin: Logos Verlag, 2011, pp. 237–272.

[18] J. M. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *Journal of the audio engineering society*, vol. 21, no. 7, pp. 526–534, 1973.

[19] M. L. Brun, "A derivation of the spectrum of fm with a complex modulating wave," *Computer Music Journal*, vol. 1, no. 4, pp. 51–52, 1977. [Online]. Available: http://www.jstor.org/stable/40731301

[20] A. Odlyzko, "The 10^{22}nd zero of the Riemann zeta function," *Dynamical, Spectral, and Arithmetic Zeta Functions: AMS Special Session on Dynamical, Spectral, and Arithmetic Zeta Functions, January 15-16, 1999, San Antonio, Texas*, vol. 290, p. 139, 2001.

[21] A. M. Odlyzko, "On the distribution of spacings between zeros of the zeta function," *Mathematics of Computation*, vol. 48, no. 177, pp. 273–308, 1987.

[22] V. Kargin *et al.*, "Statistical properties of zeta functions' zeros," *Probability Surveys*, vol. 11, pp. 121–160, 2014.

[23] M. Berry and P. Shukla, "Hearing random matrices and random waves," *New Journal of Physics*, vol. 15, no. 1, p. 013026, 2013.

[24] R. D. Morris, *Composition with pitch-classes: a theory of compositional design*. Yale University Press, 1987.

[25] M. V. Mathews and J. R. Pierce, "The bohlen-pierce scale," in *Current directions in computer music research*, M. V. Mathews and J. R. Pierce, Eds., Cambridge, MA, 1989, pp. 165–173.

[26] R. T. Dean, "Widening unequal tempered microtonal pitch space for metaphoric and cognitive purposes with new prime number scales," *Leonardo*, vol. 42, no. 1, pp. 94–95, 2009.

[27] P. Buch, "Favored cardinalities of scales," *unpublished*, 2005. [Online]. Available: https://www.researchgate.net/publication/265112276_FAVORED_CARDINALITIES_OF_SCALES

[28] F. Grond and T. Hermann, "Aesthetic strategies in sonification," *AI & society*, vol. 27, no. 2, pp. 213–222, 2012.

[29] C. Scaletti, "Sonification≠ music," in *The Oxford Handbook of Algorithmic Music*, A. McLean and R. Dean, Eds. New York, NY: Oxford University Press, 2018, pp. 363–386.

[30] K. Broughan, *Equivalents of the Riemann Hypothesis: Volume 1, Arithmetic Equivalents*. Cambridge University Press, 2017.

THE DESIGN AND EXPLORATION OF AUDITORY DISPLAY EFFECTS FOR BLIND DRIVERS IN AUTONOMOUS VEHICLES

David Dewhurst

www.HFVE.org

david.dewhurst@HFVE.org

Hyundai Motor Company Design Challenge :

"Auditory User eXperience Design for Autonomous Vehicles and Future Mobility"

ABSTRACT

This work forms a part of a wider project, in which the author is developing a system to present visual images, and other material, via sets of auditory (and tactile) display effects. The main contribution of this paper is to describe the design, and examine the effectiveness, of these effects in an automotive context, specifically in the context of blind drivers travelling in autonomous (self-driving) vehicles.

This paper also brings together and summarizes auditory display effects and techniques that have previously been reported by the author, and describes several new features. The effects are termed tracers; polytracers; drone and matrix effects; imprints; and multi-level multi-talker "focus" effects.

The paper describes the potential automotive application of such auditory display effects in:- command and control; route presentation; maps/cartography; and enhancing the journey experience of blind travelers.

Methods of presenting rectangular areas within a scene, (termed "audio previews") are described and discussed, as is the concept of a small set of effects termed a "glimpse".

The results of informal assessment sessions with a totally blind person, and two sighted people, are described.

1. INTRODUCTION

One source estimates that there are about 39 million blind people in the world; and another estimates that there are nearly 253 million people who are blind or visually impaired [1], [2]. Several attempts have previously been made to present aspects of vision to blind people via other senses, particularly hearing and touch. The approach is termed "sensory substitution" or "vision substitution".

The coming introduction of autonomous vehicles presents many opportunities for blind people. One estimate foresees fully autonomous cars accounting for up to 15 percent of passenger vehicles sold worldwide in 2030 [3].

This paper focuses on the use of auditory displays as part of the user experience of autonomous cars, mainly for blind travelers, but with application to sighted users whose visual attention is elsewhere.

1.1. Other previous work

There is often the need to convey general visual information to blind people. An existing approach is to use relief images e.g. tactile maps. While these are convenient for conveying unchanging two-dimensional images, the instantaneous production of vision substitution images is more difficult to achieve. Devices can be devised that present other senses with information that includes aspects of sight, but other senses are not as powerful, or as able to comprehend such information [4].

Work in the field dates back to Fournier d'Albe's 1914 Reading Optophone [5], which presented the shapes of characters by scanning across lines of type with a column of five spots of light, each spot controlling the volume of a different musical note, producing characteristic sequences of notes for each letter of the alphabet Fig. 1.

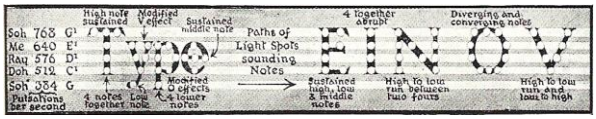


Figure 1: Optophone scanning across a line of type.

Other systems have been invented which use similar conventions to present images and image features [6], [7], or to sonify the lines on a 2-dimensional line graph [8]. Typically height is mapped to pitch, brightness to volume (either dark- or light- sounding), with a left-to-right column scan normally used. Horizontal lines produce a constant pitch, vertical lines produce a short blast of many frequencies, and the pitch of the sounds representing a sloping line will change frequency at a rate that indicates the angle of slope.

Previous work in the field is summarized in [9], [10]. Previous approaches have allowed users to actively explore an image, using both audio and tactile methods [11], [12]. BATS (Blind Audio Tactile Mapping System) presents maps via speech synthesis, auditory icons, and tactile feedback [13]. The GATE (Graphics Accessible To Everyone) project allows blind users to explore pictures via a grid approach, with verbal and nonverbal sound feedback provided for both high-level items (e.g. objects) and low-level visual information (e.g. colors) [14], [15]. An approach used by the US Navy for attending to two or more voices is to accelerate each voice, and then present them in sequence [16].

The Discrete REconfigurable Aural Matrix (DREAM) is a multi-speaker array technology, and [23] describes an approach to presenting multiple geometric shapes, including vertex highlighting; and producing "aural paintings".

Google's Lookout software allows blind users to identify information about their surroundings [2]. Microsoft's Seeing AI software allows users to touch an image on a touch-screen to hear a description of objects within an image and the spatial relationship between them [17].

(The merits of these other approaches are not discussed further in this paper.)

1.2. The HFVE system

The author's HFVE (Heard & Felt Vision Effects) system attempts to present aspects of visual images to blind people,

https://doi.org/10.21785/icad2019.012

via a rich set of audio and tactile effects, conveying images as a series of items, with the user controlling what is presented.

A major feature of the system is presenting modified speech i.e. spoken word sounds that are changed, multiplied, and moved, in order to intuitively convey the location, size, shape, and other properties, of the items they are presenting.

Another feature allows a blind person to navigate between levels of view within visual or non-visual representations, rising up levels for an overview, and drilling down levels for more detail, via, for example, a mouse wheel or dial device.

(Note that several of the features described in this paper have been reported previously [19], [20], [21], [22].)

2. SUMMARY OF AUDITORY DISPLAY EFFECTS, AND USER EXPERIENCE

In this section the auditory display effects produced by the HFVE system are summarized. Tactile equivalents will also be briefly described.

Most of the auditory display effects described below can be combined – for example imprints and tracer effects can present the same item simultaneously.

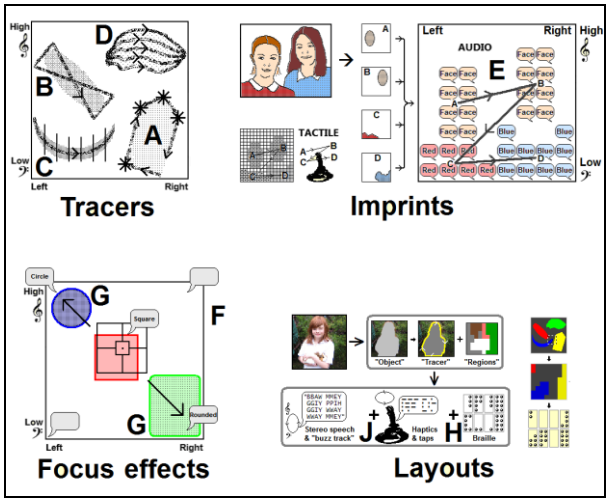


Figure 2: HFVE effect types (see text for details).

Particular item types, such as faces, text, persons, etc., can typically be presented using preferred per-item-type effect settings held by the system. For example faces and people could be presented as symbolic tracers, while blobs / areas of particular color could be presented via imprints.

The nature and aesthetics of the auditory display effects can be experienced by visiting the author's website, which includes demonstration videos [18].

The application of the effects to automotive use is described and discussed in Section 3 below.

2.1. Tracers

By smoothly changing the pitch and binaural positioning of particular sounds, speech and other sounds can be made to appear to move, whether following a systematic path, or describing a specific shape. Such moving effects are termed "tracers", and can be "shape-tracers" (A) Fig. 2, whose paths convey the shapes of items in an image [19].

In the tactile modality, tracer location and movement can be presented via a moving force-feedback device Fig. 5 that moves/pulls the user's hand and arm – in both modalities the path can describe the shape, size and location of the items.

As the system outputs both audio and tactile effects, users can choose which modality to use; or both modalities can be used simultaneously.

The system presents corners/vertices within shapes (A) Fig. 2, which are found to be very important in conveying the shape [19], [21]. Corners are highlighted via audiotactile effects that are included at appropriate points in the shape-conveying tracers, for example by momentarily stopping the tracer, or outputting a short distinct audio or tactile effect.

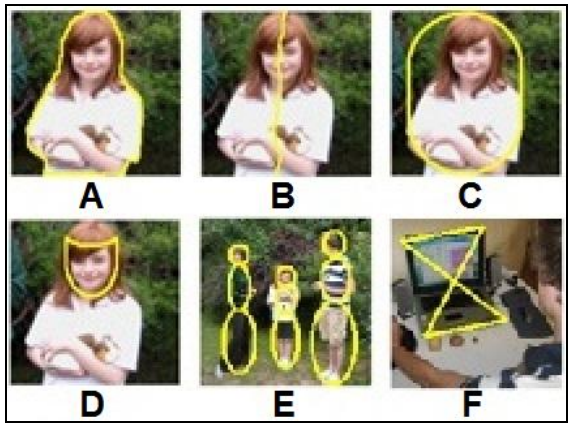


Figure 3: Shape, and symbolic, tracer paths (the yellow lines show the routes travelled by several types of tracer).

Although one possible tracer path for presenting an item's shape is the item's outline (A) Fig. 3, other paths such as medial lines (B), or frames (C), can be used. Symbolic paths e.g. for Face (D), Person (E), and Unknown (F) (and (B) Fig. 2) are found to be effective, as they present the location, size, orientation and type of object via a single tracer path.

A "drone" or "buzz" tracer, played at the same time as the speech tracer, can more clearly convey the size and shape, and present other information [20]. One effective such sound is a buzzing sound, but with a clearly defined pitch. (An additional non-speech tracer, of differing timbre, can convey distance information, if available, via pitch. Alternatively, the pitch of either the standard speech or the standard buzzing sound can convey distance information, with the other conveying height. A similar approach can be used for presenting distances for polytracers, imprints, etc., which are described below.)

"Matrix" effects (C) Fig. 2 are produced by dividing the image into several equal-width columns and/or rows, so that distinct effects can be triggered whenever the tracer moves from one such column or row to another, allowing the shape of lines to be perceived more clearly – if the tracer travels at a constant speed, the rate at which the effects are presented will correspond to the angle of slope [20].

A "Polytracer" (D) Fig. 2 uses additional tracers to present a single item. A polytracer can present non-speech tone sound tracers in a similar manner to existing optophone-like systems; or the extra tracers can also be speech, presenting the same speech sounds as the main tracer, but moving in soundspace so that their pitch and binaural location at any moment corresponds to the location of the image matter that they are representing [20].

The moving speech sounds that the voices present are each stretched or shortened as required in order that the re-pitched voices together present synchronized speech sounds.

2.2. Imprints

"Imprints" (E) Fig. 2 rapidly summarize the content of a scene via multiple stationary audio and tactile effects, using mappings similar to those used for tracer effects [21]. (In (E) Fig. 2 each speech bubble represents a voice/speech source.)

The system can step round the items of a scene, sequentially presenting imprints of the items (E) Fig. 2.

Audio imprint effects can be speech-like; or non-speech-like sounds (e.g. tone-like, musical, buzzing, humming, tapping, and/or bubbling/dynamic, sounds); or combinations of both. (The timbre can be mapped e.g. to color categories.)

Imprints produce a combined effect that may rapidly and intuitively convey the approximate extent of the item(s) being presented. Wide-ranging items produce a dispersed effect of a wide range of pitches and apparent stereophonic locations. Compact items produce a more constricted effect of fewer, or closer, voices, with a narrower pitch range.

When speech-like, the multiple voices, of different pitches and locations, but synchronized, give the impression of a group of people speaking in unison.

In an assessment session a totally blind person suggested that both speech and non-speech imprint effects should be available, and be user-controllable [21].

2.3. Multi-level multi-talker ("Focus") effects

Multi-level multi-talker effects (termed "Focus effects") (F) Fig. 2 allow several properties and items, at different levels of view, to be presented and perceived at the same time [22].

A blind user can rapidly navigate between such levels, e.g. by using a mouse wheel or a dial device, while hearing the focus effects speaking the level of view (e.g. spreadsheet cell, column, row, or block) that is currently emphasized, and at the same time being made aware of the levels above and below the current level of view, which have distinguishing effects applied (e.g. voice character, persona, etc.).

Focus effects can also be used to present property values of non-visual and non-spatial properties, for example levels of categorization and analysis, as found in academic and other fields. For example a car manual, or the Dewey Decimal system [30] could be presented and navigated round using focus effects, as described in section 3.1 below.

The system presents the items that are currently the primary focus of attention via crisp non-modified sounds, for example via speech sounds. At the same time the system presents the speech sounds for items that are not at the focus of attention, but applies a distinct differentiating effect on them, for example by changing the character of the speaker, or by applying echo or reverberation effects.

The system can artificially move the presented items (G) Fig. 2, so that the audio separation is maximized. This helps users to focus their auditory attention on the item emphasized by the system, or switch their attention to another item that is also presented but not emphasized. The user can then cause the system to highlight that other item instead.

The differentiated focus effects, for example echo and reverberation, can be applied to most of the other effect types, such as polytracers or imprints, so allowing such effects to have a faraway, hazy, unfocussed, quality analogous to the way that photographers use depth of field to accentuate focused items, with out-of-focus items also present which the observer is aware of but not directed towards.

In the visual domain, the system can produce higher-level consolidations of image content [22]. While HFVE knows how to consolidate general visual images, it does not know about other domains such as, for example, Excel spreadsheets. Instead such entities can be submitted to HFVE as client entities, for HFVE to present. For example consider the spreadsheet (A) Fig. 4. Although it could be presented as a visual-domain view i.e. as a series of patches of color and perhaps some text recognition, it is more meaningful to be able to inspect it via a spreadsheet-domain view (B), consolidating cells (Level 5) to columns and rows (Level 4), then to individual blocks (and objects such as charts and pictures) (Level 3), then to all blocks (and all objects) (Level 2), then to top level Spreadsheet (Level 1). (Level 0 gives a top-level overview of all available domain views.)

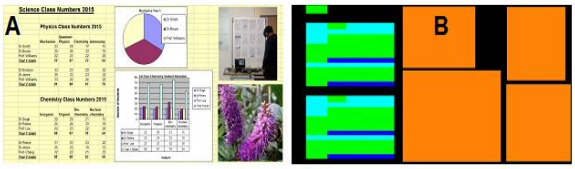


Figure 4: A spreadsheet and corresponding "ItemMap".

Such higher-level view groupings facilitate obtaining meaningful summaries/overviews of content, and help with navigating around the items of the image/entity.

In order to present externally-produced images and other entity types via HFVE, a straightforward interfacing method has been devised. This comprises submitting a standard 24-bit color bitmap (.bmp) file e.g. (B) Fig. 4 that includes all of the required basic item blobs (termed the "ItemMap" file); and a standard text (.txt) file (termed the "ItemKey" file) that describes how those blobs are marked via particular bit settings on the bitmap, and specifies how those basic item blobs are consolidated up to produce higher-level "group items". This pair of files, that fully describes the blobs of the image/entity, and how they are consolidated, can be created manually using a simple image painting application and a text editor, or can be created via an external application.

In the case of a spreadsheet Fig. 4, the ItemMap bitmap and ItemKey text file can be produced automatically via an Excel Add-In that has been developed. HFVE does not know about Excel, but processes the resultant pair of files like any other, getting item identifier bits from the ItemMap bitmap pixels, then looking up the corresponding item details (e.g. words to speak) from the ItemKey text file.

For certain entities some blobs may overlap (for example detected faces, and areas of color), and the system can use a number of bits in each pixel of the 24-bit bitmap for marking particular sets of non-overlapping blobs. Such content is resolved by the ItemKey text file, which specifies which bits are significant,-and their values for particular basic items.

Demonstration videos of the auditory display effects are available at the website: http://hfve.com. (Note that most of the effects that are described above in this section have been reported previously.)

2.4. Tactile display effects

Though not the primary concern of this paper, many of the auditory effects have corresponding tactile equivalents.

A force-feedback joystick makes an effective pointing device with which to indicate areas of the image, as it can

also be programmed to tend to position itself to one of a number of set positions, so that notch-like effects are felt as the joystick is moved, giving a tactile indication of location.

A force-feedback joystick can be programmed to allow free movement only within a restricted range, or along a lineal route. It can also be moved by the system, by it moving successive Spring condition effects, so pushing and pulling the user's hand and arm to trace out shapes (and highlight corners). Additionally it can be used to command the system via button and twist actions, and output tapping effects can present morse-like information to deafblind users (J) Fig 2.

Microsoft's Sidewinder Force Feedback joystick and Logitech's Wingman Force Feedback Mouse Fig. 5 are suitable devices, and can be controlled via Microsoft's DirectInput methods.

Figure 5: Microsoft's Sidewinder Force Feedback 2 joystick, and Logitech's Wingman Force Feedback Mouse.

Though both of these example force-feedback devices are relatively dated, bespoke new force-feedback devices could be developed for use in new automotive applications.

(Tactile braille effects (H) Fig. 2 and tactile tap effects (J) (termed "Layouts") can also be output to tactile devices by the system, as described in earlier papers [19], [20].)

The method of user interaction can be "exploring" in style, using a moving pointer to inspect a scene; or alternatively allowing the system to announce items, and the user then selecting one of the announced items for further inspection – the latter approach requiring less input from the user, and applicable to information navigation.

Furthermore, the user can tap commands onto a touch-screen or touchpad, and touch or drag over them to indicate parts of the image [22].

An optional pitched and panned buzzing sound can help to convey the location of the pointer within the image area.

2.5. Audio previews

To summarize, the project uses both tone and speech sounds, suitably modified, in order to convey visual information to blind people. This is exemplified in some very recent (incomplete) development work, which involves presenting the location, size, and aspect ratio, of a smaller rectangle within the larger square extent of the full presentation area.

It is intended that such "audio previews" Fig. 6 can optionally be used to quickly convey the location and extent (size and aspect ratio) of an item immediately prior to its details and exact shape etc. being presented via the other methods described, especially when only audio methods are available i.e. no tactile presentation. Standard height to pitch mapping, and left to right panning, is used, with stereo sound placement Fig 6. (Audio previews can be toggled on or off.)

The non-speech methods tested included using:- fixed height or sloping lead-in and trail-out tracer phases, to convey the top and base heights of the rectangle; different sound timbres for each phase; musical step changes with height change; 2-level oscillating pitches; click sounds between phases; a second sloping tracer to mirror the slope of the first; "L"-shaped representation of the rectangle; extra presentation areas to the left and right of the presentation square (for when the rectangle is located fully at the right or left edge); and variable speed, volume, and pitch range.

The audio properties such as high or low pitch start, oscillating pitch frequency, and other sound properties, can be mapped to rectangle (or other) properties.

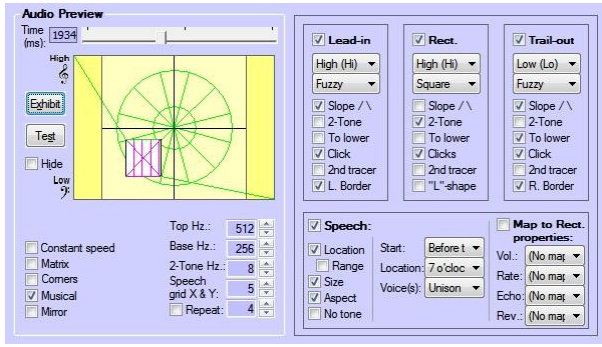


Figure 6: User interface for experimental "audio previews".

Alternatively or additionally, speech output can directly present the rectangle extent e.g. "Eight o'clock, square, medium", the speech being pitched and panned to correspond to the location of the rectangle. Optionally two voices can speak in unison or in sequence, being audio-located at opposite corners of the rectangle. The location can be presented via cartesian coordinates (e.g. chessboard-style "B3", or phonetic alphabet-style "Bravo 3", or "Top right" style); or via polar coordinates (clock positions). The volume, speech rate, optional echo and reverb effects, and other speech properties, can be mapped to particular properties.

The test rectangle can be drawn in a particular position with a mouse, or randomly positioned and dimensioned by the system, and can be hidden for test purposes. The sounds can be repeated a number of times.

See section 4 below for initial informal test results. Participants generally felt that speech was easier to use and gave immediate information, but some thought that they may be able to more quickly and intuitively interpret the tone sounds with further practice.

2.6. Glimpses

In developing the system, it was found to be effective to apply the concept of a short list of typically 4 to 8 items that are presented in a burst of a few seconds, for example as imprints stepping through the list and presenting the approximate location and size of each item. These small sets of items are termed "glimpses" Fig. 7. The concept can be used to help indicate an appropriate number of items to present at any point, whether the user is exploring an image ad-hoc, or navigating around the items in the layers of items.

The author speculates that the effectiveness of the number of items, and the timing of such glimpses, may be

related to the neuroscientific and psychological concept of a "psychological present", wherein there is a window of about 2 to 3 seconds within which your brain fuses what you are experiencing [24]. It may also relate to the well-known effect that only about 6 to 8 unrelated "chunks" of information can be comfortably handled in people's short term memory [25].

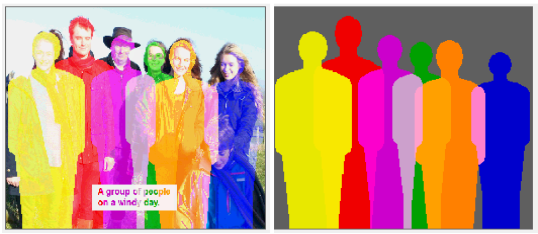


Figure 7: A "glimpse" – a set of a few "items", that are presented over an approximately 2 to 3 second period.

Furthermore, from a system development point of view such a concept leads to an effective system interface – it allows a clear separation of, and interfacing between, the item selection and navigation processing; and presenting the auditory (and tactile) effects representing the items.

Another advantage of this approach is that it provides a scalable design and allows effects to be distributed across several instances of the application, and in theory across several machines (virtual or actual). For example the system may produce an impression analogous to that of "covert attention" in vision – several instances can each present the content of separate locations i.e. the user can be simultaneously presented with data about several locations, whereby the effect known as covert attention is simulated.

3. AUDITORY DISPLAY EFFECTS FOR BLIND TRAVELLERS IN AUTONOMOUS VEHICLES

In this section the potential applications of the described auditory display effects for blind travelers in autonomous vehicles are considered and discussed.

It is assumed that any such vehicle will correspond to Level 4 or Level 5 of the SAE's automation level definitions i.e. requiring no driver attention [26].

The following application areas will be considered:- command and control; route presentation; maps/cartography; and enhancing the journey experience of blind travelers.

Automatic list to bitmap production; route presentation; and maps; will be demonstrated at the ICAD conference.

3.1. Command and control

A blind person in charge of an autonomous vehicle will often need to both give instructions to the vehicle; and receive information and feedback from the vehicle.

One way of giving and receiving such information is to use pseudo-visual representations of hierarchical multi-level structures such as menu structures, lists, etc. These are termed "ListMaps", and can be thought of as 3D explorable entities that can be automatically created from text lists.

The example of a simple car user guide Fig. 8 shows a simple text file that is automatically converted by the system into an ItemMap bitmap, and an ItemKey describing the basic items, and how they are consolidated up to group items.

This is performed by initially totaling up the content of whatever quantity is to be expressed by the area shown, for each group item – the quantity can simply be the number of basic items. Then, starting at the highest level, the image area is split into rectangular areas, each sized according to e.g. the basic item count for the group items at the highest level. Then, each such rectangular area is split further according to the next level content, until a pattern of similar-area small rectangles representing the basic items is produced, grouped according to their higher-level classifications Figs. 8 & 9.

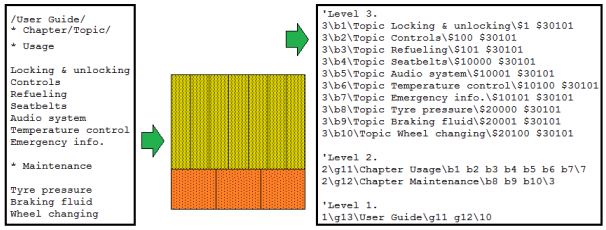


Figure 8: A simplified car user guide text file section listing, and the corresponding ItemMap and ItemKey.

Another example in an automotive context is using a similar approach to present the legs of a journey, wherein the route is structured as a set of legs, and each leg is sized to match either its distance, or estimated duration, so giving an impression of the relative distance or duration of each leg.

Complex non-visual multi-level/structured entities may also be presented as pseudo-visual/spatial representations. An example would be a full maintenance manual, that might have a complex hierarchy structure. A non-automotive example of such a structure is the Dewey Decimal classification system [30] Fig. 9. The levels might be Level 2 Class (e.g. 500 / Science & Maths) – Level 3 Division (e.g. 510/Maths) – Level 4 Section (e.g. 516 / Geometry) – Level 5 Sub-section (e.g. 516.3 / Analytic Geometry) (with Level 1 giving the entity/domain view name). The lowest level items i.e. Sub-sections can be automatically marked on a bitmap as block patterns of small rectangles (A) Fig. 9, each of a unique color shade, which can then be consolidated up through the levels to the higher-level group items (B). Then when presented as audio (and tactile) effects, the user can obtain an impression of the size and distribution of the items at each level of the entity, and navigate around them.

The basic items in the bitmap, and resultant group items, can then be presented via any of the auditory display effect types described above, with the user controlling the level of view, and then either receiving the information at any level via automatic stepping, with the user locking on the item (e.g. a chapter, in the Fig. 8 simple handbook example) that they wish to explore further when it is presented; or directly exploring the items at any level – in use, the user can freely move the pointer to find a higher level group item, lock on it, and then explore the lower level items within that item.

In this way a spatial / dimensional impression of a visual or non-visual structure can be produced.

(When moving the pointer, one option is for a different voice to start when a new item is to be announced (optionally with a different persona), but with the earlier voice continuing on at a reduced volume level, being reduced further with each subsequent item (the previous voices can also be moved to the side to keep them distinct from the new main voice). This may produce a less abrupt effect on change of item, with the previous voices gradually fading away.)

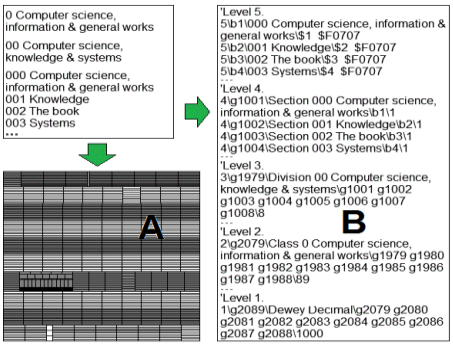


Figure 9: (Part of) the Dewey Decimal classification levels, and the corresponding ItemMap and (part of) the ItemKey.

In an automotive context, the contents of handbooks and troubleshooting guides, and journey details (as described above); as well as a car's technical settings and control and gauge values etc.; could be presented in a similar manner, and this facility could also be used by sighted people.

3.2. Route presentation

It is assumed that any autonomous vehicle will have the planned route available e.g. from an Internet mapping service [31]. The route could be presented to a blind traveler as a tracer, whether auditory, or auditory and tactile. If a force-feedback device Fig. 5 is available, then it can be locked to the route (B) Fig. 10 within the mapped area, analogous to moving a pencil tip along a shaped groove. The force-feedback device can either be moved by the system to show the route, or can be free to move, but locked to the route, so that the user can move it back and forth along the route to feel its shape, while the system announces points along the route corresponding to the current location of the device.

If a touch-screen or other touchpad is available, the user can control the position along the route by dragging back and forth across the screen of the device (or a mouse can be used).

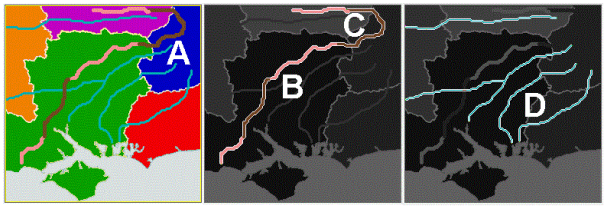


Figure 10: ItemMap with overlapping items (counties, route legs, and roads); locked on route; and locked on other roads.

The whole route (B) Fig. 10, the fraction of the route travelled so far, and/or the fraction remaining, can be presented to a blind traveler via a tracer, so giving them an auditory (and tactile) impression of each fraction of the journey. The timbre of the tracer can change between when presenting the fraction covered, and the fraction remaining. The speed of the tracer can be constant i.e. presenting the distances involved; or related to the expected speed of travel, so presenting the estimated timings. (This presentation approach could also be useful for sighted people.)

Route presentation can provide the following features:-

- Presentation of features for the point in the journey being presented, including : road name, road number, town/village name, speed limit, predicted traffic, road works, landmarks, nearby hotels, restaurants, tourist attractions, landscape (urban, forest, moorland, etc.), fuel/recharging points, etc. These can use the focus effect and selection methods described above. Higher level geographical/political regions such as city, county, or country can also be presented.

- The system can use a bitmap marked with features for possible presentation Fig. 10, in a similar manner to that described for visual and other domains in section 2.3 above. Alternatively the system can note the current coordinates of the location along the route and look up the features from other sources on the fly.

- If the route doubles back on itself (C) Fig. 10, this can be difficult for a blind person to be aware of if they are pushing a force-feedback device along the route. The system can highlight the issue via audio or tactile means. Additionally the system can automatically drive the force-feedback device though the tight corner, then return control to the user, or the user can instruct the system to do this.

- If a force-feedback device is being used then the device can display damper or friction conditions (i.e. be made harder or easier to move) depending on the speed limit and expected traffic conditions along the route. This gives an intuitive impression of likely speed of progress along the route. Corresponding audio effects can also be presented.

- Alternative routes can be presented.

3.3. Maps / Cartography

The HFVE system clearly has application in presenting maps to blind people. The maps can be geographical, or can be political, for example structured as levels showing State – Country – Region – County – Town etc. (A) Fig 10, allowing the user to explore using the methods already described, and to obtain the shape and extent of any such area.

Alternatively a user can do a simple search of any named area, and, for example, get an impression of its distance from the current location, relative size, etc. via e.g. tracer and imprint effects.

Several places could be presented simultaneously, or stepped round, so giving an auditory impression of their distance separations, and relative sizes.

Many similar cartography applications can be devised.

3.4. Enhancing the journey experience of blind travelers

As well as presenting practical information related to the journey, the system can be used to present many auditory display effects related to other aspects of travel (as well as allowing the blind traveler to access non-travel-related media, and such things as spreadsheets and structured data as described elsewhere – such uses will not be discussed here).

The system could be used to allow the blind traveler to be more aware of their surroundings along their journey. For example the system can use standard AI-related methods to present information about signage, people, etc. along the route travelled, by using text recognition and face detection respectively. For the demonstration system, open source Tesseract OCR is used for text recognition, and open source OpenCV is used for face, and motion, detection, though cloud-based services could also be used [27], [28].

Person detection is less straightforward to achieve in arbitrary situations, but the demonstration system can optionally use the simple approach of assuming that any face has a person's body below it, and that the size and distance of

the person is related to the size of the detected face, with nearer persons overlapping further-away persons. (If similar-sized faces are detected close together, then the higher-located faces will typically be for persons further away than the lower-located faces. Adjustments can also be made for age, gender, etc.)

Fig. 11 shows this process in action, with 5 faces detected, and with the resultant assumed figures overlapped appropriately (note that no statistical testing has yet been done to assess the applicability of the assumptions of this approach). The resultant figures can then be presented using the effects, for example as imprints, or symbolic tracers.

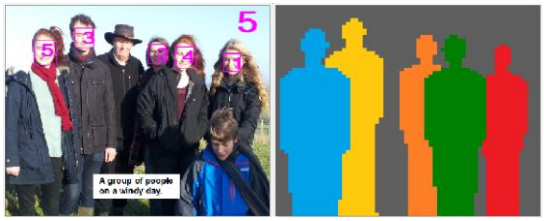


Figure 11: Human figures assumed from detected faces.

Other robust object identification methods can be used where available. The system could also make use of the large amount of Cloud-based information (e.g. OpenStreetMap etc. [31]) that is available concerning fixed landmarks etc. that may be encountered along the route of a journey, and these could be presented to the user in a similar manner.

Locked-on items

At any moment the user can lock on the item being presented (e.g. roads (D) Fig. 10). When an item is locked on, and the user moves the pointer within the area of the item, typically the items at lower- (and/or higher-) levels than the locked item can also be presented, so that the user can be aware of items in adjacent levels (or items nearby on the same level), and can switch to being locked on one of them instead. Alternatively the system can step around the lower-level items within the locked-on higher-level item, and the user can at any time lock on the item being presented.

Once an item is locked on, the subsequent interaction depends to some extent on the equipment being used to access the entity:-

Force-feedback : If a force-feedback mouse or joystick Fig. 5 is being used, the system can restrict the free movement to the area(s) of the current item/route – when pushed by the user away from the item, a spring force will attempt to push the mouse or joystick handle back to the centre or nearest part of the selected item (or to the point at which they left the item). When within the area of the item, the mouse or joystick handle will be loose/floppy and can be moved freely. The user can explore around the edge of the item with the force-feedback device, and get audio feedback at the same time.

If the item is multi-blob, e.g. a group item such as “Roads” (D) Fig. 10 or a fragmented basic item, then the user can command a jump to the next blob, and then explore that shape and content. Alternatively, with a force-feedback device the user can simply push the handle around the image and it will tend to snap to the nearest applicable blob.

Mouse : If a standard computer mouse is being used, an audio cue can signify and warn that the user has attempted to leave the area of the item/route. However the cursor pointer can be locked at the edge of the item (e.g. via a Windows SetCursorPos action), so that the user does not need to find the item again and can simply move their mouse back in the opposite direction.

Touch : If a touch-screen, or an absolute mode touchpad, is being used, then the system cannot easily restrict the physical movement of the user’s finger, so needs to directly tell the user or give non-speech cues to indicate how to move back to the locked item/route area. However users will typically be better able to recall the approximate location of the item (e.g. route) within the physical fixed area of the touch-screen or touchpad, than when using a standard relative mode mouse.

The system could use a virtual reality 360-degree camera or similar to gather images containing the distributed items that surround the blind (or sighted) traveler's vehicle, and corresponding effects then located in 3D soundspace.

Online facilities exist to provide words summarizing the content of images, so providing a top-level summary term for visual images [29].

4. ASSESSMENTS

Informal assessments of both the new (incomplete) audio preview feature, and the application of features for automotive use, were conducted with one blind participant “AB” (not his real initials), who has been totally blind since birth, and two sighted participants (“CD” & “EF”).

AB, who is very familiar with TTS speech synthesis, felt that the monotone voice (that accurately conveys height through pitch mapping) was sometimes hard to comprehend, and suggested that it should be easy to rapidly switch to one of the standard PC voices (which are more prosodic, but of less clear pitch level). This applied to both the speech used for audio preview, and for the automotive applications.

Concerning automotive applications, AB was interested in being able to easily produce a navigable hierarchical structure from a simple list, for several possible applications. He was very interested in the route presentation facility, and thought it could have application beyond automotive use.

Regarding audio preview, AB felt that while speech gave a relatively clear description, the tone sounds gave more of an impression of the presented rectangle.

CD (sighted) thought that using timbre to distinguish audio preview tracer phases was effective, and thought it should also be used to distinguish the two legs presenting the "L"-shape representation of the rectangle. She preferred the clock-face format to the location coordinates format.

CD felt that if too many special properties where applied to the non-speech sounds then they could sound confusing.

She was generally positive about the automotive applications, and liked the feature for handling sharp turns in locked routes (see section 3.2 above).

EF (sighted) was positive about the mapping and route presentation applications.

Regarding audio preview, she initially preferred speech to tone presentation, but didn't like using the phonetic alphabet (“Alfa”, “Bravo” etc.) when hearing coordinates, preferring the chessboard-style (e.g. "B2") terminology.

Overall the participants liked using polar coordinates (clock-face directions are easy to rapidly interpret).

Another approach discussed was presenting location by playing a circular tracer from the 12 o'clock position round to the location of the rectangle, then presenting the rectangle.

Other features discussed included controlling the volume of each phase; and presenting location by successively saying two numbers in the range 1 to 9, the first number representing the position of the target location within one of 3x3 squares numbered as the keys on a typical telephone handset, and with the second number representing a smaller square within the area of the first square, in a similar manner.

Participants generally felt that speech was easier to use and gave immediate information, but some thought that they may be able to more quickly and intuitively interpret the tone sounds with further practice. One thought that additional tone sounds might be helpful for exact rectangle positioning.

(All of these points require further investigation.)

5. CONCLUSIONS AND FUTURE WORK

In this paper possible applications of the HFVE system to automotive vehicles have been described and considered, with particular emphasis on applications for blind travelers.

Assessment sessions with a totally blind participant, and two sighted participants, are reported above.

Future work should include detailed evaluations, with an examination of specific tasks and approaches, detailed statistical analysis of results, and a qualitative analysis of post task interview data.

The system will be demonstrated at ICAD 2019.

6. REFERENCES

[1] World Health Organization, "Visual impairment and blindness" Fact Sheet No. 282, Updated October 2013, http://www.who.int/mediacentre/factsheets/fs282/en/.

[2] *Google Lookout*, https://blog.google/outreach-initiatives/accessibility/lookout-discover-your-surroundings-help-ai/.

[3] McKinsey, "Rethinking car software and electronics architecture", 2016. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/rethinking-car-software-and-electronics-architecture.

[4] Zahl, P.A. (Ed.) *Blindness : Modern approaches to the unseen environment.* Hafner Publishing, 1963.

[5] E. E. Fournier d'Albe, "On a Type-Reading Optophone" in *Proc. Royal Society of London. Series A*, vol. 90, no. 619 (Jul. 1, 1914), pp. 373-375.

[6] P.B.L. Meijer, "An Experimental System for Auditory Image Representations" in *IEEE Trans on Biomedical Engineering*, vol. 39, no. 2, pp. 112-121, 1992.

[7] U.S. Patent No. US 6,963,656 B1.

[8] D.L. Mansur, M.M. Blattner and K.I. Joy, ``Sound Graphs, A Numerical Data Analysis Method for the Blind," in *Journal of Medical Systems*, vol. 9, pp. 163-174, 1985.

[9] A. Edwards, "Auditory Display in Assistive Technology" in *The Sonification Handbook*, T. Hermann, A. Hunt, J.G. Neuhoff (Eds.), 2011.

[10] T. Pun et al., "Image and Video Processing for Visually Handicapped People" in *EURASIP Journal on Image and Video Processing*, vol. 2007, Article ID 25214, 2007.

[11] Roth P, Richoz D, Petrucci L, Pun T, "An audio-haptic tool for non-visual image representation" in *Proceedings of the 6th International Symposium on Signal Processing and its Applications* 2001 (Cat.No.01EX467) : 64-7.

[12] Patrick Roth, Thierry Pun, "Design and Evaluation of Multimodal System for the Non-visual Exploration of Digital Pictures". In *Proceedings of INTERACT 2003.*

[13] Parente, P. and G. Bishop, BATS: The Blind Audio Tactile Mapping System. ACMSE. Savannah, GA. March 2003.

[14] Kopeček, I and Ošlejšek, R, "GATE to Accessibility of Computer Graphics" in *Computers Helping People with Special Needs: 11th International Conference*, ICCHP 2008. Berlin: Springer-Verlag, pp. 295-302, 2008.

[15] Kopeček, I and Ošlejšek, R, "Hybrid Approach to Sonification of Color Images" in Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technologies. Los Alamitos: IEEE Computer Society, pp. 722-727, 2008.

[16] Derek Brock, Christina Wasylyshyn, and Brian McClimens, "Word spotting in a multichannel virtual auditory display at normal and accelerated rates of speech" in *Proc. of 22nd International Conference on Auditory Display(ICAD-2016)*,Canberra,Australia, 2016.

[17] *Seeing AI,* http://microsoft.com/en-us/seeing-ai.

[18] *The HFVE system*, http://hfve.com.

[19] D. Dewhurst, "Accessing Audiotactile Images with HFVE Silooet" in *Proc. Fourth Int. Workshop on Haptic and Audio Interaction Design*, Springer-Verlag, 2009.

[20] D. Dewhurst, "Creating and Accessing Audiotactile Images With "HFVE" Vision Substitution Software" in *Proc. of ISon 2010, 3rd Interactive Sonification Workshop*, KTH, Stockholm, Sweden, 2010.

[21] D. Dewhurst, "Using "Imprints" to Summarise Accessible Images" in *Proc. of ISon 2013, 4th Interactive Sonification Workshop*, Fraunhofer IIS, Erlangen, Germany, 2013.

[22] David Dewhurst and Tony Stockman, "The Design and Exploration of Interaction Techniques for the Presentation of Foreground and Background Items in Auditory Displays" in *Proc. of ISon 2016, 5th Interactive Sonification Workshop*, CITEC, Bielefeld University, Germany, 2016.

[23] Ivica Ico Bukvic, Denis Gracanin, Francis Quek, "Investigating Artistic Potential of the Dream Interface: the Aural Painting", in *International Computer Music Conference Proceedings*, Volume 2008, August 2008.

[24] Laura Spinney, "How long is now?" in *New Scientist*, vol. 225, no. 3003, pp. 28-31, 10th January 2015.

[25] G.A. Miller, "The magic number seven, plus or minus two: Some limits on our capacity for processing information" in *Psych. Review*, 63, pp. 81-93, 1956.

[26] *Self-driving car*, http://en.wikipedia.org/wiki/Self-driving_car.

[27] *Tesseract*, http://github.com/tesseract-ocr/tesseract/wiki.

[28] *OpenCV (Open Source Computer Vision)*, http://opencv.org

[29] *IBM Watson Visual Recognition service*, http://ibm.com/watson /developercloud/doc/visual-recognition.

[30] *Dewey Decimal Classification*, http://en.wikipedia.org/wiki/Dewey_Decimal_Classification.

[31] *Comparison of web map services*, http://en.wikipedia.org/wiki/Comparison_of_web_map_services

SONIFICATION WITH MUSIC FOR CYBERSECURITY SITUATIONAL AWARENESS

Courtney Falk

Infinite Machines
12948 Cantigny Way
Carmel, IN, USA
courtney.falk@infinite-machines.com

Josiah Dykstra

U.S. Department of Defense
Cybersecurity Operations
Ft. George G. Meade, MD, USA
JDykstra@LTSnet.net

ABSTRACT

Cyber defenders work in stressful, information-rich, and high-stakes environments. While other researchers have considered sonification for security operations centers (SOCs), the mappings of network events to sound parameters have produced aesthetically unpleasing results. This paper proposes a novel sonification process for transforming data about computer network traffic into music. The musical cues relate to notable network events in such a way as to minimize the amount of training time a human listener would need in order to make sense of the cues. We demonstrate our technique on a dataset of 708 million authentication events over nine continuous months from an enterprise network. We illustrate a volume-centric approach in relation to the amplitude of the input data, and also a volumetric approach mapping the input data signal into the number of notes played. The resulting music prioritizes aesthetics over bandwidth to balance performance with adoption.

1. INTRODUCTION

The SOC is the heart of cyber defense for many industry and government organizations. The tactical cyber operators and analysts who work in these environments monitor and respond to threats against their organization's mission sometimes 24 hours a day. Their work is high-value and complex, and the workforce suffers from fatigue, frustration, and high cognitive workload [1]. The SOC aims to maximize the productivity of analysts detecting and mitigating cyber events, while accounting for the human limitations of such work. Cyber defenders also work outside of SOC environments where some or most of their time may even be spent on non-security related tasks.

The economic value of effective cybersecurity can be extraordinarily high. Organizations that quantify their expected losses in terms of data breaches, productivity, or intellectual property report that they routinely lose millions of dollars. Therefore, mitigations and controls are justified by the value they bring in lowering such risk. Data-driven fiscal decisions justify a robust and comprehensive approach to cybersecurity.

Security professionals have access to a plethora of software tools and seemingly endless volume of data that can provide insights about the health and status of computer networks. The data are commonly in text and binary formats, and visualization tools can help the human analysts more easily consume and analyze the data. Unfortunately, security analysts cannot afford the mental demand to visually monitor these displays continually.

Signals of information from cybersecurity data are both discrete and continuous. Depending on the situation, a security analyst may discern the status of security by considering one or more stream of real-time events including intrusion detection alarms, user login events, and changes in network traffic volume. The work is unpredictable and dynamic.

The focus of our research is to gently aid cybersecurity situation awareness. Situational awareness takes many forms, including detection of anomalies occurring in a variety of data sources such as user logins and remote exploitation attempts. In this paper, we consider the application of music with low information density as sonification of cyber activity. We envision our sonification as background music for cybersecurity professionals or groups, including the SOC. Ideally, the music would be a pleasant and subtle experience for those unaware of its information value.

Many sonification implementations, including those for network traffic and other cyber security data, prioritize information over aesthetics. That is, they seek to very clearly convey meaning of the input data, and sometimes to maximize the amount of information conveyed in sound. The result is an information-rich application, but one ill-suited for pleasant and continuous listening.

Music may offer the ability to aid SOC analysts in an appealing and complimentary manner to their existing environment. We support and extend Vickers and Hogg's intuition that aesthetics facilitates ease of listening [2], and propose that it is possible to use music to convey information that sacrifices some information bandwidth for improved aesthetics. Instead of mapping every sound and audio feature to the data directly, this approach relies on broad musical characteristics to subtly inform the user. In time through this line of research, we aim to show that the use of subtle musical cues improves the speed and accuracy of analytical tasks and achieves greater satisfaction from users compared with other alternatives [3].

This paper makes the following contributions:

- We propose a method for mapping discrete and continuous signals of events to music, with attention to application with cybersecurity data.
- We offer a research prototype implementation of this approach and demonstrate it by sonifying 708M authentication events from nine months on a live enterprise network.
- We evaluate the information bandwidth of this technique,

https://doi.org/10.21785/icad2019.014

and offer a research agenda for continued experimentation, development, and operations.

2. BACKGROUND AND RELATED WORK

In this section, we introduce the concepts and related work in music and cybersecurity necessary to understand the subsequent work.

There is a stark division in scholarly research between sonification and music. As a communications medium, music has been tied to emotion. Film scores, for example, provide useful insights because they use music as a passive medium to introduce or reinforce information. However, film scores are most often used to communicate emotion, not concrete data. Music in film is consumed passively, as a secondary stimulus to the visual display and spoken dialogue.

Research shows that music convey meaning in data. In a 2018 survey of sonified weather data, researchers revealed that musical characteristics contribute to meaningful data perception, analysis and interpretation [4]. They also found increased engagement levels with melodies that included pitch, timbre and rhythm. The same study participants reported pronounced differences between perceived usability and aesthetics.

Music offers a canvas for communicating a wide variety of diverse data. In theory, music can be generated from any source of input data. Research by Davis and Mohammad produced a system ("TransPose") that would generate music from text taken from literature [5]. Natural language text is a special case of data because it is by its very nature unstructured. Unstructured data presents its own unique challenges that continue to be addressed by natural language processing to this data. The research in this paper focuses on structured data, specifically logs of network login events.

The SOC typically serves to monitor trends and triage security events, and to determine whether they should be escalated for in-depth analysis. As a result, analysts only require enough data and granularity to make those decisions. Data feeds come from many internal and external sources. Some feeds, such as intrusion detection systems, can continually generate more than 100,000 events each day [6]. Data sources vary from one SOC to another, and are impacted by organization size, sector, and budget. Common data sources of security-relevant information include:

- Network intrusion detection systems
- Host intrusion detection systems
- Network traffic logs (raw and summarized)
- Operating system activity and audit logs
- Server and network device logs such as web server, proxy, and DNS logs
- Security device logs, such as firewalls and anti-virus
- Malware analysis and sandbox analysis
- User and entity behavior

Researchers have suggested that detecting anomalies in network traffic "has potential value as an anomaly-detection approach include long-term, continuous listening to the sonification for real-time detection of deviations" [7]. Current approaches to sonification, such as those pursued by Axon et al., emphasize encoding as much data in the music signal as possible [8]. The benefits from such an approach are two-fold. First, the human analyst who hears the sonification receives all the data as original received. Second, there is not filtering or processing involved, which greatly simplifies the process of transforming raw data into music. Examples of this work can be heard online (`https://soundcloud.com/user-71482294`).

Sonification is a viable technique for both real-time and historical analysis. In 2005, Childs explored auditory display for monitoring real-time data. Like cyber defenders, financial traders monitor and act on data streams from text and visual displays. Using simple and sparse musically-based sonification, they reported that a commercial prototype program was "effective" using a two-note scheme [9].

Music can be an intuitive medium requiring little or no training about how network events are mapped. Research shows that both trained and untrained humans can perceive the common elements of music. These elements are pitch (which governs melody and harmony), rhythm (and its associated concepts tempo, meter, and articulation), dynamics, and the sonic qualities of timbre and texture. Further, most listeners have learned to associate musical cues with emotions. In Vivaldi's *Four Seasons*, the composer uses music cues that attempt to auralize the sights, sounds, and events of the natural seasons. Film soundtracks also use music to prompt or bolster a desired feeling along with the video. Soundtracks are static and pre-selected, synchronized with the video. We are unaware of any attempt to use soundtrack music other than for art and emotion. That is, soundtrack music is not being used to communicate data in music.

In this paper, we focus on sonification through instrumental classical music. We limit ourselves to instrumental output without vocals to avoid extraneous variable, complexity, and mental demand on the listener. We describe our output as "classical" in the tradition of western music with established principles [10]. Classical music is found to be aesthetically pleasing by the general population [11].

An important variable in the ability to use music for sonification is the rate of information transfer. We define *auditory cognitive bandwidth* as the channel capacity of information that a human listener can perceive and process from an audio stream. We measure this bandwidth as bits per second. Intuitively, active listening maximizes the auditory cognitive bandwidth compared to passive, background music. In 2009, Ramakrishnan proposed an application of information theory to sonification design that allowed quantification of information communicated by a sonification [12]. This work focused on maximum rates and did not differentiate between active and passive listening. We hypothesize, but have not yet explored, that channel capacity is reduced with passive listening. However, in an 18-subject study of simple tasks, Hildebrandt et al. revealed that using sonification to monitor a process as a secondary task had no significant effect on performance in either task [13].

3. METHOD

In this section, we introduce methods for encoding discrete and continuous signal from cybersecurity events into musical forms. Multiple different data signals, both continuous and discrete, can be encoded in such a way that they all fit in a coherent musical framework. Parseihian and Katz strive for a similar, generative and modular approach in their research to produce sonification cues that represent a physical environment [14]. A hypothetical example scenario is data center monitoring. The processing loads

of machine could provide a continuous input data signal. Events of when machines crash could provide a discrete input data signal. A system administrator could listen to the output music and quickly gain a general sense about the health of equipment in the server room.

3.1. Discrete Signals

Some cybersecurity data occurs as a discrete signal. Discrete data signals occur infrequently relative to other inputs. For example, network login events in a sufficiently large network occur frequently and regularly enough that they appear as a continuous signal. Login failures, however, are significantly less frequent, and the gaps between their signals make each event appear separate and discrete. These discrete signals should stand out from the other, competing data signals in the music and demand attention. Other examples of cybersecurity data that imitate a discrete data signal include anti-virus alerts and abrupt machine shutdowns, which may signal faulty hardware or malicious software.

We offer two approaches to encoding discrete signals into music: the intrusive noise, and harmonic tension. These two approaches are not mutually exclusive and could be combined or used for different signal simultaneously.

3.1.1. Intrusive noise

One approach to encoding a discrete event is an intrusive sound to represent the event. The sound could be either a musical or non-musical element. This noise may noticeably stand out from the ongoing music, making itself apparent. Take for instance the "BRAAAM" effect (aka the "BWONG") popularized in the movie, *Inception* [15]. The more distinctive and invasive the sound, the more likely that a listener will notice. However, overuse of the noise will cause fatigue and annoyance.

3.1.2. Harmonic tension

A second approach would be to take the music already being played at a given timestep and introduce harmonic tension. The idea is to create something that is still aesthetically pleasing while being noticeable. For instance, adding a minor seventh to a major chord produces a dominant seventh chord. In Western music, dominant sevenths create a sense of musical tension from the dominant key. The upside to this approach is that it is aesthetically pleasing and less fatigueinducing than a "BRAAAM." However, the downside is that it would not be as noticeable, and may require some amount of training on the part of the analyst in order to properly recognize its signal.

3.2. Continuous Signals

Some cybersecurity data occurs as a continuous signal. A continuous data signal is one that is present for most or all timesteps. Examples of cybersecurity data in this form include network traffic and web server logs. One way of encoding these signals is to create a continuous music stream and alter one parameter in relation to the data signal. Arpeggiating over chords is one technique to create such a musical stream.

We offer two approaches to encoding continuous signals into music: increasing and decreasing volume, and adding and removing voices. These two approaches could also be combined or used for different signal simultaneously.

3.2.1. Increasing and decreasing volume

As the data signal increases, increase the volume of the musical stream. Start at a pianissimo for the lowest level, going up to fortissimo for the highest level. This could be fatiguing for a signal that stays at high levels for long periods of time, causing the music to be loud over long periods of time.

3.2.2. Adding and removing voices

A second approach to creating a musical stream is to select multiple independent musical lines, or voices. Then quantize the data signal into equal levels. When the data signal is within the lowest level, only play that corresponding voice. As the data signal increases and crosses boundaries, add the other voices. Vice versa, the data signal decreasing removes voices. Adding and removing voices could be done in conjunction with the volume approach where the volume for each voice is set to correspond to where the data signal is in that particular quantized layer.

4. CASE STUDY

To illustrate the application of our approach, we consider a SOC task of monitoring authentication events for anomalies. We use the public network authentication dataset from the Los Alamos National Laboratory (LANL) [16]. This anonymized data set encompasses nine continuous months and represents 708,304,516 successful authentication events from real users to computers collected from the LANL enterprise network. The data are representative of other similar networks, and offer a non-critical indicator of activity on the network. The stream of login events produces a continuous data signal that varies according to the day of week and time of day, as seen clearly in Figure 1 where Monday is the first day of the week.

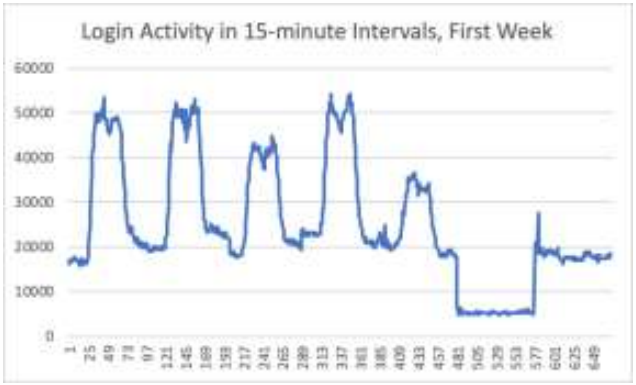


Figure 1: Graph of LANL login events ordered chronologically for the first week of time.

One drawback to the dataset is that it contains only successful login events. An analyst monitoring a computer network may want to monitor login traffic and investigate anything unsuccessful. To produce a more real-world scenario, we utilize a Poisson distribution to artificially insert unsuccessful login events.

The LANL data was quantized into 15-minute increments. This aggregation compresses the time required to analyze the data, and smooths the signal. Since each 15-minute increment is then encoded into a quarter note, the music is in common (4/4) time,

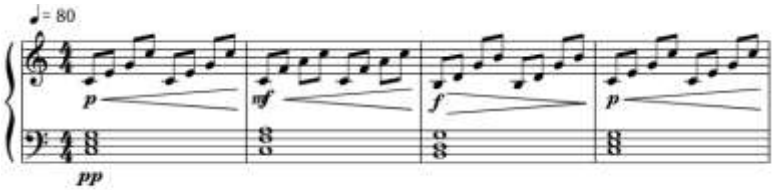


Figure 2: Sonification of a continuous data signal where differences in the input signal are encoded into the volume of the notes.

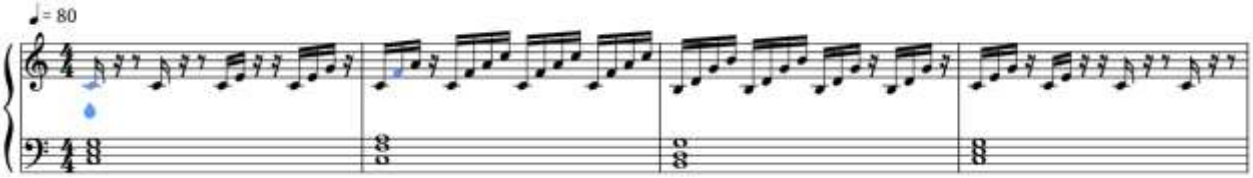


Figure 3: Sonification of a continuous data signal where the differences in the input signal are encoded into the number of sixteenth notes present in an arpeggio starting from the bottom.

each measure equates to one hour of time in the input data. Data about the user-computer pairing for each login event were abandoned, keeping only the information about the time when the login occurred. The goal was to sonify the number of logins in each 15-minute block into a musical structure that describes its state. Two techniques were applied: volume modulation, and changing the number of notes in a corresponding beat.

One of the driving goals of this research is to provide concrete artifacts that demonstrate the described techniques in action.[1] All source code is freely available under the GPLv3 license. The source code is written in Python 3 with all module dependencies being readily available via the Python Package Index (PyPI). MIDI is the generated as the output, which is playable via a wide variety of software audio players and synthesizers [17].

5. RESULTS AND DISCUSSION

This project experimented with two different novel sonification approaches. Both approaches utilized a common framework. An aesthetically-pleasing musical sequence was chosen to constrain the structure of the notes. There are numerous common chord progressions in Western music. This first system used the I-IV-V-I chord progression of a major key, which is extraordinarily common. Both approaches play the chord sequence as whole notes two octaves lower than the music from the data signal. This produces the carrier signal that tells a listener that the system is indeed working as expected even if there is a lack of data signal. Additionally, both approaches equate each 15-minute block of input data as one quarter note beat in the output music. The system outlined here has a few basic hyperparameters that could be changed to produce different results:

- Harmonic components:
  - Chords in the progression.
  - Musical key.
- Tempo components:
  - Length of each input time block.
  - Beats per minute.
  - Time signature.
- Timbre components:
  - Number of musical voices
  - Instruments/samples associated with particular voices.
- Genre/style [18].
- Length of quantization for the input data.

5.1. Loudness-centric Approach

The first sonification approach is to vary the volume of the output musical notes in direct relation to the amplitude of the input data. As the values of the input data increase, so too do the volume of the music it produces. A first pass through the data set provides the maximum signal level. Each step of the input data is then converted to a value between 20 and 127 that corresponds to its value relative to the maximum value. The values 20 and 127 were chosen because 127 is the loudest volume level for a MIDI signal and 20 is still barely audible. Figure 2 shows what the musical notation for such a volume-based encoding would look like.

The benefits of the volume-centric approach is that it guarantees a continuous, coherent musical stream. One drawback is that it may prove difficult for the human ear to distinguish subtle difference in the volume. So for a continuous data signal where minor changes are significant, this may not be the best approach to use.

5.2. Volumetric Approach

The second sonification approach encodes the input data signal into the number of sixteenth notes played in a given quarter node interval. Recall from earlier that each quarter node interval corresponds to a fiften minute block of time in the LANL login event data. Each sixteenth note is a step in an arpeggio much like the volume-based approach. But for the number-based approach, each fifteen minute interval of input data is quantized by four. The lowest 25% of the input signal would only cause the lowest note of the

[1] https://github.com/CalmLogarithm/sonification

arpeggio to be rendered while the highest 25% of the input signal would cause all four notes of the arpeggio to be rendered. So as the input data signal increases, the number of notes played increases, and due to the structure of the arpeggio, also goes higher (Figure 3).

There are two drawbacks to using the number-based approach. First, a signal that is mostly in the bottom quarter of the input range will create music that is mostly a sequence of sixteenth notes that occur on the beat. This creates a somewhat percussive effect. The second drawback is the coarse granularity because there are only four notes used in the arpeggio then the data can only be divided into four parts. There are workarounds that could improve this. One workaround is to stack multiple different voices on top of one another so that once the lowest voice has rendered all of its arpeggio then the next voice up begins rendering its arpeggio. A second workaround is to analyze the distribution of the input data signal and divide up the input value range in unequal blocks to make it more likely that more than one of each note in the arpeggio is playing, creating more of a sense of diversity in the output music.

5.3. Evaluation

We sought to maintain a low-bandwidth information channel. This implementation focused on a single input variable: network logins. At the peak, there were 125,008 login events encoded in a single 15-minute increment. That value requires a 17-bit value to encode. The music is 80 beats per minute, or 1.33 beats per second. As a result, this music communicates 22.66 bits per second. We have not yet conducted user testing for these prototypes. However, our initial assessment from listening to the sonifications is that they are both aesthetic and effective in communicating the authentication events. We invite the reader to listen to audio clips of this dataset.[2]

6. CONCLUSIONS AND FUTURE WORK

In this work, we proposed a novel sonification process for mapping data related to cyber defense into music. The resulting music prioritizes aesthetics to balance performance with adoption. The musical cues relate to notable events in such a way as to minimize the amount of training time a human listener would need in order to make sense of the cues. We demonstrated our technique on a dataset of 708M authentication events over nine continuous months on an enterprise network.

One limitation is that the software implementations linked to this paper only operates on a static data set. A fully-featured tool that is ready for enterprise use must also operate over a continuous data stream. Changes are required to the source code for it to function correctly with data of a dynamic and previously unknown range. However, we believe that the approach is applicable to any input data that can be quantized.

Future work should consider music generation via neutral nets [19]. Neural networks can be trained on a variety of musical styles and genres, creating dynamic yet coherent compositions to accompany the sonification techniques described in this paper. Due to how network data is of variable length, specific neural network architectures such as long short-term memory (LSTM) are the most applicable to this problem [20].

We intend to experiment with our approach in real SOC settings and to validate the utility and satisfaction of the music on their work. Axon et. Al have both developed a sonification approach and tested its usefulness in a realistic scenario [21]. Realistic tests should utilize network analysts as the pool of test subjects. If possible, network analysts both with and without backgrounds in music education should be used in order to test how easily the sonification techniques described in this paper are interpreted by people of varying degrees of musical skill.

We look forward to continued research, development, and operational use of this technique for improved security.

[2] https://soundcloud.com/user-679831789/sets/sonification

7. REFERENCES

[1] C. L. Paul and J. Dykstra, "Understanding Operator Fatigue, Frustration, and Cognitive Workload in Tactical Cybersecurity Operations," *J. Info. Warfare*, vol. 16, no. 2, pp. 1–11, 2017.

[2] P. Vickers and B. Hogg, "Sonification abstraite/sonification concrete: An 'aesthetic perspective space' for classifying auditory displays in the ars musica domain," in *Proc. of the 12th Int. Conf. on Auditory Display*, London, UK, 2006, pp. 210–216.

[3] S. Barrass and P. Vickers, "Sonification Design and Aesthetics," in T. Hermann, A. Hunt, and J. G. Neuhoff, *The Sonification Handbook*, pp. 145-171, Berlin: Logos Publishing House, 2011. Retrieved from https://sonification.de/handbook/download/TheSonificationHandbook-chapter7.pdf

[4] J. Middleton, J. Hakulinen, K. Tiitinen, J. Hella, T. Keskinen, P. Huuskonen, J. Linna, M. Turunen, M. Ziat, and R. Raisamo, "Sonification with Musical Characteristics: A Path Guided by User Engagement," in *24th Int. Conf. on Auditory Displays*, Houghton, MI, 2018.

[5] H. Davis and S. M. Mohammad, "Generating Music from Literature," 2014. Retrieved from https://arxiv.org/pdf/1403.2124.pdf.

[6] C. Zimmerman, "Ten Strategies of a World-Class Cybersecurity Operations Center," 2014. Retrieved from https://www.mitre.org/sites/default/files/publications/pr-13-1028-mitre-10-strategies-cyber-ops-center.pdf.

[7] B. N. Walker and M. A. Nees, "Theory of Sonification," in T. Hermann, A. Hunt, and J. G. Neuhoff, *The Sonification Handbook*, pp. 9–39, Berlin: Logos Publishing House, 2011. Retrieved from https://sonification.de/handbook/download/TheSonificationHandbook-chapter2.pdf

[8] L. Axon, J. R. Nurse, M. Goldsmith, and S. Creese, "A formalised approach to designing sonification systems for network-security monitoring," in *Int. J. on Advances in Security*, 2017, 10(1–2).

[9] E. Childs, "Auditory Graphs of Real-Time Data," in *11th Int. Conf. on Auditory Displays (ICAD)*, Limerick, Ireland. July 2005.

[10] "classical, n.2." OED Online, Oxford University Press, March 2019. Retrieved from www.oed.com/viewdictionaryentry/Entry/33881. Accessed 9 March 2019.

[11] D. K. Simonton, "Aesthetic success in classical music: A computer analysis of 1935 compositions," in *Empirical Studies of the Arts*, vol. 4, pp. 1–17, 1986.

[12] C. Ramakrishnan, "Sonification and information theory," in *Proc. of the 6th Int. Conf. on Auditory Display (ICAD)*, pp. 121–142, 2009.

[13] T. Hildebrandt, T. Hermann, and S. Rinderle-Ma, "Continuous sonification enhances adequacy of interactions in peripheral process monitoring," *Int. J. of Human-Computer Studies*, vol. 95, pp. 54–65, 2016.

[14] G. Parseihian and B. F. G. Katz, "Morphocons: A new sonification concept based on morphological earcons," *J. of the Audio Engineering Society*, vol 60, no. 6, pp. 409– 18, July 2012.

[15] K. Jagernauth, "Who Really Created The 'Inception' BRAAAM? Composer Mike Zarin Sets The Record Straight," 13 November 2013. Retrieved from IndieWire: https://www.indiewire.com/2013/11/whoreally-created-the-inception-braaam-composer-mikezarin-sets-the-record-straight-91690/

[16] A. D. Kent, *User-Computer Authentication Associations in Time*, Los Alamos Laboratory, 2014. doi:10.11578/1160076

[17] The MIDI Association, "The Complete MIDI 1.0 Detailed Specification," 1996. Retrieved from https://www.midi.org/specifications/item/the-midi-1-0-specification

[18] C. J. Carr and Z. Zukowski, "Generating Albums with SampleRNN to Imitate Metal, Rock, and Punk Bands," 2018. Retrieved from https://arxiv.org/abs/1811.06633

[19] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," 2018. Retrieved from https://arxiv.org/abs/1802.08435

[20] M. Kaliakatsos-Papakostas and A. Gkiokas, "Interactive Control of Explicit Musical Features in Generative LSTM-based Systems," in *Proc. of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ACM, September 2018.

[21] L. Axon, B. Alahmadi, J. Nurse, M. Goldsmith, and S. Creese, "Sonification in Security Operations Centres: What do Security Practitioners Think?" in *Proc. of the Workshop on Usable Security (USEC)*, February 2018.

EVALUATING THE MAGNITUDE ESTIMATION APPROACH FOR DESIGNING SONIFICATION MAPPING TOPOLOGIES

Jamie Ferguson

Glasgow Interactive Systems Sections
University of Glasgow
Glasgow, G12 8RZ, Scotland
j.ferguson.4@research.gla.ac.uk

Stephen Brewster

Glasgow Interactive Systems Sections
University of Glasgow
Glasgow, G12 8RZ, Scotland
stephen.brewster@glasgow.ac.uk

ABSTRACT

A challenge in sonification design is mapping data parameters onto acoustic parameters in a way that aligns with a listener's mental model of how a given data parameter should sound. Studies have used the psychophysical scaling method of magnitude estimation to systematically evaluate how participants perceive mappings between data and sound parameters - giving data on perceived polarity and scale of the relationship between the data and sound parameters. As of yet, there has been little research investigating whether data-to-sound mappings that are designed based on results from these magnitude estimation experiments have any effect on users' performance in an applied auditory display task. This paper presents an experiment that compares data-to-sound mappings in which the mapping's polarity is based on results from a previous magnitude estimation experiment against mappings whose polarities are inverted. The experiment is based around a simple task in which participants need to rank WiFi networks based on how secure they are, where security is represented using an auditory display. Results suggest that for a simple auditory display like the one used here, whether or not the polarities of the data-to-sound mappings are based on magnitude estimation does not have a substantial effect on any objective performance measures gathered during the experiment. Finally, potential areas for future work are discussed that may continue to investigate the problems addressed by this paper.

1. INTRODUCTION

Parameter mapping is a technique for data sonification: "the use of non-speech audio to convey information" [1]. In a parameter mapping sonification system (commonly shortened to *PMSon* for parameter mapping sonification) data values are used to manipulate acoustic parameters which facilitates the communication of the data. One of the most fundamental design challenges during the development of a PMSon system is the mapping topology - the relationship between the data parameters and acoustic parameters. In their chapter on PMSon in *The Sonfication Handbook* [2], Grond & Berger posit that "effective PMson often involves some compromise between intuitive, pleasant and precise display characteristics". However, there is little theory or evidence to guide designers toward what is the most effective acoustic parameter to convey a particular data value. Negative consequences caused by a deficit in Grond & Berger's trio of necessary characteristics can be grave in high-stakes contexts. This has been seen in noted instances of nuclear control room operators, locomotive drivers and aircraft pilots turning off auditory displays due to sounding unpleasant, or the information that they intend to convey being misleading or false [3]. Therefore, the imperative to move towards designing parameter mappings with a balance of these three characteristics is clear, yet it remains an under investigated problem.

Walker proposed the use of *magnitude estimation* as a tool which could be used by sonification designers to aid in establishing what the most effective acoustic parameter would be to represent a particular value of data [4]. Magnitude estimation maps the relationship between a sensory stimulus and its associated perceived intensity [5]. Walker's method provides two psychophysical measurements: polarity and scale. Polarity is the directional aspect of the mapping (e.g. increasing or decreasing pitch mapped to increasing temperature). Scale defines the amount of change in the acoustic parameter for a given change in the data parameter (e.g. for an increase from 10 °C to 20 °C, increase pitch by 50 Hz).

Walker used polarity as a measure of the "naturalness" of a mapping - the more unanimously participants perceived the polarity of a given data-to-sound mapping, the more "natural" this mapping was and therefore more effective as a parameter mapping topology. This methodology has been used in more recent studies for a variety of types of data and acoustic parameters [6, 7] and even beyond data-to-sound mappings into data-to-vibration mappings [8]. These studies indicate that magnitude estimation is a useful predictor of the effectiveness of a data-to-sound mapping in that it tells the researcher how unanimous the mapping polarity is amongst participants. However, there has been no research investigating the extent to which using results from these experiments to influence the design of a data-to-sound mapping has in an actual sonification task. This paper describes an experiment which investigates the effect that using parameter mapping polarities based [7] has on the performance of a simple sonification task, when compared with using parameter mappings with arbitrary polarities - in this case, the inverse polarity. The goal of this study is to establish to what extent the data from magnitude estimation experiments are generalisable when used in actual sonification tasks and therefore, further understand how to use experimental methods

 This work was funded by the EPSRC

and techniques to design parameter mappings for sonification.

2. RELATED WORK

2.1. A Brief Introduction to Parameter Mapping Sonification

There are three main aspects that must be considered when designing the mapping between a data parameter and an acoustic parameter. Firstly, there are psychophysical aspects: polarity and scale [2]. In addition to the psychophysical aspects, there are contextual factors. For example, the semiotics of the parameter mapping - what is the nature of the acoustic representation of the data? Kramer described a continuum for sound representation which ranges from analogic to symbolic, where analogic representations are more directly connected with the object (such as a Geiger counter) and symbolic representations are more abstract and indirect such as using pitch to represent temperature.

Walker & Kramer conducted the first study to investigate the effect that the choice of acoustic parameter(s) had on participant performance in a PMson system (originally presented in 1996 [9], published in 2005 [10]). They used of a number of acoustic parameters commonly used in sonification systems (pitch, onset, loudness and tempo) to convey simple data variables (temperature, pressure, size and rate) in a process-monitoring task. These parameters were split into four ensembles: *Intuitive*, *Okay*, *Bad* and *Random* based on how "natural" the designers believed a mapping to be. Results showed that the mappings which the sound designers believed to be optimal, e.g. temperature:pitch, did not result in either the most accurate or the fastest responses. Contrarily, mappings in the *Bad* ensemble yielded the fastest response time and *Random* led to the best performance.

Walker continued this line of research [4, 6], which investigated the use of the psychophysical method of magnitude estimation for systematically evaluating the perceived relationship between a data concept and an acoustic parameter in the context of sonification, with the goal of determining how "natural" a mapping is. Walker's studies obtained polarity and scale data for a number of data-to-sound mappings for basic data concepts and acoustic parameters such as temperature:pitch or pressure:tempo. Based on this data, Walker used the following criteria for evaluating a data-to-sound mapping: if a given polarity obtained a majority of all responses by participants in a block it was predicted to be a "good" polarity choice and it could therefore be predicted that the mapping itself was effective.

Present in the discussion section of these studies [10, 4, 6] is the importance of the listener's *mental model*. The "representation of some domain or situation that supports understanding, reasoning, and prediction" [11], or how they expect a data value to sound when it is sonified. A general assumption may be that an increase in a data value should be represented by an increase in an acoustic parameter, however findings from all of these previous works show that in many cases this is false. An example of this can be seen in the results from Walker's 2007 study [6] in which they found that when frequency was used to represent a value of size, more participants responded in a negative polarity than a positive. This suggests that these participants felt that "bigger" things are better represented by lower acoustic frequencies - aligning with a more physically-based mental model of sonification mapping, as larger things in the world often produce lower sounds.

Ferguson & Brewster conducted an experiment using the same magnitude estimation paradigm in which they investigated a number of additional data-to-sound mappings [7]. This study focused particularly on mappings in which both the data concept and the acoustic parameter are both generally deemed to be "undesirable" - attempting to further investigate the role of listeners' mental models in their determination of polarity and scale. The data concepts explored in this study were all semantically negative (*danger*, *error* and *stress*) and the acoustic parameters used to represent them (*roughness* and *noise*) would generally be considered undesirable from the standpoint of music or sound quality. Findings from this study suggested that for all of these mappings, the majority of participants perceived the mappings in a positive polarity - suggesting that for the data-to-sound mappings presented, an increase in a musically "undesirable" acoustic parameter like noise was perceived as conveying an increase in a semantically negative data variable such as stress. This again supports Walker & Kramer's assertion of the importance of the listener's *mental model* of the data being sonified being an important factor in how they expect a given data value to sound when it is sonified.

In a section entitled *Continuing Research Needs* in Walker's paper detailing initial investigations into using magnitude estimation for parameter mapping design [4], they posit that "the final test would always be instantiating these and other findings in more and varied sonification applications and systematically evaluating their effectiveness". However, this step has not yet been taken and the work detailed here aims to provide a starting point for this next stage of investigation into this method.

3. EXPERIMENT

An experiment was conducted to investigate if using mapping polarities based on results from a prior magnitude estimation experiment has any effect on performance during a simple auditory display task. In this experiment, a task consisting of ranking three WiFi networks based on their security level was used, in which the level of security for each network (low, medium, high) was conveyed using an auditory cue. The data-to-sound mappings and the polarities were based on results from Ferguson & Brewster [7], specifically here using the mappings of *danger:roughness* and *danger:noise* - danger in this context being how insecure or "dangerous" a WiFi network may be. In this magnitude estimation study, they found that when noise was used to represent danger, 13 of 15 participants perceived this mapping in a positive polarity (increasing noise = increasing danger). Similarly, when roughness was used, 12 of 13 participants responded in a positive polarity.

3.1. Participants

Twenty four participants took part in the study. Participants were: 12 female, 11 male, 1 non-binary, mean age = 28.2 years, SD = 6 years, 23 right-handed, 1 left-handed. All participants reported no uncorrected vision impairment and no hearing impairments.

3.2. Design

Eight conditions were investigated in which the independent variables were the acoustic parameter used and the polarity in which it was mapped to the security of the network. The polarity of each data-to-sound mapping was either based on results from [7], or were the inverse from the prior study, meaning the polarity was inverted such as increasing roughness = *increasing* danger would thus be inverted to increasing roughness = *decreasing* danger. For the sake of brevity, in this paper we will refer to all polarities based results from [7] as *aligned* and the others as *inverted*. The main dependent variable collected during the experiment include completion time, correctness of responses and NASA Task Load Index (TLX) [12]. The experiment used a within-subjects design. For this experiment, it was important to consider the order in which participants were presented with each polarity, as a participant may favour whichever polarity they were exposed to first, thus reducing the quality of the data gathered. Therefore counterbalancing was used to ensure that equal numbers of participants received each polarity first.

3.3. Stimuli

Roughness and noise were used as acoustic parameters in this study, based on the stimuli used in Ferguson & Brewster's magnitude estimation study [7]. These parameters were chosen in this prior work due to the effect of roughness on the perception of danger [13] and noise's effect on the perception of image focus [14] (with lack of focus or "bluriness" also being a semantically negative or "undesirable" data concept). This prior study used ten levels for each acoustic parameter, an example being the noise condition in that study containing ten sound cues ranging from a clean tone to total white noise. Results from another study that Ferguson & Brewster carried out using both acoustic roughness and noise to convey information suggested that participants found it difficult to interpret ten levels of these stimuli [14]. Therefore for the experiment described in this paper we reduced the number of levels to three to ensure that the task would be simple and the sound cues could easily interpreted. As in [7], each stimulus was 2 seconds in length. Each stimulus had an amplitude envelope with a 0.2 second linear ramp onset (attack) and offset (release). An amplitude envelope was included in the sound design, as an abrupt start or stop of a sound can be perceived as unpleasant [15]. All stimuli were created in the Supercollider programming language [1]. The acoustic design of each stimuli is described below.

- **Roughness**
  100 % sinusoidally amplitude modulated 1000 Hz pure-tone with modulation frequencies of 0, 11 and 70 Hz.
- **Noise**
  This condition consisted of a 1000 Hz pure tone for the first level, and equal blend of a pure tone and broadband white noise for the second level and the final level was solely broadband noise.

3.4. Procedure

The experiment consisted of four blocks - each acoustic parameter (roughness, noise) mapped in each polarity (*aligned*, *inverted*). Each block consisted of three trials. At the beginning of each condition, participants were presented with a screen which explained the acoustic parameter being used in that block and how it was mapped to each level of network security (Figure 1). In this screen, participants could use the three coloured buttons to hear the sound cues for each level of security for the given condition's data-to-sound mapping. Participants could not press the continue button until the button for each level of security was pressed at least once, however they could listen to each sound cue as many times as they needed. In each trial, participants were presented with a screen showing three WiFi networks, each with a button to play the sound cue to convey their level of security (Figure 2).

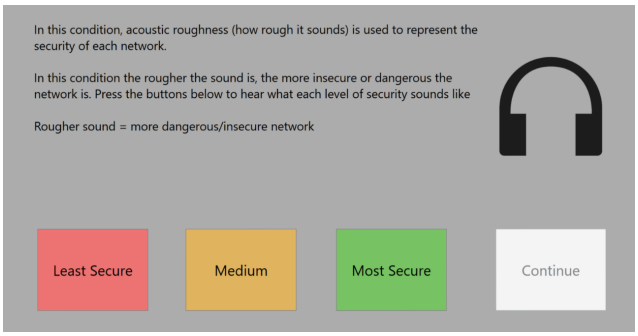


Figure 1: Condition introduction page where the data-to-sound mapping is explained, including the polarity. Three coloured buttons allow the participant to hear the sound for each level of security.

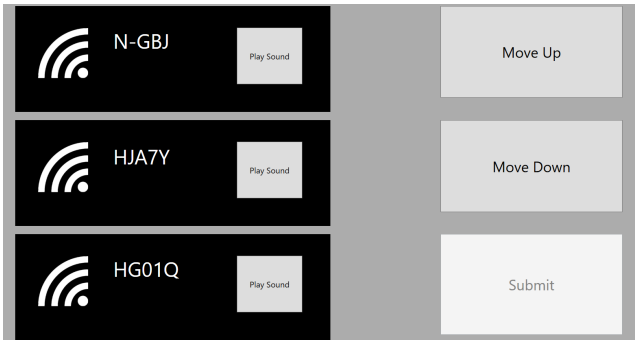


Figure 2: Screen for each trial showing three networks, each with a button to play their associated sound stimuli.

Sounds were presented using a pair of Beyerdynamic DT100 headphones. The participants were tasked with rearranging the networks based on their security (top = most secure, bottom = least secure) as conveyed by the sound cues. Similarly to the first screen (Figure 1) participants had to listen to each stimuli at least once before being able to submit their ranking, however there was no upper limit to how many times they could repeat each sound. For each condition participants completed three rankings, each time the ordering and name of the networks were randomised. The randomisation was implemented such that it was ensured that the network ordering was always mixed - ensuring the ordering wasn't correct at the beginning of each trial - thus ensuring the participant

[1] http://supercollider.github.io

had to rearrange the ordering. At the beginning of the experiment, participants completed a practice condition to allow them to familiarise themselves with the experiment. This practice condition was identical in procedure to all the subsequent conditions, except participants were required to rank the networks based on colour rather than using sound cues. Sound was omitted from the practice to ensure that participants were not influenced in any way which may have an effect their future responses. At the end of each condition, participants completed the NASA Task Load Index.

4. RESULTS

Firstly, since each ranking was completed three times, the mean completion time was calculated for each participant, for each sound/polarity combination. No statistically significant difference was found between the completion times for both polarities in the roughness ($F_{1,46}$ = 1.6, p = 0.2) and noise ($F_{1,46}$ = 0.02, p = 0.9) conditions. Of all 288 ranking trials completed, only 10 were ranked in an incorrect order. In this case, it is of interest to attempt to gain a more thorough insight into how little polarity choice impacted the participants' completion time. For example, it may be useful for an auditory interface designer to understand if carrying out a magnitude estimation experiment to evaluate a particular data-to-sound mapping is necessary for their particular use-case. Therefore, we calculated effect sizes (using the recommendations set out by the Transparent Statistics in Human-Computer Interaction Group [16]). Looking at the results for the roughness conditions, a Welch's t-test shows that the estimated difference in the means between the *aligned* and *inverted* polarities is -1774ms (95% CI: [-4601, 1053]). Cohen's d = 0.36. For the results for the noise conditions, a Welch's t-test shows that the estimated differences in the means between the *aligned* and *inverted* polarities is -149ms (95% CI: [-2505, 2206]). Cohen's d = 0.03. A Wilcoxon signed rank test found no statistically significant differences between NASA Task Load Index workloads for both polarities in the roughness and noise conditions (Table 1 shows TLX results).

Condition	Polarity	Mean (Workload)	SD (Workload)
Roughness	*Aligned*	31.1	13.6
Roughness	*Inverted*	32.2	12
Noise	*Aligned*	25.8	9.7
Noise	*Inverted*	30.1	15.7

Table 1: Summary of NASA Task Load Index results including mean workload and standard deviations

5. ANALYSIS

The small effect sizes and estimated differences between polarities for both acoustic parameters is surprising, as it would be a reasonable expectation *a priori* to assume that the mapping design in which the polarity is based on results from a previous magnitude experiment would result in a faster completion time. The estimated differences in the means for the roughness conditions is less than two seconds and less than a quarter of a second for the noise conditions. For many applications this very small difference may be acceptable, meaning that carrying out a magnitude estimation experiment to gather polarity data may not be necessary in some cases. In order to further explore this notion, we utilised the *Akaike Information Criterion* to investigate whether a model in which the completion times for both polarities are *equal* is more representative of the data gathered from this study, than a model in which the completion times for each polarity is assumed to be different. The following section provides an introduction to this method and describes its application to the results from this experiment.

5.1. A Brief Overview of the Akaike Information Criterion

As this method is uncommon in auditory display literature (with the notable exception of Frohmann et al. [17]) a brief overview of the process of using the method is given in this section. Hirotugu Akaike's information criterion (*AIC*) [18] is a method to estimate the relative quality of statistical models for a given data set. AIC estimates the amount of information lost when data is fitted to a given model, thus when comparing two potential models, the model with less information lost is more representative of the data in question. Null hypothesis testing cannot allow acceptance of the null hypothesis (in this case being "the completion times for mappings that are based either on polarity data from a prior experiment or inverted polarities are equal") however AIC can be used to reframe the question to be "is there more support for a model in which the completion times for both polarities are equivalent than one in which they are not". The following equation is used to estimate the AIC of a model [18, 19]:

$$AIC = -2log(\hat{L}) + 2k \tag{1}$$

where k is the degrees of freedom and $\hat{L}$ is the maximum value of the likelihood function of the model. To quantify the quality of each model, the raw AIC score must be converted to weighted scores. The first step is to calculate the differences in AIC for each model with the respect to the AIC of the best candidate model [19, 20]:

$$\Delta_i(AIC) = AIC_i - AIC_{min} \tag{2}$$

Where AIC_{min} is the minimum of the AIC values. This transformation causes the best model to have $\Delta_i(AIC) = 0$, while the rest of the models have positive values. The next step is to establish the relative likelihood L for each model i given the data

$$L(M_i|data) \propto exp\{-\frac{1}{2}\Delta_i(AIC)\} \tag{3}$$

where $\propto$ denotes "is proportional to". Finally, the relative likelihoods for each model are normalised to obtain weighted AIC scores for each model (w_i). Here each model's relative likelihood is divided by the sum of the likelihoods of all other models being compared, like so:

$$w_i(AIC) = \frac{exp\{-\frac{1}{2}\Delta_i(AIC)\}}{\sum_{k=1}^{k} exp\{-\frac{1}{2}\Delta_k(AIC)\}} \tag{4}$$

Finally, the weighted AIC scores can infer the best fitting model. For example, if two models: A and B are being compared, with weighted AIC scores of $w_a(AIC) = 0.6094$ and $w_b(AIC) = 0.2242$, their weighted AIC scores would be used to show that model A is around 2.7 times more likely to be a better fit for the data than model B [2]:

$$\frac{w_a(AIC)}{w_b(AIC)} = \frac{0.6094}{0.2242} \approx 2.7$$

[2] example taken from [21].

5.2. Applying AIC to the Current Experiment

As discussed in the prior sections, we use the Akaike Information Criterion to answer the question:

> *"Is there more support for a model in which the completion times for both polarities are equivalent than one in which they are not?"*

Firstly, we fit two models: a linear model in which the completion times for both polarities is forced to be equivalent (effectively treating the data as if there were only one polarity category) - henceforth written as *Equivalent Model*, and a linear model in which they are assumed to be not equal - here named *Unequal Model*. By using the process described in the previous section, the quality of these models can be compared. Table 2 shows the results of the AIC analysis.

6. DISCUSSION

From the AIC analysis results in Table 2 we can see that for both roughness and noise, the *equivalent model* i.e. the one in which the completion times for each polarity are assumed to be equivalent is the best model for the given data (1.2 and 2.7 times more so for roughness and noise respectively). The AIC results in combination with the small effect sizes reported earlier support the argument that for this task, the polarity of the data-to-sound mapping did not have a substantial effect on the time it took participants to complete the task, with the estimated effect being $\sim$1.8s for roughness and $\sim$0.15s for noise. Furthermore, the NASA TLX results also suggest similar levels of workload in both polarities.

These results are surprising as the previous magnitude estimation study [7] showed that nearly all the tested participants perceived *increasing* noise or roughness as *increasing* danger, therefore it was expected that the mappings used in this study that were based on this would result in faster completion times, however this was not found to be true. This suggests that for *simple* auditory displays using roughness or noise such as the application used in the experiment here, the polarity in which the data is mapped to the acoustic parameter does not have a substantial effect. This means that for designers working in a similar space, the expenditure of resources to carry out a magnitude estimation experiment to establish polarities may not be necessary if the design can afford the potential discrepancies in completion time as discussed earlier.

7. LIMITATIONS AND FUTURE WORK

The generalisability of data-to-sound mapping polarities obtained from magnitude estimation studies like [4, 6, 7] has yet be fully determined. This study provided insight into the how generalisable polarities obtained for two data-to-sound mappings: *danger:roughness* and *danger:noise* [7], but it is only a first step toward understanding how generalisable data from these magnitude estimation experiments are in practice. The following sections discusses some limitations of the current study and puts forward potential future work that may address them.

7.1. Difficulty of The Task

The primary limitation of this study is that participants could potentially work through a "bad" data-to-sound mapping, because the task was relatively simple - 96.5 % of rankings were completed correctly. For example, even if a participant thinks that a more natural representation of increasing danger for them is using increasing roughness to convey this increase, they may still be able to complete the task in a fairly quick amount of time using a conflicting representation (i.e. increasing danger conveyed by *decreasing* roughness) due to the relative easiness of the task. The task was intentionally designed to be easy to carry out - both to account for participants who may be new to the notion of an auditory display and so that we could begin investigating this area with a simple auditory display. Many situations where auditory displays are commonly confronted by most people are quite simple such as mobile phone notifications or in-car displays etc. so we wanted to reflect that in this study before moving onto more complex sonifications. We intend to carry out a similar study with a more complex task by using a similar auditory display but in a more cognitively demanding and potentially more ecologically valid situation - again, to attempt to reflect the fact that many auditory displays are specific in context and complex, such as used by aircraft pilots or process-monitors.

7.2. Specificity of Context

This experiment focused solely data-to-sound mappings conveying danger - specifically the danger posed by an insecure WiFi network. The previous magnitude estimation study [7] presented danger in general and contextually agnostic terms, therefore it may be useful for future works attempting to investigate the generalisability of polarities gathered from magnitude estimations to evaluate multiple contexts for a given data-to-sound mapping. For example, a sonification of a value of danger in terms of WiFi security may be perceived vastly differently than a much more severe context such as a process-monitoring sonfication system in a nuclear power station. Therefore evaluating a broader range of contexts may afford a more well-rounded view of how generally polarities and scales from magnitude estimation experiments may be applied.

8. CONCLUSIONS

The research presented in this paper presents a first attempt to investigate the effect of designing data-to-sound parameter mapping polarities based on data from a magnitude estimation experiment. We presented a study in which we compared the time it took participants' to complete an auditory display based ranking task using two data-to-sound mappings in a simple auditory display task: one mapping in which the the polarity was based on results from a previous magnitude estimation experiment and one mapping in which the polarity was arbitrarily designed - in this case inverted. Based on results from this experiment we used the Akaike Information Criterion to discuss statistically that the polarity of the data-to-sound mappings did not have a substantial effect on the time it took participants to complete a ranking. Finally, we discussed some limitations of this study and suggest some future work which may address them. This work represents a first step toward researching how data obtained from magnitude estimation experiments can be appropriately applied in real-world sonification tasks and results from this study underline the need for further research in this area.

Condition	Model	DoF	$\log(\hat{L})$	AIC	ΔAIC	$w_i(AIC)$	$\frac{w_a(AIC)}{w_b(AIC)}$
Roughness	*Equivalent*	2	-475	954.8	0	0.5449	1.2
Roughness	*Unequal*	3	-474	955.1	0.3601	0.4551	
Noise	*Equivalent*	2	-466	935.6	0	0.7294	2.7
Noise	*Unequal*	3	-466	937.6	1.983	0.2706	

Table 2: Summary of results from AIC analysis.

9. ACKNOWLEDGMENTS

The authors would like to thank David Borchers for their suggestions regarding statistics during the development of this work, in addition to the reviewers for their insightful and useful comments.

10. REFERENCES

[1] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff, R. Bargar, S. Barrass, J. Berger, *et al.*, "The sonification report: Status of the field and research agenda. report prepared for the national science foundation by members of the international community for auditory display," *International Community for Auditory Display (ICAD), Santa Fe, NM*, 1999.

[2] F. Grond and J. Berger, "Parameter mapping sonification," *The sonification handbook*, pp. 363–397, 2011.

[3] R. D. Sorkin, "Why are people turning off our alarms?" *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 1107–1108, 1988.

[4] B. N. Walker, "Magnitude estimation of conceptual data dimensions for use in sonification," *Journal of Experimental Psychology: Applied*, vol. 8, no. 4, pp. 211–221, 2002.

[5] R. Teghtsoonian, S. Stevens, and G. Stevens, "Psychophysics: Introduction to its perceptual, neural, and social prospects," *The American Journal of Psychology*, vol. 88, no. 4, p. 677, 1975.

[6] B. N. Walker, "Consistency of magnitude estimations with conceptual data dimensions used for sonification," *Applied Cognitive Psychology*, vol. 21, no. 5, pp. 579–599, 2007.

[7] J. Ferguson and S. A. Brewster, "Investigating perceptual congruence between data and display dimensions in sonification," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 611.

[8] J. Ferguson, J. Williamson, and S. Brewster, "Evaluating mapping designs for conveying data through tactons," in *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. ACM, 2018, pp. 215–223.

[9] B. N. Walker and G. Kramer, "Mappings and metaphors in auditory displays: An experimental assessment," in *ICAD 1996, Proceedings of the International Conference on Auditory Display*. Georgia Institute of Technology, 1996.

[10] ——, "Mappings and metaphors in auditory displays: An experimental assessment," *ACM Transactions on Applied Perception (TAP)*, vol. 2, no. 4, pp. 407–412, 2005.

[11] D. Gentner, "Mental models, psychology of," in *International Encyclopedia of the Social & Behavioral Sciences*. Pergamon, 2001, pp. 9683 – 9687.

[12] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[13] L. H. Arnal, A. Flinker, A. Kleinschmidt, A.-L. Giraud, and D. Poeppel, "Human screams occupy a privileged niche in the communication soundscape," *Current Biology*, vol. 25, no. 15, pp. 2051–2056, 2015.

[14] J. Ferguson and S. A. Brewster, "Evaluation of psychoacoustic sound parameters for sonification," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction-ICMI 2017*. ACM Press, 2017, pp. 120–127.

[15] P. Bergman, A. Sköld, D. Västfjäll, and N. Fransson, "Perceptual and emotional categorization of sound," *The Journal of the Acoustical Society of America*, vol. 126, no. 6, pp. 3156–3167, 2009.

[16] T. S. in HumanComputer Interaction Working Group, "Transparent Statistics Guidelines," Feb 2019, (Available at https://transparentstats.github.io/guidelines).

[17] L. Frohmann, M. Weger, and R. Höldrich, "Recognizability and perceived urgency of bicycle bells." Georgia Institute of Technology, 2018.

[18] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[19] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: A practical information-theoretic approach*. Springer-Verlag, 2002.

[20] ——, "Multimodel inference: understanding aic and bic in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.

[21] E.-J. Wagenmakers and S. Farrell, "Aic model selection using akaike weights," *Psychonomic bulletin & review*, vol. 11, no. 1, pp. 192–196, 2004.

SONIFIGRAPHER. SONIFIED LIGHT CURVE SYNTHESIZER

Adrián García Riber

Image and Sound Art,
Francesc Martí i Mora 1-B 22-3,
Palma de Mallorca, 07011, Spain
adrian@imageandsoundart.com

ABSTRACT

In an attempt to contribute to the constant feedback existing between science and music, this work describes the design strategies used in the development of the virtual synthesizer prototype called *Sonifigrapher*. Trying to achieve new ways of creating experimental music through the exploration of exoplanet data sonifications, this software provides an easy-to-use graph-to-sound quadraphonic converter, designed for the sonification of the light curves from NASA's publicly-available exoplanet archive. Based on some features of the first analog tape recorder samplers, the prototype allows end-users to load a light curve from the archive and create controlled audio spectra making use of additive synthesis sonification. It is expected to be useful in creative, educational and informational contexts as part of an experimental and interdisciplinary development project for sonification tools, oriented to both non-specialized and specialized audiences.

1. INTRODUCTION

According to Vickers' [1] distinction between auditory and sonified graphs, Sonifigrapher can be defined as a virtual synthesizer that provides non-MIDI sonified graphs through the exploration and mapping of the RGB values of user-loaded PNG files. Initially inspired by the Chamberlin and the Mellotron concept, where each key of a keyboard reproduces an analog tape pre-recorded sound, *Sonifigrapher* synthesizer works as an image sonification sampler that allows single playing and looping. The prototype has been initially developed to create sonifications from the publicly-available graphic astronomical information published at Mikulski Archive for Space Telescopes (MAST) [2,3], the simulated light curves of the Planet Hunters project from the Transiting Exoplanet Survey Satellite (TESS) [4,5], and the curves generated with the *Lightkurve* software package for Kepler & TESS time series analysis in Python [6]. However, its design can be easily adapted to any kind of graphic representation.

2. REFERENCE WORKS

In 1980, motivated by his interest in the direct synthesis of the time pressure curve [7], Iannis Xenakis started developing the Unité Polyagogique Informatique du CEMAMu (UPIC) which by the nineties, already represented a new paradigm in the use of experimental music devices for learning [8]. Created as a mouse-controlled, graphical-musical composing system [9], UPIC allowed musicians to give orders to the computer through drawings [10].

This inspiring concept underlies numerous projects focused on graphic to sound conversion and laid the foundations of the graphical open-source sequencer for digital art IanniX [11], [12]. Further from the artistic point of view and designed for the creation of multi-purpose auditory graphs, *Sonification Sandbox* by Walker & Cothran [13], provides a multiplatform MIDI-based, graph-to-sound, user-selectable sonification engine, based on previous projects *MUSE* and *MUSEART*. It was designed for a wide range of users, from novice to expert, and allows adding context to the sonified data using reference tracks. In a more specific approach, Bell3D audio-based Astronomy Education System [14] and xSonify astronomical data sonification software [15], represent two reference examples of sonification projects designed to make data-driven astronomy accessible to both visually impaired and sighted users. The first one, by Jaime Ferguson, allows users to learn about basic astronomy through surround sonifications of user-selected stars' parameters. The xSonify project from Diaz-Merced et al. [16], sonifies two-dimensional data from text files, for large data sets in frequency dimensions, trying to find visually masked correlations and patterns.

3. ABOUT LIGHT CURVES AND TRANSITS

Light curves are graphic representations of the brightness flux variations along time, observed in celestial objects. When a planet passes in front of a star, it generates a partial eclipse called transit, producing a flux decrement in its light curve which can be measured in terms of time and depth.

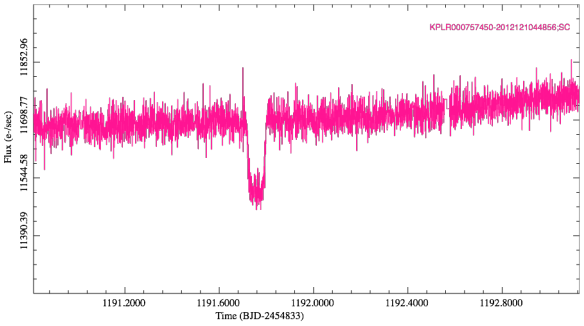


Figure 1: Enlarged view of KPLR000757450-2012121044856 light curve showing a planet transit [17]. PDCsap Bright Flux (Pre-search Data Conditioning Simple Aperture Photometry, electrons per second) [18] vs time expressed in BJD-2454833 (Barycentric Julian Date, 2454833.0 offset) [19].

In 1999, HD 209458 b, nicknamed 'Osiris', was the first exoplanet to be seen in transit around its star, opening new horizons in exoplanet characterization through the transit detection method [20]. Currently, the number of confirmed

https://doi.org/10.21785/icad2019.016

exoplanets is about four thousand and increasing every week (3972 confirmed exoplanets, 05/26/2019) [21].

As described by Seager & Mallén-Ornelas [22], the planet's orbit can be characterized by analyzing the decrease of energy presented in the light curve. Measuring time between transits, the orbital period is also obtained to complete the equations.

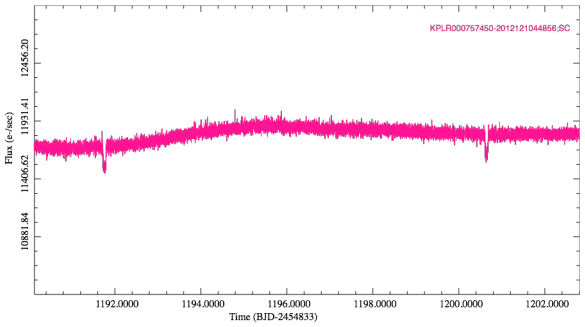


Figure 2: KPLR000757450-2012121044856 light curve showing two transits [17]. PDCsap Flux (electrons per second) [18] vs Time (BJD-2454833) [19].

In addition to transit photometry, the radial velocity of the host star is needed to determine planet's radius and mass. This method detects the oscillating Doppler shift variation in the radial velocity of the star due to the gravitation of an orbiting planet [22]. The planet's temperature and atmospheric properties can also be determined through transmission spectroscopy methods consisting of the observation of transit light curves at different wavelengths [23].

However, the detection and characterization of exoplanets is not free of challenges and many factors make it usual to generate false positives or loss of information and uncertainties. Worth mentioning that stellar systems such as Brown Dwarfs or Eclipsing Binaries can produce variations in the light curves similar to those generated by the orbiting planets. The stellar activity can also affect the observation of the exoplanet's atmosphere and inherent noise can affect the detection of the weakest planet's signals [24].

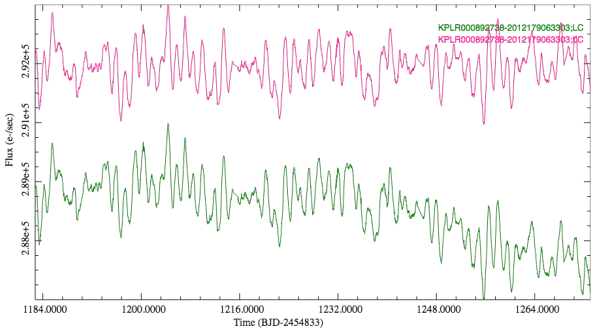


Figure 3: KPLR000892738-2012179063303. Possible Red Giant star [17]. PDCsap Flux (red) Sap Flux (green) [18] vs Time (BJD-2454833) [19].

4. VIRTUAL INSTRUMENT DESIGN

Sonifigrapher is a quadraphonic/stereo virtual instrument prototype designed for sonifying light curves. In the same way as wavetable synthesizers work, it makes use of the curves to generate filter-controlled audio spectra through additive synthesis which control variables have been mapped from the graphic representation input. Csound's [25] backwards and future compatibility compromise, together with the possibilities that its API for Python [26] and Cabbage's [27] VST plug-in exporting option provide, induced the workflow decision for the project.

Subscribing the words by Gerhard Steinke related to newly developed electronic music instruments expressing that "a great number of composers should be given the possibility of interpreting their ideas in various ways after having become familiar with the apparatus" (about his subharmonic synthesizer prototype, *Subharchord*, 1966) [28], *Sonifigrapher* can be downloaded for testing as a packed 'ZIP' file from:

https://archive.org/details/SonifigrapherMacOSX

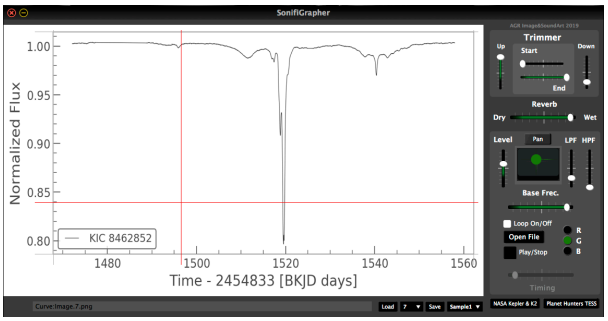


Figure 4: *Sonifigrapher* interface synthesizing the unusual light curve of Tabby's star, extracted from *Lightkurve* Python's package website [6].

A first stage of intensive testing with Marilungo's examples of CSound's image processing opcodes [29], [30] was crucial to motivate the development of *Sonifigrapher* prototype. These reference approaches made it possible to summarize the benefits and drawbacks of sonified graphs as well as to experiment with the aesthetics and accuracy aspects of the final design. McCurdy's Cabbage and CSound examples [31] were also consulted for technical resolution strategies in the final implementation.

The design of the prototype relies on the main pillars of sonification processes highlighted by Scaletti [32], Hermann [33], [34], De Campo [35], Vogt [36] or Kramer et al. [37], assuring:

- Original information communication.
- Adaptation to the language and needs of the research field.
- Systematic transformation of the input data.
- End-user control and/or interaction.
- Reproducibility.
- Possibility of validation and repetition with different input data sets.
- Integrability.

The final synthesizer has been developed to meet the following goals:

- To explore the multimodal display possibilities of CSound and Cabbage workflow.
- To provide a sonification tool for testing auditory transit detection.
- To demonstrate the added value of multimodal synergies in graphic datasets.
- To generate a cross-domain, interference-preserving mapping.

- To create a multimodal tool for approaching and/or communicating the information contained in light curves databases from different perspectives within creative, informational and/or educational contexts.
- To implement an extremely intuitive UI with almost no learning curve.
- To allow real-time operation and live performance.
- To integrate the virtual instrument in Digital Audio Workstations.

5. IMPLEMENTATION ALGORITHM

In order to generate a sound representation allowing accurate perception of graphic changes and according to the aesthetics of Experimental, Electronic and Electroacoustic music, *Sonifigrapher* uses additive synthesis with a non-quantified frequency scale generated from a user-defined base frequency. This approach makes it possible to create tonal sweeps and microtonal sounds or chords and improves accuracy in light curves' pitch tracking. As the final sonified spectrum relies on a user-defined base frequency, it is possible to adapt the graphic changes in the curves to different frequency ranges for a better perception of the sonification, or fine-tuning in creative applications. To maintain coherence with the transit detection method, the lowest flux values in the curves correspond to the highest frequencies in the sonification. In this way, when a transit is produced, a high frequency sine is reproduced facilitating its detection.

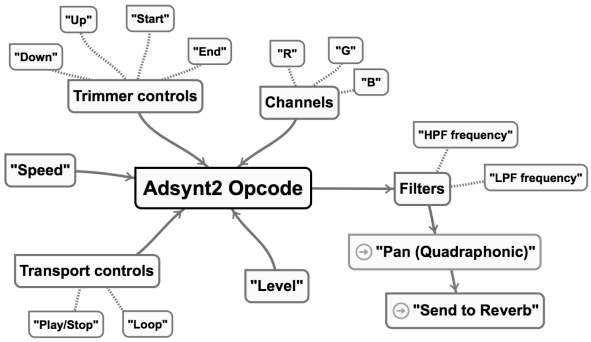


Figure 5: *Sonifigrapher* design map showing control variables and signal flow.

The core of the prototype is therefore CSound's *adsynt2* opcode [29], which performs additive synthesis with an arbitrary number of partials, not necessarily harmonics. This opcode also provides interpolation to lightly soften the most pronounced graphic transitions. *Figure 5* describes the design implementation map with all the variables used to control the sonification process. The R, G and B values of the loaded image are extracted using CSound's image processing opcodes [29] to work as input arguments for *adsynt2*. Its monophonic output is low- and high-pass filtered and sent in parallel to a quadraphonic matrix and reverberation processor to be used in creative applications. End users can select the sonified R, B or G channel input to reduce noise and focus attention.

The synthesizer also provides trimmer controls to adjust the 'start' and 'end' points of reproduction as well as the 'up' and 'down' graphic limits to avoid the sonification of non-relevant information printed in the sampled image. The speed control allows both detailed analysis and fast monitoring, if the loop playback is not enabled. The loop reproduction works on an eight seconds Mellotron-based time scale and disables timing control. All changes made to the 'trimmer', 'speed' and 'loop' controls are applied once the current reproduction is completed. High- and low-pass filters frequencies are controlled before the signal is sent to the reverberation processor to improve sound quality. A user controlled "x-y" matrix is also provided for sound allocation in a quadraphonic reproduction system. Default auto-panning configuration follows the graphic timeline bar with stereo compatibility. The 'level' fader acts over both the 'dry' and 'wet' signals by minimizing the number of controls required.

Going further on the sonification process through *adsynt2* opcode control, next figure describes the input and output arguments being used by the algorithm (referred to CSound variables). The complete source code is available in the '.csd' file included in the downloadable version of *Sonifigrapher*.

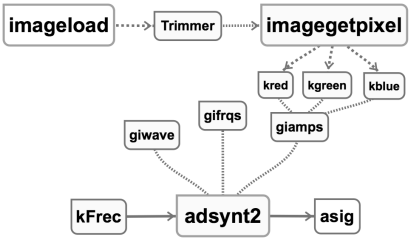


Figure 6: Design map showing CSound's *adsynt2*, *imageload* and *imagegetpixel* opcodes with input and output arguments. (Based on Marilungo's work).

Once the image is loaded, four global 'i' variables (giImageIni, giImageEnd, giVLimUp and giVLimDw), hold the trimmer options before the extraction of every pixel's bright amplitude in the x-y exploration process. These amplitudes are stored in three 'k' variables (kred, kgreen and kblue), used to control the amplitude ratios of the generated sine waves through a global 'i' variable (giamps).

If the prototype is loaded with a white background plot, these values are inverted using a threshold detection conditional statement to reduce background noise related to high bright pixel values. End users can select the RGB channel to be sonified (gkR, gkG and gkB in the source code), and adapt the base frequency (kFrec) of the synthesized spectrum to their needs. The total number of sinusoids and the frequency ratios are introduced in the opcode through two global 'i' variables (giwave and gifrqs).

The synthesized signal (asig), is passed through two cascade high- and low-pass filters, generating the mono filtered audio output (aFilt). This signal is routed to four audio output channels using two parallel *pan2* opcodes that allow front and rear panning (gaLf, gaRf, gaLr and gaRr). The reverberation effect has been implemented in an independent instrument for best audio quality and makes use of two independent processors fed by the same stereo front audio signals (gaSenL and gaSendR), enhancing the surround environment. The master level of the four output channels is controlled with a single global 'k' variable (gkAmp).

6. ACCESING THE DATABASE

To simplify the access to the light curve database, the prototype includes a 'NASA Kepler&K2' button which links to NASA's description of 'Two ways to get Kepler Light Curves' [2]. The

following instructions allow the reproduction of the light curve showed in *Figure 1*.

- Go to: http://archive.stsci.edu/kepler/data_search/search.php
- Press the 'Search' button to access the complete catalog.
- Mark the KPLR000757450-2012121044856 row in the first page and press the 'Plot marked Light curves' button at the top of the list.
- Zoom into the red curve around (1200.6964,11775.290) to obtain a transit representation like *Figure 1*.
- To visualize the orbital period showed in *Figure 2*, zoom into the red curve from (1190.1292,12980.986) to (1202.8074,10357.050).

For the sonification of the light curves:

- Press the 'Create png Image' button (just below the light curve) and save the file.
- Inside *Sonifigrafer* user interface, open the '.png' file and press 'Play'.
- Adjust the base and filter's frequencies for sound response optimization.
- The complete configuration, including a copy of the image, can be saved and recalled.

The prototype also includes a 'Planet Hunters TESS' button that links to this online classification project [4]. To sonify its light curves with *Sonifigrapher* just choose an image, save it as PNG file or make a screenshot, and load it using the synthesizer's 'Open file' button.

7. TASK-BASED VALIDATION EXAMPLES

In the second phase of this project, the Planet Hunters TESS light curves [4], [5] have been used with the double intention of testing the possibilities for exoplanet transits auditory detection and validating the prototype with the exploration of a different data set. This open collaborative project provides an interactive light-curve classification tool in which final users can mark observed transits and generate discussion forums.

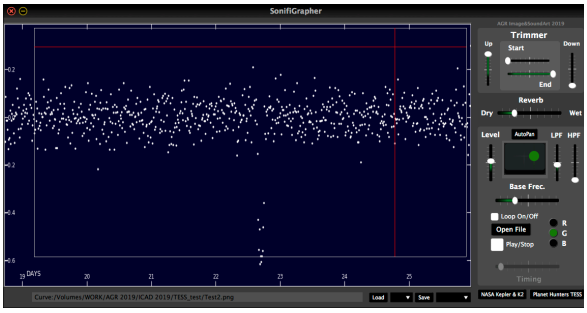


Figure 7: *Sonifigrapher* interface capture during a Planet Hunters' light curve sonification [4]. Single transit detected.

Intensive testing has been made to evaluate the synthesizer's behavior with this data set. Acting over the base frequency, trimmer controls and filters, auditory transit detection has been satisfactory achieved for the deepest transits, enhancing the graphic information with a real time easily identifiable sonic cue. Several video examples showing the effectiveness of the prototype in these situations are available at:

https://archive.org/details/transits

A more detailed auditory analysis of the curves is also possible acting over the 'start' and 'end' trimmer controls and controlling the playback speed. The smallest amplitude variations in the light curves can be perceived using higher base frequencies and slower velocities. Extremely slow single-pass playback allows bright points discrimination and the creation of chords and electric piano-like arpeggios. This sound sequences can also be repeated, creating eight-seconds loops from the graphic information between markers. Five descriptive light-curve samples are included for exploring the synthesizer. Listening to some examples is also possible in video format, via the *Sonifigrapher* download page.

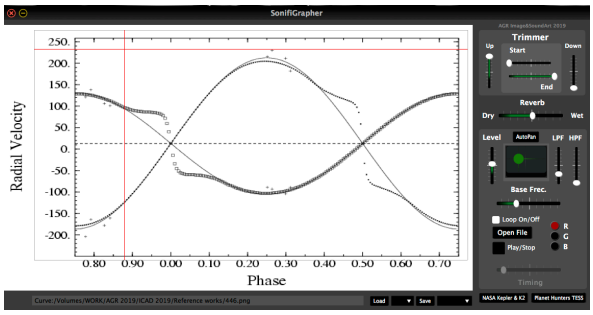


Figure 8: *Sonifigrapher* interface capture during the sonification of CX Aqr star. Radial velocity vs phase [38].

For final testing and prototype validation, the Catalog and Atlas of Eclipsing Binaries (CALEB) [39] and the CSI 2264 CoRoT's light curves [40] have also been used. A user experience study based in this last catalog is expected to be the third phase of the synthesizer's development project with the intention of providing useful information from both specialized and non-specialized users.

8. CONCLUSION

Although CSound is not an image-processing oriented programming language, it allows the development of multimodal interactive software tools that can bring sonification closer not only to scientific specialized audiences but also to students at any level and in any field of knowledge. "Pairing data sonification with data visualization is a synergistic tool for augmenting both visualization and sonification, and can highlight connections or correlations between variables" Scaletti, 2018 [32].

Cabbage's VST exporting options, in conjunction with CSound's never-ending programming and processing possibilities, establish a multidisciplinary natural connection between science-oriented sonifications and music creation environments that opens the possibilities of both disciplines and allows constant feedback between them.

If we look at education and music as relevant parts of our everyday life, it seems important to increment the efforts invested in the development of new multimodal and interdisciplinary sonification tools designed to bring science closer to people. Sonification can represent any kind of information and its unique communication and engagement capabilities could be considered and included when developing institutional education materials to normalize the use of sonifications in the next generation of digital natives. As highlighted by Quinton et al. [41], teaching how to listen seems crucial in the acceptance of sonification as a data analysis tool.

On the other hand, and according to Scaletti's distinctions between sonification and music [32], the use of sonified data as a sound source for the creation of original music material represents an open field with full potential to build a bridge between the sonification processes and general public. Data-driven virtual instruments' development projects provide the possibility of acting in both public-private, known-unknown and interactive-fixed areas of Scaletti's Sonification Space, generating environments of interest for the sonification community to explore.

9. REFERENCES

[1] Vickers, P. (July 2005). Wither and wherefore the Auditory Graph? Abstractions and Aesthetics in *Auditory* and *Sonified* Graphs, *Proceedings of the 11th International Conference on Auditory Display. (ICAD),* Limerick, Ireland.

[2] https://www.nasa.gov/kepler/education/getlightcurves

[3] http://archive.stsci.edu/kepler/data_search/search.php

[4] https://www.zooniverse.org/projects/nora-dot-eisner/planet-hunters-tess

[5] https://tess.mit.edu/

[6] https://docs.lightkurve.org/

[7] Georgakki, A. May (2005). The grain of Xenakis' Technological thought in the Computer Music research of our days, *Proceedings of the International Symposium Iannis Xenakis*, Athens, Greece.

[8] Nelson, P. (1997). The UPIC system as an instrument of learning. *Organised Sound, 2*(1), 35-42.

[9] Xenakis, I. (1992). *Formalized Music*. *Thought and Mathematics in Music*. Hillsdale, NY: Pendragon Press.

[10] http://www.centre-iannis-xenakis.org/cix_upic_presentation?lang=en

[11] Jacquemin, G., Coduys, T. and Ranc, M. (May 2012). IanniX 0.8. *Journées d'Informatique Musicale (JIM)*, Mons, Belgium.

[12] IanniX software, accessed March 2019: https://www.iannix.org/en/

[13] Walker, B. N. and Cothran, J. T. (July 2003). *Proceedings of the 2003 International Conference on Auditory Display,* Boston, USA.

[14] Ferguson, J. (2016). Bell3D: An Audio-based Astronomy Education System for Visually-impaired Students. *CAPjournal*, No.20, pp35.

[15] Diaz Merced, W. L. (2013). *Sound for the exploration of space physics data.* (Doctoral Thesis). University of Glasgow.

[16] Diaz-Merced, W. L., Candey, R.M., Brickhouse,N., Schneps, M., Mannone, J.C., Brewster, S. and Kolenberg, K. (2012). Sonification of Astronomical Data. *New Horizons in Time-Domain Astronomy Proceedings IAU Symposium No. 285, 2011. R.E.M. Griffin, R.J. Hanisch & R. Seaman, eds.*

[17] http://archive.stsci.edu/kepler/condition_flag.html

[18] https://keplergo.arc.nasa.gov/PyKEprimerLCs.shtml

[19] http://archive.stsci.edu/kepler/manuals/archive_manual.pdf

[20] https://www.nasa.gov/feature/jpl/20-intriguing-exoplanets

[21] https://exoplanetarchive.ipac.caltech.edu/

[22] Seager, S. & Mallén-Ornelas, G. (2002). A Unique Solution of Planet and Star Parameters from an Extrasolar Planet Transit Light Curve. Retrieved from: https://iopscience.iop.org/article/10.1086/346105/fulltext/

[23] Alapini Odunlade, A. E. P. (2010) *Transiting exoplanets: characterization in the presence of stellar activity.* Doctoral Thesis. University of Exeter.

[24] Winn, N. J. (2010). *Transits and Occultations*. Retrieved on March 2019 from: https://arxiv.org/abs/1001.2010

[25] CSound software, accessed March 2019: http://www.csounds.com/

[26] CSound and Python API, accessed March 2019: http://floss.booktype.pro/csound/c-python-in-csoundqt/

[27] Cabbage software accessed March 2019: http://cabbageaudio.com/

[28] Steinke, G. April 1966. Experimental Music with the Subharchord. Subharmonic Sound Generator, *Journal of the Audio Engineering Society*, vol.14, no. 2, pp. 141.

[29] Vercoe, B. MIT Media Lab et al. *The Canonical Csound Reference Manual*. Retrieved from: http://www.csounds.com/manual/html/

[30] Boulanger, R. (Ed.) (2000). *The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming.* Cambridge, MA, USA: MIT Press.

[31] McCurdy, I. accessed March 2019, http://iainmccurdy.org/

[32] McLean, A. and Dean, T. (Editors) (2018). *The Oxford Handbook of Algorithmic Music,* Oxford University Press, New York, USA, ch.21.

[33] Herman, T. June. (2008). Taxonomy and definitions for sonification and auditory display. *Proceedings of the 14th International Conference on Auditory Display*. Paris, France.

[34] Herman, T., Hunt, A. y Neuhoff, J. G. (Eds.) (2011). *The Sonification Handbook*. Logos Verlag, Berlin, Germany.

[35] De Campo, A. (February 2009). *Science by ear. An interdisciplinary Approach to Sonifying Scientific Data.* (Thesis). University for Music and Dramatic Arts, Graz.

[36] Vogt, K. (June 2010). *Sonification of Simulations in Computational Physics.* (Thesis). Institute for Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria.

[37] Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J. H., Miner, N. and Neuhoff, J. (2010). Sonification Report: Status of the Field and Research Agenda. *Faculty Publications, Department of Psychology. Paper 444*. Retrieved from: http://digitalcommons.unl.edu/psychfacpub/444

[38] http://caleb.eastern.edu/model_display.php?model_id=446

[39] http://caleb.eastern.edu/

[40] https://irsa.ipac.caltech.edu/data/SPITZER/CSI2264/lcurves_corot.html

[41] Quinton, M., McGregor, I. and Benyon, D. (June 2018). Investigating Effective Methods of Designing Sonifications. *Proceedings of the 24th International Conference on Auditory Display. (ICAD),* Michigan, USA.

10. ACKNOWLEDGMENT

This research has made use of the NASA/ IPAC Infrared Science Archive, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

Special thanks to Helen Gräwert for reading data driven projections and to Ruth Capó Mesa for listening to the stars.

EXPLORING SONIC PARAMETER MAPPING FOR NETWORK DATA STRUCTURES

Brian Hansen, Leya Breanna Baltaxe-Admony, Sri Kurniawan, and Angus G. Forbes

Computational Media Department
University of California, Santa Cruz
Santa Cruz, California, USA
{brmhanse,bbaltaxe,skurnia,angus}@ucsc.edu

ABSTRACT

In this paper, we explore how sonic features can be used to represent network data structures that define relationships between elements. Representations of networks are pervasive in contemporary life (social networks, route planning, etc), and network analysis is an increasingly important aspect of data science (data mining, biological modeling, deep learning, etc). We present our initial findings on the ability of users to understand, decipher, and recreate sound representations to support primary network tasks, such as counting the number of elements in a network, identifying connections between nodes, determining the relative weight of connections between nodes, and recognizing which category an element belongs to. The results of an initial exploratory study (n=6) indicate that users are able to conceptualize mappings between sounds and visual network features, but that when asked to produce a visual representation of sounds users tend to generate outputs that closely resemble familiar musical notation. A more in-depth pilot study (n=26) more specifically examined which sonic parameters (melody, harmony, timbre, rhythm, dynamics) map most effectively to network features (node count, node classification, connectivity, edge weight). Our results indicate that users can conceptualize relationships between sound features and network features, and can create or use mappings between the aural and visual domains.

1. INTRODUCTION

A network data structure is an arrangement of data into interconnected groupings of information (nodes) according to relationships between groupings (edges). Network data structures are an integral component of our daily lives. Social networks facilitate personal and professional communication. The internet, a network of linked documents, is a ubiquitous utility used in nearly every facet of contemporary life. Transportation networks, such as subway maps, are used by millions of people who rely on these networks for their daily commute. Similarly, in data science, the "hairball"— a densely connected network— has become the dominant icon for the information age, describing the need for analysts to invent new methods to untangle the complex relationships between data points [1].

Sonification has the potential to play a key role in illuminating relationships present in network data structures. In social networks, a sonic queue could signify a group of friends or social connections with common interests and be assigned a unique *earcon* [2]. With respect to website hierarchies, there is often a lack of topological orientation present when an individual visits a web page. A sonification could give the user a sense of place within the website topology, facilitating more accurate and relevant navigational decisions. Sonification has already been used to help passengers interpret navigational systems. Japanese composer Minoru Mukaiya has composed over 100 unique jingles for different train stations throughout Tokyo which are played each time a train leaves the station. Each jingle acts as an earcon, conveying a range of information. For instance, a crescendo and rising pitch in the Shibuya station departure song represents the train's uphill journey to the next platform. The melodies themselves are strongly mnemonic and reinforce the passenger's awareness of their location. Moreover, the jingle for each station along a route can be concatenated to form a coherent song, providing confirmation to a passenger where they are headed and at what point of the journey they are in [3].

Despite existing examples of and speculative uses for network sonification, there is a lack of research on the ability of users to create useful mappings between network elements and sonic parameters. In this paper, we present an initial investigation into how network data structures could be effectively sonified. Our contributions include: *a)* a characterization of the challenges unique to network sonfication (Sec. 3); *b)* a formulation of initial hypotheses that incorporate these challenges (Sec. 4); *c)* the results of two qualitative pilot studies (Secs. 4.1 and 4.2) that assess user interpretation of sonic parameters mapped to network features; and *d)*, a delineation of themes extracted from user responses that identify representational elements that many users expect, and that point to potentially useful avenues for future study (Sec. 4.2.3). Our results indicate that users can create an effective mental model of a sonified network structure. Furthermore, we find that users can make meaningful associations between network features and sonic parameters.

2. BACKGROUND AND RELATED WORK

Network data structures are commonly represented by node-link diagrams consisting of nodes and edges, where the nodes are positioned either to facilitate readability, or based on a particular grouping criteria, and the edges depict a connective relationship between the nodes [4]. This approach was popularized by Jacob Moreno as early as 1932, in which he formalized graphical characteristics to represent actors and relationships in social networks [5]. Visual characteristics used in Moreno's drawings include arrows

https://doi.org/10.21785/icad2019.055

to indicate the direction of nodal connections, colors for multiple layers of nodes and nodal connections, shapes of nodes to communicate characteristics of social actors, and variations in the nodal location to emphasize structural features of the data.

Network graphs are often multivariate in nature, and mapping parameters to particular sounds has the potential to enhance the interpretation and analysis of network features. Prior work in sonification has shown success in this regard, as the properties of audio allow the presentation of multiple dimensions without information overload for users. This is demonstrated in one of the earliest studies of sonification by Pollack and Ficks [6], who evaluate mappings of multidimensional data onto sound. They measure information transmitted to subjects as the sum of the number of bits in each correctly identified dimensional level and find that multidimensional audio displays outperform uni-dimensional displays. Yeung [7] presents a parameter mapping between seven sonic parameters and seven chemical variables, utilizing two pitches, loudness, damping, direction, duration, and rest (silence) to represent the structure of a chemical. Upon hearing the sonification, classification occurs with a 90% success rate before training, and increases to 98% after training.

3. CHALLENGES

Given that sound has shown to be successful in representing multivariate data, our goal is to determine sonic parameter mappings appropriate for network data structures. The relational aspects among nodes and edges pose a unique challenge for our sonification. Specifically, this includes representing nodal connectivity, location, and orientation. In addition, it can be advantageous to have an overall impression of the network architecture, allowing for the acquisition and recall of the structural topology.

The size of the network structure is also of important consideration. A very large network with numerous nodes and a highly complex edge topology may yield a sonification that is too acoustically saturated for interpretation. In this case, the sound may be best utilized to convey the gist of the overall data structure [8]. Alternatively, if examining a network structure that is either very small or at a highly localized level, a sonification may saliently communicate the low level details present in the data.

3.1. Connectivity

In representing network connectivity, fundamentally we need to determine the best sonic representation for a connection between two nodes. This could be accomplished numerous ways including via musical texture, articulation, or the production of sound effects. For example, if two nodes are represented as sequential tones separated by time, then the melodic transition between them could represent the presence or absence of a connection. A legato or glissando articulation between them would indicate a nodal connection, while the presence of silence a detachment. Alternatively, two connected tones could sound simultaneously producing a harmonic texture. Further, a sound effect could be introduced between the tones signifying their connection. For example, a synthesized Doppler effect could indicate not only that tones are connected but also give the impression of the connective direction.

The problem of representing nodal connectivity becomes increasingly difficult when considering the addition of numerous nodes and the combinatorial possibilities for their connections. Connections may be unidirectional or bidirectional. Simultaneous connections may exist among the nodes as one-to-many or many-to-one mappings. Groups of nodes can be chained together generating a higher order of connectivity. The ensemble of possibilities has the potential for yielding a highly saturated sonic display that could be very difficult to interpret. Thus, a sonification of network data structures must carefully consider how the multiple types of connections are displayed.

3.2. Location and Orientation

Nodal location and orientation plays an important role in depicting nodal relationships and an overall network architecture. This poses a unique challenge for data sonification because nodes are commonly represented in a geospatial context, with their location and relative orientation displayed in two or three dimensional space, and exclusive of a relationship to time. These facets pose a challenge for our sonification because the perception of sonic events depends primarily on the temporal domain, where the majority of sonifications involve the presentation of data as a sequence of sonic events over time.

Although the vast majority of sonifications involve time based representations, there has been successful exploration using sound to communicate spatial data. For example, Flowers, Buhman, and Turnage used the dimensions of frequency and time to display 2D scatter plots of data [9]. In addition, Alty and Rigas devised the tool AudioGraph that paired notes to represent geospatial data points. In their representation, timbre indicated a particular axis and frequency the distance along that axis [10, 11].

Further, in the area of cartography, Schito and Farikant utilize parameter mapping sonification to represent digital elevation models, where the sonic parameter of pitch was shown to be the most successful in accurately interpreting sonic displays. Krygier proposes a set of nine "sound variables"— including sound source location, loudness, pitch, register, timbre, duration, rate of change, order, and envelope— that could be used to represent spatial data [12]. Krygier paralleled his approach to the semiotic system for graphics previously established by the cartographer Bertin [13, 14].

Questions remain as to how effective such a geospatial sonification can be. In comparison between sonic and visual mappings, visual representations of spatial data are much more accurate. In a visual map, data points can be precisely plotted on an x, y coordinate plane and their location can be clearly comprehended. Compared to source localization of sound, a listener's notion of the sound object's position is much less accurate [15].

As a possible solution, the temporal characteristics of sound could be mapped to geospatial metrics. For example duration, measured in a unit of time such as beats per minute, could represent the distance to a sonically positioned object. In addition, meter, which is often conceived as a one-dimensional grid, can be used to quantify distance. Rhythm could also be used to quantify distance and could be expanded to multiple dimensions via the notion of polyrhythms, where this could help represent distances in a multi-dimensional space.

Although a network data structure is often times presented spatially, the spatial relationships presented may not exist in reality. For example, social network visualizations may show nodes of individuals and their interconnectivity organized into groupings on a two dimensional space. In reality, the orientation of the individuals (that is, the physical location of the individuals in the real world) has no relationship to nodal orientation displayed on

a graph. The importance of the spatial representation is unique to data visualization, in that it acts as a means to communicate data groupings. When it comes to sound, groupings can be represented in a similar fashion, perhaps by assigning a unique timbre or register to the data points.

Regardless of the approach, it is important for us to distinguish which features of our sonification require geospatial precision and which do not. If certain features can not be precisely represented via sound, then the question may be to what extent can the information be conveyed? For example, given the geospatial placement of a sound source, we may not be able to determine its precise distance from us, but we can at least perceive that it is either close or far. Additionally, given the placement of multiple sources, we may at a minimum be able to determine which source is relatively closest or furthest. If we must accept that geospatial relationships cannot be perceived accurately, then we may be required to deemphasize this feature in the sonification.

3.3. Topological Impression (Acquisition and Recall)

It is important to perceive the overall impression of a network topology. The minutiae of relationships within network structures are distinct and complex. Taken in aggregate, the impression of the overall structure is an identifying principle that signals the general relationships among its more detailed components [16]. As such, a higher level impression of the structure enhances our ability for knowledge acquisition and recall, allowing us to more efficiently identify, categorize, and compare network data sets.

Musical structures, such as a melodic phrase, are similar in that a lower level of multivariate sonic information is encapsulated in the higher level structure. When a person hums a tune, they are referencing an abundance of parametric data with an elaborately organized collection of pitches, durations, onsets, rhythms, articulative effects, and dynamics. All of these features are efficiently encoded within the musical structure which can be commonly recalled to a high degree of accuracy.

We look to take advantage of this facet via a successful parameter mapping, where upon transforming a network data set into a higher level musical structure, the network features will be more efficiently acquired and recalled. To accomplish this, we looked to utilize the musical genre of a *jingle*. Most commonly employed in advertising, jingles exhibit strong mnemonic qualities, facilitating learning and recall [17]. Yalch presented experiments with jingles where it was shown that they are highly effective in low-exposure advertising (low frequency of slogan occurrence), thus demonstrating the strong recall qualities of the genre [18]. Jingles are "catchy," consisting of simple musical phrases that are easily sung and can be recalled with a high degree of accuracy. The compositional makeup of a jingle supports recall through its frequent use of the pentatonic scale (the most universally utilized scale) and the 4/4 time signature (the most common musical meter), its brevity (usually no more than two bars in length with a highly limited set of melodic notes), and by being registrally within a nominal singable range.

4. USER STUDIES

In this section, we describe two pilot studies that each explore parameter mapping between jingles and small networks. To address the challenges in determining which musical parameters are best suited to represent elements of network data structures, we consider the following questions related to nodes, edges, and recall:

Nodal content

- How many nodes are present in the representation?
- How does the representation convey nodal position?
- How does the representation convey qualitative aspects of the nodes?

Edge content

- How many edges are present in the representation?
- How does the representation convey connectivity between edges and nodes?
- How does the representation convey qualitative aspects of the edges?

Mnemonic strength of the representation

- How easily and accurately can knowledge of the representation be obtained and recalled?

Based on our survey of the literature and our own experience conducting research in information visualization and data sonification, we formulate the following hypotheses:

Hypothesis 1 (H1): Subjects can conceptualize a general representation of a network graph upon hearing a musical example.

Justification: A mental construct is formed when a person is presented with a sonic stimulus. Musical structures, being sonic stimuli, consist of sonic elements that are grouped into specific relationships. As network data structures exhibit similar features, they can be conceptualized by a musical representation.

Hypothesis 2 (H2): Subjects can conceive specific and meaningful correspondence between musical and graphical features of a network structure with only minimal exposure to the musical excerpt.

Justification: Musical phrases, in particular jingles, have demonstrated a strong mnemonic quality allowing a person to acquire and recall structural details to a high degree of accuracy. With the musical structure of a jingle internalized, subjects can recall and continuously reference the structure, allowing them to formulate meaningful correspondence with a network graph. Due to the mnemonic quality of a jingle, this can be accomplished with minimal exposure to the musical source.

Hypothesis 3 (H3): When presented with a mapping between a musical example and network graph, subjects can identify correspondence between musical and graphical features.

Justification: Sonification has a demonstrable history of success with respect to parameter mapping. In such cases, subjects are able to associate musical and graphical parameters.

Hypothesis 4 (H4): Subjects have a preconceived notion of how musical parameters correspond to features of a network graph.

Justification: Musical features are described in a qualitative and quantitative fashion. This implies musical perception consists of associative structural descriptors. As such descriptors exist prior

to knowledge of a parameter mapping to network data structures, subjects will have a preconceived notion of how musical features ought to correspond to features of a network graph.

Hypothesis 5 (H5): Subjects can consistently justify their association of musical examples with network graphs when mapping musical parameters to network structural features.

Justification: Effective parameter mapping sonifications have demonstrated a high degree of accuracy and consistency when sonic parameters intuitively map to particular data elements. The presence of highly salient parameter mappings generates consistency, allowing mappings to be preserved over numerous domains.

To validate these hypotheses, we conducted two pilot studies to obtain qualitative feedback on the relationship between musical and network structure representations. For each study, we composed a set of jingles containing musical elements with the potential to correspond to features of a small, highly localized network structure. Each jingle emphasized a specific musical parameter, such as pitch, melody, harmony, rhythm, timbre, or dynamics. For example, one jingle highlighted dynamics, containing notes that are either quiet or loud. Other jingles presented harmonic vs. monophonic content, while yet others contained obvious drastic changes in timbre among melodic tones. As we describe below, each pilot study included a range of tasks in which subjects were presented with jingles and asked to make correspondences between the musical elements and network features. Their feedback gives us insight into sonic parameter mapping and provides initial empirical evidence about which feature correspondences are meaningful.

4.1. First Pilot Study

Our first pilot study aimed to discover if individuals could conceptualize networks as sound and to explore an individual's mental model of a sonified network. We were also curious to see if individuals could intuitively understand the mapping between precomposed jingle-network pairs. The study consisted of three tasks and was conducted as follows:

4.1.1. Participants

This pilot study was conducted with six individuals - five males and one female. Four were graduate students in music, one was a graduate student in computer science, and one was a professor of music. All participants had a strong background in both music and technology, but had varying experience working with abstract network representations. An introduction to networks was presented in order to ensure that everyone was familiar with fundamental network concepts.

4.1.2. Methodology

To begin, participants were shown examples of common network visual representations (an organization chart, a bus map, and an abstract node-link network diagram), preparing them to participate in three tasks. A detailed explanation of network features, including core concepts such as connectivity, edge weight, and node value, was given between Task 1 and Task 2 so that the immediate inclination of participants in Task 1 would not be influenced. A post-study survey was conducted to obtain participants' musical experience and familiarity with networks.

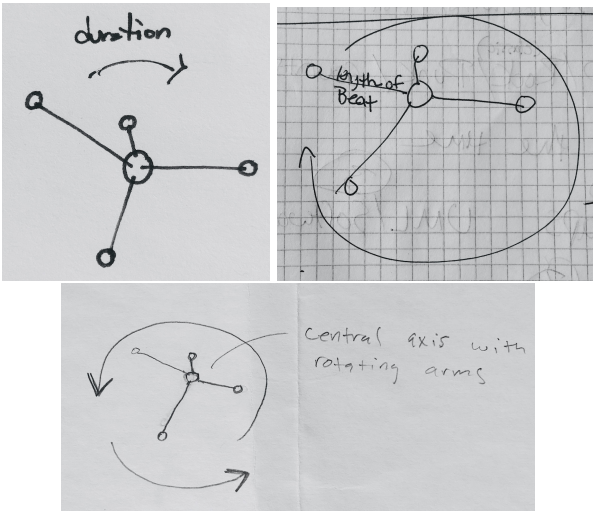


Figure 1: This figure shows examples from selected participants' feature mapping for a graph and jingle pair from Pilot Study 1, T3.

Task 1 (T1) - Drawing a Network:
The first task contained two subtasks (Subtask 1 and Subtask 2). In each, we played a jingle for our subjects. The jingle presented for this task exhibited common characteristics of the genre (see Sec. Topological Impressions): it was limited to 2 measures of 4/4 time at a tempo of 120 beats per minute, totaling 4 seconds in duration. The texture of the jingle was purely monophonic, comprised of an easily singable melody confined to a nominal register, with a total number of notes limited to 9 or less. This jingle placed particular emphasis on dynamic contrast among the notes.

T1-Subtask 1 (S1) - Memory: Upon hearing the jingle, subjects were then instructed to draw a picture of a network structure that resembled what they heard and justify their reasoning. In this task, we aimed to test knowledge acquisition and memory, and thus asked them to listen to the jingle only once before drawing a graph.

T1-Subtask 2 (S2) - Detail: In the second task, they were instructed to listen to the jingle from Task 1 as many times as they wished and revise their original drawing as desired. This task was designed to see if individuals could conceptualize and generate a meaningful mapping of sonic parameters to visual features of a network.

Task 2 (T2) - Drawing Network Features:
The second task was focused on the number of nodes present and their connectivity. After an explanation of core network concepts, we asked the subjects to draw another network graph based on a new jingle they had not yet heard. The jingle played in this task was similar to that employed in T1, but placed particular emphasis on timbral contrast. We emphasized that they should consider the number of nodes and their connectivity in the drawing and in their justification. This task aimed to uncover sonic inclinations for particular network features.

Task 3 (T3) - Identifying Correspondences between Network Features and Sonic Parameters:
In the third task, we presented a precomposed pair— a graphical representation of a simple network structure and a corresponding jingle which we composed to "match" it. The outer nodes of the graph, shown in Figure 1 were sonically represented by each pitch in a clockwise manner with the arc length corresponding to the duration of each note. This was done to establish the position and orientation of each node in the graph. In this representation, the central node acted as a connectivity hub for the other nodes and was sonically treated as a tonal center (tonic). Nodes that appeared above this hub sounded higher in register, while nodes below sounded lower in register. Nodal distance from the hub was portrayed by pitch, where a higher or lower pitch relative to the tonic, indicated distance extremity. Without revealing this chosen mapping, we asked the subjects to label the network concepts on the provided graph with what they felt was represented by musical concepts from the audio. This task aimed to discover whether individuals could interpret a preconceived mapping without guidance, and uncover any points of friction in our mapping.

4.1.3. Results

In this study we observed individuals create their own graphs corresponding to provided music passages and interpret existing music/graph pairs. This shows us that these individuals were capable of conceptualizing a mapping between sound and network representations, confirming H1.

In T1, 5 out of 6 participants were able to generate some graph after hearing the jingle once. 3 participants were able to provide noticeably more detail to their graphs in T1-S2 with many listens than in T1-S1 with few. Some participants listened all the way through the jingle multiple times before drawing, while others paused throughout playback to draw. Each participant did seem to value the allowance to go back through and listen again, and willingly took time to play the jingle until they felt comfortable with their interpretation. This refutes H2.

In T1 and T2 we found that participants had a tendency to draw networks that closely resemble musical notation. This was unexpected, but reveals the power of a known structure on an individual's mental model. As shown in the example in Fig. 2, almost all participants map pitches as nodes occurring over time with a height corresponding to pitch. One participant had a hard time breaking from this musical structure at all. This may be a result of the subjects' musical background, but music representations such as sheet music and MIDI scrollers are commonplace, and this inclination might be present in anyone who has seen them. Responses to T2 indicate that subjects were thinking about specific parameter mappings. However, the act of drawing graphs for two graph concepts was not an effective way of quantifying the parameter mappings.

For T3, 5 of 6 individuals intuitively mapped both the pitch duration to arc length. 5 of 6 mapped note order to circular node position, but only 3 of 6 correctly identified the mapping to be clockwise (see Fig. 1). However, no participant was able to grasp how pitch corresponded to cardinal node position. Thus H3 is partially supported. This is a promising finding, but indicates that more work is required to identify a set of meaningful mappings. Another interesting finding from T3 is individual's curiosity about the mappings. After the study was completed, participants engaged in a discussion of what they thought the mappings were, and continued inquiry with the experimenter even after the study ended. Some feedback we received was that one participant could have created drawings and mapped features better if they had been told what the mappings were to begin with. From these interactions it seems likely that once mappings have been identified, participants would be able to interpret the sonic data better when the mappings are provided.

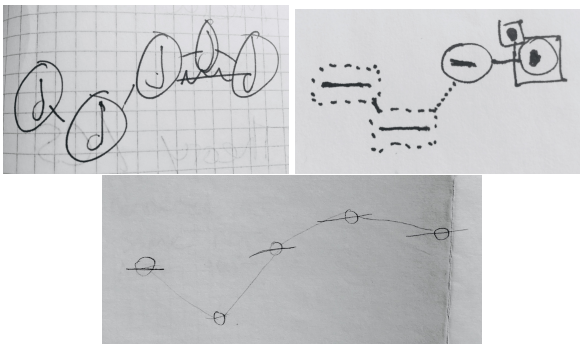


Figure 2: This figure shows selected examples of the graphs which participants drew to correspond with the jingle played in Pilot Study 1, T2. These selected graphs closely parallel music notation, where notes are "plotted" pitch vs. time.

Overall, we concluded that indeed correspondences can be made between musical parameters and network features. However, at the conclusion of this first pilot study, it was not clear to us which parameter mappings were the most effective. This led us to develop a second pilot study that aimed to determine more precisely what constitutes a salient parameter mapping (H4 and H5).

4.2. Second Pilot Study

The second pilot study aimed to find a specific mapping of musical features to network features. As in Pilot Study 1, we again asked users to listen and respond to jingles. Pilot Study 2 consisted of three tasks, followed by a general survey. The study was conducted as follows:

4.2.1. Participants

The study was carried out with a convenience sample of 26 undergraduate participants with a range of musical and computational experience. All participants were students in a course offered within our university's engineering department. This group included 7 females, 17 males and 2 non-binary individuals. 22 of 26 participants have had past experience playing an instrument or singing, with 9 of the 22 consider themselves to be currently practicing musicians. The university majors of this group were (including two students with double majors): Computer Science (9), Art & Design: Games and Playable Media (9), Cognitive Science (6), Computational Media (2), Technology and Information Management (1), and Theater (1).

4.2.2. Methodology

To begin, a presentation on network and music foundations was given to help solidify participant's understanding of each by

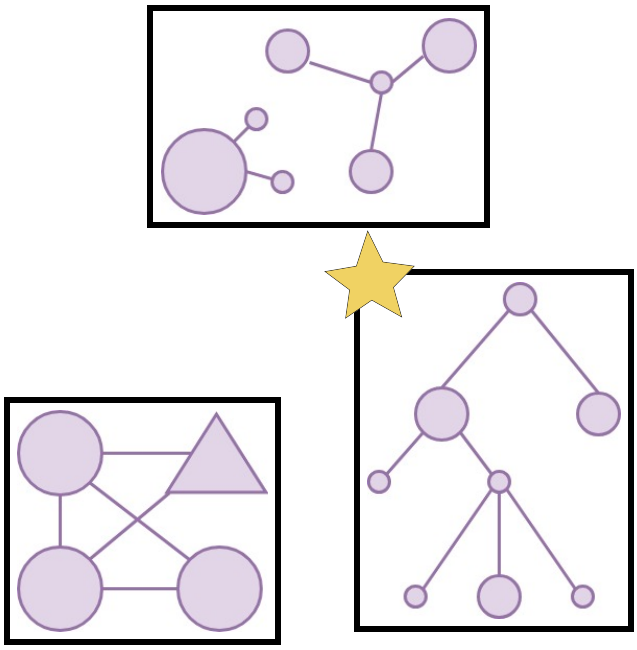


Figure 3: This figure shows examples of the networks used for Pilot Study 2, T2. The network which was most frequently selected to match the jingle played in T2 is starred.

playing different variations of sounds on a speaker in a classroom setting. Key terms for each space were defined during this presentation. Following the introductory material, participants were asked to participate in two tasks.

T1 - Mapping Sonic Parameters to Network Features:
The first task presented the subjects with a range of different network features, including: the number of nodes, how the nodes are connected, strength of the connectivity, and nodal shape. We then asked the subjects to choose which musical feature best represented given network features and instructed them to justify their answer. They were limited to select from the musical features of pitch, melody, harmony, rhythm, timbre, and volume and were confined to choose only one musical feature to represent each correspondence in the feature mapping.

T2 and T3 - Matching a Network Diagram to a Jingle:
In each of tasks two and three, subjects were presented with three unique graphical representations of a network structure. They were then played a jingle and asked to select which of the three graphs most closely matched the musical excerpt they heard. They were also asked to justify their choice. For this study, each jingle was conceived of and composed independently of any concept of a visual representation. This was done to avoid introducing bias in parameter mapping. The jingles were composed similarly to those of Pilot Study 1. However, the jingle in T2 placed particular emphasis on harmonic dyads integrated into the melody, and the jingle in T3 placed particular emphasis on a timbral effect placed on specific melodic notes.

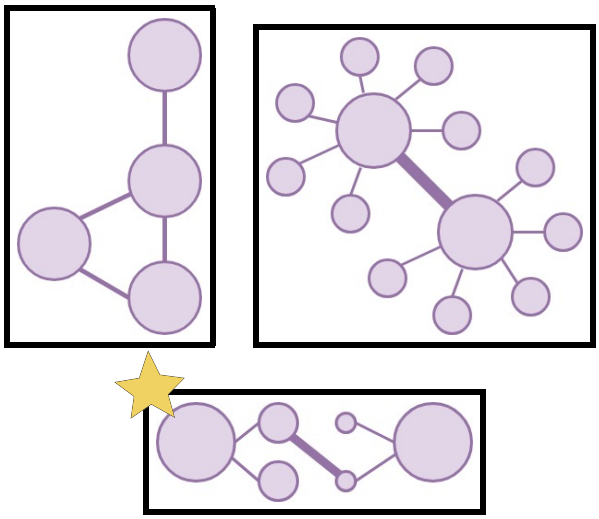


Figure 4: This figure shows examples of the networks used for Pilot Study 2, T3. The network which was most frequently selected to match the jingle played in T3 is starred.

4.2.3. Results

The responses of the 26 individuals were recorded and analyzed using grounded theory [19], where three researchers individually identified emerging themes from the collected data. We report the number of occurrences of each unique response for all short answer prompts which shows general trends for parameter mapping. No further statistical analysis has been conducted due to the small sample size.

We found quantitatively from T1 that subjects have the strongest preconceived notion of mapping for the features connectivity, weight, and shape, which supports H4. However, users were not as effective at mapping the number of nodes. Thus H4 is partially supported. The top scoring mappings for each of the four features is as follows: 21 of 26 participants chose timbre to represent node shape, 15 of 26 participants chose volume to represent edge weight, 13 of 26 participants chose melody to represent node connectivity, and 9 out of 26 participants chose rhythm to represent the number of nodes present. (The remaining values can be seen in Table 1.)

We found from T2 and T3 that although participants were decisive about what musical features should represent graph concepts, those choices did not necessarily remain consistent. We expected individuals to strongly hold to their preconceived notions as exhibited in T1. However, there were many inconsistencies between individuals conceptual choice in T1 and their actions when relating graphs to music in T2 and T3. Further, even when our subjects articulated a reasonable rationale for mapping a parameter to a network feature, those mappings would change when presented with different sounds. Also, in some cases multiple musical parameters were cited as justification for a singular graph, and in other cases a singular musical parameter was cited as justification for multiple graphs. An example of this variability can be seen under the “Note Count” theme, where Participant 18 chooses two possible mappings for note count. Thus, H5 was not fully supported, as each user’s parameter mapping may not be consistent across varying contexts. This indicates a possible level of adapt-

	# of Nodes	Connectivity	Weight	Shape
Duration	4	1	4	0
Rhythm	9	3	0	0
Pitch	6	1	0	4
Harmony	2	7	3	0
Melody	2	13	1	1
Volume	2	0	15	0
Timbre	0	1	1	21

Table 1: This table shows the number of votes for particular musical feature mappings recorded for T1 of Pilot Study 2, including: the number of nodes, whether or not nodes were connected to one another, the edge weight of these connections, and the shape of the nodes.

ability of a user's mental model when relating graphs to audio data. Although our original expectation was not met, the implication that users have the ability to flexibly intuit new mappings appropriate for particular contexts can serve as an advantage, as it allows for a broader range of parameter mapping choices when designing a network sonification.

We were also able to identify several themes in participants justifications which are independent of their graph selection in T2 and T3. Those themes are linearity, character, and note count. Selected quotes and analysis of each theme are detailed below.

Linearity:
A desire for network sonifications to have a notion of "start" and "finish" that aligns to the time dimension of the audio was the strongest theme. This trend was noticeable in Pilot Study 1, where many of the network drawings resembled MIDI musical notation— a well known representation for displaying pitches over time. In addition, this concept was present in participants' answers, regardless of their selection in T2 and T3. This theme also emerged as the most selected node-link diagram for both tasks, as seen in Figs. 3 and 4. In each case the most frequently selected diagram was the most linear of the group, portraying no cycles and exhibiting an obvious location for the "start" and "finish" for each group. Quotes from subjects related to this theme include:

"Notes . . . lasted the longest at the start and the end. The picture . . . when viewed left to right, has larger circles at the start and the end."

- Participant 12 referencing Fig. 4, bottom.

"[When viewed] from the top down the number of nodes will match what is playing [over time]."

- Participant 1 referencing Fig. 3, bottom right.

Character:
Most individuals chose to focus on specific features of the given networks and audio. However, some individuals chose to focus on the overall "feeling" or "character" of the music. Participants created a narrative for the network diagram or related it to known objects. Aside from choosing a specific feature mapping, it may be important to consider how the sonification supports the overall character of the network. Quotes from subjects related to this theme include:

"[The graph] look[s] like those instruments that have beads attached to them that bounce off the drum . . . this [matches] the plentiful, trailing, low volume notes. . . "

- Participant 18 referencing Fig. 4, top right.

"[The] graph is [like a] father and son playing catch. The music gives off that vibe too: peaceful, calm, happy"

- Participant 21 referencing Fig. 4, bottom.

Note Count:
Participants exhibited a tendency to try to match the number of notes in a jingle to a graph feature. The number of notes seems to be a salient feature that individuals want to find specific meaning for. Note count was not present as a musical feature in the mapping step and needs to be explored further as a potentially powerful feature in network sonification. Two examples of note counting follow:

". . . There were 5 notes/chords played which might correspond to number of connections"

- Participant 20 referencing the jingle in Pilot Study 2, T3.

"There were 5 notes played and this graph has the closest number of nodes. There are also 5 links in the graph. . . one link could [be represented by] one note"

- Participant 18 referencing the jingle in Pilot Study 2, T3.

"The beginning sounds like a harmony [of 2 notes] which is held and then changes to one note, converging again to 2 notes. . . "

- Participant 11 referencing the jingle in Pilot Study 2, T2.

5. CONCLUSIONS

Our initial results show that individuals are able to conceptualize a mapping between audio and network representations (H1). Due to our convenience sample, all findings from these studies cannot be said to be universal— our sample size was small and all participants were affiliated with the university. However, we believe that these initial findings can be validated in future studies of network sonification, as well as to inspire the development of network sonification tools.

From our first pilot study, we found that individuals' mental models were influenced by their previous exposure to various music notation (plotted pitch vs. time). We did not find that jingles were memorable enough for individuals to fully grasp a network from only one listening (H2). In this study, we also found that individuals could effectively reason about the mappings between a precomposed network and jingle pair, but that they were not able to confidently identify all feature mappings (H3).

In our second pilot study, we found that individuals could justify feature mappings of unrelated networks (H3), and that they did have some preconceived notions of what network and musical features should correspond to one another (H4). These notions were however somewhat flexible as shown by inconsistency between some of their mappings (H5). More specifically, Pilot Study 2 showed that node connectivity, edge weight, and node shape correspond most popularly to melody, volume, and timbre respectively. Three important themes with respect to participants' mapping justifications were also identified: linearity, character, and note count.

Both studies show that users can create associations between musical and network structures. The results showed attempts made by subjects to formulate specific relationships between the musical and graphical details. For example, in T3 of Pilot Study 2, subjects who selected the second graph primarily did so because they associated the sound effect present with the graphical image. Comments described the sound as "rippling" or "twinkling," and provided a precise description of how this quality represented the nodal structure in the graph. Although this is a promising indication, more work needs to be done with a larger sample size and more rigorous analysis to conclusively identify parametric relationships at a more detailed level. We plan to devise and test a more robust set of parameter mappings and confirm their validity in a future study.

Although a sonification can serve as an aid to the visual representation of a network data structure, ideally the sonic representation would stand on its own. If fundamental aspects of the network could be represented in the audio domain, this could free the visual domain to encode other data elements, or could lead to network displays that are more readily accessible by visually impaired populations. Results from Pilot Study 2 appear to support such segregation, as the preconceived notions that subjects had about a parameter mapping were not reflected by their actual associations selected. Thus, it appears that a person's mental model of a sonic representation is flexible and may be liberated from a pre-existing visual anchor. In the future, we plan to conduct studies to further explore relationships between the purely conceptual elements of a network structure and sonic parameters.

Materials and data from the conducted pilot studies are attainable via the project repository located at `https://github.com/CreativeCodingLab/SonifyingNetworksData/`.

6. REFERENCES

[1] A. D. Lander, "The edges of understanding," *BMC biology*, vol. 8, no. 1, p. 40, 2010.

[2] D. Oswald, "Non-speech audio-semiotics: A review and revision of auditory icon and earcon theory," in *Proc. ICAD*, 2012.

[3] Hermesauto, "The man behind Japan's train departure melodies," Apr 2018. [Online]. Available: https://www.straitstimes.com/asia/east-asia/the-man-behind-japans-train-departure-melodies

[4] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner, "Visual analysis of large graphs," *Computer Graphics Forum*, vol. 30, no. 6, pp. 1719–1749, 2011.

[5] L. C. Freeman, "Social network visualization, methods of," *Computational Complexity: Theory, Techniques, and Applications*, pp. 2981–2998, 2012.

[6] I. Pollack and L. Ficks, "Information of elementary multidimensional auditory displays," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 155–158, 1954.

[7] E. S. Yeung, "Pattern recognition by audio representation of multivariate analytical data," *Analytical Chemistry*, vol. 52, no. 7, pp. 1120–1123, 1980.

[8] T. Hermann, A. Hunt, and J. G. Neuhoff, *The sonification handbook*.

[9] J. H. Flowers, D. C. Buhman, and K. D. Turnage, "Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples," *Human Factors*, vol. 39, no. 3, pp. 341–351, 1997.

[10] J. L. Alty and D. I. Rigas, "Communicating graphical information to blind users using music: the role of context," 1998.

[11] M. Brittell, "Seeking a reference frame for cartographic sonification," in *Proc. ICAD*, 2018.

[12] J. B. Krygier, "Sound and geographic visualization," in *Modern cartography series*. Elsevier, 1994, vol. 2, pp. 149–166.

[13] J. Bertin, *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin Press, 1983.

[14] J. Schito and S. I. Fabrikant, "Exploring maps by sounds: using parameter mapping sonification to make digital elevation models audible," *International Journal of Geographical Information Science*, vol. 32, no. 5, pp. 874–906, 2018.

[15] T. Nasir and J. C. Roberts, "Sonification of spatial data," in *Proc. ICAD*, 2007.

[16] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The Craft of Information Visualization*. Elsevier, 2003, pp. 364–371.

[17] M. Alexomanolaki, C. Loveday, and C. Kennett, "Music and memory in advertising: Music as a device of implicit learning and recall," *Music, Sound, and the Moving Image*, vol. 1, no. 1, pp. 51–71, 2007.

[18] R. F. Yalch, "Memory in a jingle jungle: Music as a mnemonic device in communicating advertising slogans." *Journal of Applied Psychology*, vol. 76, no. 2, p. 268, 1991.

[19] A. Strauss and J. Corbin, "Grounded theory methodology," *Handbook of qualitative research*, vol. 17, pp. 273–85, 1994.

TEXT-DRIVEN MOUTH ANIMATION FOR HUMAN COMPUTER INTERACTION WITH PERSONAL ASSISTANT

Yliess HATI *Francis ROUSSEAUX* *Clement DUHART*

Leonard de Vinci
Pole Universitaire, Research Center
Paris La Defense, 92916, France

URCA CReSTIC
Moulin de la Housse
Reims, 51100, France

MIT MediaLab
Responsive Environments Group
Cambridge, 02139, USA

yliess.hati@devinci.fr francis.rousseaux@univ-reims.fr duhart@mit.edu

ABSTRACT

Personal assistants are becoming more pervasive in our environments but still do not provide natural interactions. Their lack of realism in term of expressiveness and their lack of visual feedback can create frustrating experiences and make users lose patience. In this sense, we propose an end-to-end trainable neural architecture for text-driven 3D mouth animations. Previous works showed such architectures provide better realism and could open the door for integrated affective Human Computer Interface (HCI). Our study shows that such visual feedback improves users' comfort for 78% of the candidates significantly while slightly improving their time perception.

1. INTRODUCTION

Recent developments in the Artificial Intelligence (AI) community – more precisely in Deep Learning – re-enhanced affective computing with the promise of new communication layers between human and machine. This research area explores how computers can sense, analyze, generate, and express affect features as humans do. Sense and analysis of users' emotional states received much attention in the research community, especially for facial expression recognition [1], body gesture recognition [2], speech recognition, and natural language processing [3]. Usually, the identification of complex human affect expressions requires the use of multimodal frameworks or data fusion techniques [4], and the current state of the art allows the software to be responsive to the user's emotional states. However, affective Human Computer Interface (HCI) would also benefit from giving this kind of abilities to the computer. On the one hand, the computer should be able to generate an internal affective state in response to the user's interactions. On the other hand, it should be able to express it more naturally and realistically. Therefore, the user's biases regarding computer interaction could be reduced and would allow more credible communication loop-back between human and machine. Several contributions have studied the use of emotions during interactions with virtual agents, especially in negotiation [5, 6]. At our best knowledge, all of them used hard-coded fixed sets of expressions, whereas human's emotions is a continuous space. Our work aims to provide visual and acoustic expression abilities to computers. Figure 1 presents our workspace setup where users can interact with a personal assistant using our text-driven mouth framework. In this contribution, we propose an end-to-end architecture for voice synthesis with its associated mouth animations. This work is a preliminary step toward affect controls.

Figure 1: Personal Assistant based on our text-driven mouth animation system in a workspace setup.

2. RELATED WORK

Over the last decade, facial animation received considerable attention from industries and research communities for reducing their production cost and their fidelity. At the early days of this filed, facial animations were made by hand and required the expertise of professional animators. The 3D models were first rigged, weighted and then animated frame by frame. This process is time-consuming and varies with the animators. More recently, the use of performance capture allowed to semi-automatize facial animations. Facial movements are recorded with specific software and hardware by tracking markers on human actors in real-time. The collected data is then transferred to the 3D character model. The results are highly dependent on the actor morphology and appearance and require intensive cleaning [7, 8, 9]. Nowadays, thanks to the recent enhancements in Deep Learning, video and audio based approaches are gaining popularity.

The overall process requires the system to produce mouth ani-

https://doi.org/10.21785/icad2019.032

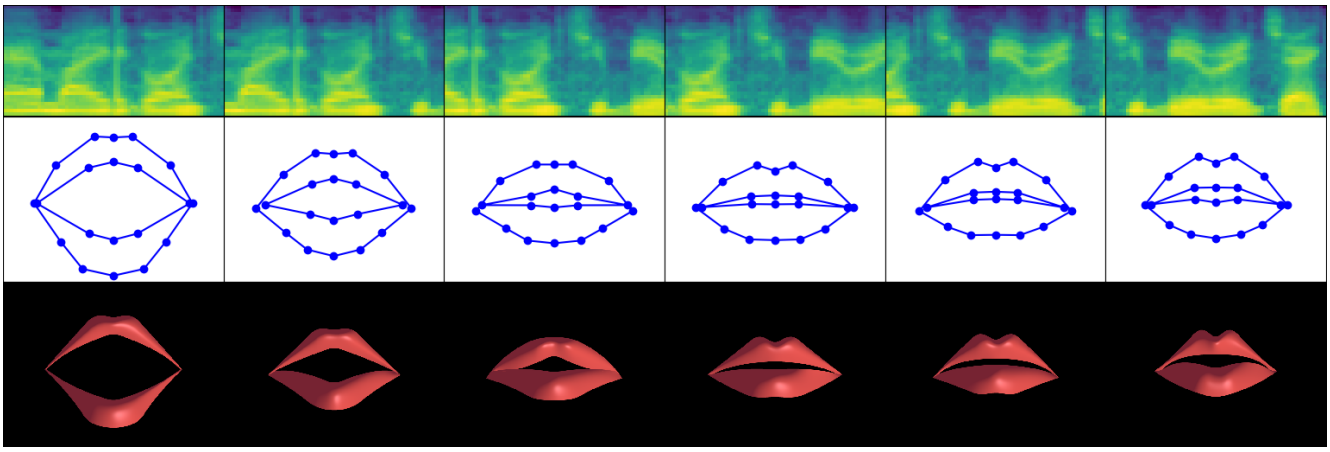


Figure 2: The first line represents the successive 40ms Mel-spectrograms computed from a testing video audio track with the corresponding inferred 3D mouth landmarks on the second line which have been used to generate the final rendering on the last line.

mations synced with the speech. One method is to define a dictionary of visual mouth forms called visemes. Then, the system needs to learn how to associate those visemes with the right phonemes and smooth the result using interpolation. Such approach discretizes the visual and acoustic spaces to bind them. Early works proposed the use of Hidden Markov Model [10, 11] whereas modern ones are using Deep Learning techniques [12, 13, 14, 15]. One limitation is the lack of expressiveness in the mouth articulations, limiting the realism of such generated animations. Moreover, audio and phonemes alignment is a complicated task and is not always feasible.

Nowadays, the Deep Learning community has been able to solve similar issues in translation. Instead of discretizing the input and output space to find bindings, the system learns how to map those spaces together directly. Hence, results are consistent in the output space. In the animation field, recent contributions present significant improvements using such approaches known as audio-driven or speech-driven facial animation. Such frameworks allow to synthesize facial movements and can include an emotional dimension [16, 17]. Unfortunately, the authors did not share their dataset, limiting reproducibility for future contributions.

Recently, ObamaNet from Rithesh and al. introduced the first approach for text-driven mouth animations [18]. Using text as input, their model can generate an audio waveform and its photo-realistic lip-sync frames. Their results are impressive and can lead to future approaches. However, this method is not suitable for other applications requiring the control of a 3D character model.

Our contribution includes:

- A methodology for the creation of text-driven mouth animation datasets from a video bank, including audio. First, the speaker's face is automatically detected and cropped from the video. Then, facial 3D landmarks are identified and projected into an invariant space. Finally, the mouth's key points are extracted and normalized, providing natural synchronicity with the audio.
- A new Deep Learning pipeline for text-driven mouth animation following best practices from state of the art. The speech synthesis module has been replaced by more recent contributions. We also introduce a new module to focus on 3D landmarks controls instead of 2D rendered frames.
- An experiment evaluating how our proposed generated mouth visuals can impact users' comprehension during a listening test and their realism during a blind test compared to landmarks extracted from real videos.

For research reproducibility, we will soon publish our generated dataset as well as our source code with public access.

3. ADMA DATASET

In this section, we describe the methodology used to generate our dataset ADMA-TED for Audio-Driven Mouth Animation (ADMA) TED based on the Lip Reading Sentence (LRS3-TED) dataset [19]. This dataset is composed of 400 hours of TED and TEDx English talks videos distributed over 4004 videos for training and validation, and 451 for testing. Each video varies from 1 to 6 seconds and is cropped on the speaker's face with a 224 by 224 pixels resolution. This dataset has been chosen for its diversity and quality, such as the faces are almost always visible continuously. The rejection rate is lower than 1% and concerns recordings with microphone glitches or where the face is not visible enough. For each video, we provide 20 normalized 3D mouth landmarks every 40 ms to ensure continuous face tracking during the sentence pronunciation, as illustrated in Figure 2.

3.1. Mouth Landmarks

Face Alignment Network (FAN), proposed in Bulat and al. [20], is a popular model for face landmark inference on pictures. It has been applied on the entire LRS3-TED dataset at 25 FPS rate, providing 68 3D face landmarks for each frame as illustrated Figure 3.

Only mouth landmarks are extracted, as illustrated in Figure 4. They are then projected into a head rotation invariant and normalized space. We defined the top of the nose and the chin as our y axis, both eye's landmarks as our x axis and their cross product as our z axis. These axes define the face referential. Using the transformation matrix TM in Eq. 1, the mouth's landmarks are projected into a new front-facing referential. These landmarks are normalized between $[0;1]$ with the center of the mouth located at 0.5 in each axis.

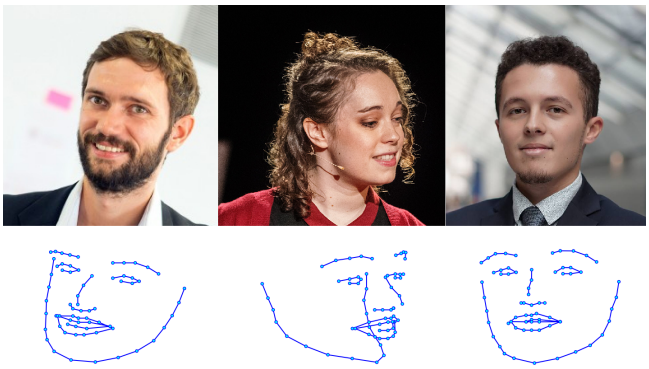


Figure 3: Examples of extracted landmarks in 3D coordinates using FAN algorithms on cropped faces.

$$TM(x,y,z) = \begin{bmatrix} x_0 & y_0 & z_0 & 0 \\ x_1 & y_1 & z_1 & 0 \\ x_2 & y_2 & z_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

Finally, as people possess different types of mouth, identity is removed by computing the median lips thickness over the entire dataset to reduce this bias. Depending on given thresholds, some videos are rejected for being out of the maximum and minimum range of the mouths' opening and closing.

3.2. Audio Processing

Each video's audio channel from LRS3-TED is sampled at $16kHz$. Their Mel-spectrogram are computed using *32 frequency bands*, *128 hop length*, and *512 window size*. Each extracted mouth is associated with a centered *64 window size* of the Mel-spectrogram and filtered to remove high-frequency glitches. Each datum of the dataset contains a spectrogram with its associated mouth landmarks. The dataset is composed of 1,032,219 elements for training and 35,473 for testing.

4. END-TO-END NEURAL NETWORK ARCHITECTURE

Our end-to-end neural network for text-driven mouth animations is based on the architecture proposed by Rithesh and al. [18]. As state of the art has improved over the years, we have upgraded the Text-to-Speech (TTS) module. Our neural module for audio-driven mouth 3D landmarks regression is connected to the end of the processing pipeline, as illustrated in Figure 6. The 20 3D mouth landmarks are used in a 3D engine to compute mouth's vertices and normals frame by frame using 3D spline interpolation, as shown in Figure 5. The final rendering is achieved with the use of a Phong shader.

4.1. Mouth Landmark Regression

This section details the mouth landmarks regression module in charge of estimating mouth 3D landmarks coordinates conditioned on 40 ms Mel-spectrograms from speeches.

For this task, we considered two aspects: the mouth landmarks 3D spatial positions and their dynamics over time. Hence, our model contains three stages. The first stage learns frequency correlations in a compact feature maps representation over time. Then, the second stage determines the correlation dynamic between these step representations. Both stages use convolution layers with rectangle kernels such as time is considered as a spacial dimension instead of a sequence, allowing faster training and inference performances. Stride is preferred over pooling layers to keep as much information as possible. Finally, these compact features are used to train a Mean Squared Error (MSE) regression to estimate the 3D normalized landmarks positions.

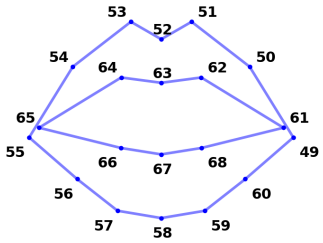


Figure 4: Illustration of the 20 points composing the mouth skeleton with their corresponding FAN annotations.

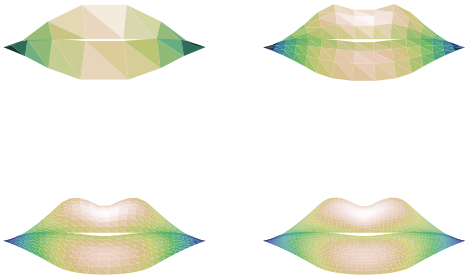


Figure 5: 3D Spline interpolation applied on the 20 mouth landmarks. Each image represent a different interpolation resolution. From left to right and top to bottom each 3D plot respectively correspond to interpolation factors of (8, 1), (16, 4), (32, 8) and (64, 16) with horizontal interpolation first and vertical interpolation second.

4.2. Text-Driven Mouth Skeleton

Rithesh and al. [18] proposed the first contribution of end-to-end neural architecture for text-driven lip-syncing. Their work inspired our proposed architecture. We upgraded their architecture with state of the art contributions and extended it from 2D space coordinates to 3D ones. We replaced the Char2Wav module by Tacotron2 [21] to convert text input into a Mel-spectrogram and WaveGlow [22] for phase reconstruction.

4.3. Neural Architecture Pipeline

Our proposal is an end-to-end neural architecture for text-driven mouth animations composed of different modules. Each one of

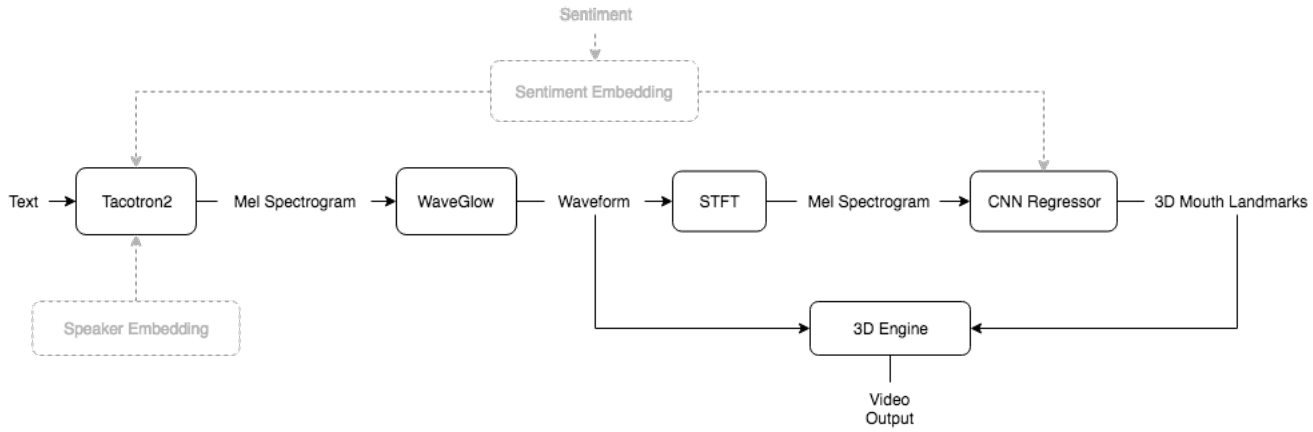


Figure 6: The overall architecture is composed of successive neural modules. The modules Taoctron2 and WaveGlow are in charge of the text to speech transformation. The following ADMANet module computes the 3D mouth landmarks animation based on the speech waveform. Finally, a rendering engine generates the final animation with its corresponding audio file.

those is a crucial step because of cumulative approximation errors throughout the entire model. During development, we tried different approaches for the pipeline. One of them was to connect our CNN regressor directly to the Tacotron2 output. However, the output resolution was not sufficient to allow our model to learn the mouth landmarks properly. The Mel-spectrogram use by our architecture is computed using an STFT on the waveform produced by WaveGlow, as illustrated in Figure 6, to allow control on its resolution.

Layer	Kernel	Stride	Outputs	Activation
Frequency Domain				
Conv2D	-	-	1x64x32	-
Conv2D	1x3	1x2	72x64x16	ReLU
Conv2D	1x3	1x2	108x64x8	ReLU
Conv2D	1x3	1x2	162x64x4	ReLU
Conv2D	1x3	1x2	243x64x2	ReLU
Conv2D	1x2	1x2	256x64x1	ReLU
Time Domain				
Conv2D	3x1	2x1	256x32x1	ReLU
Conv2D	3x1	2x1	256x16x1	ReLU
Conv2D	3x1	2x1	256x8x1	ReLU
Conv2D	3x1	2x1	256x4x1	ReLU
Conv2D	4x1	4x1	256x1x1	ReLU
Mouth Landmark Regression				
Dropout 0.5	-	-	256	-
FC	-	-	256	Sigmoid
FC	-	-	256	Sigmoid
FC	-	-	$20*3=60$	Linear

Table 1: ADMA-Net architecture is composed of three stages. First one is in charge of learning frequency correlations at each time step whereas the second stage learns their dynamics according to the incoming Mel-spectrogram. Finally, the last stage learns a regression between these frequency dynamic feature maps and the mouth landmark positions.

4.4. Training

Our model has been trained for 5 hours over 1000 epochs using a single *Nvidia 980M 8Go GPU*. Using the Adma optimizer with a $1e^{-3}$ *learning rate* and $(0.9, 0.999)$ *betas*, we achieved an error of $1e^{-3}$ on the *trainval* set and an error of $2e^{-3}$ on the *test* set using a simple MSE as our objective function.

5. EVALUATION

To asses the quality of our results, we conducted a blind user study. We wanted to evaluate the realism of our generated mouth animations and the impact of visual feedback on user's time perception and comprehension when listening to a potential personal assistant.

- The realism of our artificial mouth animations is evaluated by comparing them to mouths generated by a landmark tracking system applied to real videos. We considered two scenarios. In the first one, these two video categories have to be distinguished independently. In the second one, both classes have to be discriminated by pairs.
- The impact of visual feedback on user experience is evaluated through a listening and comprehension test. In this test, the user answers questions about recordings with or without mouth animations. The user must also estimate the speech's length which provides indications of its patience level.

5.1. Setup

The evaluations have been conducted in an open space environment on our experimental table presented in Figure 1. Candidates are confronted alternatively to recordings with and without the mouth animations over our different experiments presented as following.

The 24 candidates sampled for this experiment are randomly selected among English-speakers, including native ones, based on their exposure to 3D animations in the form of movies, videos games and modeling independently of their age and gender. We defined exposure as three categories: rare, casual, and daily. Finally, to avoid any bias in the results, professional animators and 3D modelers are excluded from the candidates.

	Age		Exposure to 3D animations		
	18-25	*26-59*	*Rare*	*Casual*	*Daily*
Man	8	4	2	2	8
Woman	4	8	4	6	2
Total	12	12	6	8	10

Table 2: Study candidate composition.

5.2. Realism Evaluation

This evaluation estimates how our system can fool candidates. For a given audio recording, candidates need to identify and classify artificial mouth animations from tracked ones. In this sense, we confronted them to two experiments.

- Candidates are exposed to ten independent mouth animations from 3 to 6 seconds in random order and have to assign them to the appropriate category. Each candidate took 2 to 5 minutes to complete the task. In Figure 7, the median success rate among all profiles is around 50% with a standard deviation of 14.2%. Results show that candidates are not able to correctly identify the videos as being part of one of the two classes.
- To extend our experiment further and assess the robustness of the results, users are then exposed to five pairs of 3 to 6 seconds of videos from each category and have to distinguish them side by side. This task took 3 to 7 minutes to complete for each candidate. In Figure 8, candidates are still not able to discriminate the two mouth animation generation methods.

According to Figure 7, and Figure 8, gender and age do not seem to impact candidates' ability to solve the identification and classification tasks. Contrary, the exposure criteria tend to help the users. Therefore, as illustrated in Figure 9, both experiments confirm that our model can produce mouth animations realistic enough to fool human candidates.

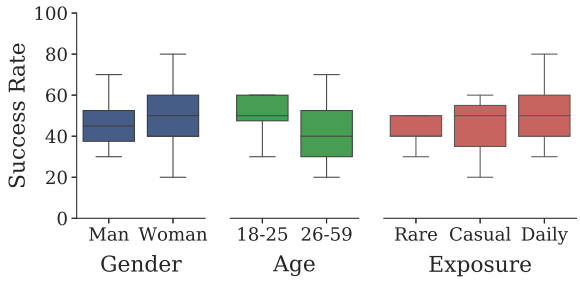


Figure 7: Success rate during blind independent identification test of artificially generated and human tracked mouth animations. Results are displayed per independent discrimination criterion: gender (in blue), age (in green) and exposure to 3D animations (in red). Each box-plot describes the maximum, first quartile, median, third quartile and minimum of the candidates' success rate.

5.3. Comprehension Evaluation

Visual feedback is capable of affecting the human's comprehension capacity. To evaluate the impact of our mouth animations on comprehension, we realized two experiments. Sixteen audio recordings – from 3 to 18 seconds randomly placed in 20 seconds tracks – with or without visual feedback, were shown to the user who needed to answer a comprehension question with four choices. In the first one, the user is asked to answer a question concerning the content of the speech. In the second one, an answer is shown to the user who needs to find the correct item among four possible questions, allowing to perform cross-validation on the visual feedback impact. We expect our mouth animations to alter candidates' cognition.

Results of this evaluation do not show any significant difference when visual feedback is provided or not, thus for any candidate profile.

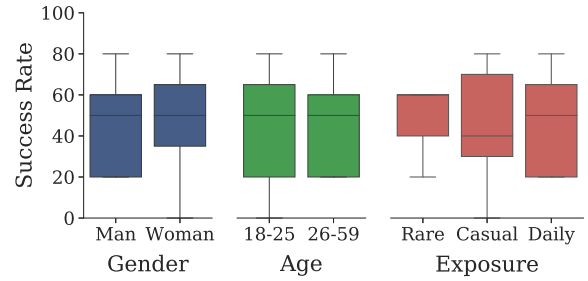


Figure 8: Success rate during blind side by side discrimination test of artificially generated and human tracked mouth animations. Results are displayed per independent discrimination criterion: gender (in blue), age (in green) and exposure to 3D animations (in red). Each box-plot describes the maximum, first quartile, median, third quartile and minimum of the candidates' success rate.

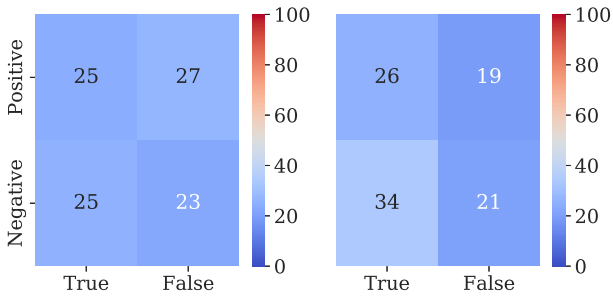


Figure 9: Confusion matrices during blind tests for respectively independent identification (on the left) and side by side discrimination (on the right) of artificially generated and human tracked mouth animations. Each value have been transformed into percentages.

5.4. Time Perception

Time perception can be an indicator of a user's patience during listening tests. Our generated mouth animations are expected to increase people's ability to measure this component by providing visual feedback. To this end, candidates were asked to estimate the duration of recordings during the comprehension tests. The audio recordings from the comprehension evaluation task are independently displayed one by one with or without visual feedback. Both comprehension and evaluation tasks took 7 to 10 min to complete in total.

In Figure 11, results show a small difference between candidates estimation errors with and without visual feedback. Compared to audio only, visual feedback slightly improves candidates ability to estimate the recordings' duration. For more details, we provide a histogram for each question Figure 10. These histograms show that visual feedback tends to increase candidates accuracy for the duration estimation task. Answers are less sparse and more centered around the right one than with audio only.

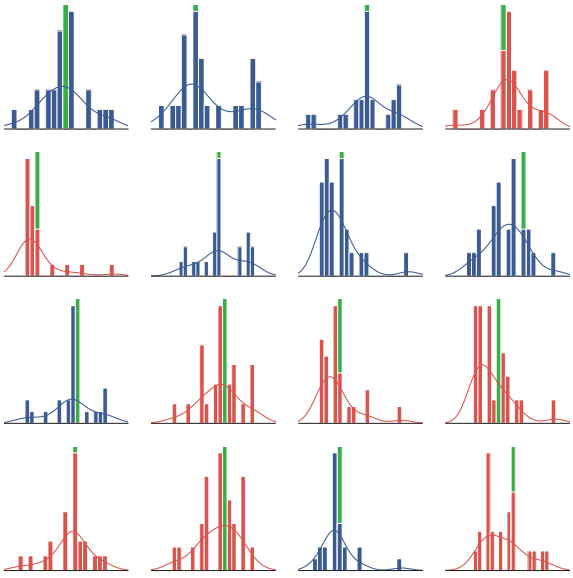


Figure 10: Histograms of speech duration estimation for audio only (in blue) versus audio with visual feedback (in red). Each histogram corresponds to a different speech. Speech duration varies from 3 to 18s. Original answer for the speech duration (in green) and histogram tendency curve can be observed on the graph.

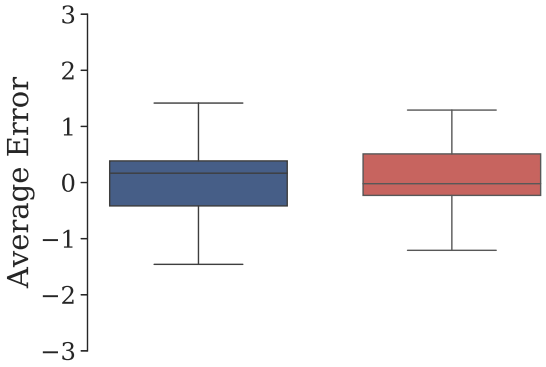


Figure 11: Average error of the speech duration estimation by the candidates during comprehension test between audio only (in blue) and audio with visual feedback (in red). Each box-plot describes the maximum, first quartile, median, third quartile and minimum of the candidates' duration estimation difference from the right answer.

5.5. Interactivity Comfort

The ultimate perspective of this work is interactivity comfort with personal assistants. This contribution focuses on the interest of visual feedback during conversations between users and personal assistants. Hence, it has been challenging to evaluate how visual feedback improves interactions independently of the personal assistant's limitations. In this sense, we let candidates rate their use of a Google Assistant with and without our ADMA interface. The average rating shows 78% of the candidates estimated that comfort is significantly improved with visual feedback, independently of their gender, age or exposure to 3D animations.

5.6. Synthesis

Our proposed evaluation shows that the generated mouth animations combined to speech synthesis are realistic enough to fool, at list partially, human candidates. If initial expectations considered visual feedback as an approach to more natural HCIs, our study could not allow concluding quantitatively either qualitatively such assumptions. However, we observed that the candidates have a better estimation of the recording's length in the presence of mouth animations. Hence, we can assume that visual feedback improves time perception and comfort independently of the gender, age, and exposure to 3D animations.

6. CONCLUSION

In this work, we present an end-to-end Text-driven Mouth Animation model. Contributions in generative animations traditionally benefit the entertainment sector for video games and animated movies. We explore such technology to the use of virtual avatars for personal assistants. Visual feedback often enables more natural ways to interact, especially when combined with audio. Our focus in this work is the generation of text-driven mouth animations for such personal assistants to improve the user experience.

To this end, we proposed a methodology for the creation of datasets dedicated to text-driven 3D mouth animations. This approach does not require the use of performance capture and can easily be extended to future applications. We also developed an end-to-end neural network model combining state of the art in speech synthesis and a custom neural regressor inspired by previous work trained on our dataset. This model allows the creation of 3D mouth animations with its associated speech audio conditioned on unique text input. We also provided an evaluation set to estimate the realism of the generated mouth animations and their value in terms of improvement for HCI. We conclude that the level of realism is good enough to fool human candidates, whereas it does not help in solving cognition task of comprehension. The interest seems limited to improving user comfort.

6.1. Possible Applications

We consider some application domains in which our model and study can be relevant for future developments.

- Video Games and animated movies could use our text-driven mouth animations to animate their characters or as placeholders for preview animations proving a better sense of the final rendering. Compared to performance capture, our system does not require the use of sophisticated software and/or hardware

that some could not afford. Our proposal also provides consistent results and does not depend on the artist's animation skills.

- Art museums and exhibitions could use text-driven mouth animations to renew how visitors can experience art galleries. Text-driven portraits or even text-driven description cards could bring interactivity to traditional audio guides. Such technology could be extended with other visual mouth renderings instead, the one presented in this work.
- People suffering from Autism spectrum disorder (ASD) could also benefit from such technology. Controlled robotic interaction has proven to be less frustrating among these individuals and is used as a new form of therapy. Children with ASD are better at integrating audiovisual feedback compared to real people during social interactions. As our work focus on the importance of visual feedback for the creation of less frustrating natural interactions, our model could benefit those robots and enhance their ability.

6.2. Limitations

Major limitations observed in our work are directly inherited from deep learning sensitivity to the data it is trained on and its heavy computation costs.

- The visual realism fidelity of our mouth animations seems affected by the lack of horizontal movement. Our model does not seem to capture the horizontal mouth movements as good as the vertical ones. To tackle this issue, we plan to retrain our regressor network using different regularization techniques.
- The computation cost of our pipeline is limited by the requirement of a powerful GPU. It cannot be suitable in some applications requiring limited physical space, power consumption, and network constraints. Hence, two aspects of the pipeline could be improved. On the one hand, WaveGlow, used in the TTS module, is a bottleneck for being able to integrate such project on Sytem On a Chip (SOC) devices and could be optimized by different approaches such as batch inference. On the other hand, our visual rendering engine is limited by the mouth's 3D vertices and normals computation occurring on each frame and could be optimized by re-targeting the 3D landmarks on a 3D mouth model instead.
- The unique voice used in our Speech Synthesizer can induce biases and is not suitable for applications requiring different gender, race or age characteristics or even neutrality. We plan to address this issue by retaining and conditioning each module of the TTS on a speaker latent variable using a multi-speaker dataset.

6.3. Future Work

One extension of this work is the introduction of an emotional and prosody context to the entire system. The emotional context of the speaker influences lips movement, speech tones, and styles. We think avatars and personal assistants should be able to express this kind of features and would result in better interactions with humans. Previous works show that a sentiment and prosody embedding or latent context variable can leverage such controls over TTS and Facial Animations independently [17, 23]. We want to extend our work using this approach of a learned latent variable as a way to influence the whole system in its integrity.

7. REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *CoRR*, vol. abs/1804.08348, 2018. [Online]. Available: http://arxiv.org/abs/1804.08348

[2] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[3] F. Weninger, M. WÃűllmer, and B. Schuller, *Emotion Recognition in Naturalistic Speech and Language - A Survey*. John Wiley & Sons, Ltd, 2015, ch. 10, pp. 237–267.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98 – 125, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517300738

[5] C. M. de Melo, P. Carnevale, and J. Gratch, "The effect of expression of anger and happiness in computer agents on negotiations with humans," in *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, ser. AAMAS '11. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 937–944. [Online]. Available: http://dl.acm.org/citation.cfm?id=2034396.2034402

[6] ——, "The effect of virtual agents' emotion displays and appraisals on people's decision making in negotiation," in *Intelligent Virtual Agents*, Y. Nakano, M. Neff, A. Paiva, and M. Walker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 53–66.

[7] W.-C. Ma, M. Lamarre, E. Danvoye, C. Ma, M. Ko, J. von der Pahlen, and C. A. Wilson, "Semantically-aware blendshape rigs from facial performance measurements," in *SIGGRAPH ASIA 2016 Technical Briefs*, ser. SA '16. New York, NY, USA: ACM, 2016, pp. 3:1–3:4. [Online]. Available: http://doi.acm.org/10.1145/3005358.3005378

[8] A. Smith, M. Sanders, C. A. Wilson, S. Pohle, W.-C. Ma, C. Ma, X.-C. Wu, Y. Chen, E. Danvoye, J. Jimenez, and S. Patel, "Emotion challenge: building a new photoreal facial performance pipeline for games," 07 2017, pp. 1–2.

[9] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," in *ACM SIGGRAPH 2011 Papers*, ser. SIGGRAPH '11. New York, NY, USA: ACM, 2011, pp. 77:1–77:10. [Online]. Available: http://doi.acm.org/10.1145/1964921.1964972

[10] D. Cosker, D. Marshall, P. L. Rosin, and Y. Hicks, "Speech driven facial animation using a hidden markov coarticulation model," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1 - Volume 01*, ser. ICPR '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 128–131. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.851

[11] L. Xie and Z.-Q. Liu, "Speech animation using coupled hidden markov models," vol. 1, 01 2006, pp. 1128–1131.

[12] Y. Zhou, S. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *CoRR*, vol. abs/1805.09488, 2018. [Online]. Available: http://arxiv.org/abs/1805.09488

[13] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. M atthews, "A deep learning approach for generalized speech animation," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 93:1–93:11, Jul 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073699

[14] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Trans. Graph.*, vol. 24, no. 4, pp. 1283–1302, Oct. 2005. [Online]. Available: http://doi.acm.org/10.1145/1095878.1095881

[15] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 127:1–127:11, Jul 2016. [Online]. Available: http://doi.acm.org/10.1145/2897824.2925984

[16] H. X. Pham, Y. Wang, and V. Pavlovic, "End-to-end learning for 3d facial animation from raw waveforms of speech," *CoRR*, vol. abs/1710.00920, 2017. [Online]. Available: http://arxiv.org/abs/1710.00920

[17] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 94:1–94:12, Jul 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073658

[18] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *CoRR*, vol. abs/1801.01442, 2018. [Online]. Available: http://arxiv.org/abs/1801.01442

[19] T. Afouras, J. Son Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv e-prints*, p. arXiv:1809.00496, Sept. 2018.

[20] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," *arXiv preprint arXiv:1712.02765*, 2017.

[21] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[22] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," *arXiv e-prints*, p. arXiv:1811.00002, Oct. 2018.

[23] Y. Lee, A. Rabiee, and S. Lee, "Emotional end-to-end neural speech synthesizer," *CoRR*, vol. abs/1711.05447, 2017. [Online]. Available: http://arxiv.org/abs/1711.05447

DATA-DRIVEN AUDITORY CONTRAST ENHANCEMENT FOR EVERYDAY SOUNDS AND SONIFICATIONS

Thomas Hermann

Ambient Intelligence Group
CITEC, Bielefeld University
Bielefeld, Germany
thermann@techfak.uni-bielefeld.de

Marian Weger

Institute for Electronic Music and Acoustics (IEM)
University of Music and Performing Arts
Graz, Austria
weger@iem.at

ABSTRACT

We introduce Auditory Contrast Enhancement (ACE) as a technique to enhance sounds at hand of a given collection of sound or sonification examples that belong to different classes, such as sounds of machines with and without a certain malfunction, or medical data sonifications for different pathologies/conditions. A frequent use case in inductive data mining is the discovery of patterns in which such groups can be discerned, to guide subsequent paths for modelling and feature extraction. ACE provides researchers with a set of methods to render focussed auditory perspectives that *accentuate inter-group differences* and in turn also enhance the *intra-group similarity*, i.e. it warps sounds so that our human built-in metrics for assessing differences between sounds is better aligned to systematic differences between sounds belonging to different classes. We unfold and detail the concept along three different lines: *temporal*, *spectral* and *spectrotemporal* auditory contrast enhancement and we demonstrate their performance at hand of given sound and sonification collections.

1. INTRODUCTION

The human auditory system is an amazing information data processor that has both phylogenetically and ontogenetically shaped to make sense out of the sounding world around us [1, 2]. Thus it is tuned for characteristics of sound as we encounter it in the world, be it music, language, soundscapes or interaction sounds, and it provides a mapping from sound space to meaning, i.e. it enables us to extract relevant information from the sounds. Sonification connects to these perceptual resources by providing a transformation of data into sound such that listening will in turn allow us to learn about the patterns in given data [3, 4]. Let's assume we are given a collection of data sets from patients under different conditions. Ideally, sonifications will represent the data so that meaningful differences in the data are perceivable. However, this requires that sonification designers know the pattern already before creating the sonification. While this may be true for sonifications that communicate information, it is not given in the case of exploratory data analysis [5, 6] where the goal lies in the discovery of hidden/unexpected patterns and which is inductive in nature. Hence by applying any given sonification method we will likely get sounds where a meaningful systematic difference may be not at all perceivable, or strongly masked by other less informative parts. Likewise in sound-based machine diagnostics or in auscultation, the overall sounds may include some features helpful for discrimination, yet they may be masked by acoustic elements that only aggravate discrimination. Luckily we are equipped with powerful perceptual skills for source separation and auditory focussing, yet these also have their limits. In summary, an inevitable problem both in sonification for inductive data mining and in real-world exploratory investigation is that relevant structures can be inaccessible as they are masked by irrelevant noise.

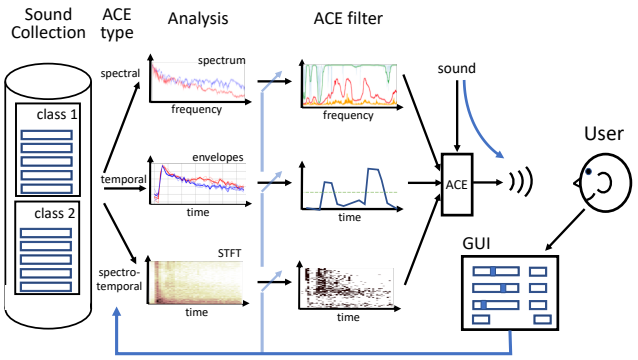


Figure 1: Auditory Contrast Enhancement improves inter-stimulus differences between sounds or groups of sounds.

Individual training, i.e. to rely on auditory learning alone, can empower listeners to better extract information in difficult situations, e.g. car mechanics become experts in associating sound patterns to engines condition, same as trained physicians learn what to attend to in auscultation to diagnose certain heart and chest problems. However, there is another issue: such implicit knowledge will be difficult to communicate to others, we lack a kind of *pointer into auditory structures* compared to the visual modality where we can more easily point our finger and thus share what we regard as relevant. The ACE presented in this paper, with its interactive controls will also serve as a novel kind of 'adjustable pointing device' that can direct novel listeners' attention to the relevant patterns in a complex sound/sonification, and thus help to better deal with the subjectivity of listening which still hinders scientific uses.

The trend in state-of-the-art modern diagnostics in our computer age, however, is to completely abandon the direct sensorial

https://doi.org/10.21785/icad2019.005

contact with the raw data in favour of machine learning and AI to classify data and communicate the results in clear language. In a way, purely machine-learning based diagnosis 'throws the baby out with the bath water' by taking the humans and their domain expertise and broader knowledge out of the loop. This leaves the analyst out of touch with the details on which the classification is based, and it is prone to the risk of false positives and false negatives. However, it would also be suboptimal to leave the potential of machine learning unused. So our approach aims at an enhanced *human-machine cooperation*: (i) machine learning can provide a data-driven enhancement of sounds according to criteria derived from given (e.g. labeled) data. (ii) users can optimise the sounds interactively to further increase their contrast, (iii) this might trigger new ideas about relevant patterns and recursively lead to finer differential diagnosis, and result in suitable enhancement settings for practical applications.

We define *Auditory Contrast Enhancement* (ACE) as a system that transforms given input sound signals into enhanced output sound signals, which facilitates their perception and hence improves the conveyance of the underlying information. We differentiate between two types of ACE.

intra-stimulus contrast refers to the strengths of peculiarity of a single stimulus. It is conceptually similar to the visual domain where contrast refers to the degree to which areas of an image differ in luminance [7, p. 169]. Likewise spectrotemporal contrast can be enhanced in sound. This topic is extensively elaborated in our companion paper [8].

inter-stimulus contrast refers to systematically perceived differences between stimuli from two (or more) groups (A,B,...) accessible and assessed via their A-B comparison. Methods for this ACE will be introduced below.

The following differences further motivate to split the topic of ACE into two papers: as intra-stimulus ACE does not depend on any other data, it can enhance structure from an unfolding sound in real-time, and thus it can serve as a non-parametric post-processing plugin for interactive exploration practises such as percussion & auscultation, and be used in auditory augmentation and blended sonification [9, 10]. In contrast the inter-stimulus ACE depends on given samples at hand of which the detailed processing is crafted. The processing is then applied to either the given input samples, or could also be applied to other independent samples. Interactive uses of data-driven ACE for interactive applications is not so much a focus of this paper, but could be a promising continuation that merges both works.

For inter-stimulus ACE, we distinguish two special cases: (i) *supervised ACE learning* refers to a situation where a number of stimuli are given with their known attributes (e.g. class label), and (ii) *unsupervised ACE learning* has to base the ACE solely on a set of given sounds without knowledge of a ground truth interpretation. The paper mainly unfolds supervised ACE learning and only sketches concepts for unsupervised ACE learning.

Section 2 will formally introduced ACE followed by the presentation of ACE methods in Sec. 3. An implementation of the methods in python will be shown in Section 4. Section 5 will introduce a number of sound and sonification collections and demonstrate the ACE types at hand of these. This will lead to the discussion and conclusion.

Sound and sonification examples are provided as supplementary material via the following DOI: 10.4119/unibi/2935744.

2. DATA-DRIVEN AUDITORY CONTRAST ENHANCEMENT

We define Auditory Contrast Enhancement (ACE) as a system that transforms given input sound signals into enhanced output sound signals, which facilitates their perception and hence improves the conveyance of the underlying information.

We further define *inter-stimulus ACE* as a data-driven method that optimises the enhancement processor from a collection of sound recordings where sounds exhibit systematic differences. We distinguish the supervised learning situation where the correct label or attribute is known and the unsupervised learning situation where no such labels exist. We have the case of classification problems if the label is binary[1], or the case of regression, if a continuous variable describes the variation.

We first unfold ACE at hand of a collection of samples with a binary class label. With this focus, the two main goals of inter-stimulus ACE are (i) to enhance our ability to discriminate sounds that belong to different classes and (ii), to eliminate those parts (spectral, temporal or spectrotemporal) of the sounds that don't contribute to their discrimination. These goals should be reached under the following given conditions:

- the ACE should converge with increasing amount of training data.
- after a training phase, the ACE should be applicable to any previously unseen test data, and thus generalize beyond seen data.
- ACE application should yield a sound that still is an analogic (raw) representation of the underlying stimuli, according to Kramer's continuum [3].
- sound discrimination has priority over structural integrity: it is acceptable if resulting sounds are modified to become even not recognizable as the sounds before ACE application, as long as contrast is increased.

Practically, data-driven ACE is a software method to manipulate given sound signals consisting of parts: (i) a module to analyze a given collection of sound samples, either labeled or unlabeled resulting in a set of ACE features that allow to discriminate between relevant and irrelevant parts of the signal if it comes to perceive inter-stimuli differences. (ii) a module to apply the ACE features to a sound signal. (iii) a user interface that enables users to interactively select the ACE method and adjust any parameters involved in the transformation.

For (i) and (ii), we introduce Spectral ACE, Temporal ACE and the Spectrotemporal ACE in the subsequent sections. The relevant parameters for (iii) will be introduced along. The graphical user interface will be described in Sec. 4.

2.1. Problem Statement and Sound pre-processing

First let's formally introduce some nomenclature. We assume that a sound signal $s[n]$ is given which contains a sequence of sound events that clearly stand out from background noise. We assume that we have $m = m_1 + m_2$ sound events where m_i is the number of events belonging to class i. W.l.o.g. we can assume the sounds to be ordered. We limit our discussion first to binary classification settings, i.e. $i \in \{1, 2\}$. For example consider the case of judging

[1] more generally: discrete, yet in this paper we limit the treatment w.l.o.g. to binary classification problems

whether a wall is hollow or solid behind the wallpaper: we could have 10 impact sounds for knocking on the solid and hollow wall each as our collection.

As a first step we have to extract the individual sound events from the input signal. The onsets need to be properly aligned, at least for some of the ACE methods introduced below to work well.

To this end we compute the signal root mean square (RMS) of 1 ms analysis windows and accept it as event onset if a silence threshold, e.g. -20 dB, is exceeded. The end of an event is defined by the RMS staying below that threshold longer than a given silence time. Events are extended left and right with some milliseconds to make sure no transients are lost. The resulting events are $s_i^{(1)},\ \ i \in \{1, \dots m_1\}$ and $s_i^{(2)},\ \ i \in \{1, \dots m_2\}$. Furthermore, at this time we truncate all sounds to the smallest common duration. Alternatively it would be possible to use zero padding of shorter sounds. As another option, a set of sound files with proper alignment can be directly loaded.

Note that the unsupervised ACE learning will only have a single set of events to work with and no further label, but this is left for Sec. 6.

3. METHODS FOR CONTRAST ASSESSMENT

In this section we introduce three approaches for measuring contrast between groups of sounds: *Spectral contrast* ignores temporal patterns and identifies frequencies at which the groups differ. *Temporal contrast* only evaluates the temporal evolution of a signal and identifies temporal segments at which the groups differ. Finally *spectrotemporal contrast* assesses systematic differences from the time-frequency analysis of signal collections using the Short-term Fourier transform (STFT).

3.1. Spectral Contrast

Yang et al. define spectral contrast as "the decibel difference between peaks and valleys in the [magnitude] spectrum" [11]. This definition, however, refers to contrast within a sample and is what we explore in the companion paper [8]. Here, we have to rethink the notion of contrast from the viewpoint of perceptual contrast between juxtaposed sounds $s_1^{(1)}, s_1^{(2)}, s_2^{(1)}, s_2^{(2)}, \dots$

Obviously we gain perceptual spectral contrast if we attenuate those frequencies at which the two collections of sound do not differ, and if we boost those frequencies at which they do. The spectral ACE thus simply becomes a filter.

This method will work well for instance with sounds whose spectral profile is rather constant. For instance, impact sounds such as hitting a kettle with a stick are characterized by a relatively stable spectrum determined by the physical invariance of the kettle shape. The initial excitation quickly excites a set of rather stable partial tones that decay with time. In contrast, if sounds exhibit substantial spectral changes over time, such as in a piece of music, spectral contrast will be a less usable.

Practically, we compute the one-dimensional discrete Fourier Transform for real input using the FFT algorithm. The complex-valued spectra $S_{j,k}^{(c)}$ for all given events j within class c at frequency cell k are stored for analysis as column vectors in a matrix $X^{(c)}$. The matrix $Y^{(c)} = |X^{(c)}|$ holds the spectral magnitude for all frequencies (in rows) and for all sounds (in columns).

For a given frequency (i.e. row k) the values of $Y_{*,k}^{(c)}$ represent the spectral energy in class c. We assume these values to be independent samples of an underlying unknown distribution. Under the null hypothesis H_0 that there is no difference between group $c = 1$ and $c = 2$ we can ask how likely it is that we observe the empirical means

$$\mu^{(c)} = \frac{1}{m_c} \sum_{i=1}^{m_c} Y_i^{(c)} \tag{1}$$

Under certain conditions, the normalized difference

$$t = \frac{|\mu^{(1)} - \mu^{(2)}|}{\sigma_{\text{err}}} \tag{2}$$

would be student-t distributed, allowing to compute the p-value, i.e. the probability of type-1 error of erroneously concluding a systematic difference while there is none. Hence statistical testing can help to identify if there is enough evidence to assume the spectral energy to be systematically different, or whether observed differences could be simply a product of random sampling.

Let's not engage in deeper statistical interpretation of the value of t and instead use the t-value simply as calibrated indicator for differences: t is simply the difference of the means in multiples of the joint samples' standard error. This is a useful criterion to adopt for spectral contrast. And regardless of any assumptions on the underlying distribution or statistical interpretation we can simply compute the vector $\vec{t}$ of t-values for all rows of $(Y^{(1)}, Y^{(2)})$.

The t-vector is the point of departure for defining the spectral ACE filter. Specifically we introduce two filters:

nonlinear spectral ACE Here we apply a nonlinear transfer function to $\vec{t}$ so that low values are drawn to 0 and large values soft clip to 1. We propose the transfer function

$$T_k = \tanh(g_{nl} \cdot |t_k|)^o \tag{3}$$

for all frequencies k using a user-adjustable nonlinear gain g_{nl} and order o as exponent for supressing frequencies that do not likely contribute to inter-group differences. Before filtering, T is normalized to maximum 1.

median-filtered t Here we first apply a median filter of user adjustable size r to the sequence $|t_k|$ and define the filter as

$$T_k = \text{median\_filter}(|t_k|, r)^o \tag{4}$$

for all frequencies k, again normalizing T to maximum 1. The median filter smoothes the spectral resonances which can be rather sharp, resulting in strong ringing after the inverse FFT. The median filter thus improves the temporal structure of the ACE-filtered sounds.

As T_k has the same dimension as the initial spectral vectors we can obtain the Spectral-ACE-filtered signal by

$$s_{\text{ace}} = \text{irfft}(\vec{T} \cdot \vec{S}) = \text{irfft}(\vec{T} \cdot \text{rfft}(s)) \tag{5}$$

where '$\cdot$' refers to the elementwise product of the two vectors. Note that S needs to be resampled if applied to signals s of different lengths. However, shorter input signals s can be zero-padded so that the available $\vec{T}$ works.

3.2. Temporal Contrast

Sound evolves in time and the temporal evolution of a sound's amplitude is a feature in which sounds can be different. For example two physical objects may differ in their internal damping and thus impact sounds with the objects may lead to different amplitude falloff over time, maybe so faint that we might overhear it.

Another example is cyclical machine sounds, e.g. from a printer or engine where wear and tear might change the friction and thus sound level over the cycle, which in turn would result in subtle difference compared to the sounds of new machines. With temporal contrast we aim at accentuating such moments in time.

To define temporal ACE for inter-stimulus contrast, let's quickly summarize the essential idea of spectral ACE, in order to define temporal ACE in analogy. In spectral ACE, we searched in *spectrum* for evidence that energies are different and accentuated those where a systematic difference was likely and removed the other frequencies. Likewise, for temporal ACE, we can search *along the time axis* for evidence that the energies are different and accentuate those times where a threshold evidence is exceeded, and remove all other times. This translates into two questions: how to define energy difference for a given time, and how 'to remove time'. For the first issue, we see that an instantaneous energy does not exist and is only defined in a short time span. The RMS of the signal is a good estimator. Practically, we use a triangle window of size 256 (i.e. 6 ms at 44100 Hz sampling rate) with a stride of 128. Figure 2 depicts the individual envelopes of 10 impact sounds (5 per group 'on wood' and 'on metal' each). It is visible that there are times at which the amplitudes differ systematically. The thick lines show the mean envelope of the two groups.

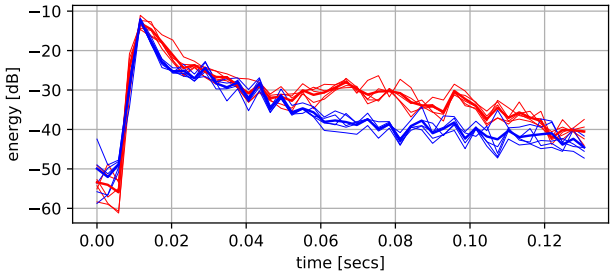


Figure 2: Energy envelopes of 10 impact sounds: 5x 'wood' (red), and 5x 'metal' (blue), the thick lines represent the mean energy over time. It can be seen that there are times at which the values differ systematically.

With these signals we can compute – in analogy to the frequency k in spectral ACE now for each time window n – the vector $\vec{t}$ of t-values

$$t_n = \frac{|\mu_n^{(2)} - \mu_n^{(1)}|}{\sigma_{\mathrm{err},n}} \tag{6}$$

which quantifies the class mean difference in multiples of their pooled standard error.

As to the second question, how to 'remove time', our first attempt was to suppress the energy for those times where t is low. However, this resulted in sparse sounds where much time was wasted with silence in unnecessarily long A-B-A-B comparison sequences. It makes sense to literally 'remove time' as the term indicates and cut and concatenate only those temporal segments in which differences stand out. This results in much shorter sounds, faster to compare and evaluate. Note, however, that this may make a signal completely different and even incomprehensible, for instance if the rhythm matters. However, it saves time if the interest is to discriminate groups.

As soft form of cutting time, instead of a binary decision, we considered a temporal warping that plays signal parts faster as they contribute little and slower as there is more difference. However, this has not been fully tested yet, and may even be irritating as it distorts the temporal structure further.

To decide which time segments to keep, we do not need a nonlinear transfer function as used in spectral ACE, as such a function should be monotonous anyway and thus wouldn't affect the result of a simple threshold operation apart from warping the threshold values as such. Thus we merely select times by taking all windows where $t_n > \theta_{\text{t-ACE}}$. As a rule of thumb, values around 3 would correspond to a 1% chance that the observed difference occurs randomly without significant differences in the means. However, take this with a grain of salt, as in this method a large number of t-tests are computed and no Bonferroni nor other correction is done, so a proper statistical interpretation is not possible.

Furthermore note that a proper temporal alignment and signal normalization is crucial for this method to give meaningful results: a slight shift of the signal in time would create large differences, which of course are not relevant, likewise would a set of louder or more quiet sounds between classes. We currently normalize sound events for peak amplitude 1, yet we see that this is not very robust to outliers. A normalization for the overall event energy as integral over time might be more meaningful in such situations. For stationary/cyclical sounds we recommend to establish temporal alignment from correlation analysis between signal energy amplitudes and choose the lag that yields maximum value.

Note furthermore that the t computation may yield NaN if means are exactly the same, which we replaced by zero for subsequent ACE computation and plotting.

3.3. Spectrotemporal Contrast

The previous two approaches have derived their evidence for systematic differences between the two groups of stimuli from spectral (resp. from temporal) energy alone. Spectrotemporal ACE combines both approaches, yet not in a sequential cascade-style fashion but by directly deriving the ACE criterion from a spectrotemporal analysis of the signals, commonly known as spectrograms. The Short-term Fourier transform (STFT) generates a representation where time windows of the signal are spectrally analyzed and thus a 2D-array of complex numbered activity within all time/frequency cells is computed. The magnitude of these values are the basis for the spectrogram. Fig. 3 depicts the mean arrays for all instances in the impact sounds on wood and metal. As these resolve both spectrum and time they are a more infor-

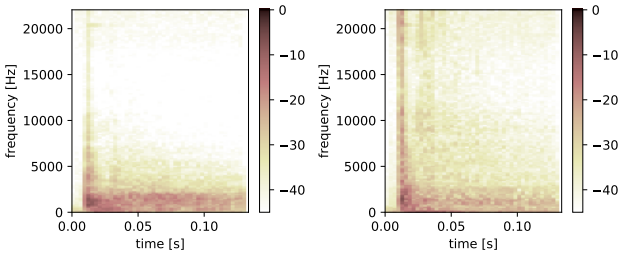


Figure 3: Mean magnitude-STFT-levels (in dB): plots for the impact sound classes wood (left) and metal (right)

mative source for evidence of systematic difference between two given sound collections.

For our signals at sampling rate 44100 Hz, an FFT size of 256 and a temporal stride of half the window size, i.e. 128, is used, using a cosine bell (Hann) window. The resulting spectrograms for example j in class c, named here $S_j^{(c)}[k,n]$ are functions of the frequency cell k and time segment n.

In analogy to the previously introduced ACE approaches we here compute as source for evidence of systematic inter-stimulus variations

$$t_j^{(c)}[k,n] = \frac{|\mu_{k,n}^{(2)} - \mu_{k,n}^{(1)}|}{\sigma_{\text{err},k,n}} \tag{7}$$

with help of the intra-class means of the STFT magnitudes at each given $[k,n]$. Figure 4 depicts the resulting t-array values for the two given vowel sounds in Fig. 3. Obviously the differences in

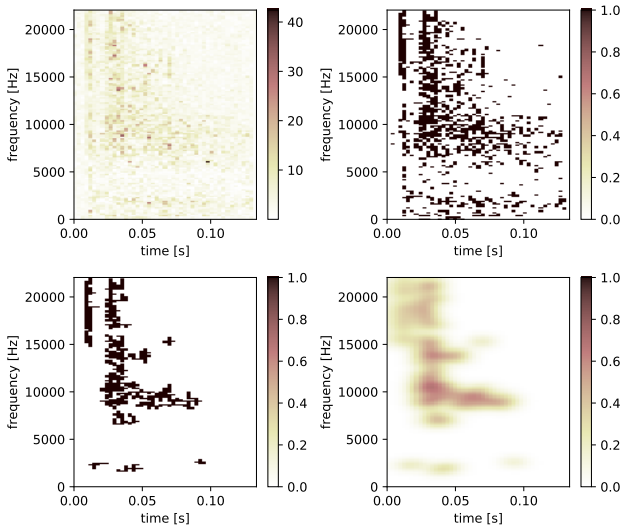


Figure 4: Plot of the 2D array of STFT cell-wise t analysis for wood vs metal impact sounds (whose means are depicted in Fig. 3. Differences in hf-signal stand out.

location and extent of formants are correctly analyzed.

Same as with temporal ACE, a good temporal alignment of any sounds in the two groups is required for the analysis to yield usable results. Such an alignment is easily obtained in the case of impact sounds due to the defining initial transient, and for sonifications where of course the sonification time is well controlled and known. It is less clear how to apply spectrotemporal ACE to stationary, patterned sounds such as cycling machine sounds, yet the same heuristics suggested for temporal ACE can be applied, to shift stimuli onsets in search of the least overall RMS of the t array.

As for the ACE, we proceed in analogy to spectral ACE. We derive a (now spectrotemporal) weighting array $\mathbf{w}$ for all time-frequency cells of the sounds and obtain the enhanced signal by applying the inverse STFT to the weighted STFT array

$$s_e[n] = \text{ISTFT}(\mathbf{w} \cdot \text{STFT}(s[n])) \tag{8}$$

For the weighting array we can take, in analogy to the spectral ACE, a nonlinearly warped t-array, for instance as

$$\mathbf{w}[k,n] = \begin{cases} 1 & \text{if } t[k,n] > \theta \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

which is shown Fig. 4 in the upper right for a threshold $t = 6$.

It turns out that due to the statistical nature of the many tests isolated pixels (cells in the STFT) are frequently supra-threshold. To remove these while retaining all larger blobs we can apply morphological operations from computer vision, namely a binary opening operation followed by a binary erosion and a binary propagation.

The result of these operations is shown in Fig. 4 in the lower left plot. To soften the weighting mask we furthermore apply a gaussian filter (scipy.ndimage.gaussian_filter) to blur with a user-controllable bandwidth σ resulting in a filter as depicted in Fig. 4 (lower left) for $\sigma = 2.5$. Note that σ is in units of pixels and equal bandwidth for spectral and temporal smoothing is currently taken.

Finally, the spectrotemporal ACEd signal is mixed with the original signal, so that any isolated enhancements are better contextualized through the original audio signal.

Note that different from temporal ACE, here no time is removed yet. This would be an additional and optional operation which would require a further criterion for excluding a time frame. The conservative approach would be to exclude only those time segments that do not have a single entry in the ACE mask after erosion. Instead of integrating this into the spectrotemporal ACE, however, an alternative procedure would be simply to cascade the temporal ACE described before.

4. IMPLEMENTATION

We implemented the ACE with python using numpy.ndarrays for audio signal representations and scipy functions to compute spectrum, STFT, t-values, and to apply morphological operations. As standard operations on audio signals is a bit tedious with plain python/numpy/scipy, a dedicated python audio coding package named pyA has been implemented by the first author. It will be made public on github and described elsewhere. With pyA, the necessary operations can be written in a very visible pythonic coding style.

For interactive testing, we developed a graphical user interface within the Jupyter ipython environment as shown in Figure 5. Basically, the user can specify audio files that include the sequence of sounds. Ideally these are separated by some silence or background noise so that the peak finder can identify them. The current code assumes q examples for class 1, followed by any number of examples for class 2. The GUI depicts in the upper row of plots the input signal, here showing 10 impact sounds on a table, five on wood, five on aluminum. The right panel depicts the results of the event finder, blue for class 1 and red for class 2. As the depicted GUI is for spectral ACE, the panel below shows the spectral mean of all q class 1 (red) and class 2 (blue) signals. Note that a truncation to common lengths is applied before this analysis. It can be seen that there are systematic differences. The plot below shows the spectral t analysis, which peaks at different frequencies. The following GUI elements allow to select ACE subtype and parameters, their changes causing an update of the bottom plot of the spectral ACE filter for weighting the original signal before re-synthesis. While much of that can probably be hidden from users in automatic ACE modes, it was helpful to inspect the details for development.

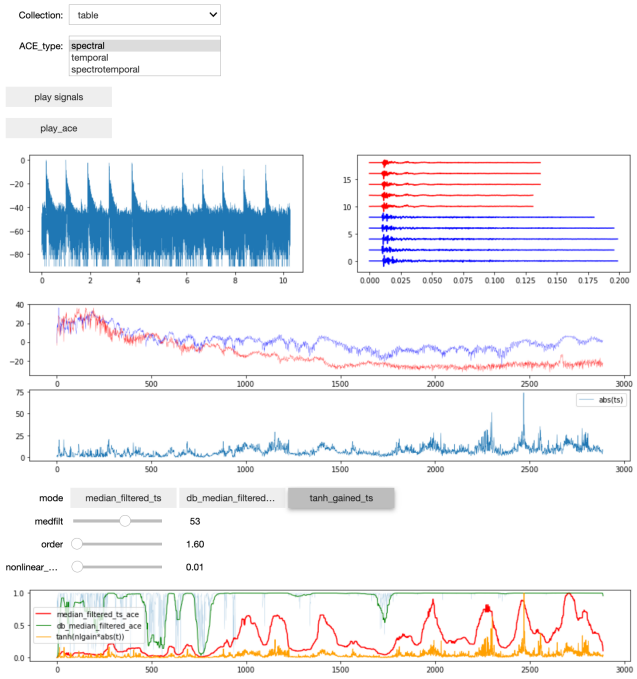


Figure 5: python GUI screenshot: users can select predefined sound collections and the ACE type, then play either original or enhanced sounds. Plots below provide some insight into differences between groups and allow to control ACE-type specific parameters.

5. ACE DEMONSTRATIONS

This section demonstrates the previously introduced ACE types at hand of some examples using different collections of sounds and sonifications.

5.1. Impact sounds

Impact sounds contain a lot of information usually in a very short time of few hundred milliseconds. They convey the material, the object size and resonances, friction and inner properties such as composition. For example, knocking on a half-filled bottle will sound different if filled with water or rice. We see further potential for ACE for auscultation via percussion on the human body.

Here we show ACE performance for the following sound collections (sound examples available on 10.4119/unibi/2935744):

table a collection of impact sounds from hitting a wooden table or on an aluminium laptop with a plastic ball pen used as mallet.

finger snaps a collection of finger snaps from one person using middle vs. ring finger

SC3 klank+noise a collection of noisy synthesized resonator banks using Supercollider's 'Klank' UGen with added Brownian noise, where the two classes differ in two detuned resonance frequencies.

5.2. Continuous sounds

Continuous and cyclical sounds are frequent in machines or rhythmic motor patterns such as sonification of swimming or repetitions of physiotherapeutic practices. For demonstration we here use

vowel a collection of vowel sounds with slightly different articulation place so that it is difficult to distinguish them

music a collection of modifications of a music signals where generally noise is added but temporally localized gains were applied as systematic difference.

5.3. Enhancing the discrimination of impact sounds

Let's start with the 'table' collection and listen to the original impact sound collection in S1.0 [2]. We can clearly perceive the differences without problem, so let's test how spectral ACE will accentuate them. With low order using the median-filtered-ts, we get sound example S1.1 which features very salient differences in response in high frequencies, while low frequencies are attenuated as there is no difference indication between the two groups. This aggravates generally as the exponent parameter 'order' is increased until only narrow ring resonances remain, see S1.2 and S1.3. The nonlinear spectral ACE which uses a $\tanh()$ warping of the absolute t-values is not as radical and leaves more of the low-impact parts intact, depending on the nonlinear gain parameter. However, because of the lacking median filtering, the resonances are very sharp, resulting in long ringing, so that the resonances have almost no damping and thus amplitude differences in these between the two groups stand particularly out at the expense of perceptibility of transients. Next example is S1.4, where a logarithmic (dB) mapping from the ACE feature to gain was applied, but on $(1-p)$-values. This maintains more of the original structure yet accentuates the differences quite well. Note that the same ACE filter is used identically for all 10 sounds, and is also capable to process new not-before-heard sounds and delivers an equally salient perceptual contrast.

Temporal ACE is not as good as expected – however, the physical merely exponential decay doesn't yield so pronounced differences. Sound example S1.5, however, shows at least how focussing on the segments of highest differences results in a notable shortening (and thus speeding up) of collection review.

The spectrotemporal ACE (examples S1.6– S1.9) combine advantages of spectral and temporal ACE in that they are capable to attenuate now spectrotemporal uninterestingness and thus render inter-stimulus differences more salient. S1.6 – S1.7 use the thresholded t-value array directly as filter, whereas S1.8 – S1.9 make use of larger σ to smooth the ace-filter.

The following sound examples are for the finger snap sound collection. Listen first to example S2.0 for the original collection. Examples S2.1 – S2.4 apply spectral ACE with different parameters, while example S2.5 uses the spectrotemporal ACE. The extreme transience of the sounds makes it hard for the STFT-based approach to resolve enough structure, for the same reason the temporal ACE delivers rather poor results.

5.4. Enhancing and Isolating structure from noise

The synthetic Klank sound collection (original sounds in S3.0) features a large amount of noise, capable of masking subtle dif-

[2] All sound examples are provided with description on `https://doi.org/10.4119/unibi/2935744`

ferences between the groups. Apparently, spectral ACE is well capable to attenuate most noise channels, more radical as the order parameter as exponent is increased, audible in examples S3.1 to S3.4. At extreme setting, only the two frequencies which were actually changed between the class 1 and 2 in this collection remain, showing clearly that spectral ACE was successful in finding and highlighting these. Example S3.5 shows that spectrotemporal ACE performs less well here as noise level fluctuations over time result in equal energy over time at relevant frequencies, giving raise to significant temporal structure.

So far spectral ACE seems to have some advantages, but unfortunately it strongly affects the temporal structure due to long resonances.

5.5. Formant changes in continuous signals

The next set of examples test ACE with speechlike sounds. The sound collection 'vowel' features articulatory sounds with slightly different articulation place (listen to example S4.0 for the unmodified sounds). The first 5 are an 'a' as in 'bar' followed by an 'ä' like in 'bear' but tried to articulate more similar to the first vowel. Perceptually they can be quite easily discerned, so let's see how ACE is able to boost the differences. Example S4.1 is the spectral ACE using the $\tanh()$ weighting with low order and low nonlinear gain. It has already shown to give long ringing for resonant frequencies. We hear that the temporal structure is largely lost, but the contrast between class 1 and 2 increases clearly. The same happens with the median-filtered-t-values mode (example S4.2). Here, the differences at high frequencies are strongly enhanced, resulting in audible differences. However the low frequency content did not pass through and thus the original formant structure is barely perceivable. Yet we argue that this doesn't matter if the focus is on classifying sounds as belonging to either the one or other class. In comparison, spectrotemporal ACE (Sound example S4.3) is rather useless and only creates a rougher and noisier version of the sound. The reason for that might be that the STFT number of samples per segment is low with only 256, thus resulting in a poor spectral resolution of 22050 Hz / 256 = 100 Hz, which is perhaps not high enough to distinguish formant differences between 'a' and 'ä'. Temporal ACE did not help at all, so we skip a sound example.

5.6. Enhance Multivariate Time-Series sonifications

Finally we test the ACE on sonifications. A frequent data type are multivariate time-series, such as EEG, ECG, EMG or motion capture sensor streams. For the example here we created a parameter mapping sonification of the building dataset [12] of hourly consumption of electrical energy, hot water, and cold water, time of day, outside temperature, outside air humidity, solar radiation and wind speed for 175 days, all variables scaled to arbitrary units in $[0, 1]$.

Since the purpose of this example is to test how ACE would enhance differences between groups, we created a rather straightforward parameter-mapping sonification of all variables as amplitudes of oscillators tuned to quart-spaced fixed frequencies. We chose 5 sunny days in summer for class 1 (days 13-17 in the building1 dataset) and 5 sunny days in late autumn, starting from midnight on day 141 in the dataset, skipping cloudy days as seen in the measurements of solar radiation, so that the groups are more homogenous. Each day is sonified in about 250 ms from midnight to midnight. The raw sonifications can be heard in sound example S5.0. As the frequencies are constant for each stream, spectral ACE can be expected to provide a good enhancer for systematic activation differences. In fact, examples S5.1 and S5.2 show that the differences in energy in the highest frequency oscillators are significant enough to constitute difference and are thus accentuated. We expected a better contrast in the solar radiation profiles as these are quite different in the different seasons. However, spectral ACE can't see their variation over time and only uses time-free spectral energies, so any differences here do not stand out. That is different in temporal ACE which accentuates certain parts of the signals, see examples S5.3. However, different from intuition which would rather expect differences over daytime to be accentuated, the temporal ACE pronounces differences before sunset and after dawn. The reason is likely that with 0 solar radiation, one variance source within those time windows is reduced, making those times appear more different than those times where solar radiation contributes to variance within the samples of each class. So temporal ACE does work, yet not necessarily as expected. It is a starting point but needs more research to design it to be more sensitive to changes in relevant structures occurring in sonifications. Finally a spectrotemporal ACE example is provided as examples S5.4

5.7. Detecting modifications in longer sound clips

The last example is a set of sounds where a snippet of music was systematically amplified or attenuated at different locations in time in the two classes. Listening to S6.0, the original collection of 2×3 events per class makes clear that it takes a long time to review many sounds, 2.5 s per sound each. Temporal ACE reduces these sounds by removing all those time segments where no systematic differences in amplitude between the two groups can be found, as evaluated by the t-value of the two samples. In turn, the resulting sound is shortened to about 250 ms, depending on the t-threshold, allowing a much faster review of the sound examples S6.1 and S6.2. Also, the differences between the groups becomes clearer: a boosting of the second chunk while attenuating the first between class 1 and class 2.

6. DISCUSSION

We have introduced *data-driven ACE* as a method to automatically modify audio signals so that a contrast between given classes becomes more salient. As first step, we presented spectral, temporal and spectrotemporal ACE and gave exampled for a number of sound collections with systematic differences between two groups. While most sound examples gain perceptual contrast, particularly the spectral ACE has subjectively proven most useful to help discriminating sounds into the two classes. One problem with the current approach is that still the method depends on a number of parameters that cannot be automatically chosen easily, so it requires the human in the loop to tune parameters for good results. Yet this might be acceptable as this would only be required once, e.g. for the designer of a tool to enable machine diagnostics-by-listening. However, it would be certainly nice to integrate some good heuristics for automatic parameter selection, e.g. from testing all parameter combinations on a grid and applying a machine listening based contrast assessment to choose useful initial settings.

Temporal contrast was demonstrated to work, and it is useful to reduce long sound signals into short 'difference thumbnails'

which only present those parts in which differences may lurk.

Interestingly the ACE modifications can be applied to any input signal: spectral ACE independent on signal duration, temporal ACE independent on sampling rate, only spectrotemporal ACE requires matching duration and sampling rate. That means that an ACE trained to enhance contrast between two extremes such as different pathologies reflecting in chest tones, or different materials behind the surface while knocking on a wall, can also be applied to any signal with unknown label. It will be a useful experiment to measure how ACE will reshape the accuracy of classification particularly for sounds that are on the continuum between the two extremes used for ACE training. As a preliminary first test for this, we applied the ACE on a continuous transition of vocal sounds continuously varying from 'a' to 'ä'. Sound examples S4.4 and S4.5 show the original and the enhanced version. We see that more research is needed and more experience needs to be gained with ACE. One possible study could be to ask subjects to assign ACEd sound examples to classes. The mean of ratings for stimuli along the connecting line between class 1 and 2 would probably be somewhat sigmoidal without ACE, yet it should move towards a steeper sigmoidal function with proper ACE.

The data-driven approach to ACE can be extended to work in *unsupervised learning* settings. Consider we have a collection of sounds yet no class label. Assume further that there are systematic differences in the sound features. If intra-class variation is lower than inter-class variation, the first eigenvector of the feature data set covariance matrix (i.e., the first principal axis) should be aligned to the line connecting the centroids of the two clusters. A useful ACE could be derived from that information alone. For instance assume that the features would be the magnitude spectrum components. Then the PCA vector $\vec{u}$ would show certain positive or negative elements. If we take the ACE to be 0 for those frequencies where $|u_i| < \theta$, i.e. is smaller than a threshold θ and 1 else, we would filter out those frequencies that do not change much along the main variance axis of the data. In turn, the remaining frequencies will become more salient. This and more refined approaches for extending data-driven ACE to unsupervised learning remain subject of future research.

7. CONCLUSION

Auditory Contrast Enhancement has been introduced in this paper as a method to process sound in general, and sonifications in particular with the goal to facilitate the perception of relevant sonic differences between selected groups of sound, e.g. created under a different condition. We have focussed on the three special cases of spectral, temporal and spectrotemporal contrast and introduced data-driven enhancements that transform sound in a systematic and reproducible way. The presented ACE processors provide a supervised-learning method yet instead of merely reporting the results as text, they provide an interactive sound manipulation method to better use the human-built-in listening skills to distinguish patterns in data. ACE is capable of removing signal components that apparently do not contribute to any differences between selected groups, and of actively boosting signal parts where differences between groups are likely. In consequence, the resulting signal is less prone to masking. We believe that ACE can serve as a widely applicable sound post-processor for many situations where sounds are perceived in the listening mode of diagnostics or exploration. A thorough psychophysical validation of the method will be required to optimize the methods further, yielding suitable control parameters and interfaces that establish ACE as a standard plug&play post-processing component for sonification tool chains.

8. ACKNOWLEDGMENT

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Thanks to the developers of python/numpy/sciepy/matplotlib/jupyter for their amazing tools.

9. REFERENCES

[1] R. Fay and A. Popper, "Evolution of hearing in vertebrates: The inner ears and processing," *Hearing research*, vol. 149, pp. 1–10, 2000.

[2] E. Hester, "The evolution of the auditory system: A tutorial," *Contemporary Issues in Communication Science and Disorders*, vol. 32, pp. 5–10, 2005.

[3] G. Kramer, Ed., *Auditory Display - Sonification, Audification, and Auditory Interfaces*. Addison-Wesley, 1994.

[4] T. Hermann, "Taxonomy and Definitions for Sonification and Auditory Display," in *Proceedings of the 14th International Conference on Auditory Display (ICAD 2008)*, P. Susini and O. Warusfel, Eds. Paris, France: IRCAM, 2008.

[5] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[6] T. Hermann, "Sonification for Exploratory Data Analysis," PhD Thesis, Bielefeld University, Bielefeld, Germany, Feb. 2002.

[7] A. M. Colman, *Oxford Dictionary of Psychology*, 3rd ed. Oxford University Press, 2009.

[8] M. Weger, T. Hermann, and R. Höldrich, "Real-time Auditory Contrast Enhancement," in *Proceedings of the 25th International Conference on Auditory Display*, Newcastle, U.K., 2019.

[9] T. Bovermann, R. Tünnermann, and T. Hermann, "Auditory Augmentation," *International Journal on Ambient Computing and Intelligence (IJACI)*, vol. 2, no. 2, pp. 27–41, 2010.

[10] R. Tünnermann, J. Hammerschmidt, and T. Hermann, "Blended Sonification: Sonification for Casual Interaction," in *ICAD 2013 - Proceedings of the International Conference on Auditory Display*, 2013, pp. 119–126.

[11] J. Yang, F.-L. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Communication*, vol. 39, no. 1-2, pp. 33–46, 2003.

[12] L. Prechelt, "PROBEN1 - a set of neural network benchmark problems and benchmarking rules," Universität Karlsruhe, Karlsruhe, Tech. Rep. 21/94, 1994.

SOUNDSCAPE CLOCK: SOUNDSCAPE COMPOSITIONS THAT DISPLAY THE TIME OF DAY

Abdullah Ismailogullari

University of Hamburg
Institute for Systematic Musicology
Neue Rabenstraße 13
20354 Hamburg, Germany
a.ismail@mailbox.org

Tim Ziemer

University of Hamburg
Institute for Systematic Musicology
Neue Rabenstraße 13
20354 Hamburg, Germany
tim.ziemer@uni-hamburg.de

ABSTRACT

This paper presents an ambient auditory display that communicates the time of day. Four soundscapes represent different quadrants of the clock. Auditory icons divide the quadrants into three parts that represent hours, and four partitions that represent every quarter of an hour. The auditory display is little intrusive and only informative to those who are privy to its principles. Suitable application areas are offices where staff can derive the time from the soundscape, while customers stay unaware and may only enjoy the calm, auditory nature scene. To experience the calm ambient character of the auditory display we suggest you to play the demo while reading the paper: https://tinyurl.com/y4yd8zkh .

1. INTRODUCTION

Receiving the natural soundscape in urban areas and especially inside buildings can be complicated and sometimes even impossible. Natural soundscapes deliver information about what is happening around us. They consist of daily and seasonal rhythms and patterns [1, pp. 76 – 77]. These patterns provide information like about the time of day, for instance the singing of a rooster in the morning or the seasonal appearance of various animals and their distinct sound. We suggest to generate soundscape compositions that can deliver information like natural soundscapes do.

Mostly, people are used to audible alarms as auditory displays, like the ringing of a phone, that grab our attention immediately. In contrast to that, ambient auditory displays [2] are non-intrusive and should not distract the listener. Hence, careful sound design is necessary to be informative, but at the same time pleasant, unintrusive and not attention-capturing, annoying or even stressful. In this paper, we present an ambient auditory display, that combines soundscape compositions and auditory icons [3] for peripheral awareness of daytime. Staff privy to the auditory display principles can derive the rough time of the day from the sound scene, and extract more precise time information when consciously paying attention. At the same time, the auditory display is non-intrusive. Customers stay uninformed and may enjoy the nature soundscape, while employees are not distracted by the sounds.

2. BACKGROUND AND RELATED WORK

Studies suggest that the impact of nature sound improves the recovery after stress compared to noise [4]. Nature sounds have been shown to mask annoying background sounds in offices [5]. Hui Ma and Shan Shu found that soundscapes had a stronger impact on psychological restoration in simulated open-plan offices compared with visual scenes and both continuous and intermittent had positive effects [6]. Birdsongs have been found to influence perceived naturalness, annoyance, and pleasantness of road traffic noise environments with low sound levels [7]. Fountain sounds sometimes reduced perceived loudness of traffic noise and birdsongs significantly enhanced soundscape pleasantness and eventfulness [8]. Water sound have been found to enhance urban soundscapes the most [9] and the pleasantness of water fountains in public spaces was positively correlated with their temporal variability [10]. So besides perceiving nature sounds as pleasant [11], they offer an opportunity to protect from a less pleasant and potentially distracting effect of environmental noise [12].

Many auditory display studies are concerned with intrusiveness. Fredrik Kilander and Peter Lönnqwist present a concept that uses audio cues to deliver information like incoming mail [13], [14]. The concept is meant for personal and shared space. For the election of sound cues, they attach much importance to intrusiveness of sound. They propose sounds that are easily recognisable and natural to minimise intrusion. Ralf Jung and Tim Schwartz use a location based approach to deliver personalized information [15]. Their concept bases on background music, which is meant to stay peripheral and avoid too much attention. Within the music they embed audio cues consisting of user specific instruments. These are rhythmically and melodically adjusted to the background music, and directed to the position of the user, by playing it through the closest speaker. Hanna Zoon, Saskia Bakker and Berry Eggen present with Chronoroom Clock [16] a study on sonification of time. Their design provides the time of day based on the location of a sound source. It consists of many small piezo speakers on the wall, that are placed in a circle. Each plays a sound for a certain time, then the sound is played by the next speaker. Its design is similar to a usual wall clock, which makes it easy to interpret.

3. SOUNDSCAPE DESIGN

People tend to check the time of the day frequently to stay aware. Sometimes it is needed immediately and it is critical for the next

https://doi.org/10.21785/icad2019.034

activity. Sometimes it is less urgent but still important. This need leads to checking out the time on devices with visual displays. In some situation this behavior can be seen rude by other people, for example when being at a meeting. Even in everyday situations, the possibility of being aware of the time without taking out the mobile phone or turning the body from the area of attention appears attractive. By making the time audible the information can be sensed without moving.

3.1. Sonification of Time

A 12-hour clock represents the time of the day. This 12-hour period is divided into four three-hour quadrants. Each quadrant is subdivided into the concrete hour. Hours are divided into quarters. This hierarchy reflects the degree of consciousness necessary to interpret the detail of the daytime: Peripheral awareness is necessary to interpret quadrants of the day. Conscious counting is necessary to interpret the exact hour. However, counting from one to three should still allow task sharing, so interpreting the hour should not distract the staff too much from other tasks. Additional focus to the timing of events is necessary to interpret the quarters of an hour. At this detail level deriving the time may be as distracting as looking at a watch. Nonetheless, concentrating on sound may seem less interruptive and impolite to the customer than looking at a watch.

3.1.1. Quadrants of the day

Four Scenes represent quadrants of the day. Scenes are different compositions of natural soundscapes. Example scenes are illustrated in Fig. 1, these are Seashore, Rustling Leaves, River and Bonfire. We made the order of the scenes appropriate to time of the day. It starts with a calm and gentle Seashore Soundscape. Rustling Leaves is more restless. River then is more calm but steady. The last scene is made for late evening/night. The Bonfire scene is appropriate, as it suggests comfort and a homely feeling. This scene also contains the sound of a cricket to imply nighttime. The sound compositions contribute to the comfort of the customer and for the staff the scenes serve to mask noise and reduce stress, while delivering a rough estimate of the daytime that needs little attention to interpret.

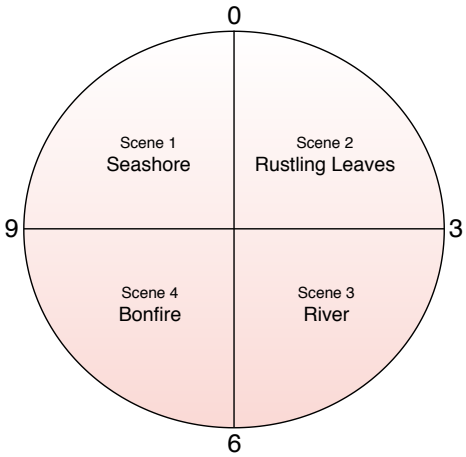


Figure 1: Example of scenes that consists of four different soundscape compositions.

3.1.2. Hours

Hours are represented by a bell sound as an auditory icon. The sound represents the hour within the quadrant, i.e., the clock strikes one to three times. Assuming the example in Figure 1 one bell strike represents 9, 12, 3 or 6 o'clock, two bell strikes 10, 1, 4 or 7 o'clock and three bell strikes 11, 2, 5 or 8 o'clock.

3.1.3. Quarters of an hour

In a way, the constellation of bell strike and birdsong can be considered as an earcon: The temporal distance of birdsong to the bell sound indicates quarters of an hour. Figure 2 illustrates the four temporal settings at which the birdsong is played. The bell icon in Figure 2 represents the temporal position of the bell sound. A birdsong that starts shortly before the bell sound means "quarter to half past". A birdsong shortly after means "half past to quarter to". The other two settings are temporally more distant to the bell sound. The one that is played before the bell sound indicates "sharp to quarter past". Whereas the other one indicates "quarter to sharp". The temporal distance of a birdsong to a bell sound can be 1 to 5 seconds.

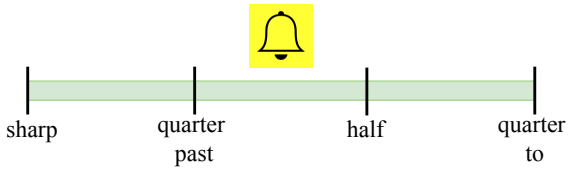


Figure 2: Temporal settings at which the birdsong is played.

3.2. Implementation

As part of this study, four scenes were created with Pure Data [18]. Those compromise scenes with rustling leaves, seashore, small river and bonfire. A possibility for the creation is to record natural soundscapes und playback the sampled sound. We decided to use Granular Synthesis [19] and also generate sound according to Procedural Audio [20]. With Granular Synthesis sampled sound can be manipulated and new sound is created by resynthesis. By Procedural Audio we mean the creation of sound models, like a wind-model to create wind sounds, that are controlled by physically meaningful parameters, like speed of wind.

We developed software that provides many parameters to design the compositions and the mapping of information on the fly. We also do not need hours of recordings to avoid repetitiveness. Because these scenes are designated to be played over weeks or months, we think the risk of repetitiveness may become an issue. With Granular Synthesis we only need recordings up to a minute length. By random variation of typical granulation parameters [19] like playback rate (Pitch and tempo shift), amplitude, grain length and sample position, we get an output that is enormously diverse. Alongside that we developed models of wind, fire and animals that are based or inspired by examples from Andy Farnell [21, pp. 327 – 649]. These models provide many possibilities to modify the sound. By randomizing the parameters we get a constantly evolving and changing output. The bell sound is created by adding sines with different Amplitude-Envelopes (Additive Synthesis). The birdsong is created with Frequency Modulation Synthesis. The parameters of the Frequency Modulation and the shape

of the Amplitude-Envelope are randomized, so that every birdsong has a unique character.

3.3. Example

A demo of the soundscape clock is available at `https://tinyurl.com/y4yd8zkh` . The demo is 12 minutes long and contains three minutes of each of the soundscape compositions that are illustrated in Fig. 1. For the purpose of demonstrating the variations of the bell strike-birdsong patterns, we showcase different times for every minute. The demo starts with the seashore scene at 9 o'clock sharp. The other times that are displayed are listed in the description to the demo. The scenes are not repetitive except for the 0.1 Hz amplitude modulation of the wind model. This model is used in the second and third scene of the demo. Its spectrogram is plotted in Fig. 3. The spectral distribution is held between brown and pink noise, which can mask high-frequency sound and is more pleasing than noise with more high-frequency energy [22], [23].

4. DISCUSSION

In this paper, we presented a concept for an ambient auditory display. Like other studies [13], [14], [15], [16], our design is highly concerned with intrusiveness of sound. We use soundscape compositions as a framework within which information is placed and displayed to the user. It is intended to run in the background. Information that is placed inside is supposed to be well integrated, so that it becomes a part of the composition. Such a design can offer information in a way that Weiser and Brown introduce as *calm technology* [24]. "Calm technology engages both the center and the periphery of our attention, and in fact moves back and forth between the two" [24, p. 81].

Our concern is to provide an overall awareness of time without creating too much attention and disturbance. This goal is achieved by embedding auditory icons as part of a soundscape composition. A design choice for the icons could be the use of sounds that are plausible in the scene, like a typical animal sound to be found in a natural soundscape. The icon for displaying the minutes, the birdsong, is such a choice. Then again, noticeability of the sounds seems also important to consider. For example when birdsong is already a component of the scene, adding another as an auditory icon can be confusing. Bell sound on the other hand is a very recognizable sound, so there is a risk that it grabs too much attention. Besides the choice of the sounds, we think that psychoacoustic metrics [25] like sharpness, roughness, loudness and tonality are important to consider for these goals. In our case we smoothen the amplitude rising at the beginning of the bell sound to reduce the sharpness of the attack.

We chose to display the auditory icons once every minute. The challenge is to be not too disturbing by a high recurrence rate but still be around when a user needs to know the time. A frequent appearance of the icons can blend out other parts of the composition. When it becomes too foregrounded, it can potentially ruin the composition, because it is not being perceived as an authentic soundscape. Once a minute is possibly already too short. Our choice is made up on the reflection of the situation that a user needs to know the time and waits for the icons to appear. By using a period of more minutes, the waiting could quickly become annoying.

We want to keep the display of time easy, so that it is mentally not challenging. By using the sequence of scenes to already deliver the quadrants, the information that is delivered by our auditory icon for hours is simplified. The number of bell sounds that has to be counted is restricted to a maximum of three. Counting needs particular attention, which is counterproductive for our goals. The displaying of the minutes is also kept easy to reduce consciousness necessary to interpret. As a result our display of minutes is very rough.

By using soundscape compositions we show an opportunity for auditory displays that can also enhance the auditory field of rooms like workspaces. We think that the idea to cover information by using natural soundscape composition is a valuable contribution, because it keeps our auditory icons less intrusive. Additionally, soundscapes are known for having positive effects and there is a requirement to generate them for places such as workspaces.

5. CONCLUSION AND FUTURE WORK

This paper presented a concept of how to use soundscape compositions as auditory displays. We use it to deliver the time of day. Other information that is workplace related, or personal, can be incorporated in the composition. For example mail income or, like mentioned by Kilander et al., the coworker presence, by assigning everyone to a certain birdsong could be displayed [13]. Delivering personal information at places that are used for public could be difficult because in our concept information is not directed to a chosen person. The use of our concept should depend on the character of the information. If the information is a case of emergency, using birdsong in a natural soundscape composition may be less suitable.

To deliver quadrants we use different soundscape compositions. The order of these is in our case intended to be appropriate to the time of day. The election and order of these could also depend on the season, the room and location they are played in or the taste of the user. Besides, we think that scenes do not have to be specific to certain real places. Because people are used to the blending with human made sounds, we also think the choice of sound for auditory icons can be diverse without seeming inappropriate.

We mentioned that natural soundscapes consists of daily and seasonal rhythms and patterns. A prospective research could be a close examination of these pattern. We could gather details on how natural soundscapes provide information, possibly about the time, season and weather, and how these affect humans. In this manner, we could generate soundscapes that work equivalent to the real ones. From this examination, we could also gather a better understanding that helps to design appropriate sonifications to provide various information. For such an attempt, our implementation in Pure Data is very suitable, because every component is synthesized or processed, and therefore offers parameter that potentially enables each to provide information. Possibly the whole generated soundscape can provide information.

For future work the concept as shown has to be evaluated by users. Our use of different compositions in a sequence and the design of the auditive clock have to be tested. In a basic evaluation case that tests the readability of the clock, subjects would first be taught about the meaning of the components. Afterwards the compositions would be played to them whereby they have to write down the time that is provided by it in succession. Besides that, it is also important to evaluate the intrusiveness of our concept in a real life case study, for example at a workplace. To examine whether the auditory display is in fact ambient to those who are not privy to the sound principles, and informative to those who are, a

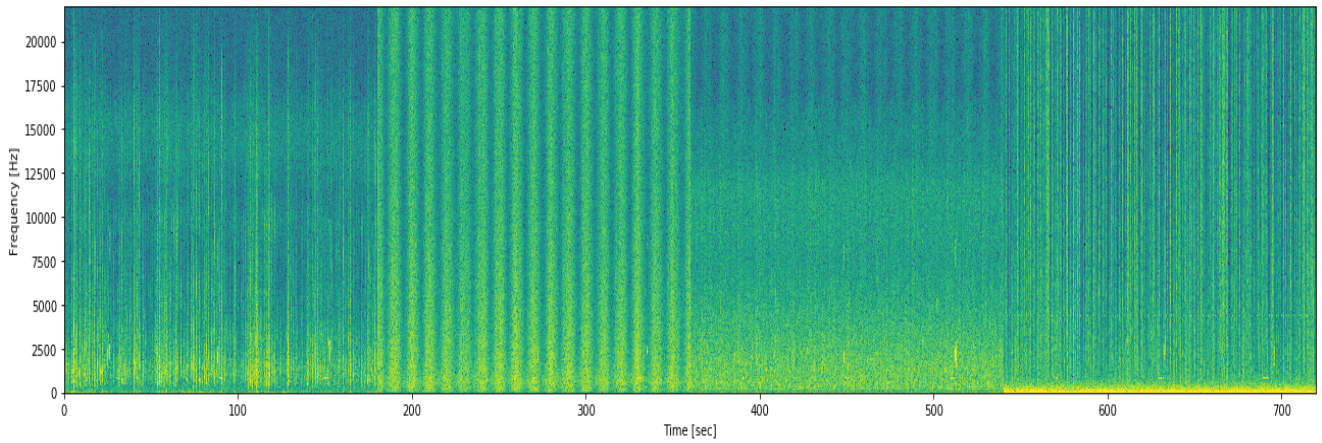


Figure 3: Spectrogram of a 12-minute soundscape clock demo including four different scenes.

test installation and questionnaire as in [26] could be employed. In this way, many choices we made, such as the components or the period of time, could be analyzed.

6. REFERENCES

[1] B. Truax, *Acoustic Communication*. Noorwood, New Jersey: Ablex Publishing Corporation, 1984.

[2] S. Ferguson, "Sonifying every day: Activating everyday interactions for ambient sonification systems," in *Proceedings of the 2013 International Conference on Auditory Display (ICAD 2013)*, Lodz, Poland, July 6-10 2013.

[3] W. W. Gaver, "Auditory icons: Using sound in computer interfaces," *Hum.-Comput. Interact.*, vol. 2, no. 2, pp. 167–177, June 1986.

[4] J. J. Alvarsson, S. Wiens, and M. E. Nilsson, "Stress recovery during exposure to nature sound and environmental noise." *International Journal of Environmental Research and Public Health*, vol. 7, no. 3, pp. 1036–1046, March 2010.

[5] A. G. DeLoach, J. P. Carter, and J. Braasch, "Tuning the cognitive environment: Sound masking with "natural" sounds in open-plan offices," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, p. 2291, 2015.

[6] H. Ma and S. Shu, "An experimental study: The restorative effect of soundscape elements in a simulated open-plan office," *Acta Acustica united with Acustica*, vol. 104, no. 1, pp. 106–115, 2018.

[7] Y. Hao, J. Kang, and H. Wrtche, "Assessment of the masking effects of birdsong on the road traffic noise environment," *The Journal of the Acoustical Society of America*, vol. 140, no. 2, pp. 978–987, 2016.

[8] B. D. Coensel, S. Vanwetswinkel, and D. Botteldooren, "Effects of natural sounds on the perception of road traffic noise," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. EL148–EL153, 2011.

[9] J. Y. Jeon, P. J. Lee, J. You, and J. Kang, "Perceptual assessment of quality of urban soundscapes with combined noise sources and water sounds," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1357–1366, 2010.

[10] M. Rdsten Ekman, P. Lundn, and M. E. Nilsson, "Similarity and pleasantness assessments of water-fountain sounds recorded in urban public spaces," *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 3043–3052, 2015.

[11] K.-C. Lam, L. Brown, L. Marafa, and K.-C. Chau, "Human preference for countryside soundscapes," *Acta Acustica united with Acustica*, vol. 96, no. 3, pp. 463–471, May 2010.

[12] J. Errett, E. Eileen Bowden, M. Choiniere, L. M Wang, E. Ryherd, E. , and W. , "Effects of noise on productivity: Does performance decrease over time?" in *AEI 2006: Building Integration Solutions - Proceedings of the 2006 Architectural Engineering National Conference*, vol. 1. American Society of Civil Engineers (ASCE), March 2006, pp. 221–228.

[13] F. Kilander and P. Lönnqvist, "A weakly intrusive ambient soundscape for intuitive state perception," in *Continuity in Future Computing Systems*, J. Doherty, M. Massink, and M. Wilson, Eds. Oxford: The Central Laboratory of the Research Councils, 2001, pp. 70–74.

[14] ——, "A whisper in the woods - an ambient soundscape for peripheral awareness of remote processes," in *Proceedings of the 2002 International Conference on Auditory Display (ICAD 2002)*, Kyoto, Japan, July 2-5 2002.

[15] R. Jung and T. Schwartz, "Peripheral notification with customized embedded audio cues," in *Proceedings of the 13th International Conference on Auditory Display (ICAD 2007)*, Montreal, Canada, June 26-29 2007.

[16] H. Zoon, S. Bakker, and J. Eggen, "Chronoroom clock : peripheral time awareness through sound localization," in *Proceedings of the 17th Annual Conference on Auditory Display (ICAD 2011)*, Budapest, Hungary, June 20-24 2011.

[17] M. Puckette, "Pure data," in *International Computer Music Conference*, San Francisco, USA, 1996, pp. 224–227.

[18] J. P. Leonard, "Granulation of sound in video games," in *AES 41st International Conference*, London, UK, February 2-4 2011.

[19] A. Farnell, "Procedural audio theory and practice," in *The Oxford Handbook of Interactive Audio*, K. Collins, B. Kapra-

los, and H. Tessler, Eds. New York: Oxford University Press, 2014, pp. 531–540.

[20] ——, *Designing Sound*. London, England: The MIT Press, 2010.

[21] M. Gardner, *Fractal music, hypercards and more.* New York, United States: W.H. Freeman and Company, 1992.

[22] T. Ziemer, H. Schultheis, D. Black, and R. Kikinis, "Psychoacoustical interactive sonification for short range navigation," *Acta Acustica united with Acustica*, vol. 104(6), pp. 1075–1093, 11 2018.

[23] M. Weiser and J. S. Brown, "The coming age of calm technology," in *Beyond Calculation: The Next Fifty Years of Computing*. New York, NY: Springer New York, 1997, pp. 75–85.

[24] H. Fastl and E. Zwicker, *Psychoacoustics. Facts and Models*, 3rd ed. Berlin, Heidelberg: Springer-Verlag, 2007.

[25] E. Brazil and M. Fernstrm, "Investigating ambient auditory information systems," in *Proceedings of the 13th International Conference on Auditory Display (ICAD 2007)*, Montreal, Canada, June 26-29 2007.

SONIFYD: A GRAPHICAL APPROACH FOR SOUND SYNTHESIS AND SYNESTHETIC VISUAL EXPRESSION

Woohun Joo

Virginia Polytechnic Institute and State University,
Human Centered Design,
Blacksburg, VA, USA
joowh@vt.edu

ABSTRACT

This paper describes *Sonifyd*, a sonification driven multimedia and audiovisual environment based on color-sound conversion for real-time manipulation. *Sonifyd* scans graphics horizontally or vertically from a scan line, generates sound and determines timbre according to its own additive synthesis based color-to-sound mapping. Color and sound relationships are fixed as default, but they can be organic for more tonal flexibility. Within this ecosystem, flexible timbre changes will be discovered by *Sonifyd*. The scan line is invisible, but *Sonifyd* provides another display that represents the scanning process in the form of dynamic imagery representation. The primary goal of this project is to be a functioning tool for a new kind of visual music, graphic sonification research and to further provide a synesthetic metaphor for audiences/users in the context of an art installation and audiovisual performance. The later section is a discussion about limitations that I have encountered: using an additive synthesis and frequency modulation technique with the line scanning method. In addition, it discusses potential possibilities for the future direction of development in relation to graphic expression and sound design context.

1. INTRODUCTION

This project starts from an idea of what sound graphic design makes. In 2014, I completed my master thesis project titled "Transition between Color and Sound" at Rhode Island School of Design. This became a starting point of my color-sound study. My thesis project demonstrated how graphics can be transferred into sound, however, in terms of sound design aspect, I founded some issues. The lightness to amplitude mapping for an additive synthesis design was not very noticeable. For example, if timbre is determined by 1000 pixels and lightness of each pixel is mapped into an amplitude of each partial, it is not easy to recognize the differences over amplitude changes.

I have been exploring a synesthetic design approach [1] based on a color-sound conversion in order to go beyond the previous issues I mentioned and designing a functioning platform that possibly can be fully utilized as a tool for multimedia, audiovisual and musical expression with an immersive real-time projection of the image scanning process.

The ultimate goal of this research is to stretch traditional approaches of visual design/music, and to extend our understanding of multi-sensory experience design to cultivate visual and musical aesthetics.

For better understanding of how this system can be presented, my installation work, *Demo1 installation*, is attached. I have enhanced its functionality by applying a control interface, code based graphic presentation (instead of using a still image file extension such as JPEG) and more complex color-sound mappings coming with an amp modulation and FM-based sub-oscillator.

Sonifyd consists of three components: the first canvas that displays graphics/images, the second canvas that shows an image scanning visualizer meaning the image scanning processes, and lastly the sonification engine. Processing [1] codes are given on the basis of *Sonifyd*'s graphic component and the graphics will feed into MaxMSP[2] to create sound. Both components are controlled via an OSC [2]-based control interface, such as changing colors and the position of the shapes.

The creation of the codes written in Processing and variable changes via the OSC controller given motivates the user to explore a unique and spontaneous sound creating environment. During this procedure, the original graphic sources transferred to MaxMSP produce a dynamic visual feedback reflecting the image scanning process. This visual feedback provides more intuitive sound making experience and allows instant monitoring between graphics and sonic character; this scanning process will be further discussed in Section 3.

I applied a line scanning method that captures data in line with an invisible scan line. This is a visceral way to interpret 2D graphics because this scanning method can be understood in a similar way to a regular DAWs' approach; for example, Ableton Live's clip view has a scan line moving left to right to read MIDI data or audio contents.

In terms of the mapping between colors and sounds, it is firstly fixed within a particular range (in an octave). However, the range of the color-sound mapping can be extended or shortened according to a user's preference. This flexible nature of the system may seem chaotic at first, but users will acquaint themselves with repetitive uses like the way we memorize words from other languages. The mapping between color and sound will be described in Section 3.3.

[1] https://processing.org/

[2] https://cycling74.com/

https://doi.org/10.21785/icad2019.045

2. BACKGROUND AND RELATED WORK

The conversion of graphic into sound, or also known as graphic sound synthesis [3], with a concept of synesthetic design culminated in my master thesis in 2014. Since then I have been developing this sonification inspired multi-sensory design approach from there. The works by famous audiovisual artists such as Ryoji Ikeda, Carsten Nicolai, Olaf Bender and Brian Eno, although their works are intended to be sound/media art rather than sonification, have turned into my artistic inspiration and motive power for my ongoing color-sound research. I further saw the potential of this multi-sensory design approach as an assistive device with Neil Harbisson's Eyeborg[1]. There was another noticeable auditory display scenario working as an assistive tool. Khan et al. [4] developed an exercise platform for visually impaired individuals.

The multi-modal character of synesthetic phenomena has attracted much of academic attention, creating various types of an art form embracing visual and sound in fine art, multimedia and sound field. These approaches have a long history that is outside the range of this paper. However, it is relevant to briefly review precedent works where sound character is applicable graphically. There must be a close connection with early analog sound-on-film techniques and glass disc type instrument, starting from Piano Optophonique[2] and a Russian composer Arseny Avraamov's the first hand-drawn, animated, and ornamental soundtrack. An engineer Evgeny Sholpo developed a photosynthesizer called Variofon [5] with Rimsky-Korsakov's support as well. Under the Multzvuk group in the Soviet Russia, founded by Arseny Avraamov, Nikolai Voinov, Nikolai Zhelynsky and Boris Yankovsky carried out research on ornamental sound tracks. The incredible ANS synthesizer [5][6] by Evgeny Murzin, visualizes sound and vice versa. Oramics [7] is a notable example because it turns a particular shape into pitch, timbre and intensity of sound through 35mm film. In art, these techniques were covered in the works of famous visual music pioneers Oskar Fischinger and extended to Norman McLaren, John Witney, James Davies and Evelyn Lambart. For digital platforms, in the early 90s, Hyperupic [8], an image to sound transducer, was introduced by Chris Penrose as a continuous improvement of Xenakis' UPIC system [9]. Metasynth[3], a successor of Hyperupic [8], offered a unique sound making environment where we paint sound. Monalisa Application [10] interprets binary codes as sound. These examples drove us to explore new sonic experiences. In installation art, Scott Arford's Static Room [4], Granular Synthesis's Lux [5] and Noisefields by Steina and Woody Vasulka[6] are worth noting here. In addition, Atau Tanaka showed the transition between a photograph and sine wave in Bondage [11]. Shawn Greenlee used a digital microscope to scan hand-drawn paintings in his work Impellent along with his graphic waveshaping technique [12]. When it comes to image scanning methods, Probing [13] and Raster Scanning [14], termed by Yeo, are representative scanning frameworks. *Sonifyd* took a similar approach with Probing [13] to allow users to determine the location in line if they want.

Levin's audiovisual performance tool [15] is worth noting because this work is based on free-form image sonification for the real time performance that the direction seems quite similar to my approach. A Stead et al. designed a graphic score system [16] that can be interpreted via cellphone camera and a multi-touch input instrument called ILLUSIO [17] that reads hand drawn sketches for musical expression was introduced. These works exemplify how graphic elements can be associated with sound manipulation. An iOS synthesizer app NAKANISYNTH [18] turns hand-drawn sketches into sound waves and creates amp envelope curves. More recently, a visual score scanner called CABOTO [19] was published in 2018. The creator of CABOTO developed visual scores for music composition as well as scanning system for the visual score.

In music industry, the latest software like Audiopaint[7], Coagula [8], Kaleidoscope [9] and Photosounder [10] can be considered as close relatives to *Sonifyd* because they all are sound design softwares controlled by images. However, my approach was to focus not only on sound design but on the dynamic visual representation to provide multi-sensory experience. The scanning method that I used can be understood as a scanned synthesis [20], however, my algorithm is not targeting to adopt a physically inspired model.

3. IMPLEMENTATION

Processing [21], Syphon [22], MaxMSP, and Lemur [23] are utilized for implementation of this work. Processing sketches travel to MaxMSP for both sonification and scanning process visualization via Syphon, an open-source real time video sharing application. *Sonifyd*'s MaxMSP patch supports two projectable screens so that the system can display both Processing sketches and their corresponding line scanning process. (See Figure 1.)

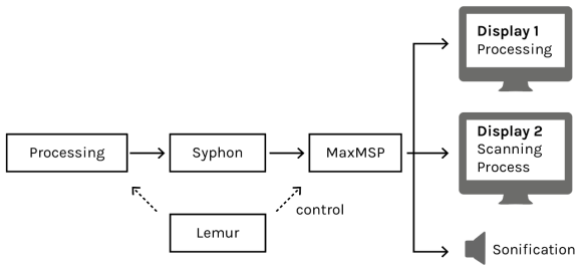


Figure 1: Schematic diagram of *Sonifyd*

3.1. Graphics

Sonifyd adopts an open source programming language Processing for the visual component. Processing is widely used among artists as well as interaction/visual designers thanks to its greater accessibility and ease of use. Processing has been a long time favorite when it comes to computer

[1] https://www.ted.com/speakers/neil_harbisson
[2] https://baranoff-rossine.com/optophonic-piano/
[3] http://www.uisoftware.com/MetaSynth/
[4] www.recombinantfestival.com/2017/project/scottarford/
[5] www.epidemic.net/en/art/granularsynthesis/proj/lux.html
[6] www.vasulka.org/Videomasters/pages_stills/index_42.html
[7] http://www.nicolasfournel.com/?page_id=125
[8] https://www.abc.se/~re/Coagula/Coagula.html
[9] https://www.2caudio.com/products/kaleidoscope
[10] https://photosounder.com/

programming for artists [24][25][26]. I pursue minimalistic graphic design expression in connection with my color-sound sonification strategy because minimalism in graphic design maximizes efficiency of visual communication and this approach can be applied for the same purpose, efficiency of sonic communication. Inbar et al. [27] proved that the minimal graphic expression recorded high acceptance rate among people to information visualization. The definition of sonification is 'the use of nonspeech audio to convey information [28].' I applied Bauhaus's [29] famous design philosophy "Form follows function". That is to say that sonification strategy can be reinforced by minimalism and my work can be characterized by simplicity of my graphic expression in both visual/sound aesthetics and communication viewpoint.

3.1.1. Interactive Images

I have tested a series of the minimalistic graphics shown in Figure 2 with the current version of *Sonifyd*. For instance, a mono color background with mono color basic shapes and a gradient color background with both mono/gradient colored shapes are examined; basic shapes here include an ellipse, rectangle and triangle. Section 5 includes more details about my future plan regarding these interactive images.

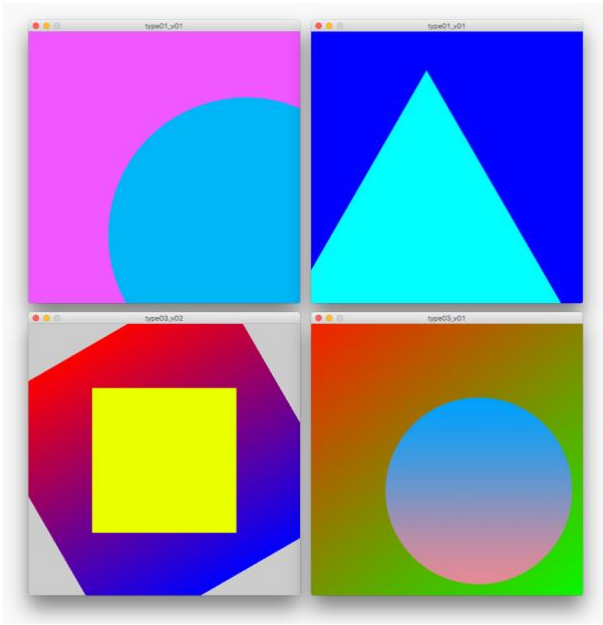


Figure 2: Graphic examples written in Processing

3.2. Image Scanning

MaxMSP's **jit.gl.syphonclient** can mirror Processing sketches and import them to **jit.matrix** in real-time. This system has three **jit.matrix** and two **jit.window** objects for both sonification and visualization purpose. The main **jit.matrix** displays Processing sketches onto **Window1** with the maximum screen size 512x512. The system further offers downsampling capability that leverages built-in Jitter matrix manipulations; as the number of the pixels is reduced, this naturally changes the frequency resolution.

The first **jit.matrix** is divided into two other **jit.matrix** objects for sound representation and showing the scanning process through **Window2** (See Figure 3). The Max's built-in video delay object called **vz.delayr** is connected between these two jit.matrix objects. If this delay object is activated, it simultaneously changes the timbre of the sound; more details will be discussed in section 3.3.4.

There is an invisible scan line moving in either vertical or horizontal position of the screen. 512 pixels per each row/column will determine tone color in connection with the sonification engine. The second **jit.matrix** that displays the scanning process has two arguments called **srcdimstart** and **srcdimend** (See Figure 3), so according to where the scan line is positioned, the corresponding pixels will be shown at the first row/column of **Window2** and the pixels will be stretched to the last row/column of the second screen (See Figure 4). For more flexibility of the visual effect, this system allows users to change the degree of the stretchiness. The bigger the degree is, the smoother the visual transition will be. I attached the images captured when the scanning process was in motion; this includes two stretchiness examples (See Figure 4). This is to provide intuitive experience design that makes people feel a close relationship between images and sounds with the real time scanning procedure as a dynamic imagery presentation.

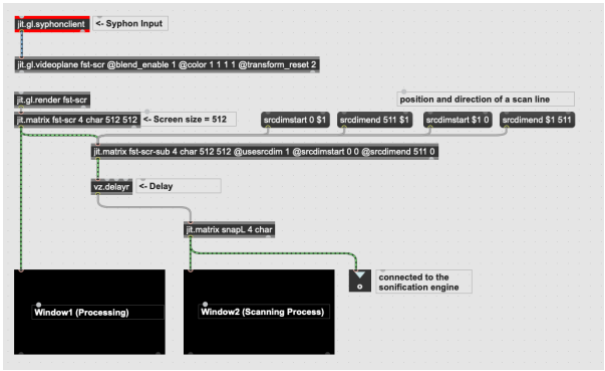


Figure 3: MaxMSP objects for two displays

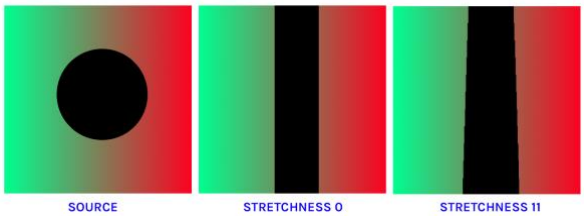


Figure 4: Scanning visualization

3.3. Color-Sound Synthesis Technique and Mapping

The concept of sound design underlies additive synthesis technique. This is mainly because the scan line moves row by row or column by column in parallel with the pixel size of the screen. Additive synthesis is an excellent way to reconcile all pixel data at where the scan line is located. For example, FormSound [30], an additive synthesis-based platform, shows clear tonal changes based on the number of the particles. *Sonifyd* converts 512 pixels in line into sound. Each axis of the screen represents time and oscillator. Software called AudioPaint [31] uses a similar tactic because this software converts pictures into sounds with an additive synthesis

technique; one line of the picture serves as an oscillator. But AudioPaint is not real time based and *Sonifyd* has a difference color-sound mapping strategy.

3.3.1. Color-sound mapping

I applied HSL color system to interpret color values into sound. Within ICAD community, *Hue Music* [32] was discussed and hue values from 2D image create timbre. In *Sonifyd*, hue values present frequencies of each partial, and saturation and lightness control pitch shifting. Amplitude of each partial is fixed with the value 0.8(0.0-1.0) to hear each frequency component at the same amplitude level; the attached video example *Demo1-Installation* supports saturation and lightness mapping that are responsible for an octave and gain control; however, there was no significant tonal changes observed.

Hue mapping represents microtonal scales in an octave (See table 1) and the range is 369Hz(F#) to 739Hz(F#); this implies that if two hues are adjacent, beat tones can be heard. It is known that the spectrum of visible light can be within an octave [33]. This is the system's default setting that can be easily adjusted with the minimum and maximum values for more wider timbral flexibility. However, I do not recommend going beyond 5000Hz because it is known that a pitch perception is not accurate above 5KHz [34][35]. The color-sound relationship here is not the main focus at this stage because *Sonifyd* is primarily designed for audiovisual instrument. As musical instrument, the wider the frequency range is, the more tonal flexibility the users experience.

This system's additive synthesis technique has been implemented by the Max object called **oscillators~**, developed by CNMAT, UC Berkeley.

Color	**Sound**
Hue (0.0-1.0)	Frequency of each partial (369Hz to 739Hz)
Saturation (0.0-1.0) *1.0 is neutral	Pitch Shift B (0 to 2 octaves) *This is because when saturation goes down, the lightness goes up.
Lightness (0.0-0.5) *0.5 is neutral	Pitch Shift A(0 to -2octaves)
Lightness (0.5-1.0) *0.5 is neutral	Pitch Shift B(0 to 2octaves)

Table 1: Color-sound mapping (the main oscillator)

3.3.2. Pulse wave amplitude modulation

For the attached video *Demo1-Installation*, black and white represent the lowest and highest frequency depending on the color-sound mapping. Instead, black and white in *Sonifyd* generate a pulse wave. These colors have no effect on timbre but modulates an amp envelope of the main oscillator (See the wavetable in Figure 5 and refer to LFO in Figure 6). The black area from the right image(B) will turn into 0 and the wavetable modulates amplitude. In Figure 5 below, the red line shows the invisible scan line and the right image(B) represents the scanning window. The idea of using black and white colors to modulate the amplitude comes from graphic design metaphor that black and white signify empty space (like rest in music). If I draw black or white grid/stripe, the image will play a role as a step sequencer.

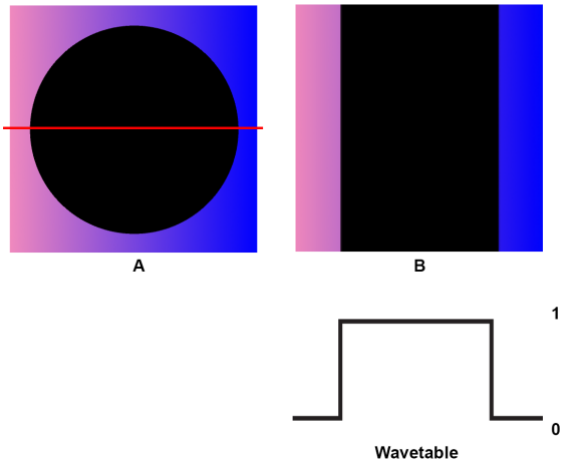


Figure 5: Pulse wave from black and white

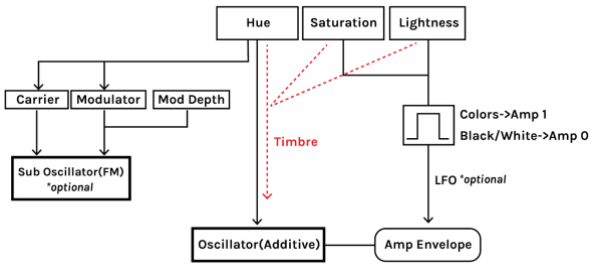


Figure 6: Sound Design

3.3.3. Sub-oscillator

Sonifyd comes with an optional sub-oscillator action to enhance sound character and increase the density of the sound. This is to improve the drawback of an additive synthesis. For example, an additive synthesis makes no tonal difference between the image A and B because they both consist of the same frequency components (See Figure 7).

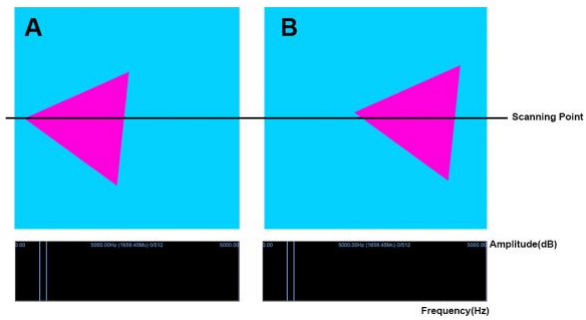


Figure 7: Images that have the same frequency component.

The sub-oscillator is based on frequency modulation synthesis and the carrier frequency goes two octaves below based on the lowest frequency of the main oscillator. However,

this mapping is customizable according to users' preferences; it goes down from one to three octaves.

Hue values draw a wavetable for the modular frequency of the sub-oscillator (See Figure 6) and the frequency of the modular frequency is determined by the lowest frequency of the main oscillator; the modular frequency is three octaves lower than the lowest frequency of the main oscillator. This sub-oscillator can be activated or deactivated.

FM	**The frequency range of the main oscillator**
Carrier Frequency	Two Octaves below from 369Hz (the lowest) *Customizable
Modulator Frequency	Three Octaves below from 369Hz (the lowest) *Customizable
Modulation Depth	Fully Flexible (between 0 to 1000)

Table 2: Color-sound mapping (the sub-oscillator)

3.3.4. Audio reverb and video delay

Sonifyd has an optional audio reverb and video delay effect. If a video delay is engaged, it makes a visual reverberation and naturally creates additional color tones (See Figure 8). This means that it consistently increases the number of partials of the synthesized sound. The amount of the video delay corresponds to the amount of audio reverb. My future plan regarding sound and visual effect is described in Section 5.

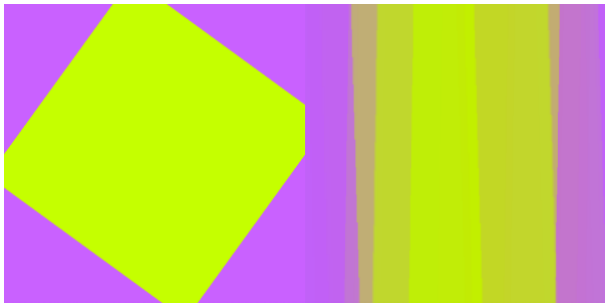


Figure 8: A delay effect that is applied into the scanning system.

3.4. Control Interface

Processing sketches can be accessed with Lemur for the real time practice/performance. Lemur is an iphone/iPad MIDI/OSC compatible controller interface to control graphic attributes such as shapes, size, position or colors. Further, Lemur allows users to switch the scanning direction between horizontal and vertical, and increase/decrease scanning speed in MaxMSP. In Figure 9, a user can change RGB values of background/shape colors, rotate shapes, scale the size of the shapes, and move between different Processing sketches if multiple Processing windows are running at the same time. Using touchable interface as a controller is not a new technology these days. As *Sonifyd* is developed mainly for real-time audiovisual performance, a control interface is needed. Hardware MIDI controller may work for this purpose, but it is difficult to customize. Lemur is the most reasonable choice to access *Sonifyd* because OSC-based controllers like Lemur or TouchOSC[1] provide more flexible graphical control interface. A brief example of the interactive scenario (Demo3) is shown in the video sample I linked.

Figure 9: Lemur control interface

4. DESIGN CRITERIA

Sonifyd is firstly designed for my own multimedia performance tool. In this paper, I aim to introduce my audiovisual instrument and explain how I designed it. The current version of *Sonifyd* is still at the early phase of my longer research. Considering this, I performed self-evaluations focusing on sound design and controllability perspective. First, the more color values the scan line has, the richer sound this system will create. Since gradient colors contain a wider sound spectrum range in comparison to plain colors, a pleasing sound with the smooth visual and sonic transition can be heard if gradient colors are shown. This is a natural character of an additive synthesis technique. Second, Syphon causes a little communication delay between MaxMSP and Processing. The delay does not give rise to a critical issue so far, however, it will speed up the system if there is a way to bring Processing into MaxMSP without Syphon. I have posted this issue on the online community, but clear solution does not exist yet. Third, additive synthesis is not appropriate to distinguish the position of the shape. For example, a magenta triangle on the left side of the screen and another magenta one on the right side of the screen with the same background color will create the same timbre (See Figure 7). This is why I applied the sub-oscillator described in section 3.3.3. It significantly helps to solve this issue, but another question arose. This is about how many times these modulations must be triggered along with the speed of the scanning process; further study of this issue will be required. As I previously mentioned in section 3.3.3., this modulation frequency is determined by the lowest frequency of the main oscillator. However, the sound differences between two images in Figure 7 are subtle. If I can find out another useful way to clarify these differences with the same scanned synthesis I used, *Sonifyd* project can take another step forward and I will be soon ready for auditory icon study [36][37][38] I want to explore.

[1] https://hexler.net/software/touchosc

5. CONCLUSION AND FUTURE WORK

Sonifyd that has been applied an additive sound synthesis technique and wavetable-based sound modulation [12][39] determines timbre and creates various tonal characters by scanning images. In addition, since there is no absolute connection between color and sound frequencies, *Sonifyd* provides an improvisational and experimental sound design environment with the customizable mapping interface. This study was a good starting point and found a great potential that will lead me to more advanced types of instrument to cultivate a unique sound making and immersive synesthetic audiovisual experience. Figure 10 shows how I exhibited the previous version of *Sonifyd* using an immersive projection space.

This section lists my future plan to expand the functionality and develop other variant platforms. First, different sound synthesis techniques, FM synthesis, subtractive synthesis, and wave shaping synthesis, will be applied to design the main oscillator. Second, further experiments on reverb and delay will be performed to see how these spatial sound effects can be utilized as both sonification variables and visualization effects. Third, Processing sketches will be widely expanded following graphic design themes such as the sound of grid system, the sound of geometric shapes, etc. These themes will be designed for both interactive audiovisual performance and media art installation purpose. Fourth, I will compare each graphic-sound mapping model with the same visual themes and analyze the results with a question of what sound synthesis strategies best represent a gradient circle on a mono color background for instance. This approach will lead me to investigate how the minimalistic graphics can act as visual scores as well. Fifth, *Sonifyd* will support multiple Processing sketches running all together and each sketch will be multi-layered to form an ensemble; different sonification strategies can be applicable to them. Lastly, a tempo sync function that is compatible with other hardware/software synthesizers, DAWs and step sequencers can be necessary to enhance the capabilities that allow more practical, flexible, and wider musical expression as a novel sonification-based instrument. The scanning method that I used has lots of precedents (See Section 2) and I may need to consider to use another scanning method such as a spot-mapping method that SPOTTY [40] and Voice of Sisyphus [41] used. Solving questions listed above is my absolute priority.

Figure 10: Installation using an immersive projection space.

6. REFERENCES

[1] M. Haverkamp, Synesthetic design: Handbook for a multi-sensory approach. Walter de Gruyter, 2012.

[2] M. Wright and A. Freed, “Open SoundControl: A New Protocol for Communicating with Sound Synthesizers.,” in ICMC, 1997.

[3] C. Roads and J. Strawn, Graphic Sound Synthesis. The computer music tutorial, vol. 32, no. 6. MIT press, 2003.

[4] R. A. Khan, M. Jeon, and T. Yoon, “‘Musical Exercise’ for people with visual impairments: A preliminary study with the blindfolded,” 2018.

[5] T. Y. Levin, “‘Tones from out of Nowhere’: Rudolph Pfenninger and the Archaeology of Synthetic Sound,” Grey Room, vol. 12, pp. 32–79, 2003.

[6] S. Kreichi, “The ANS synthesizer: Composing on a photoelectronic instrument,” Leonardo, pp. 59–62, 1995.

[7] J. Hutton, “Daphne Oram: innovator, writer and composer,” Organised Sound, vol. 8, no. 1, pp. 49–56, 2003.

[8] C. Penrose, “Hyperupic.” 1992.

[9] H. Lohner, “the UPIC system: A User’s Report,” Comput. Music J., vol. 10, no. 4, pp. 42–49, 1986.

[10] K. Jo and N. Nagano, “Monalisa:" See the Sound, Hear the Image",” Proc. 8th Int. Conf. NIME, pp. 6–9, 2008.

[11] A. Tanaka, “Biomuse to bondage: Corporeal interaction in performance and exhibition,” in Intimacy Across Visceral and Digital Performance, Springer, 2012, pp. 159–169.

[12] S. Greenlee, “Graphic Waveshaping,” in Proceedings of the International Conference on New Interfaces for Musical Expression, 2013.

[13] W. S. Yeo and J. Berger, “A framework for designing image sonification methods,” Proc. 11th Int. Conf. Audit. Disp. (ICAD 2005), pp. 323–327, 2005.

[14] W. S. Yeo and J. Berger, “Raster Scanning-A New Approach to Image Sonification Sound Visualization, Sound Analysis and Synthesis,” Proc. ICMC 2006, pp. 309–314, 2006.

[15] G. Levin, “Painterly Interfaces for Audiovisual Performance,” MIT Media Lab., pp. 1–151, 2000.

[16] A. G. Stead, A. F. Blackwell, and S. Aaron, “Graphic Score Grammars for End-Users,” New Interfaces Music. Expr., pp. 196–200, 2011.

[17] J. Barbosa, F. Calegario, V. Teichrieb, G. Ramalho, and G. Cabral, “Illusio: A Drawing-Based Digital Music Instrument,” NIME ’13 Proc. 2013 Conf. New Interfaces Music. Expr., pp. 499–502, 2013.

[18] K. Nakanishi, P. Haimes, T. Baba, and K. Kushiyama, “NAKANISYNTH: An Intuitive Freehand Drawing Waveform Synthesiser Application for iOS Devices,” Proc. Int. Conf. New Interfaces Music. Expr., vol. 16, pp. 143–145, 2016.

[19] R. Marogna and J. Van Stolberglaan, “CABOTO : A Graphic-Based Interactive System for Composing and Performing Electronic Music,” pp. 37–42, 1930.

[20] R. Boulanger, P. Smaragdis, and J. ffitch, “Scanned Synthesis: An introduction and demonstration of a new synthesis and signal processing technique.,” in ICMC, 2000.

[21] B. Fry and C. Reas, “Processing. org,” Process. org, 2010.

[22] “Syphon.” [Online]. Available: http://syphon.v002.info/. [Accessed: 27-Mar-2019].

[23] “Lemur – Liine.” [Online]. Available: https://liine.net/en/products/lemur/. [Accessed: 27-Mar-2019].

[24] H. Bohnacker, B. Gross, J. Laub, and C. Lazzeroni, Generative design: visualize, program, and create with processing. Princeton Architectural Press, 2012.

[25] M. Pearson, Generative Art. Manning Publications Co., 2011.

[26] D. Shiffman, The Nature of Code: Simulating Natural Systems with Processing. Daniel Shiffman, 2012.

[27] O. Inbar, N. Tractinsky, and J. Meyer, “Minimalism in information visualization: attitudes towards maximizing the data-ink ratio.,” in ECCE, 2007, vol. 7, pp. 185–188.

[28] P. C. Gregory Kramer, Chair; Bruce Walker and Terri Bonebright; Perry Cook; John Flowers; Nadine Miner; John Neuhoff, “The Sonification Report: Status of the Field and Research Agenda.” [Online]. Available: http://www.icad.org/websiteV2.0/References/nsf.html. [Accessed: 27-Mar-2019].

[29] L. Moholy-Nagy, The new vision: fundamentals of Bauhaus design, painting, sculpture, and architecture. Courier Corporation, 2012.

[30] S. Park and W. Joo, “FormSound: A particle formation-based audiovisual interface,” in 2017 ICMC/EMW - 43rd International Computer Music Conference and the 6th International Electronic Music Week, 2017.

[31] “Nicolas Fournel » AudioPaint.” [Online]. Available: http://www.nicolasfournel.com/?page_id=125. [Accessed: 27-Mar-2019].

[32] D. Payling, S. Mills, and T. Howle, “Hue music-creating timbral soundscapes from coloured pictures,” 2007.

[33] “Pitch/Frequency Related to Color.” [Online]. Available: http://wagneric.com/audiocolors.html. [Accessed: 27-Mar-2019].

[34] C. Roads and J. Strawn, Perception of Frequency. The computer music tutorial. MIT press, 1996.

[35] E. D. Schubert, Psychological acoustics, vol. 13. Hutchinson Ross Publishing Company, 1979.

[36] K. van den Doel et al., “Geometric shape detection with soundview,” 2004.

[37] M. Cooley, “Sound+ image in computer-based design: learning from sound in the arts,” 1998.

[38] E. D. Mynatt, “Designing with auditory icons,” 1994.

[39] C. Roads and J. Strawn, Waveshaping Synthesis. The computer music tutorial. MIT press, 1996.

[40] G. Evreinov, “Spotty: imaging sonification based on spot-mapping and tonal volume,” 2001.

[41] I. Conference and A. Display, “VOICE OF SISYPHUS : AN IMAGE SONIFICATION MULTIMEDIA INSTALLATION Ryan McGee , Joshua Dickinson , and George Legrady Media Arts and Technology University of California , Santa Barbara,” Int. Conf. Audit. Disp., pp. 141–147, 2012.

‘MUSIC OF THE PEOPLE’: MUSIC FROM DATA AS SOCIAL COMMENTARY

Rob King

Music Dept., Durham University
Palace Green
Durham, DH1 3EP, United Kingdom
robert.g.king@durham.ac.uk

ABSTRACT

Data-music reflects the ubiquity of data in modern society. Composers have not engaged widely with the opportunities opened up by this, despite the chance to overcome a gulf between academic art music and social engagement. Their reluctance might be traced to the challenge of reconciling abstract data and concrete sound, in political implications, and in technological barriers in computer music. The present paper argues that socially relevant music composition for the 21st century can adopt a programme of sonification grounded in politically acute data. As examples of such practice, two compositions are discussed founded upon US and UK social data sets, and realised via the SuperCollider programming language. The consequences for the composer of new music are further discussed from political and musicological angles, with the ‘purpose’ of writing such music analysed from the perspective of various commentators.

1. INTRODUCTION

There are more data collected on humans today than at any period in history. These data are given freely, by people answering questionnaires and ‘liking’ photographs, but are also harvested by corporations logging sites and products we look at online, or by government-sponsored security firms watching us through the lenses of countless cameras. It is the responsibility of global citizens to be aware of the data collected on them, and yet too often we are keen to ‘accept all terms and conditions’ without proper scrutiny [1]. Presenting data as art can provide a fresh opportunity to examine this datopia, with the potential to plant questioning seeds in the minds of those whom the data concern. Composersan opportunity to engage with and critique contemporary society; in an internet-based world where clicks are currency, an artistic representation of society’s greatest global social and economic force can offer food for thought for the demographics from whom the data are collected [2].

Moreover, there exists a disconnect between society and classical music – concert attendance is decreasing, record sales low, and contemporary art music perceived to be deliberately difficult and arcane [3, 4]. By no means does this paper suggest a ‘solution’ to this ‘problem’, but it does propose a method by which contemporary art music can reconnect with a wider populace.

There have long been efforts to integrate data into composition work. These stem from the efforts of composers to treat particular data as compositional material, such as Iannis Xenakis and stochastic functions, Charles Dodge and geomagnetic data, or Natasha Barrett and weather systems, often with the aid of computers (in Xenakis’ case, his first access to computers was in 1962). Xenakis viewed the computer as a means by which he could write mathematically driven music - he was not, as I am, interested primarily in the computer as a sociological phenomenon, and its use as a signifier of the information age in the creation of art [5]. Social, political, environmental, and personal data are often rich and interesting sources, to be sure, but carry with them much extramusical baggage. Xenakis’ music was not based on societal phenomena; his sonifications were mathematical, platonic conceptions. Indeed, there exist comparatively few composers writing about political issues by using data from humanity.

One composer who does treat sociological data is R. Luke DuBois. He views his work as primarily political, as opposed to technological, “using the tools of a particular established medium to critique it” [6]. It is this politicised approach to writing music using data that this paper seeks to consolidate. Other musicians have further refined the field to sonifying *social* data; for instance, De Campo and Egger De Campo in 1998, using data on executions in the USA since 1977 [7].

The case studies examined in this paper do not present data in a manner in which the data are legible as information, pace the model of Vickers and Hogg [8]. The music presented has had certain compositional processes applied to it, so that the data forms a basis for the music; the translation of data-parameters to musical ones has been undertaken, primarily, as a musical endeavour, with musical considerations. As such, sonifications of this sort have a slight tendency towards *ars informatica,* but the decisions of the composer mean that they retain the critical element that classifies them as *musical* - that they are *organised* by a composer, first and foremost. This is a valid measure; art is a response to stimuli, but those responses are filtered through the personality of the artist. Thus, it is reasonable for someone who is creating art to manipulate that art in line with their views on the world; to do so in the field of data science, or to claim artistic representations of data are scientific would be unjustified.

Presented in this paper are a number of works for a combination of computer and live performers. Incorporated are considerations about the sources of, practical employment of, and political implications of the use of data in composition. The works are complemented by SuperCollider code examples, and are further contextualised by reference to Vickers & Hogg’s model, and yet further to the wider political context in which they have been created [8].

https://doi.org/10.21785/icad2019.007

2. COMPOSITIONAL EXAMPLES

There exist, as has been discussed, more data than ever on the subject of humanity and its habits. For the purposes of this paper, data has been selected from verifiable sources, such as the Federal Bureau of Investigation in the United States of America, or the Office for National Statistics in the United Kingdom (for the first and second case studies respectively). An examination of potential future projects using data from *unverified* sources is given in Future Directions (section 3), below. Sound examples of both case studies are available at: soundcloud.com/robking-2.

2.1. Case Study #1

In the first piece, data from gun license background checks carried out in the US between November 1998 and January 2019 were sonified. This sonification involved mapping state by state monthly gun license checks (represented in fig.1 on the *y*-axis) to pitch.

The passage of time, represented in the graph by the *x*-axis, is mimicked by the composition (in which, instead of years, the data is presented in a matter of minutes). The choice of presentation timescale is the most critical artistic decision to be made, alongside whether to present the data with minimal compositional interference, hoping that the audience receive a performance as close to *ars informatica* as possible, or whether to react to the data emotionally, and imbue the character of the resulting piece with this reaction.

In this case, the decision was made to have the perceived tempo of the piece increase over time. This complemented the data - generally, there was an upward trend in gun sales over time, so this was matched by the increase in perceived tempo. It also made the piece sound much more complex as it continued - this, too, was a deliberate choice, as the generally recognisable cyclicity of the first c.10 years of data becomes disrupted and disfigured, especially around the time of a large upwards spike in 2013.

The piece features a clear accelerando effect, as the duration of the notes decreases following a quadratic function. The *reason* for such an accelerando is the composers' perception of the quickening frequency of national disasters in the USA brought about by gun violence around the Autumn and Winter of 2017.

It is not only more compelling to highlight the increased frequency of gun license background checks; it is in fact necessary. This is in accord with Cardew's assertion that "[the composer] must demand works that relate directly to the issues and struggles and preoccupations of the present […] [H]e must stringently criticise such works from the point of view of both form and content, with the aim of building up their strength. He should do this conscientiously […]" [9]. As such, the piece draws not only on the data, as published by the FBI, but the reporting in the media that mass shootings in the USA are becoming more frequent [10], and attempts to highlight this in a sonic fashion.

By outputting a MIDI File from SuperCollider based on this reading of the data, and introducing this MIDI to Sibelius, a score was made. This score was designed to be played by a pianist, but, due to its complexity, this is a difficult task. This is an anticipated side-effect, and is more informed by compositional theories surrounding New Complexity as a genre of contemporary music, than it is by computer science

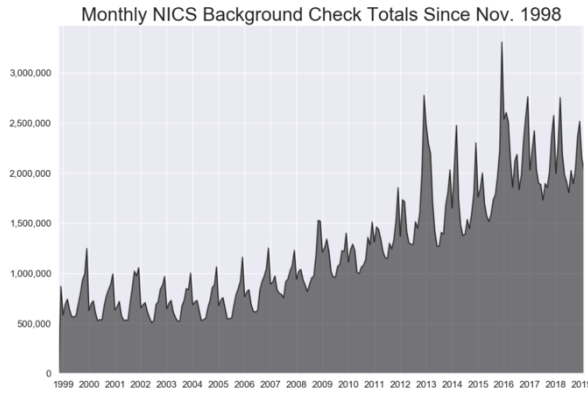


Fig.1: NICS Background Check Totals Since Nov. 1998 [11].

or data analytics[1], and is intended to make the experience of performing the work stressful for the performer. Simply put, the intention is to depict a difficult, contentious subject in a difficult manner with a difficult aesthetic, in line with the composer's views on said subject, so as to align the work as *ars musica* above *ars informatica*.

Lastly, to reference Vickers & Hogg's other axis (concrete-abstract), it should be noted that, by translating the MIDI data to a score designed to be performed by solo pianist, the work is immediately driven towards the abstract end of the scale, where the sounds generated during composition do not represent the sounds of the data. For instance, a less abstract composition, using the same gun licensing data, could use general MIDI sound 128 - the sound of a gunshot. Alternatively, gun license background checks could be sonified with the sound of a cash register (as the checks are required whenever a customer wishes to purchase a firearm), panned left or right depending on longitude data from the state in which the check was carried out.

Analysis of compositional method of this type gives composers the opportunity to write music, in the knowledge of where their work ranks in terms of abstraction, and therefore gives them tools with which they might self-assess their composition and its nature. The utilisation of a *concret-abstrait* axis also encourages composers to think about the sounds they use in sonifications, thereby giving them an opportunity to address the allegation (addressed in the introduction) that their music is wilfully obscure.

2.2. Case Study #2

This piece is formed from self-report data published on loneliness in young people in Britain, aged between 16 and 24 [12]. These data are presented in a more detailed manner than those in the previous example, in that they are *not* presented as a measure of one group over time; there are three groups (those who often, sometimes/occasionally, or rarely/never feel lonely, respectively) which are presented as percentages of the population. Instead of presenting these groups over time, the data are from a single survey in May and June 2018. They are further categorised in the downloadable file by gender, race, and other relevant markers.

[1] Although, this complication makes the connection between those from whom data is harvested and the music that derives from these data more difficult to establish. This is discussed further in section 3.

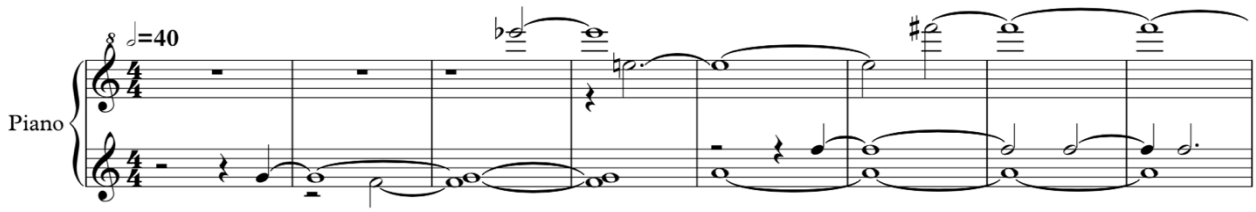


Fig. 2: An excerpt from the score generated for solo pianist; note the long durations and sparse texture.

This means that, unlike Case Study #1, there are no audible trends over time. Similarly, there cannot be a presentation of the data using 'soundalike' MIDI, as there is no real-world sound associated with the data. Indeed, it could be said that loneliness is associated with an absence of sound. This is where considerations as a composer must come to the fore - how can a piece based on loneliness data achieve a clear aesthetic, when there is no way that the audience could notice trends in the data, due to the way that they have been collected and collated?

This ambiguity is intentional, and aimed at potential criticism of music generated by sonifying data. Specifically, it addresses the comprehensibility and audibility of data, and presents the notion that the individual data need not be understood for the listener to understand the overall aesthetic of the piece as music [13]. Whilst the data may not be legible, at a datum-by-datum level, the overall affect of the piece can still be communicated by other factors involved in the presentation of the material.

This is the reason for the final format of the piece: the simultaneous presentation of music performed by laptop and by solo pianist. The solo pianist is (necessarily) alone on the stage - a decision taken, not because of its implication in the data, but because of the compositional desire to communicate loneliness. This is yet another reason why, when presented 'in a vacuum' sonified data can appear obscure, and its meaning oblique, but when taken holistically, a concert setting can communicate the crux of the work. Other decisions have been taken to communicate the lonely aesthetic still further - decisions that the data could not possibly have dictated. For instance, the pianist is directed to face away from the audience.

Making choices to communicate the essence of the piece is also the reason for compositional choices made in the musical material. Where no meaningful impression of the data is left by its translation to pitch, it is the job of the composer to communicate 'beneficial' meaning to the audience, in order that it might 'raise their consciousness' [9].

The choice of an increasing tempo, as per the first case study, was eschewed in favour of a more 'lonely' aesthetic, with irregular beats, making a tempo difficult to determine for the listener. Thus, the piano solo score is sparse, the feeling of *being without* a tempo contributing, once more, to the effect of isolation. See fig. 2 for an example from the score given to the pianist - note the long, sustained, comparatively infrequent notes.

By the same token as the decisions surrounding tempo, it would not have made sense to present the material as one long, continuous line (as in the first case study) as it is not organised chronologically. There are three columns of data presented, recorded simultaneously from three self-identifying groups (those who never, sometimes, or often feel lonely). Thus, these are sonified simultaneously, so as to better reflect the manner in which the data were presented originally.

The choice of instruments has been discussed already, but the individual design of these instruments is also a key consideration. There are two synthesising instruments, created in SuperCollider, which have been designed so as to mimic the timbre of a piano with a few key differences. Whilst the lonely pianist plays the data derived from those who often or always feel lonely, the first synthesised piano performs that which is derived from those who sometimes or occasionally feel lonely. The SuperCollider plugin 'MdaPiano' was used, and its parameters manipulated to be comparatively close to that of a real piano (the parameters of decay and reverb, for instance, could be modified in a performance to more closely mimic the live piano). The second sound sonifies the data from those who identify as rarely or never feeling lonely. Thus, it is designed to sound somewhat different from a piano. It retains, for example, a quick attack, but its decay and resonance values make it sound dissimilar to a piano over the course of the envelope of a note. It is also different from MdaPiano in that it is formed from a combination of harmonics.

These three instruments are tiered in such a hierarchy as a play on the already-established bias towards a visual presentation of data over the sonic. The piece acts as a kind of thought-experiment for the listener. The synthesising pianos present similar material to the lone pianist (again, so that no pulse can be easily identified, nor any tonal centre, nor thematic material), thereby leading the audience to the main difference between the sounds - their timbre.

The hierarchy is conceived so that the most piano-like sound (the *real* piano) is given the most attention - both in terms of staging and in terms of the complexity and 'realism' of its envelope. The sound dedicated to the data from those who only sometimes feel lonely is the next-most similar to a piano, as a metaphor for the implied empathy with the pianist. By the same token, the second (most unrealistic) synthesised piano is symbolic of the concept of it having the least in common with the live pianist. Its sounds are noticeably dissimilar to the real piano, metaphoric of the idea that the audience should empathise with the live pianist the most, out of the three sounds.

The 'disembodiment' of the two synthesised piano sounds intentionally shifts focus away from them. The simultaneous presentation of sonic and visual stimuli is further discussed later, but the removal of the visual cue of a performer from the synthesised piano sounds in this case reframes the solo pianist as the primary source of engagement in the performance.

The overall effect is designed to encourage the audience to associate with the pianist more than the synthesised sounds, in order to present a figurative 'cure' for the loneliness the pianist feels. This is achieved by the selection of timbres and staging cues; the data, in their raw form, could not communicate this, so it is incumbent on the composer to do so instead. The composer devises a musical response to data, not a scientific sonification, and there is therefore a degree of justifiable 'artistic license'.

3. WIDER PERSPECTIVES

The use of data in modern society is universal; in science, marketing, politics, and industry, its use is critical. Contemporary art, however, does not reflect this ubiquity. Political music has existed for countless years, but given the comparatively recent emergence of data as a force in the digital age, a composer must now reconcile themselves with the consequences of using data that is, at its core, derived from humanity. Musicologists have critically interrogated the composition of music using techniques derived from traditional Western genres, and indeed the place of abstract music as an art form has been problematised, by commentators such as Cornelius Cardew and Susan McClary.

There are a panoply of different projects that data use of this sort could produce. One of particular interest is that of the installation. Installations offer a number of exciting opportunities for composers and artists, in that they can transcend the *musica/informatica* axes of Vickers & Hogg, by presenting abstract sound and highly representative data simultaneously, as in work of Ryoji Ikeda (for instance, his audiovisual work *datamatics [ver. 2.0]* [14].

Furthermore, there still exists a bias toward trusting visually displayed data more than sonified [15]. Simultaneous presentation of visual and sonic data provides an opportunity to the viewer to strengthen the bond between what they see and what they hear [16]. Ikeda himself touches on this, commenting on *datamatics [ver. 2.0]*: "I like the invisible phenomena in sound. Data you can see as a result on the display monitor, but the concept of data is so abstract you can't touch it." [17].

As a composer, though, I am concerned with presenting data in an artistic and immersive manner. It is known that, when one of the senses is neglected, "information and meanings derived, and the affective engagement invoked, will be decreased; everything from realism to user satisfaction, from dimensionality to ease of use, will suffer unacceptably." [15]. Installations, by offering a parallel representation of sonified and visualised data, can combat pro-graphic bias in the public's eye. They can provide a more stimulating artistic experience, where the communication of data and the audience's emotional response are both improved. This can lend installations an element of the sublime, where the constituent parts of the work meld together for an overall effect, 'greater than the sum of its parts' [18].

If the aim is art, there are those who have employed algorithms to aid creation and tailored the outcomes according to their needs [19, 20]. A purely formalist approach is promoted by those concerned with the scientific sonification of data, but is not always adopted by artists - certainly not in the instances of the case studies above - and this raises ethical questions. Is artistic interpretation of data, or the alteration of a sonic product of said data, essentially a lie? Provided a disclaimer that the product is artistic as opposed to scientific or factual is given, there is no ethical dilemma, but there are always considerations like these to be taken into account when using (potentially contentious) data from human subjects.

However, if the aim is to report on data to the wider public, via the vehicle of the media, for instance, there are undercurrents and tensions. The idea of collecting verifiable data in the case studies has been discussed, but what happens when those data are not available, or are corrupted? 'Fake news', and its relationship with music and musicology, are topics for a different paper,[2] but it can still be discussed how the presentation of fake news might be undertaken by a composer.

As previously mentioned, data that are sonified are inherently less trusted than those that are visualised. How could this distrust be brought to bear in a composition? The approach could be somewhat similar to the example of the second case study - by using unfamiliar or 'mock' instruments, that are somehow alien to the listener. However, whilst most listeners might not be familiar with the nuances of every possible synthesiser configuration, many people could be reasonably expected to have come into contact with abstract synthesised sounds due to their widespread presence in popular music. For the production of a piece based on fake news, a more radical direction is proposed: rather than a new instrument that may not be noticed, it is suggested that fake news is sonified alongside verifiable data by using two opposed tuning systems. For instance, due to its almost universal application in Western music since the mid 19th century, verified data could be sonified using just intonation, whereas other, uncommon or unique tuning systems could be sought for the fake news. This would create a parallel between the sonic dissonance created, and the cognitive dissonance we experience when we hear a 'fact' based on false evidence.

Wider than these positions, though, are the reasons *why* a composer might find themselves drawn to writing data-music. Susan McClary has argued that composers have been very frequently inspired by abstract concepts of 'talent' and 'genius', even more so than the other art forms [23]. Contrary to this position, though, is that which this paper takes: that music is "essentially a human, socially grounded, socially alterable construct" [23].

A literal interpretation is acceptable: that one's compositions are rooted in the societal history and mutual consciousness of humanity. However, on a more practical level, the idea that music is deliberately mystified by producer and consumer alike, is the stance against which this paper stands. Reification of music above the real-life circumstances in which it is created is, in the mind of Cardew at least, something to be fought against, as it does not "relate directly to the issues and struggles and preoccupations of the present, and lead the way forward to a better society" [9].

4. CRITICISM

Individual drawbacks have been highlighted in the case studies concerning legibility of data, and the use of abstract, musical sounds in sonifications. There is wider debate about the field of sonification as it currently stands, and its drawbacks. For example, it has been argued that, in spite of efforts like ours to create socially responsible pieces of music, sonification is an expensive undertaking [18]. That being said, I contend that, actually, the key components –a computer and an internet connection, and some software with which to sonify data– are no longer prohibitively expensive. SuperCollider is free, and 9/10 UK households have internet access [23]. The issue is slightly more complex than this, however, as the software required for sonification requires a certain amount of time and learning to understand - privileges which are not afforded to everybody. There is more work to

[2] For instance, there has been a recent spike in 'fake' albums being 'released' by popstars, later to be discovered to be fan-made hoaxes [22].

do on interactive software, to "encourage casual exploration of sonification", so as to improve access to the field [24].

More pressing than this, the issue of how to interpret data that have been sonified comes to the fore. These concerns mainly break down into two areas: cognitive or perceptual abilities of the listener, and musical or aesthetic considerations taken by the sonifiers.

By mapping data to the range of a piano proportionally, the second case study addresses one of the issues of perception: that, should this decision not have been taken, the resulting frequencies will be outside of a human's hearing range [15]. Similarly, being able to discriminate between sounds as individual events is important if those sounds represent discrete data points. The first case study accelerates to such an extent that, whilst data are audible as individual points at the beginning, the overall effect towards the end of the piece is more aimed towards presenting a trend than an array of individual data. Moreover, there exist, and may never exist, no agreed methods for sonification in the wider community, such that a level of interpretation or familiarity with the material or method are required in order to get the most detail from a sonification [24, 15].

Musical considerations, too have been highlighted as areas of contention. Vickers and Hogg have claimed that sonifications and music are the same thing, depending on perception [8], and have highlighted how attention to musical factors (realistic timbre or familiar tonality, for example) can aid in the communication of data as sound. In both case studies above, musical factors have been brought to bear (such as the dramatic *accelerando* effect in the first, or the sparse texture in the second), but these are compositional choices, not derived from information inherent in the data. I contend, though, that the decision to present the pianist as alone on the stage (in the second case study), facing away from the audience, are artistic choices that, whilst not *demanded* by the data, are considerations that are *in line with* the nature of the data.

The two case studies herein are evidently not purely scientific sonification - they do not communicate quantitative, discrete data to an audience. However, through the implication of musical and extramusical compositional decisions that have been taken, a degree of the original meaning of the data is still possible to pick up, from the perspective of the listener. Where, then, does this leave the case studies in the wider context of data-music? The use of the term 'data-music' is not by accident: it is contended that, for cases where music is derived from data (rather than 'pure' sonification), 'data-music' is a better term, as it accounts for the music of composers who have "turned to sonification as a seed or driver for their work" [13]. By doing so, it allows for the scientific pursuit of the representation of data in sound, whilst establishing a parallel space for composers looking to the field for inspiration in their choices as artists.

5. CONCLUSION

This paper has explored the engagement of contemporary composers with data in their music, and how commentators have reconciled various usages with the abstraction of musical art. This paper has gone beyond the sonification of data, and its categorisation on an 'information-music' continuum, and has discussed the use of installations and other ways in which composers can use this continuum to generate ideas for new compositions, or alter current compositions. The limitations of writing music based on data are also acknowledged, in terms of communicating the underlying nature of those data, or decoding the 'meaning' behind the data.

Moreover, the political reasons for writing music based on data (informed by the arguments of composers and musicologists) have also been discussed. Music based on theories of tonal and post-tonal composition have their place in the canon, yet, as has been expressed, data-music is still an emergent art - this paper has sought to outline ways in which data can be used to write music, and, furthermore, to compose whilst cognisant of contemporary issues in musicology, and in wider political theatres.

6. ACKNOWLEDGMENT

With thanks to Nick Collins and Eric Egan for supervision in matters musical and political. I am also indebted to Erin Johnson-Williams, for her assistance in developing my musicological thought.

7. REFERENCES

[1] J. A. Obar, and A. Oeldorf-Hirsch. "The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services." *Information, Communication & Society*, pp.1-20, 2018.

[2] D. Szetela, "Customers Now: Profiting From the New Frontier of Content-Based Internet Advertising." Bloomington, IN: iUniverse Incorporated, 2009.

[3] S. M. Price, "Risk and Reward in Classical Music Concert Attendance: Investigating the Engagement of 'Art'and 'Entertainment'Audiences with a Regional Symphony Orchestra in the UK." Diss. University of Sheffield, 2017. [Online]. Available: http://etheses.whiterose.ac.uk/16628/1/Sarah%20M%20Price%20-%202017%20-%20Risk%20and%20Reward%20in%20Classical%20Music%20Concert%20Attendance.pdf

[4] R. Sessions, "How a 'difficult' composer gets that way." *New York Times* 89, 1950.

[5] I. Xenakis, "Formalized Music." Stuyvesant, NY: Pendragon Press, 1992.

[6] C. Flego "R. Luke DuBois 'Music into data::Data into music", 2019. [Online]. Available: https://www.academia.edu/38372818/Luke_DuBois_Music_into_data_Data_into_music_._Approaching_interactive_media_art?source=swp_share.

[7] A. de Campo, and M. Egger de Campo, "Sonification of Social Data." In *Proceedings of the 1999 International Computer Music Conference,* ICMC, 1999.

[8] P. Vickers, and B. Hogg, "Sonification abstraite/sonification concrète: An 'aesthetic perspective space' for classifying auditory displays in the ars musica domain." in Proceedings of the International Conference on Auditory Display, London, 2006.

[9] C. Cardew, "Stockhausen serves imperialism, and other articles: With commentary and notes." Latimer New Dimensions, London, 1974.

[10] A. Reynolds, “Are Mass Shootings Becoming More Frequent?”, Cato Institute, 2018. [Online]. Available: https://www.cato.org/blog/are-mass-shootings-becoming-more-frequent.

[11] “Montly Nics Bacground Check Totals Since Nov. 1998”, 2019. [Online]. Available: https://github.com/BuzzFeedNews/nics-firearm-background-checks/blob/master/charts/total-checks-all.png

[12] D. Snape, and S. Manclossi, “Loneliness in children and young people”, Office for National Statistics, UK, 2018. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/lonelinessinchildrenandyoungpeople.

[13] P. Vickers, “Sonification and Music, Music and Sonification”. In *The Routledge Companion to Sounding Art*. Routledge, London, pp. 135-144, 2016. [Online]. Available: http://nrl.northumbria.ac.uk/24597/.

[14] R. Ikeda, et al. “Ryoji Ikeda: Datamatics.” Charta, 2012.

[15] B. N. Walker, and M. A. Nees. "Theory of sonification." In *The Sonification Handbook,* pp.9-39, 2011. [Online]. Available: https://sonification.de/handbook/download/TheSonificationHandbook-HermannHuntNeuhoff-2011.pdf.

[16] N. Sagiv, R. T. Dean, and F. Bailes, *Algorithmic synesthesia*. na, 2009.

[17] M. Barnes, “Ryoji Ikeda - music for percussion + datamatics [ver. 2.0]”, Barbican, London, 2018. [Online]. Available: https://www.barbican.org.uk/digital-programmes/ryoji-ikeda-music-for-percussion-datamatics-ver-2-0.

[18] T. Rutherford-Johnson, “Music After the Fall: Modern Composition and Culture Since 1989.” Univ of California Press, 2017.

[19] P. Doornbusch, “Composers’ views on mapping in algorithmic composition,” Organised Sound, vol. 7, no. 2, pp. 145– 156, 2002.

[20] A. McLean and R. T. Dean, The Oxford Handbook of Algorithmic Music. New York, NY: Oxford University Press, 2018.

[21] S. McClary, “The blasphemy of talking politics during Bach Year” (1987) in *Reading Music: Selected Essays*, Ashgate, pp.13-62, 2007.

[22] A. X. Wang, “A Fake Rihanna Album Climbed the Music Charts this Weekend”, *Rolling Stone,* 2019. [Online]. Available: https://www.rollingstone.com/music/music-news/fake-rihanna-album-charts-803144/.

[23] C. Prescott, “Internet access – households and individuals, Great Britain: 2018”, Office for National Statistics, UK, 2018. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2018#9-out-of-10-households-have-internet-access.

[24] G. Kramer, et al., "Sonification report: Status of the field and research agenda." 2010. [Online]. Available: http://icad.org/websiteV2.0/References/nsf.html.

SPEECH COMPANIONS: EVALUATING THE EFFECTS OF MUSICALLY MODULATED AUDITORY FEEDBACK ON THE VOICE

Rébecca Kleinberger, George Stefanakis, Sebastian Franjou

MIT Media Lab
75 Amherst Street, MA, Cambridge
rebklein@media.mit.edu, stefanag@mit.edu, sfranjou@mit.edu

ABSTRACT

Changing the way one hears one's own voice, for instance by adding delay or shifting the pitch in real-time, can alter vocal qualities such as speed, pitch contour, or articulation. We created new types of auditory feedback called Speech Companions that generate live musical accompaniment to the spoken voice. Our system generates harmonized chorus effects layered on top of the speaker's voice that change chord at each pseudo-beat detected in the spoken voice. The harmonization variations follow predetermined chord progressions. For the purpose of this study we generated two versions: one following a major chord progression and the other one following a minor chord progression. We conducted an evaluation of the effects of the feedback on speakers and we present initial findings assessing how different musical modulations might potentially affect the emotions and mental state of the speaker as well as semantic content of speech, and musical vocal parameters.

1. INTRODUCTION

This work seeks to assess how different musical feedback modulations might affect the general mental state of the speaker, semantic content of speech, emotions in vocal tonalities and vocal parameters of musicality. Modulated Auditory Feedback uses digital signal processing to transform the way someone hears their own voice. Modulated Auditory Feedback has documented effects on how someone speaks in terms of speed, articulation, and fluency. For example adding a short delay to the voice can lead to prolongation of vowels, repetition of consonants, increased intensity of utterance, and other articulatory changes [1, 2]. A short delay (20-150ms) can help people who stutter become more fluent [3] but a longer delay (higher than 200ms) can lead to jammed speech [4].

In recent years, the research community has investigated the possible effects of altering vocal auditory feedback for regulation of emotions [5, 6]. In these studies, modulated feedback is used covertly to make the voice sound more calm, sad, happy or fearful by manipulating formants, overall pitch, and by adding filters. The researchers then established the effects on the subjects measured through self-reported emotions and levels of stress. Our approach consists of producing aesthetic musical manipulation of the voice instead of covert intonation and testing the effects on the speakers fine-tuned ability to shape their voice and speech. Musically Modulated Auditory Feedback is a new approach that creates real-time musical transformations of the voice, for instance by generating guitar chords accompanying the rhythm and pitch variation of the voice, or by creating several pitch shifted versions of the voice and layering them in real-time to create a choir-like harmonization of the voice. We conducted a study to assess whether specific Musically Modulated Auditory Feedbacks can induce particular effects and modulate emotional content from the voice, in addition to affecting vocal parameters. Our objectives are twofold: first, we are interested in studying the potential regulatory effects of music when woven into voice. Second, we wish to bring more awareness to the intrinsic musicality present in everyday speech and explore possible research applications based on perceiving the spoken voice as an inherent musical signal. These applications range from infant-directed speech and language acquisition to speech pathology and aphasia re-education. Such research could also show useful for music composer or could lead to new tools for phonologists to characterise human speech. We present the background supporting our inquiries in terms of neurology, research on music and emotion, and self-perception theory. Then we present the study design and detail the data analysis and results.

2. BACKGROUND

2.1. Musicality of everyday speech

Speech is one of humanity's richest and most ubiquitous forms of communication. Its richness lies in the combination of linguistic and nonlinguistic information. Musicality is a crucial nonlinguistic component of speech, incorporating the tempo and rhythm of the speaker along with the pitch variation and unique texture of vocal sound. In casual everyday speech, individuals possess a unique musicality, rhythm and melodic style. In 1954, urban folklorist and sound archivist Tony Schwartz proposed the idea that "there is music in everyday speech, and often a kind of poetry in the way people talk" [7]. In our work, we aim to increase awareness of the beauty and diversity of musicality in our everyday experience of voices. Vocal, non-verbal behaviors such as prosody, tone, loudness, breathiness, accent, pitch envelope, and tempo are all parameters that are most often unconsciously controlled when speaking, but they implicitly convey a great deal of information. For instance, pitch intervals can reveal changes in mood [8] or hormone levels [9], tempo information can be a marker of depression [10]. Prosody and especially pitch accentuation can also be used to modify semantic content [11].

Our system creates different types of musical layers on top of the spoken voice by extracting existing musicality from speech and aims to bring more awareness on this intrinsic musicality.

https://doi.org/10.21785/icad2019.035

2.2. Music and Emotion

The influence of music on emotion is not a novel concept. As early as 350 B.C., Aristotle characterized different musical modes by the emotions they evoked [12], and throughout the classical age of music, the "feel" of a piece was often married to more objective qualities like tempo and chord. In terms of valence, minor keyed pieces and melodies are traditionally associated to sad, nostalgic or morose atmospheres, including Chopin's Funeral March and Mozart's Requiem. On the other hand, major keyed-pieces are classically associated with joyful, strong and uplifting athmospheres, including Mendelssohn's Wedding March and Rossini's William Tell Overture. Whether some innate qualities of the major and minor tonalities informed theory and popular opinion, or vice versa, is a philosophical inquiry which is not to be dwelled upon, but the popular social perception of the major and minor chords, for hundreds of years in the western tradition, has been as that the former is classically joyful, while the latter is often considered sorrowful [13]. Of course there are exceptions; many pieces of music exist which do not follow this categorization. Furthermore, the perceptions of individual pieces can vary widely from person to person. The famous paper published by Hevner et al. in 1935 elucidates the various affective qualities of the major and minor musical modes [14]. The author claims that major is dynamic, more natural and fundamental than minor, and "expresses varying degrees of joy and excitement." She goes on to assert that "[the major] sounds bright, clear, sweet, hopeful, strong, and happy," while the minor "expresses gloom, despair, sorrow, [and] grief," and is "mournful, dark, [and] depressing." Many theories and studies have supported this notion of musical modes having intrinsic emotional connotations implicit within them, and several support the idea that music can indeed evoke strong emotional responses in listeners [15, 16].

Although the findings that minor chords have a negative valence effect have been presented in many prior work on music emotion, as of the time of this writing, we haven't found any prior work assessing effects of the use of minor vs major keys when interactively woven into spoken voice. In this work, we are proposing a step toward assessing unconscious effects of auditory musical transformation of speech.

2.3. Self-Perception Theory

In his self-perception theory, Daryl Bem [17] postulates that individuals come to know their own attitudes, emotions, and other internal states partially by inferring them from observation of their own overt behaviors. He argues that internal cues are "weak, ambiguous or uninterruptible", and that we often have to rely on external cues to understand our own behaviors the same way an outside observer would.

This theory suggests that it is partly by monitoring the way we overtly express our emotions that we infer our emotional state and attitudes. Multiples studies support this theory, by showing that forcing the outside symptoms of an emotion can reinforce said emotion in the subject [18]. Similar results have been obtained for vocal expression of emotion: subjects asked to imitate vocal patterns associated with specific emotions (eg. laughter) reported their emotions being affected accordingly [19]. The previously mentioned studies involve active cooperation from the subject, but further studies have found similar effects in cases where patients didn't have to consciously adjust their behavior, or weren't even aware of anything being modified. Subjects who heard their voices processed in real-time to make it sound as if they were happy, sad, or afraid experienced changes in tension and self-reported positivity usually associated with the experience of the corresponding emotion. This suggests an influence of the perception of the subjects own voices alone on their emotions despite them not even noticing any modification of their voices [6]. Similarly, participants whose voices were modified to sound calmer and fed back to them in real-time during relationship conflicts reported feeling less anxious than those having unmodified feedback [5]. These studies suggest that emotions can be regulated by feeding back modified version of a speaker's voice in real time even if the modification is not consciously detected.

In our work, we explore this field by modifying the subjects' fed back voices to match purely musical expressive features. Links between prosodic and musical emotional features have been suggested, such as the use of the interval of a minor third for affects of negative valence for both speech and music [20]. By mapping the fed-back voice to musical attributes considered happy or sad we hypothesize similar emotional responses to those induced by previous non-musical manipulation.

2.4. Neural Basis

A large body of work conducted on neural control of speech has been accumulated in Frank Guenther's book of the same name. A key idea presented in the book is that of neural auditory feedback control, which is operated by means of a feedback/feedforward mechanism. In this scheme, it is suggested that fluent speech is dependent on fluid, logical, sensory feedback streaming back to the speaker. It is for this reason, Guenther asserts, that delayed auditory feedback results in a range of dysfluent behavior, up to and including complete cessation of speech [21]. The importance of auditory feedback in speech production has been further proven by studies on the effect of modified real-time and delayed feedback on speech and sustained vowel sounds. It was found that modification of the fundamental frequency (F0) of the feedback voice produces a compensatory opposing shift in the pitch of the resultant sound for both sustained vowels and speech [22, 23] due to brain overcompensation. Formant shifts in feedback have also been found to produce compensatory changes in the spectral characteristics of the voice [24], even when participants were consciously informed of the modifications and instructed not to compensate [25]. Thus it appears that auditory feedback plays a crucial role in speech production, to the point where it sometimes cannot be ignored even if the speaker is consciously trying to combat its effects. The neurological basis of our study is to interfere with the encephalic speech-feedback mechanism by overlaying the stream of one's own raw voice, using musical modulations. The goal is to monitor the alterations in the resulting feed-forward mechanism of new speech being produced. We also seek to analyze the semantic nature of speech produced when the backward-fed vocal audio is substantially altered in either major or minor chord progressions.

2.5. Measure of Musical Parameters in Speech

It can be difficult to assess and characterise the musicality of speech. The question is so polemic that quite often, researchers assess the level of musicality by asking experts with extensive music training to subjectively rate vocal sound samples. In Music, Language and the Brain [26], Aniruddh D. Patel distinguishes musical and linguistic sound systems in the way they carry pitch, timbre, rhythm and melody. One way to assess pitch is through the

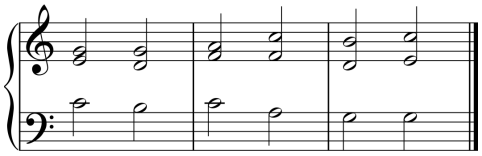
analysis of the mean pitch (Pm) of a vocal sample. Pm provides information about the fundamental frequency (F0) of a subjects voice. Males with lower voice will have smaller F0 thus lower Pm. Level of melodicity can be very roughly accessed through the measurement of the pitch standard deviation (Psd) of a vocal sample. Psd gives cues about the pitch envelope in speech: the lower the Psd in a given phrase, the more monotone and concentrated around the main pitch, the voice will be. Contrary to a lot of musical systems and instruments that present a fixed timbre, speech is also fundamentally a system of organised timbral contrast, as timbral variation in vocal sound is the basis of phoneme production. In addition, on account of the shape of formants, subtle vocal timbral variation is what allows us to distinguish the voice of different speakers. Timbre in speech can be measured with different parameters such as jitter, breathiness, or harmonic-to-noise ratio (HNR expressed in dB). HNR is a more global way to see timbre as it indicates the energy concentration of the sound around the main pitch. HNR represents the degree of acoustic periodicity. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise. And a HNR of 20dB means that 99% of the energy of the signal is in the periodic part. Singing voices have higher HNR than spoken voices [27]. Pm, Psd and HNR are used in this study as measurement of variation of musical parameters of speech. In this work, we did not intend to measure rhythm in speech partially as it is used in the generation of the pseudo beats in speech companions.

3. SPEECH COMPANIONS

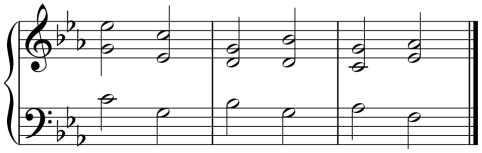
We created new types of auditory feedback called Speech Companions that generate live musical accompaniment to the spoken voice. The Speech Companions used for this study are based on a type of active harmonizer. An harmonizer is a pitch shifter that combines the shifted pitch version with the original sound to create a two or more notes harmony. Our system combines the original vocal signal with two extra layers creating a musical chord. A constant harmony chord being played in a sustained manner can create a very dull effect. In live or studio music production, harmonizers are often controlled manually by a keyboard that changes chords to make it more reactive. For our study, we wanted the feedback to react to the inherent rhythm of speech. Our system triggers a new chord, from a predetermined set, at each pseudo-beat of speech.

Pseudo-beats are triggered at near-regular intervals determined by minimum delay and natural attacks in the voice. Sound attack corresponds to onset or peak in the intensity of the sound signal. After each chord change, the system counts down the chosen delay in milliseconds and then waits for the next speech onset to generate the next pseudo-beat controlling the next chord change. When chords are changed at a regular interval, the feedback seems very static and creates a ticking clock effect that can feel stressful and alter the natural speech rhythm. By using the pseudo-beat method, we ease the chord variation into the organic speech tempo to respect the built-in musicality of speech. The system was implemented using Max MSP 8 for pseudo-beats detection and with MHarmonizerMB for Reaper64 to create the harmonization.

The system randomly draws a chord to harmonize from a predetermined chord progression - either major or minor. The chord progressions were chosen to unambiguously convey the key and mode regardless of which order the chords were played in, as they were to be fed to the subjects in random order. The key of C was chosen, and the chords are in the modes of C ionian (major) and aeolian (natural minor) (see Figures 1.a and 1.b). Although commonly used by classical composers, the harmonic minor was avoided as the augmented second interval can sound jarring or exotic to western listeners. This interval is usually avoided by following proper voice leading rules, but this wasn't possible due to the random order of the chords. The aeolian or natural minor mode, commonly found in popular music, was chosen instead to bypass this problem. The chords are voiced in the mid-range so that the harmonized feedback would not sound too distant in pitch from the normal voices of most subjects. The range and spread of the chords were kept comparable (see Figure 1). The major chords are triads, while the minor chords are sometimes enriched to convey more tension and sadness.



(a) Major chord progression



(b) Minor chord progression

Figure 1: Chord progression used for for the major (a) and minor (b) mode of our study

Figure 2 illustrates the use of the pseudo-beat to trigger changes in the MIDI track that always last longer than a minimum delay and are ultimately triggered by speech attacks from the raw voice. The result generates harmony changes in the processed voice (middle track) that exhibit different spectrum peaks than the raw voice. In this case each chord lasts a minimum of 3000ms but can extend longer if no attack is detected. The volume was kept the same for all participants and was loud enough to mask the actual voice. We hypothesise that such feedback might affect the valence of the speaker as well as the musicality of their speech.

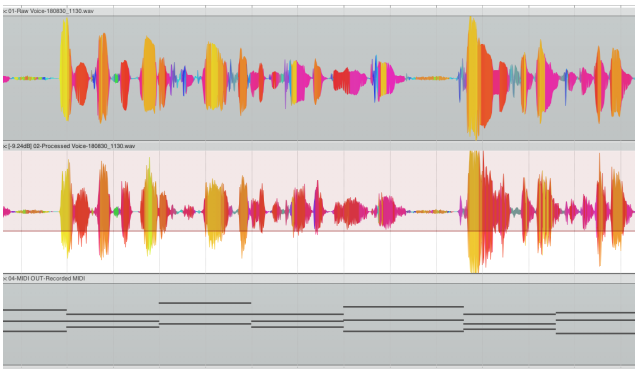


Figure 2: Illustration of the Speech Companion in use: attacks in the raw voice (top track) trigger the midi chords (bottom tracks) that control harmony changes in the processed voice (middle track)

4. STUDY DESIGN

4.1. Participants

The institutional review board approved this study, which was registered as COUHES protocol no.1802248976. The sample comprised 20 adults (11 women and 9 men). There were two groups: one group of 10 adults received the "major scale" condition, and the other group of 10 adults received the "minor scale" condition. No compensation was offered to the participants. The study was organized over 5 days, in which we measured respectively 1, 3, 6, 5 and 5 participants. The settings were identical throughout the 5 days in terms of environment, microphone settings, audio loudness, and lighting.

4.2. Study Setup

The study was conducted in a soundproofed room to reduce background noise. We used a Countryman E6 directional ear-set microphone and a Babyface RME Pro audio interface connected to a computer to record the voice and a pair of Bose SoundSport earphones to provide audio feedback. The SoundSport are very open (i.e. let outside sound in) which allowed the interactions between the subject and the researcher to remain natural. The researcher giving the instructions also wore a pair of SoundSport to monitor the quality of the feedback heard by the subject. The loudness of the feedback was set just loud enough to effectively cover the speakers voice without sounding unnaturally loud.

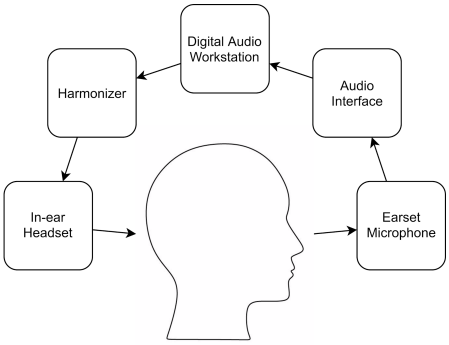


Figure 3: System Flow

4.3. Method

The study was composed of two phases (baseline and musical feedback) each containing the same three tasks (reading, mood assessment and storytelling). Subjects were initially fitted with in-ear headphones and a microphone. During phase 1, subjects did not hear any feedback through the headphones but still had to wear them to get accustomed to it in preparation for phase 2.

- Task 1 is a reading exercise to normalise the subjects mood to neutral at the start of the study. To this end, we use a adapted version of the Velten mood induction process (Velten MIP) method [28]. As we want to induce a neutral mood to all participants, we ask them to read a series of 15 trivial and factual statements which carry no emotional load extracted from the 50 sentences used in Velten MIP version used by Isen and Gorgoglione [29]. This reading task aims at initiating the same neutral common ground for each subject.

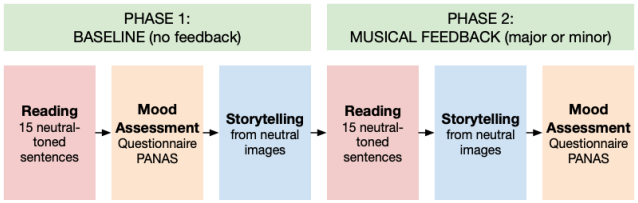


Figure 4: Order of the Study

- Task 2 consisted of filling out a short mood questionnaire to measure self-reported affect. We used the Positive and Negative Affect Schedule (PANAS) methodology [30]. The PANAS was chosen for its robustness, replicability, and widespread use, to allow for easy comparison with other works. To limit demand effects, it was issued in digital form where only one question was visible at a time. This prevented the subject from seeing the whole questionnaire and influencing to global results by correlating their answers to several questions.
- In task 3, subjects were shown four images from the IAPS image database and asked to construct a narrative loosely based on the images. IAPS is a database of images for experimental investigations of affect [31]. Each of the chosen images were in the valence range 4.5-5.5, signifying emotional neutrality. A scale score of 1 on the valence portion of the IAPS image scale means unhappy, while 9 means happy. Images with neutral-evoked emotions could go either towards the more joyous, or the more depressed, semantically and tonally.

For the entirety of phase 1, no audio feedback was played through the headphones, though audio from the microphone was being actively recorded. This initial neutral portion of the protocol was used as a baseline to evaluate the effects of the musical modes.

The study then entered phase 2, where the subject had to repeat the same three tasks while hearing the Musicalised Modulated Feedback. For each subject, either the minor or major chord harmonizer was tested and each subject listened to their voice modulated at a volume sufficiently high so as to mask their own voice.

- For this phase, task 1 was composed of 15 new neutral sentences to read.
- for task 2, the subjects were given four different images from the IAPS image database, from which to generate a new story.
- and for task 3, the subject was asked to fill a new randomised PANAS to fill out.

In phase two tasks 2 and 3 are switched compared to phase one as we wanted to give more time to the subject to get used to the feedback before measuring their self-reported mood in order to get a better sense of the change of mood induced by the study.

The musical modulations were then turned off and we asked the subjects their best guess about the purpose of the study to determine if they were aware that their mood and tone were being investigated. Indeed, past research has shown that results of studies on affect might be skewed or unintentionally affected if subjects are aware that their mood is being monitored [32]. At the end of the experiment we then verified that all the participants had remained unaware that the study was about affect and we informed them of the actual objective through a short debriefing session and asked not to divulge it to other potential participants.

5. DATA ANALYSIS

The collected data were processed into three categories: the self reported PANAS result were processed into numerical data. The stories generated (two per subjects) were analysed in two different ways: as text to assess semantic content, and as speech audio sample to assess changes in vocal affect and musicality.

5.1. Self-Reported Affect

The PANAS questionnaire was completed by the subject twice: once as part of the baseline evaluation, and once at the end of the musical-feedback task. The questionnaire gives us scores for positive affect (PA) and negative affect (NA), that are subtracted to obtain a valence score V normalize between -1 and 1. To limit the variations due to differences in initial mood between subjects, we analyzed the variation in valence induced by the experience by subtracting the valence prior and post study. This allowed us to only take into account mood changes from baseline induced during the study. These change in valence was then compared between the minor scale group and the major scale group.

5.2. Semantic Content

To analyze the semantic content of the speech, the audio recordings of the constructed narrative based on the pictures from IAPS in tasks 1.3 and 2.3 were all transcribed to text using Dragon NaturallySpeaking [33], and the text outputs were then reviewed manually and corrected to assure accurate transcription of speech. These text transcriptions were processed using the SentiWordNet database which scores words based on their positivity and negativity [34]. For each subject, we compared the difference in average positive, negative, and total scores from the SentiWordNet analysis between the baseline story and the story invented by the subject while hearing musical feedback.

5.3. Emotion Analysis from Vocal Intonations

Emotional vocal qualities were analyzed using the speech emotion recognition software OpenVokaturi [35]. OpenVokaturi is a Software Development Kit developed by Vokaturi to provide analysis of the basic emotions from speaker's vocal intonations. It is worth noting that the SDK is presented as having an accuracy on classification of only 66.5 % which highly limits the validity of the results [36]. Vokaturi provides percent likelihoods for neutrality, happiness, sadness, anger, and fear. Each speech audio sample was analyzed using the OpenVokaturi pretrained model. Scores for positive and negative affect were constructed by way of a weighted sum (Positive Affect = Happiness; Negative Affect = (Anger + Fear = Sadness) / 3), in a similar fashion to the PANAS's way of summing different positive and negative reported emotions to construct positive and negative affect [30]. We then took the differences between the scores for speech segments produced under the musically modified feedback and those produced under normal feedback conditions. To mitigate the effects due to subject particularities, we considered the relative change in affect between the baseline phase and the musical feedback phase rather than absolute affects

5.4. Vocal Parameters

We used Praat [37] for the analysis of vocal and musical parameters of speech. For the speech samples of the narative generated by subjects in phase one and two, we extract mean pitch (Pm), pitch standard deviation (Psd) and harmonic-to-noise ratio (HNR) of the voiced sections of speech. A vocal sound is said to be "voiced" when it originates from the vocal chord and not only from air leaving the lips (e.g. all vowels and diphthongs are voiced, consonants can be either voiced or unvoiced). The analysis parameters were set in Praat as followed: pitch was computer by autocorrelation between 44 and 400Hz with an octave jump cost of 3.5 on voice sections defined with a silence threshold of 0.05 and a voicing threshold of 0.25 and a voice/unvoice cost of 0.15. Detected pitch were also visually validated by researchers. For this section, we hypothesise that whatever the mode (major or minor), speech from phase 2 might have different Pm, Psd and HNR than speech from phase 1.

6. RESULTS

We report findings on data comparing changes in valence between the major scale and the minor scale group as well as changes of vocal parameters (Pm, Psd and HNR) induced for both group by the experience. All t-tests were preceded by an F-test to determine whether the samples should be assumed to have equal or unequal variance and the relevant paired t-test was then run accordingly. The significance level of all tests was set to p = 0.05

6.1. Results from Self-Reported Affect

We hypothesized that the minor mode would induce a more negative mood, and that the major mode would induce a more positive mood in the subjects. This was evaluated in three different ways. The first was self reported mood by means of a digital version of the PANAS questionnaire. We observed trends concurring with our hypothesis as the average in valence change was higher for the major scale group (3.3%) than for the minor scale group(1.2%) However a two-tailed T-test didn't show statistical significance. It is interesting to note that both groups general mood seemed to slightly increase after the study (with major mode increasing more than minor mode) which might be due to the surprise and novelty effect

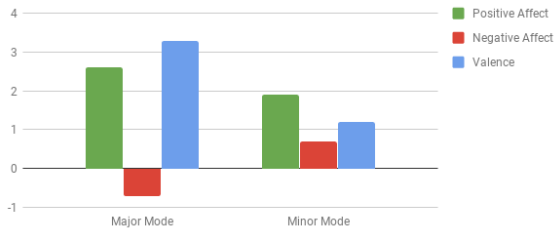


Figure 5: Difference in self-reported positive and negative affect

6.2. Results from Semantic Content

The semantic score analyses conducted on major and minor chord progressions centered around positive, negative, and valence word scores, which give holistic, normalized, numerical attributes of the degree to which the words spoken by a subject leaned more towards positive or negative speech. The valence score was calculated as the sum of the positive and negative scores. We used the Natural Language Toolkit (NLTK) [38] to obtain these scores, and the text was obtained from subjects image narratives, from phases 1 and 2.

We computed the differences in semantic scores from phase 1 to 2 of the study and then compared these across major and minor modes. We used a two-tailed t-test on the valence results as well as on the positive and negative results, and while our results didn't show statistical significance, they still present the expecting trends. We specifically observed that subjects from the minor group had a negative score difference (difference in holistic evaluation of negative words from phase 1 to phase 2), on average almost 6 times higher than those with the major mode; one-tail two-Sample t(18) = -1.0, p = 0.33 >0.05. Still, we cannot reject the null hypothesis with respect to semantic results.

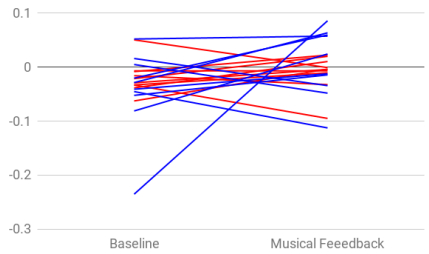


Figure 6: Evolution of semantic score valence (normalised between -1 and 1) between baseline and musical feedback for all participants the blue lines represent the subjects from the minor group and red lines represent subjects form the major group

6.3. Results from Emotion Analysis from Vocal Intonations

The third portion of analysis was comparison of the major and minor groups in terms of emotions extracted from the voice. To accomplish this, we used the speech emotion recognition software Vokaturi. As in the previous analyses, the speech used was obtained from subjects image narratives, from phases 1 and 2. We grouped the normalized Vokaturi data into three areas: positive affect, negative affect, and valence. In accordance with our hypothesis, the negative affect score was found to be significantly greater for subjects subjected to the minor mode compared to those subjected to the major mode. The two-tailed t-test, t(18) = -2.68, p = 0.015 <0.05, agrees with this finding and thus we can reject the null hypothesis here. We also localized this difference to vocal parameters indicating sadness and anger, which implies significantly that the minor mode heightens these emotions in the speaker.

Furthermore, we found that valence, or the difference between positive and negative affect scores, increased on average by over 5 times more for those who had the major mode versus those who had the minor mode; Two-Sample t(18) = 2.76, p = 0.013 <0.05. This serves to show that those who listened to the major mode feedback were much more vocally positive than negative, as compared to those with the minor mode. Although not significant, observed trends also suggest that the major mode increases happiness and positive affect in speakers. The significance of these results should also take into account the relatively low accuracy of the OpenVokaturi tool.

6.4. Results from Vocal Musical Parameters

When analyzing vocal musical parameters, we hypothesized that regardless of key (major or minor), speech from phase 2 might have different Pm, Psd and HNR than speech from phase 1.

A paired-samples two-tailed t-test was conducted to compare Pm between baseline and musical feedback conditions. There was no significant difference in the Pm between baseline (M=154.6Hz, SD=45.6Hz) and musical feedback (M=156.6Hz, SD=47.2Hz) conditions; t(19)=-0.80, p = 0.430 >0.5. This indicates that fundamental frequencies didn't change much in speakers with or without feedback.

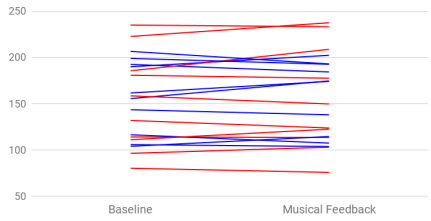
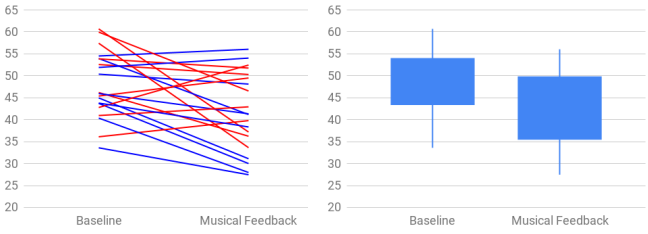


Figure 7: Evolution of mean pitch (in Hz) between baseline and musical feedback for all participants (blue lines for subjects in the minor group and red lines for subjects in the major group

However, significant differences were observed when running a paired-samples two-tailed t-test to compare Psd between baseline (M=47.9Hz, SD=7.5Hz) and musical feedback (M=41.8Hz, SD=9.0Hz) conditions; t(19)=3.024, p = 0.0069 <0.05. This result indicates that speakers became slightly more monotonous and pitch envelopes were less accentuated when hearing musical feedback. We might have expected that musical feedback would make subjects more melodic but instead it seems that as melodic and harmonic matter was added to their speech, they became more conservative in terms of accent, pitch contours and melody in their own produced speech.

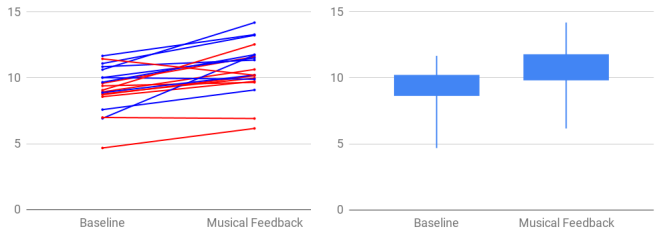


(a) (b)

Figure 8: Pitch standard deviation evolution (in Hz) between the baseline and the musical modes (blue lines for minor group and red lines for major group) (a) and for the entire population (b)

Finally, significant differences were also obtained when running a paired-samples two-tailed t-test to compare HNR between baseline (M=9.2 dB, SD=1.7dB) and musical feedback (M=10.7dB, SD=1.9dB) conditions; t(19)=-5.0, p = 0.000087 <0.05. This indicates that the spoken voice becomes more singing-like with a more precise and accentuated pitch.

Those two results indicate that in terms of timbre, the spoken voice becomes more music-like but in terms of pitch envelope, the speaker becomes more cautious and conservative. This could indicate that the subjects were distracted and further explorations should assess that potential element. This could also indicate that they were paying more attention to listening and integrating their own voice as music rather than language.



(a) (b)

Figure 9: Harmonic-to-Noise ratio evolution (in dB) between the baseline and the musical modes for each participant (with blue lines representing the minor group and red lines representing the major group) (a) and for the entire group (b)

7. DISCUSSION AND FUTURE WORK

When analyzing the data for possible valence and musical effect of musically modulated auditory feedback, we have observed some preliminary results suggesting a trend in the expected direction: self reported valence became more positive for subjects hearing the major mode than for those hearing the minor mode, though not to a statistically significant extent, on account of the small sample size. Analysis of semantic content of speech also didn't show significant results, suggesting that, if present at all, cognitive mood change due to major or minor chords is marginal. However, our study showed significant changes in vocal emotionality and in vocal musicality with a higher harmonic-to-noise ratio and lower pitch standard deviation. This suggests that the feedback makes peoples voice more song-like while reducing their pitch envelope and changes their vocal (but not verbal) emotional content. Additional studies would be necessary to better understand these effects and the factors contributing to them.

This exploratory work presents several limitations both in the context and format of the study. Relatively small sample size and possible order effects are elements that have to be addressed in our future studies. The next stage of the work will also include a different type of baseline where the subject hears their voice amplified at the same loudness without any modulation. Another possible comparison could be with a mode where subjects hear music unrelated with their speech, thought previous studies have indicated that this might create a high level of distraction. Indeed, being subjected to music has been shown to be detrimental to short term memory and cognitive tasks such as reading or processed word writing [39]. In our study, it seemed that the modulated feedback didn't affect to a large extent the ability of the subjects to speak and concentrate. Our subjects seemed sometimes slightly less cognitively and vocally fluent with the feedback but to a lesser degree than one would expect with background music at the same loudness. However, it might still be interesting in the future to assess the level of distraction induced by the system and see how distraction might be reduced when musical stimuli are responsive to user input compared to non-interactive stimuli such as background music.

Further investigations are required to also test factors such as novelty effects, social dynamic related bias, or task induced variability. The musically-altered feedback was quite novel and unusual for many, and some subjects found it, at first, to be amusing or intriguing. Such reactions would tend to indicate an initial boost of positive affect that could then skew the results and mitigate the expected variations, especially in the minor direction. In future explorations, we are also interested in comparing the reaction with Speech Companions when made to adjust to the subject's natural voice range, and see if the system can be made to blend even more with the natural musicality of the voice compared to imposing externally defined musicality onto it. Finally, in this study, the vocal modifications were made obvious, and the subjects were informed of the presence and general characteristics of the modifications. We believe it would be of interest to determine whether a more subtle modification (eg. lower feedback volume) would have comparable, enhanced or reduced effects, similar to the way both pitch shift compensation has been studied for both uninformed [22, 23] and informed [25] subjects. Finally, due to the number of people surveyed and the time frame of the study, we did not include group with no feedback as a control, but in future research we hope to test additional subjects, of which some will not hear any feedback and some will only hear background music while they speak. These extensions would help to buttress the findings of this study.

In terms of real-life applications, we are currently exploring the potential of musically modulated auditory feedback in different contexts, from assistive tools to increase fluency for adults who stutter, to systems for students musicians to better connect with music, to tools for composer who want to explore the musicality of the voice. We also believe that adaptation of such system might be beneficial in new sorts of practices as a possible therapy or relaxation assistant, due to their potential effectiveness in modifying both various voice characteristics and perceived emotion.

8. CONCLUSION

In this study, we created a new type of digital audio manipulation to generate real-time manipulation of the voice through Musically Mediated Auditory Feedback. Classification results significantly indicate that such feedback might alter voice quality and emotion valence detected from voice tonalities. Significant changes in vocal timbre and pitch variation were observed showing the potential to affect speech musicality at a subconscious level.

This early exploration proposed original ways to manipulate the voice in real-time as a way to potentially affect internal mental and physical processes in speakers. By musically altering the way people hear their own voice, we also aim to raise questions about the existing underlying effects of musicality already present in the voice and its reinforcing potential in terms of enhanced emotional regulation, self-awareness, and musicality, in the context of everyday speech.

Speech is one of the richest and most ubiquitous modalities of communication used by human beings. Its richness lies in the combination of linguistic and nonlinguistic information it conveys. Musicality is one of the most crucial nonlinguistic components of speech; it includes the tempo and rhythms of the speaker as well as the pitch variation and unique texture of the vocal sounds. Abstracting musicality from a speech in real time presents several challenges, but explorations in the domain of musically modulated speech and feedback could open doors to to explore real-life situations where the music of speech impacts speakers or listeners such as in the contexts of infant-directed speech, language acquisition, human-animal communication, speech pathology, aphasia re-education, or even music learning and musical composition.

9. REFERENCES

[1] A. J. Yates, "Delayed auditory feedback." *Psychological bulletin*, vol. 60, no. 3, p. 213, 1963.

[2] G. Fairbanks and N. Guttman, "Effects of delayed auditory feedback upon articulation." *Journal of Speech & Hearing Research*, 1958.

[3] J. Kalinowski *et al.*, "Stuttering amelioration at various auditory feedback delays and speech rates," *International Journal of Language & Communication Disorders*, 1996.

[4] G. Fairbanks, "Selective vocal effects of delayed auditory feedback." *Journal of Speech & Hearing Disorders*, 1955.

[5] J. Costa *et al.*, "Regulating feelings during interpersonal conflicts by changing voice self-perception," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 631.

[6] J.-J. Aucouturier *et al.*, "Covert digital manipulation of vocal emotion alter speakers emotional states in a congruent direction," *Proceedings of the National Academy of Sciences*, vol. 113, no. 4, pp. 948–953, 2016.

[7] T. Schwartz *et al.*, "Interview with tony schwartz, american hörspielmacher," *Perspectives of New Music*, 1996.

[8] S. K. Blau, "Musicality of speech changes with mood," *Physics Today*, vol. 63, pp. 16–17, 2010. [Online]. Available: http://physicstoday.scitation.org/doi/10.1063/1.4797228

[9] D. R. Feinberg *et al.*, "Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice," *Hormones and behavior*, 2006.

[10] Y. Yang *et al.*, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, 2013.

[11] V. K. R. Sridhar *et al.*, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *Proceedings of Speech Prosody*. International Speech Communication Association Campinos,, Brazil, 2008.

[12] S. Halliwell *et al.*, *Aristotle's poetics*. University of Chicago Press, 1998.

[13] R. G. Crowder, "Perception of the major/minor distinction: I. historical and theoretical foundations." *Psychomusicology: A Journal of Research in Music Cognition*, 1984.

[14] K. Hevner, "The affective character of the major and minor modes in music," *The American Journal of Psychology*, vol. 47, no. 1, pp. 103–118, 1935.

[15] J. A. Sloboda, "Music structure and emotional response: Some empirical findings," *Psychology of music*, 1991.

[16] K. R. Scherer, "Expression of emotion in voice and music," *Journal of voice*, vol. 9, no. 3, pp. 235–248, 1995.

[17] D. J. Bem, *Self Perception Theory*, 1972.

[18] P. M. Niedenthal, "Embodying emotion," *science*, vol. 316, no. 5827, pp. 1002–1005, 2007.

[19] E. Hatfield and C. Hsee, "The impact of vocal feedback on emotional experience and expression," 1995.

[20] M. E. Curtis and J. J. Bharucha, "The minor third communicates sadness in speech, mirroring its use in music." *Emotion*, vol. 10, no. 3, p. 335, 2010.

[21] F. Guenther, *Neural Control of Speech*, MIT Press, Ed., 2016.

[22] T. A. Burnett *et al.*, "Voice f0 responses to manipulations in pitch feedback," *The Journal of the Acoustical Society of America*, 1998.

[23] ——, "Voice f0 responses to pitch-shifted auditory feedback: a preliminary study," *Journal of Voice*, 1997.

[24] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation of speech i: Compensation and adaptation," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 2, 2002.

[25] K. G. Munhall *et al.*, "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 384–390, 2009.

[26] A. D. Patel, *Music, language, and the brain*. Oxford university press, 2010.

[27] E. L. Stegemööller *et al.*, "Music training and vocal production of speech and song," *Music Perception: An Interdisciplinary Journal*, vol. 25, no. 5, pp. 419–428, 2008.

[28] E. Velten Jr, "A laboratory task for induction of mood states," *Behaviour research and therapy*, 1968.

[29] A. M. Isen and J. M. Gorgoglione, "Some specific effects of four affect-induction procedures," *Personality and Social Psychology Bulletin*, vol. 9, no. 1, pp. 136–143, 1983.

[30] D. Watson *et al.*, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, 1988.

[31] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, pp. 39–58, 1997.

[32] R. Westermann, K. Spies, G. Stahl, and F. W. Hesse, "Relative effectiveness and validity of mood induction procedures: A meta-analysis," *European Journal of social psychology*, vol. 26, no. 4, pp. 557–580, 1996.

[33] "Dragon Naturally Speaking, howpublished = https://www.nuance.com/dragon.html, note = Accessed: 2019-03-26."

[34] S. Baccianella *et al.*, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010.

[35] Vokaturi. Vokaturi. [Online]. Available: https://developers.vokaturi.com/getting-started/overview

[36] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, p. 40, 2012.

[37] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[38] E. Loper and S. Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[39] P. Salamé and A. Baddeley, "Effects of background music on phonological short-term memory," *The Quarterly Journal of Experimental Psychology Section A*.

A DESIGN GUIDE-LINE OF AUDITORY DISPLAY FOR ELECTRIC APPLIANCE

Takanori Komatsu

Meiji University,
4-21-1 Nakano,
Tokyo, 1648525, Japan
tkomat@meiji.ac.jp

Eiji Hayashi

Carnegie Mellon University,
5000 Forbes Ave,
Pittsburgh, PA 15213, United States
ehayashi@cs.cmu.edu

ABSTRACT

The auditory channel is important for communication between computers and users because of its properties, such eye-free communication and strong attention grabbing properties. However, interpreting the meanings of sounds is not a trivial task. Users have to learn and memorize the mapping between sounds and their meanings for each device. Therefore, as the number of devices increases, this becomes challenging for users. To mitigate the challenge, it is desirable to use sounds that users can understand intuitively. Thus, investigating the intuitiveness of sounds is of significant interest. In this work, we investigated 2,012 sounds consisting of 48 earcons, 80 auditory icons, and 1,884 beep sequences through a series of user studies using Amazon Mechanical Turk as well as a lab study that validated the results of the Mechanical Turk studies. The results provided a guideline for designing sounds that users might understand more intuitively.

1. INTRODUCTION

The auditory channel is important for computers to communicate information to users [20]. Although the development of graphical user interfaces allows computers to communicate more and more information to users through the visual channel with a higher bandwidth, the auditory channel still has its advantages over the visual channel, such as eye-free communication and its strong attention grabbing properties [6]. This is especially true for mobile phones. Because there are many cases where mobile phones initiate interaction with users, such as when push notifications are received from servers, if users are not looking at the displays, communicating information through the visual channel is infeasible. Therefore, devices have to rely on the auditory channel first to communicate. Simple devices, such as digital audio recorders or microwaves, offer other examples of devices that rely on the auditory channel in communicating with their users. These devices, in most cases, have very limited visual displays, such as single line displays or even just a few LEDs. Yet, these devices have multiple informational states that they have to communicate to users.

The environment is typically full of sounds played by multiple devices. However, interpreting the meaning of these sounds is not a trivial task. Most users would have experienced challenges like these: “I am sure I heard a short melody from the next room, but I could not figure out which appliance played that sound” or “My washing machine beeps during operation, but I do not understand what it means.” These problems are due to the fact that communication via sounds strongly relies on users’ knowledge [9]. Users have to learn the mapping between a sound and its meaning. However, as the number of devices around us increases, learning and memorizing the mapping for each device is becoming increasingly difficult for users. Therefore, it is of great importance to use intuitive sounds when communicating informational states to users so that users can interpret the meaning of these sounds without intense learning.

In this paper, we investigated the intuitiveness of sounds used by electric appliances. Specifically, we evaluated 2,012 auditory signals consisting of 48 earcons [1,3], 80 auditory icons [8,23], and 1,884 beep sequences in terms of their intuitiveness. We adopted crowd-sourcing to make evaluating the large number of sounds feasible, as opposed to prior pieces of work that investigated small sets of sounds [5,9,18]. We conducted a series of three user studies consisting of four tasks by using Amazon Mechanical Turk to evaluate the intuitiveness of the sounds as well as a lab study where we validated the results of the Mechanical Turk study to compensate for its low internal validity.

Through these three user studies, our paper makes four contributions. First, it provides a novel case study where we evaluated sounds through crowd-sourcing. Second, it provides empirical data about what sounds are used by electric appliances to communicate with users. Third, it also provides empirical results regarding the intuitiveness of sounds including beep sequences that have not been investigated intensively in existing works. Finally, we provide a guideline design on the basis of empirical data about what sounds should be used to communicate which informational states.

2. RELATED WORK

There have been several studies on the intuitiveness of simple auditory signals in conveying information to users. There are two types of sounds, earcons [1,3] and auditory icons [8,23], that have been investigated thoroughly in existing work.

Blattner et al. [1] defined earcons as “nonverbal audio messages used in the user-computer interface to provide information to the users about some computer object, operation or interaction,” and Brewster et al. [3] further stated that “earcons are abstract musical tones composed of short, rhythmic sequences of pitches with variable intensity, timbre and register.” Brewster et al. [4] also stated that, because of their flexibility, earcons could be easily designed to extend any object, operation, and interaction by means of their proposed guidelines. However, it could be difficult to

https://doi.org/10.21785/icad2019.002

design sounds appropriate for communicating certain informational states to users because of a lack of concrete guidelines that explain the relationships between informational states and earcons.

Gaver [8] introduced the concept of auditory icons. Gaver defined auditory icons as everyday sounds that conveyed information about computer events through analogy with everyday events. For example, the sound of shattering dishes could represent the drop of a virtual object into a virtual recycle bin. Gaver eventually argued that these auditory icons are an intuitively accessible way to use sounds to give information to users.

There have been many studies that have compared earcons with auditory icons in terms of their effectiveness [9], e.g., learnability or memorability [2,7]. While many studies have reported that auditory icons were generally perceived as easy to learn [2,7,18] and quicker in understanding [5], some have also reported that earcons were more pleasant and appropriate for actual applications than auditory icons [22]. Although the existing work has demonstrated that earcons and auditory icons effectively convey information to users, both have their limitations. With regard to earcons, one limitation is the arbitrary relationships between sounds and the information communicated by the sounds. Because of the arbitrary mappings, users in one study had to memorize the mappings to understand the meaning of the sounds correctly [23]. With regard to auditory icons, metaphoric mappings were not always easy to find [16]. Thus, it is difficult to design appropriate auditory icons for all informational states that computer systems have to communicate to users.

Currently, some electric appliances that can play rich sounds use earcons and/or auditory icons when interacting with users. However, most appliances still use rather simple auditory signals like beep sounds. One reason is that various international or domestic standards organizations have published standards for such simple auditory signals for the visually impaired or elderly. The American National Standard Institute (ANSI/INCITS 389-393), the International Organization for Standardization (ISO 11429), and the Japanese Industrial Standard Committee (JIS S0013) are representative standards organizations that deal with auditory signals. These organizations determine standards like “its pitch should be more than 250 Hz and less than 2,000 Hz” and one beep should indicate “start” and two beeps “finish.” However, the relationship between signals and events is unintuitive [16]. This suggests that the design guidelines for such auditory signals are not clear either. Moreover, up to now, little study has been done to compare the effectiveness or intuitiveness of beep sounds with those of earcons or auditory icons.

German architect Ludwig Mies van der Rohe adopted the motto “less is more” to describe his aesthetic approach to arranging the numerous necessary components of a building to create an impression of extreme simplicity by enlisting every element and detail to serve multiple visual and functional purposes. Recently, a similar design concept has become popular in HCI studies [19]. Recent electric appliances can present rich information to users through their high-resolution displays or stereo sound systems. However, providing too much information could overwhelm users’ cognitive resources [13,17]. Thus, more work has been started with a focus on simple ways of communicating information [10,11].

Harrison et al. [10] experimentally showed that the various blinking patterns of the small LEDs of mobile phone were interpreted differently by users, and these patterns succeeded in informing users of the informational states of mobile phones, such as low-battery and the presence of notifications. Similarly, Harrison et al. [11] also proposed Kinecticons, graphical icons with simple motions that can convey various informational states to users. In terms of simple auditory signals, Komatsu et al. [14] proposed artificial subtle expressions (ASEs) for intuitively notifying users of artifacts’ internal states (specifically, their confidence level).

Therefore, it is worthwhile to investigate whether various patterns of simple auditory signals (sounds like beep sounds) could inform users of the informational states of appliances as well as blinking LEDs or Kinecticons.

3. METHOD

In this paper, we conducted three user studies by using Amazon Mechanical Turk to investigate the intuitiveness of sounds. In the first user study, we asked participants to report the names of electric appliances or devices that use sounds to convey information. In the second study, we asked them to list the informational states that these devices expressed by using sounds. Finally, in the third study, we investigated the mapping between the informational states extracted in the second study and 2,012 different sounds by using Amazon Mechanical Turk, and we validated the results in a lab study.

In terms of the adequacy of crowdsourcing experiments, Komarov et al. [15] already reported that “there were no significant differences between the two settings (lab experiment and MTurk experiment) in the raw task completion times, error rates,” so we assumed that our experimental setting was a reasonable one.

4. USER STUDY #1: EXCRACTING DEVICES

In the first user study, we created a human intelligence task (HIT) that asked each Amazon Mechanical Turk worker to list 15 electric appliances that used sounds to communicate their states. We paid $0.05 for each completed HIT. Although most electric appliances use sounds to communicate some meaning, we intended to extract those sounds of which people recognized the usage. This allowed us to collect appliances that use sounds to express various states rather than limited states, such as power on and off.

Table 1. Seven representative appliances in rich and simple sound groups

Rich Sound Group	Simple Sound Group
Mobile phones	Microwaves
Laptops	Refrigerators
Desktop computers	Washing machines
Televisions	Cars
Alarm clocks	Ovens
DVD players	Coffee machines
Music players	Doors

We collected 690 electric appliances listed by 46 workers in 7 days. The workers listed 146 unique electric appliances in total. Then, the two authors individually categorized all of the appliances into two categories: appliances capable of playing complicated sounds, such as melodies (rich sound group) and those capable of playing only simple sounds, such as beep sounds (simple sound group). Other than one

disagreement, all of the categorizations by the two authors were agreed upon. We also solved the disagreement through discussion. We thought that there could be differences between appliances in these two categories. If devices in the former category used simple sounds to communicate states, designers intentionally chose the simple sounds rather than other possible rich sounds. In contrast, in the latter category, designers were forced to choose simple sounds because the devices in this category could not play complicated sounds. This difference potentially affected what sounds were used to convey states in these devices.

For each category, we extracted 7 devices that were listed by participants more than 10 times (Table 1). These 14 devices were used in user study #2.

5. USER STUDY #2: EXCRACTING INFORMATIONAL STATES

In this step, using Amazon Mechanical Turk, we extracted the informational states that the 14 devices chosen in the first user study communicate to users via the auditory channel. We asked workers to list up to 10 artificial sounds that a specified electric appliance played to express informational states. In the task, we explicitly defined artificial sounds to mean sounds or brief melodies played by electrical appliances as indicators. We also explained that the artificial sounds did not include mechanical noise, such as the seek noise from hard disk drives, nor recorded music for listening to, such as songs stored in music players. For each artificial sound, we asked the five questions shown in Table 2 to obtain the characteristics of the artificial sounds in detail. We paid $0.10 for each completed HIT.

We collected 700 responses in total (50 responses for each of the 14 devices chosen in the first study). In total, 384 unique Mechanical Turk workers completed this task in 14 days. Furthermore, 700 responses listed 1,785 descriptions of sounds. Of the 1,785 descriptions, 156 were descriptions for mobile phones, 97 for microwaves, 194 for laptops, 95 for refrigerators, 183 for desktop computers, 153 for washing machines, 122 for televisions, 139 for cars, 120 for alarm clocks, 116 for ovens, 118 for DVD players, 96 for coffee machines, 105 for MP3 players, and 91 for doors.

Table 2. We asked participants to list up to 10 artificial sounds played by a given device chosen in the first study and to answer these questions for each artificial sound.

#	Questions
1	What do you believe this sound is attempting to communicate?
2	Is the sound repeating?
3	If repeating, how long (in seconds) is one cycle?
4	If not repeating, how long (in seconds) is the sound?
5	Is the sound a sequence of beeps or a melody?

On the basis of the 1,758 answers for question 1, we consolidated the informational states that the workers thought the sounds were trying to convey. Consequently, we obtained eight informational states that the electric appliances listed in Table 1 communicate to their users via sounds. We did not find substantial differences between the appliances in the rich and simple sound groups in this process. The consolidated informational states were used in user study #3 in which the mapping between sounds and informational states was investigated (Table 3).

Table 3. Extracted 8 informational states by consolidating 700 responses from workers about the states that electric appliances are trying to communicate via sound.

#	Informational States
1	The device acknowledges your input.
2	The device is reporting that there is a message or a notification.
3	The device is reporting that there is a warning, an alert, or an error.
4	The device is turning on, booting up, or warming up.
5	The device is sleeping, suspended, or hibernating.
6	The device is thinking, computing, or processing.
7	The device is ready to execute a task, a process, or a command.
8	The device completed a task, a process, or a command.

We also investigated the characteristics of the artificial sounds on the basis of workers responses to other questions (Table 4). The results show that most sounds used to communicate informational states are beep sequences. For all the devices we tested, we found statistically significant differences between the ratios of responses that reported that the sounds were beep sequences and those that reported that the sounds were melodies ($p < 0.01$ in Fisher's exact test). In fact, workers reported that melodies are used only for specific cases, such as ringtones on mobile phones or computers booting up/shutting down. These results indicate that, although the number of devices capable of playing rich sounds has increased recently, many devices still use simple sounds to convey informational states to users.

Table 4. Workers' descriptions of artificial sounds used to convey informational states. They reported that 66.6% of the sounds were beep sequences.

Is the sound a sequence of beeps or a melody?		Is the sound repeating?	
Beeps	1,186	Yes	782 (43.8%)
Melodies	543 (30.4%)	No	983 (55.1%)
If not repeating, how long (in seconds) is the sound?		If repeating, how long (in second) is one cycle?	
Beeps	2.68 sec (SD=2.60)	Beeps	4.58 sec (SD=6.34)
Melodies	3.27 sec (SD=3.17)	Melodies	8.02 sec (SD=10.7)

The answers to the other questions also characterize what sounds are used to communicate informational states (Table 4). Roughly half the sounds repeat a sequence multiple times. The average lengths of the sequences are 4.58 seconds for beeps and 8.02 seconds for melodies. Similarly, the average lengths of non-repeating sounds are 2.68 seconds for beeps and 3.27 seconds for melodies. We used these data in designing the sounds used in user study #3.

6. USER STUDY #3: MAPPING BETWEEN STATES AND SOUNDS

Finally, we investigated the mapping between the states (Table 3) and sounds. We investigated 2,012 sounds consisting of 48 sounds composed as earcons, 80 sound effects as auditory icons, and 1,884 beep sequences. Essentially, for each sound, we asked multiple Amazon Mechanical Turk workers to choose one of the informational states that they thought a sound was trying to convey after listening to it. Then, we analyzed the distributions of workers' responses. If the distribution for a sound was skewed to one state, this indicated that workers were likely to interpret the sound as an intuitive indication of the state. In this regard, we compared composed sounds, sound effects, and beep sequences. Furthermore, we provide the results of a

qualitative analysis on the relationships between the compositions of sounds and the states chosen by workers.

We evaluated 2,012 sounds in 3 rounds. In the first round, for each beep sequence, we asked 10 Amazon Mechanical Turk workers to choose one of the informational states that a given sound was trying to communicate. Through this process, we excluded beep sequences that were not intuitive for workers to interpret. In the second round, we asked 50 workers to evaluate the 48 composed sounds, 80 sound effects, and 238 beep sequences that passed the first round. Finally, in the third round, we conducted a lab study to verify the results from the second round.

6.1. Sounds

We used 2,012 sounds consisting of 48 composed sounds, 80 sound effects, and 1,884 sequences of beeps.

For the composed sounds, we hired two music composers and asked them to design sounds that represented the eight informational states (Table 3) according to earcon design guidelines [4]. Both had at least 5 years of experience in composing music on computers, and each composed 24 earcons (3 composed sounds for each of the 8 states). We paid them $120 each. Hereinafter, we refer to the 48 sounds as composed sounds.

For the sound effects, we downloaded 200 sound effects from a web site where royalty-free sound effects are distributed. After downloading the effects, the 2 authors chose 10 sounds that were likely to be related to each of the states in Table 3. In total, they chose 80 sound effects. We added the sound effects to our investigation to mitigate one limitation regarding the composed sounds. Although we believed our composed sounds had reasonable quality, the quality relied on the music composers' skills. Thus, to compensate for this limitation, we added sound effects that were made by various people with different skill levels.

The third type of sounds was the beep sequences. In generating the sequences, we took an exhaustive approach. Essentially, we generated all possible beep sequences under four constraints: the number of beeps, length of a beep, pitch of a beep, and gaps between beeps. Under these constraints, we generated 1,884 beep sequences. We refer to this set of sounds as beep sequences. In the following, we further describe the constraints and how we generated the sequences.

6.1.1. Length of sequences

According to the results of user study #2, the average length of the indicators (non-repeated sounds) was about three seconds (Table 4). Although there were no explicit guidelines for the length of composed sounds and sound effects, the examples of such sounds were mostly within three seconds [12]. Thus, we decided to limit the length of sequences to three seconds to make the length comparable to other sounds.

6.1.2. Length of beeps

One set of guidelines for earcons [4] showed that the sound length should not be less than 0.0825 seconds. Furthermore, the length of a single beep is mostly shorter than one second. Thus, we decided to limit the length of a beep sound to either 0.1 or 0.5 seconds.

6.1.3. Number of beeps

Because we decided to limit the length of the sequences to 3 seconds and a beep sound could be 0.5 seconds, we limited the number of beep sounds in a sequence up to three (i.e., one, two, or thee) considering that there should be some pauses between the beep sounds in a sequence.

6.1.4. Pauses

Because we decided to use up to three beep sounds in a sequence, there could be two pauses between the beep sounds. Manipulating these gaps could have affected how people perceived the sequences. The guidelines for earcons [4] showed that a 0.1-second gap between sounds was recognized by users as where one sound finishes and another starts. Thus, we decided to use two different lengths for the pauses: 0.1 and 0.5 seconds.

6.1.5. Pitch

We decided to use three different pitches: high (1,200 Hz), medium (850 Hz), and low (500 Hz). Using the three different pitches allowed three beep sounds in a sequence to have different pitches. These frequencies were chosen on the basis of the results of Edworthy et al. [7] and Roy [21], who reported, "The warning indication shall be a steady alarm/horn with a frequency of 800 Hz," "the clear indication shall be a bell or simulated chime tone with a frequency of 1,200 Hz," and "the pitch should be no lower than 250 Hz."

6.1.6. Played once or repeated

Additionally, because the results of user study #2 showed that half the sounds were repeated sounds, we played the sequences either only once or three times.

In summary, there were six different beep sounds (three different pitches and two different lengths) and two different lengths of pauses between the beep sounds (short or long). The sequences consisted of one to three beep sounds and pauses between them. Additionally, the sequences were played either only once or three times. All of these combinations gave us 1,884 sequences of beeps. These 1,884 beep sequences consisted of 12 sequences with 1 beep: 6 possible beeps and 2 possible ways of playing the sequences (i.e., played once or repeated), 144 sequences with 2 beeps: 6 possible beeps, 2 possible pauses, 6 possible beeps, and 2 possible ways of playing the sequences, and 1,728 sequences of 3 beeps: 6 possible beeps, 2 possible pauses, 6 possible beeps, 2 possible pauses, 6 possible beeps, and 2 possible ways of playing the sequences.

6.2. Evaluation Method

As described in the previous section, we gathered 48 composed sounds, 80 sound effects, and 1,884 beep sequences. To investigate how people map these sounds to the eight states extracted in user study #2, we created a task using Amazon Mechanical Turk. In the task, workers could play a given sound by using a user interface (Figure 1).

In the first question, the HIT asked workers to transcribe a four-digit number read verbally in English. This question validated whether the workers could play sounds and paid

reasonable attention to the HIT. After that, there were questions about the mapping. In the questions, the workers were instructed to play a sound, and, then, to choose one of the eight informational states (Table 3) that they felt the sound tried to convey while imagining that their mobile phones played the sound. We chose mobile phones because the results of user study #1 indicated that most people said that mobile phones used auditory signals to communicate states. Alternatively, the workers could also choose "The device is reporting something not included in this list" if they felt the sounds represent a certain state that was not included in the list, or they could choose "The device made a random sound that does not have any meaning" if they felt the sound did not mean anything. The orders of the choices were randomized.

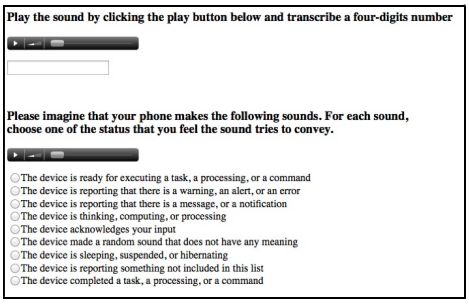


Figure 1. Screenshot of our investigation system

In one HIT, we asked the workers to evaluate five sounds one by one. We paid $0.10 for each completed HIT. The combinations of the five sounds were randomized. Because we needed a large amount of responses from the workers, we did not limit the workers to a specific region.

To evaluate the 2,012 sounds efficiently, we conducted two rounds of evaluation with Amazon Mechanical Turk and one validation in a lab study. In the first round, we asked the workers to answer questions regarding the 1,884 beep sequences. Because the beep sequences were generated by using an exhaustive approach, there were likely to be many sequences that were difficult to interpret. Therefore, we did the first round to eliminate such sequences. In the first round, we asked 10 workers to choose the state for each sound. Thus, each sound received 10 responses regarding the states that the workers thought the sounds tried to convey. Then, we eliminated sounds if the distributions of the responses were not skewed. For instance, if, for a beep sequence, 10 responses were evenly distributed among 10 choices, this indicated that the sound was not interpreted consistently. Thus, we eliminated such sequences in the first round to reduce the number of sequences. The sequences that passed the first round were further evaluated in the second round. All composed sounds and sound effects were also evaluated in the second round because we had a relatively small number of sounds for them. In the second round, we asked 40 workers to answer the same question as that in the first round; we asked them to choose a state that they thought a given sound tried to convey. Then, we analyzed the distribution of the workers' responses to investigate the intuitiveness of these sounds.

Finally, we validated the results obtained in the second round in a lab study where we asked 10 participants to evaluate the sounds that showed statistically significant results in the second round.

6.2.1. First round: eliminating beeps difficult to interpret

In the first round, we collected 18,840 responses (10 responses for each beep sequence). Seventy-four unique workers completed the task in 5 days. Out of 18,840 responses, we removed 475 responses because workers failed to transcribe the four-digit numbers correctly. Most sequences had 10 responses for each; however, because we removed 475 responses, some had 8 or 9 responses. Thus, we normalized the difference by dividing the number of responses in which a state was chosen by the number of total responses given to a sequence.

For each sequence of beeps, we focused on the states with the highest ratio of responses. Intuitively, if the ratio is high, it indicates that the workers agreed that a sequence meant a certain state and that it is easy to interpret, while, if the ratio is low, it indicates that the workers did not agree and that it is difficult to interpret.

There were three peaks with ratios around 0.2, 0.3, and 0.4 (Figure 2). Because we asked the workers to choose 1 of 10 choices, a few of the workers would have made the same choices by chance. This would have caused the peaks around 0.2 and 0.3. Thus, we decided to put a threshold at 0.4 and eliminated the sequences with a ratio smaller than 4.0. As a result, we extracted 238 beep sequences, which we further evaluated in the second round.

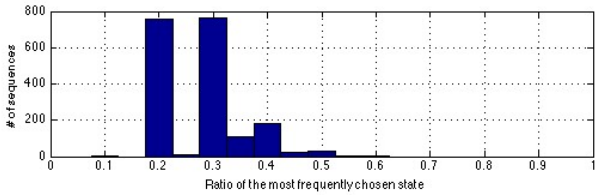


Figure 2. Graph shows distribution of mode responses divided by total number of responses for each sequence of beeps

6.2.2. Second round: comparison between sound types

In the second round, we evaluated 366 sounds that included 48 composed sounds, 80 sound effects, and 238 beep sequences. We collected 40 responses from workers for each sound in the same way as the first round. In total, we collected 14,640 responses. The task was completed by 114 unique workers in 3 days. Each worker evaluated 219 sounds on average. No worker evaluated the same sound more than once. The workers took 88 seconds to complete one HIT (i.e., transcribing a four-digit number and tagging five sounds) on average. We excluded 735 responses because workers did not transcribe the four-digit numbers correctly. We also removed 25% of the responses from the HIT that took less than 45 seconds to complete because this duration was too short to complete this HIT. In the following, we analyze the rest of the 10,406 responses.

Table 5 shows the distribution of the workers' responses. To calculate the distribution, first, we normalized the responses for each sound because each sound had a slightly different number of responses due to the removal of the low quality responses. More specifically, for each sound, we calculated the ratios by dividing the number of workers who chose a specific state by the total number of workers who evaluated the sounds. Then, we calculated the averages of the ratios to obtain the ratios shown in Table 5. We will use the distribution as a baseline for further analysis.

We then conducted 2 × 2 chi-square tests for each sound and each state to evaluate whether there was a statistically significant difference between the responses given to the sounds and the expected numbers of responses on the basis of a baseline. If at least one number in the four cells was smaller than five, we used a chi-square test with Yate's correction to evaluate the sound-state pair instead of the standard chi-square test. We regarded the differences as statistically significant when $p < 0.01$. We had to be careful when interpreting the results. Because we had 2,928 (366 sounds multiplied by 8 states) tests, we expected to have 29.3 combinations become statistically significant by chance. However, still, we were able to analyze overall trends because the statistical significances observed by chance were randomly distributed over all combinations.

Table 5. Distribution of workers' responses aggregated for all of 366 sounds

#	States	Rates
1	The device acknowledges your input.	0.098
2	The device is reporting that there is a message or a notification.	0.108
3	The device is reporting that there is a warning, an alert, or an error.	0.130
4	The device is turning on, booting up, or warming up.	0.091
5	The device is sleeping, suspended, or hibernating.	0.075
6	The device is thinking, computing, or processing.	0.127
7	The device is ready to execute a task, a process, or a command.	0.084
8	The device completed a task, a process, or a command.	0.099

Table 6. Numbers of sounds that had statistically significant ($p < 0.01$) differences between observed responses and baseline

State	Composed	Sound Effects	Beep Sequences
1	6 (12.5%)	12 (15.0%)	1 (0.4%)
2	1 (2.0%)	1 (1.3%)	3 (1.2%)
3	0 (0.0%)	0 (0.0%)	16 (6.7%)
4	8 (16.7%)	1 (1.3%)	2 (0.8%)
5	1 (2.0%)	0 (0.0%)	3 (1.2%)
6	0 (0.0%)	0 (0.0%)	15 (6.3%)
7	0 (0.0%)	0 (0.0%)	5 (2.1%)
8	0 (0.0%)	0 (0.0%)	6 (2.5%)
Total	16 (33.3%)	14 (17.5%)	51 (21.4%)

Table 6 shows the number of sounds that had statistically significant differences in the 2 × 2 chi-square tests for each state. The table clearly indicates that the workers interpreted the three sound types with different trends. The workers interpreted the composed sounds mostly as the acknowledgement of user inputs (state 1) or indications of system boot-up (state 4). Similarly, the workers mostly interpreted the sound effects as the acknowledgement of user inputs (state 1). In contrast, the workers interpreted the beep sequences as warnings (state 3) or indications of processing (state 6). Additionally, some beep sequences were interpreted as indications of executing a task (state 7) and of completing a task (state 8).

6.2.3. Intended states and interpretations

As mentioned, the composed sounds used in this study were composed by music composers to convey one of eight states (states 1 to 8). Similarly, the sound effects were selected by the researchers to represent one of the eight states. We investigated the relationships between the states that the sounds were supposed to convey and the states that the workers chose as interpretations of these sounds. The results indicate the ease or difficulty of composing/choosing sounds that convey intended states to users.

Table 7. Confusion matrix between informational states that sounds were composed/chosen to convey and informational states that workers interpreted

		Interpreted States							
		1	2	3	4	5	6	7	8
Intended States	1	**5**	0	0	0	0	0	0	0
	2	0	**1**	0	0	0	0	0	0
	3	0	0	**0**	1	0	0	0	0
	4	0	0	0	**3**	0	0	0	0
	5	0	0	0	2	**0**	0	0	0
	6	0	0	0	2	0	**0**	0	0
	7	0	0	0	0	1	0	**0**	0
	8	1	0	0	0	0	0	0	**0**

(a) Confusion Matrix of Composed Sound

		Interpreted States							
		1	2	3	4	5	6	7	8
Intended States	1	**5**	1	0	0	0	0	1	0
	2	0	**0**	0	0	0	0	0	0
	3	2	0	**0**	0	0	0	0	0
	4	1	0	0	**1**	0	0	0	0
	5	1	0	0	0	**0**	0	0	0
	6	0	0	0	0	0	**0**	0	0
	7	2	0	0	0	0	0	**0**	0
	8	1	1	0	0	0	0	0	**0**

(b) Confusion Matrix of Sound Effects

Table 7 shows the confusion matrixes for the sound-state pairs with statistically significant differences in the 2 × 2 chi-square tests. The rows and the columns denote the states that the sounds were intended to convey and the states that workers interpreted, respectively. The numbers in the cells denote the number of sounds. For instance, the bottom-left cell in Table 7 (a) shows that one composed sound that was intended to convey state 8 was interpreted to mean state 1. The numbers in the diagonal cells denote the sounds for which the workers' interpretations were the same as the intended informational states.

The composed sounds that were intended to convey state 1 (acknowledgement of user inputs) were mostly interpreted correctly. However, the other composed sounds were interpreted as state 4 (indications of turning on, booting up, or warming up) regardless of the intended states. Similarly, most sound effects were interpreted as state 1 (acknowledgement of user inputs) regardless of intended states.

These results gave us important design implications. Although rich sounds, such as composed sounds and sound effects, are expressive, they may not be as intuitive enough as we think for users to interpret their meanings. In contrast, users could more intuitively interpret simple beep sequences that conveyed some states. Therefore, when designing sounds, designers should choose appropriate sound types on the basis of the information that they intend to convey by using the sounds.

6.2.4. Third round: validation

As we already mentioned, we expected to have 29.3 combinations become statistically significant in the second round because we tested 2,930 combinations by using $p < 0.01$ as a threshold. We believe that the combinations that became statistically significant by chance were distributed

randomly across all combinations and that they would not have affected the analyses of general trends. However, to investigate the relationships between the sounds and users' interpretations of these sounds, we further validated the combinations (Table 6) that were statistically significant in the second round in a lab study, which would have higher internal validity than studies using Amazon Mechanical Turk.

We recruited 10 university students (8 males and 2 females). Their ages ranged from 21 to 25 with a mean age of 22.7. We paid $5 each for their participation. In the study, we asked participants to listen to the 81 sounds listed in Table 6 one by one and to rate the sounds. The participants were asked to rate a given sound for each state by using a 5-point Likert scale in terms of how strongly they agreed or disagreed that a sound conveyed a state (five denoted strongly agree and one denoted strongly disagree). Consequently, we obtained 648 ratings (8 ratings for each sound) from one participant. The orders of the sounds were randomized. The study took about one hour to complete.

We analyzed the data by using a one-way ANOVA to investigate whether the participants were likely to interpret the sounds as shown in the second round (within-participant design, treating eight informational states as independent variables and the ratings as dependent variables). Table 9 shows the sounds that had one specific informational state with a significantly higher average rating than all of the other seven states. These states were the same as those that were statistically significant in the second round. This ensured that these sounds were likely to be interpreted as indications of specific information with a high confidence.

7. DESIGN GUIDELINE FOR AUDITORY SIGNALS

Table 9 shows the relationships between sounds and their interpretations. On the basis of these results, we extracted a design guideline for auditory signals that communicate four informational states for which we found sounds with statistically significant differences.

- **State 1: Acknowledgement**: Two sound effects with quite short durations of less than 0.08 sec were interpreted as state 1. This indicates that users are likely to interpret sounds with very short durations as indications of the acknowledgement of user inputs. This also could explain why no beep sequences were interpreted as acknowledgement because the shortest beep sounds were 0.1 sec in our design.
- **State 3: Warning/Alert**: Eleven out of the 12 beep sequences extracted in the second round with ".H," "_H," ".h," or "(h" elements were interpreted as state 3. Thus, the beep sounds with the higher frequency in the middle of beep sequences were interpreted as indications of warning/alert.
- **State 4: Turing On/Booting Up**: All four composed sounds in melodies 5.0 sec long were interpreted as state 4. Operating systems, such as Microsoft Windows or macOS, or smartphones use rather long melodies to indicate turning on/booting up. Prior exposure to these devices would have led users to interpret the melodies as indications of state 4.
- **State 6: Processing**: All four beep sequence sounds that include at least two elements among "_," "M," or "L" were interpreted as state 6. Thus, utilizing beep sounds with medium or lower frequencies with a longer duration and longer interval in the beep sequences would be interpreted as state 6. It seems that the combination of longer sounds without high-pitched sounds and longer intervals are interpreted as relaxing situations (not like "warning/alert").

Table 8. Notations representing beep sequences

Notation	Meanings
H, M, L	Long beep sounds (0.5 sec) with high (1200 Hz), medium (850 Hz), and low (500 Hz) frequencies
h, m, l	Short beep sounds (0.1 sec) with high (1200 Hz), medium (850 Hz), and low (500 Hz) frequencies
.	Short pause (0.1 sec) between beep sounds
	Long pause (0.5 sec) between beep sounds
[ ]	Sequence played only once
( )	Sequence repeated three times

Table 9. Twenty-two sounds that succeeded in indicating specific informational states (with hyperlinks)

State	Sounds
1	2 sound effects: 0.06 sec, average F0: 880 Hz, 0.08 sec, 2,000 Hz, sounds like high-pitched ringing
2	None
3	12 beep sequences: [M_M.h], [H.H], [H.H.H], (l.H.h), (L.H_H), (L.H.m), (m.H.M), (M.m.M), (M.H.H), (h_H.l), (h.M), (H_H.m)
4	4 manually composed sounds: these sounds were melodies of about 5.0 sec
5	None
6	4 sequences: (L_M.l), (M_m_m), (h.L.L), (h.M_h)
7	None
8	None

Thus, to convey the above four informational states, we recommend utilizing the above guideline for preparing a specific melody or beeps for various electric appliances.

For the other four informational states, no sounds showed statistically significant differences in the average ratings compared with the other seven states in the validation round. However, many devices use auditory signals to convey these informational states to users. For instance, it is common to notify users that they have received e-mails by using sounds. Our results indicate that these auditory signals are less likely to be intuitive for users to interpret. Thus, devices have to communicate more information via other channels, such as text shown on a display, to compensate for the lack of intuitiveness in the auditory signals.

8. CONCLUSION

In this paper, we evaluated the intuitiveness of sounds through crowd-sourcing. By using crowd-sourcing, we explored a much larger design space, including beep sequences, than in the existing work. In the first user study, we extracted 14 devices that used the auditory channel to communicate informational states to users on the basis of 690 responses. In the second study, we collected 1,785 descriptions of sounds used to communicate informational states in the 14 devices. We then consolidated the descriptions into eight informational states that were frequently communicated via the sounds. Afterwards, in the third study, we investigated the intuitiveness of 2,012 sounds consisting of 48 composed sounds, 80 sound effects, and 1,884 beep sequences. More specifically, we asked Amazon

Mechanical Turk workers to listen to the sounds and choose one of the states that they felt the sounds represented. On the basis of an evaluation of 33,480 responses that we collected in a series of two Amazon Mechanical Turk studies, we found that the beep sequences were good at communicating notifications of warnings and status updates indicating that systems are processing commands, whereas the sound effects were mostly interpreted as indications of systems booting up, and very brief sounds were mostly interpreted as indications of acknowledgement. Finally, through a lab study, we validated the results from the user studies conducted with Amazon Mechanical Turk to provide a guideline for designing sounds used in electric appliances to communicate four informational states.

We designed our studies carefully; however, there are some limitations. In our studies, we asked workers and participants to imagine that a mobile phone made a sound. However, in practice, interpretations of sounds could depend on prior contexts. For instance, if a user started a task and heard a sound, s/he would interpret the sound as an indication of completion. We still believe that there would be many cases where users have to interpret sounds with little context, especially for mobile phones and computers because there are many background processes or push notifications on these devices. Nevertheless, the effects of context need to be further investigated. Finally, although we generated 1,884 beep sequences by using an exhaustive approach, the search space was still limited by the constraints that we set in generating beep sequences. There is a potential to improve the intuitiveness of beep sequences by modifying other properties.

Although this work still leaves unanswered questions, such as how we can design intuitive sounds for the other four states (states 2, 5, 7, and 8 in Table 6), this study presents an interesting methodology for evaluating sounds as well as a novel exploration in a large design space of beep sequences.

9. REFERENCES

[1] Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M. Earcon and Icons: Their Structure and Common Design Principles. In Proc. of SIGCHI Bull 21, 1, ACM Press (1989), 123-124.

[2] Bonebright, T. L. and Nees, M. A. Memory for Auditory Icons and Earcons with Localization Cues. In Proc. of ICAD 2007 (2007), 419-422.

[3] Brewster, S. A. Using Non-Speech Sounds to Provide Navigation Cues. ACM Transactions on Computer-Human Interaction 5, 2, ACM Press (1998), 224-259.

[4] Brewster, S. A., Wright, P. C., and Edwards, A. D. N. Experimentally Derived Guidelines for the Creation of Earcons. In Adjunct Proc of HCI 95, Huddersfield, UK (1995).

[5] Bussemakers, M. P. and De Hann, A. When It Sounds Like a Duck and It Looks Like a Dog: Auditory Icons vs. Earcons in Multimedia Environments. In Proc. of ICAD 2000 (2000), 184-189.

[6] Cohen, M. H., Giangola, J. P., and Balogh, J. Voice User Interface Design, Addison-Wesley, MA, USA (2004).

[7] Edworthy, J. and Hards, R. Learning Auditory Warnings: The Effects of Sound Type, Verbal Labeling and Imagery on the Identification of Alarm Sounds. International Journal of Industrial Ergonomics 24, 5 (1999), 603-618.

[8] Gaver, W. W. The SonicFinder: An Interface That Uses Auditory Icons. Human-Computer Interaction 4, 1 (1989), 67-94.

[9] Garzonis, S., Jones, S., Jay, T. and O'Neill, E. Auditory Icon and Earcon Service Notifications: Intuitiveness, Learnability, Memorability and Preference. In Proc. of SIGCHI 2009, ACM Press (2009), 1513-1522.

[10] Harrison, C., Hsieh, G., Willis, K. D. D., Forlizzi, J., and Hudson, S. E. Kinecticons: Using Iconicgraphic Motion in Graphical User Interface Design. In Proc. CHI'11, ACM Press (2011), 1999-2008.

[11] Harrison, C., Horstman, J., Hsieh, G., and Hudson, S. E. Unlocking the Expressivity of Point Lights. In Proc. of SIGCHI 2012, ACM Press (2012), 1683-1692.

[12] Hermann, T., Hunt, A., and Neuhoff, J. G (eds). The Sonification Handbook, Logos Publishing House, 2011.

[13] Keller, J. M. Development and Use of the ARCS Model of Instructional Design. Journal of Instructional Development 10, 3 (1987), 2-10.

[14] Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K. and Nakano, M. Artificial Subtle Expressions: Intuitive Notification Methodology for Artifacts. In Proc. of SIGCHI 2010, ACM Press (2010), 1941-1944.

[15] Komarov, S., Reinecke, K., and Gajos, K. Z. Crowdsourcing Performance Evaluations of User Interfaces. In Proc. CHI'13 (2013), 207-216.

[16] Kramer, G (eds). Auditory Display - Sonification, Audification, and Auditory Interfaces, Addison-Wesley (1994).

[17] Krug, S. Don't Make Me Think!, New Riders, CA, USA (2005).

[18] Leung, Y. L., Smith, S., Parker, S., and Martin, R. Learning and Retention of Auditory Warnings. In Proc. of ICAD'97 (1997), 288-299.

[19] Maeda, J. The Laws of Simplicity, The MIT Press, MA, USA (2006).

[20] Nass, C. and Brave, S. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship, The MIT Press, MA, USA (2005).

[21] Roy D. Patterson. Guidelines for the Design of Auditory Warning Sounds. Institute of Acoustics, 11(5):17-25

[22] Sikora, C. A., Roberts, L., and Murray, L. Musical vs. Real World Feedback Signals. In Proc. of SIGCHI 1995, ACM Press (1995), 220-221.

[23] Walker, B. N. and Kramer, G. Mappings and Metaphors in Auditory Displays: An Experimental Assessment. ACM Transaction on Applied Perception 2, 4, ACM Press (2005), 407-412.

DISCLOSING CYBER ATTACKS ON WATER DISTRIBUTION SYSTEMS. AN EXPERIMENTAL APPROACH TO THE SONIFICATION OF THREATS AND ANOMALOUS DATA

Sara Lenzi

Density Design Lab, Design Department,
Politecnico di Milano
Via G. Candiani, 72
2058 Milano, Italy
sara.lenzi@polimi.it

Ginevra Terenghi

Density Design Lab, Design Department,
Politecnico di Milano
Via G. Candiani, 72
2058 Milano, Italy
ginevra.terenghi@mail.polimi.it

Riccardo Taormina

iTrust
Singapore University of Technology and Design,
8, Somapah Raod
487372 Singapore
riccardo_taormina@sutd.edu.sg

Stefano Galelli

Pillar of Engineering Systems and Design
Singapore University of Technology and Design,
8, Somapah Raod
487372 Singapore
stefano_galelli@sutd.edu.sg

Paolo Ciuccarelli

Art + Design Department
Northeastern University,
360 Huntington Ave.,
Boston, MA, 02115
p.ciuccarelli@northeastern.edu

ABSTRACT

Water distribution systems are undergoing a process of intensive digitalization, adopting networked devices for monitoring and control. While this transition improves efficiency and reliability, these infrastructures are increasingly exposed to cyber-attacks. Cyber-attacks engender anomalous system behaviors which can be detected by data-driven algorithms monitoring sensors readings to disclose the presence of potential threats. At the same time, the use of sonification in real time process monitoring has grown in importance as a valid alternative to avoid information overload and allowing peripheral monitoring.

Our project aims to design a sonification system allowing human operators to take better decisions on anomalous behavior while occupied in other (mainly visual) tasks. Using a state-of-the-art detection algorithm and data sets from the Battle of the Attack Detection Algorithms, a series of sonification prototypes were designed and tested in the real world. This paper illustrates the design process and the experimental data collected, as well results and plans for future steps.

1. INTRODUCTION

Water supply systems are experiencing a transition from physical to cyber-physical systems: networked devices (smart sensors, industrial computers, telemetry units…) are used for monitoring and control purposes in order to increase the reliability and controllability. At the same time, though, these devices expose water plants (a key infrastructure in any country) to cyber threats. Research has recently focused on understanding the potential impacts of cyber-attacks and designing appropriate countermeasures. [1]

1.1. The Anomaly Detection Algorithm

The sonification project presented in this paper is based on an algorithm specifically designed for detecting and localizing cyber-attacks against water distribution systems. The algorithm builds on a Deep Learning model that is able to replicate the patterns of all hydraulic processes observed within a distribution system. In particular, the model is first trained using data pertaining to normal operating conditions. When fed new data, the model is likely to poorly reproduce data containing anomalous patterns, such as those induced by cyber-attacks, resulting in higher reconstruction errors. These errors are then analyzed in real-time to identify the anomalies related to cyber-attacks.[2]

The current approach for representing data produced by algorithm for consumption by a final user typically relies on visual, analytical diagrams. These traditional representations,

https://doi.org/10.21785/icad2019.044

displayed in real time on multiple screens and demanding continuous visual attention might be not suited to take an appropriate, informed response in a critical situation due to several reasons. Among these, the need for the operator to perform multiple visual tasks at a time and the visual channel overload caused by the introduction of a visual layer dedicated to cyber-threats as discussed in [3]. Thirdly, anomaly detection algorithms are still exposed to a certain degree of inaccuracy which results in the production of false alerts, with false positive alerts highly impacting on daily operations [2]. Furthermore, data compromised by an attack may differ very little from healthy data and therefore their visualization might be misleading.

1.2. Application of sonification to real-time process monitoring

Sonification [4] has been discussed in the past decades as a successful means of representing data in real-time process monitoring in the fields of medical applications [5]; financial [6]; security [3]; industrial production [7]. Specific characteristics of sound that make it a suitable candidate for the representation of real-time information have been extensively illustrated [8] and the specific area of process monitoring has also been extensively reviewed by recent literature. [9]

Cyber-security seems to be a promising field for the application of sonification to real-world situations. We already know that sound allows for peripheral monitoring [10] while leaving the center of our attention to visual tasks, also preventing information overload on the visual channel [11]. Additionally, human beings are very prompt at detecting changes in acoustic patterns [8], an added value of in anomaly detection tasks.

Despite promising results though, real-world applications are still far from being a recognized practice in academia, let alone in Industrial Control Systems. To the best of our knowledge, there is no offer of sonification tools available on the market to complement standard visualization dashboards. We believe that the main reason for this delay is the lack of prototype-based extensive experimental results involving real users in real settings. As repeatedly pointed out by the research community [12] [9] [13], the lack of applied and experimental results for sonification systems still represents one of the biggest weaknesses of a field that has shown and is showing a promising growth in public interest.

2. CYBER-ATTACKS TO WATER DISTRIBUTION SYSTEMS: DEFINING A USE-CASE FOR SONIFICATION

Artificial Intelligence is destined to gain predominance in the near future. In almost every aspect of our daily lives, we are immersed in an unprecedented mass of information we collect from the world around us: a continuous flow of data whose intricacies require an artificial intelligence able to work at a non-human scale to support humans in the task of collecting, organizing, and making-sense of it. On the other hand, hyper-reliance on automated systems and a *techno-chauvinistic* enthusiasm [14] tend to hide the fact that, in order to understand and make use of the information provided by AI systems, we still need to translate it into human – scaled knowledge.

Design is charged with the task of facilitating such translation. [15] Through design artefacts, final users are put in the condition to leverage their unique, sophisticated human experience in order to integrate machine systems and machine thinking into everyday life. We understand the representation of data as a design process aiming to transform data into knowledge [16]. Through preliminary research, the designer delineates a narration where the specific means for representing data are not pre-determined but are the result of specific design constraints. In this sense, we do not see sonification as an independent means of data representation able to reach a universal validity. The use of sound intervenes in the continuum from data to knowledge in all those cases where, based on preliminary analysis, it adds a value to the data representation process.

The following paragraph illustrates the process we conducted to delineate a specific use case for sonification for anomaly detection in water distribution systems.

2.1. Design constraints

During preliminary analysis, a series of constraints for the design of the sonification were identified. A non-exhaustive list includes:

- We consider sound as an integration, and not as a substitute, for the visual display of information in control rooms;
- As such, we do not intend for sound to represent and report all the information currently reported by the embedded visualization tools;
- The user experience takes into account the current state of a real context of usage. Questions related to the type of sound diffusion system we could use will have to be answered. For example: will users be willing to use headphones? The answer will constraint the design, for example excluding the usage of sound spatialization;
- Data are represented in real-time. As the current data resolution is capped at one hour, a sound will be designed to play every hour. This will exclude, for the time being, the option of a continuous sonification, a choice that would imply other design considerations on exposing users to a continuous background sound; [17] [18]
- In the absence of solid evidence on the advantages of the use of tuned sounds (or music) versus non tuned sounds, we will produce different versions of the prototype in order to gather first hand results.

2.2. Defining the use-case

2.2.1. The Context

The sonification represents data on cyber-attacks to the digital components (e.g., sensors, PLCs) of a water distribution network causing anomalous hydraulic processes—for example, low water pressure at the consumers' nodes caused by the intentional malfunctioning of a digitally-controlled pumping station. Currently, there are not engineering practices specifically dedicated to understanding in real time when a water network is under cyber-attack. An attack would mainly be identified as such during a subsequent forensic investigation. In case of an attack, control room operators can only see that the plant is presenting anomalous or faulty behavior in one or more of its components (tanks, water pipes, pumps, valves).

The anomaly detection algorithm object of the present project was developed to be integrated into the daily operations of the water plant. The algorithm would run in real-

time in a computer hosted in the control room, and it would trigger an alert in case of an anomaly.

2.2.2. The User

Our user is typically an engineer expert in managing water-related infrastructures. At present day, the intervention of a security operator in the event of an alert is manual: he/she acknowledges the alarm and manually proceeds to further checkups of the system, for instance, opening the current software visual interface to access a detailed report on the behavior of a specific component; controlling other parts of the system that might be connected to the alarm; run a in situ manual check, and so on. A protocol is in place for the management of emergencies, as well as a detailed procedure provided by the developers of the system to help differentiating among types of alerts. During a normal operation day, the user would distribute his/her time among different tasks, ranging from reading historical data, compiling reports, receiving phone calls, talking to colleagues, monitoring the real time status of the system both in terms of quality and quantity of water. In a medium-sized water plant, a system of about nine screens will display to two operators all the necessary information, making most of their current tasks visual, thus requiring the focus on their attention.

2.2.3. The Objectives

Information is conveyed to the operator to allow him/her to gain awareness on the status of the system at all times. In this specific case, the anomaly detection algorithm for cyber-threats makes it possible for the operator to discriminate between anomalies generated by faults of the systems and anomalies generated by external, malicious intrusion. This second type of anomalies are extremely hard to identify due to their specific nature, notably, that cyber-attacks are able to interfere with existing monitoring systems to deliberately introduce false information to fool the operators. [2] Once information on a possible cyber-attack reaches the operator, he/she will use it to take action, be it cross check it against data from other software; run a manual check of the plant; analyze historical data; launch a full-scale alarm; dismiss it as non-relevant.

2.3. Preliminary hypothesis

As mentioned, at this stage of development, machine learning detection mechanisms, such as the one adopted in this study, are still subject to a considerable rate of false (mainly, false positive) alerts. As a consequence, they can lack reliability in a real context of usage. In a private communication, an operator of a water plant control room reported that as the rate of false positive alert reaches up to one per day, the only solution for maintaining an efficient work flow is to keep the alarm system off. As a participant to our experiment, and former operator, puts it, "*In a water plant we do not trust machines*".

We hypothesize that this feeling of mistrust is essentially due to a design issue in the choice and implementation of the data-to-knowledge process. We consider our user a human operator with a thorough, sophisticated, long acquired knowledge of the system she/he manages. We therefore hypothesize that a well-designed relationship with the data should put the operator in the position to leverage his/her field experience to limit the impact of the algorithm's errors on the decision-making process. Such a well-designed relationship should grant a better understanding of causality between events (i.e. recognize and act in face of a real attack or dismiss alerts a false alarm), while taking advantage of an artificial intelligence-led system able to identify and make prediction on the specific nature of cyber-attacks.

2.4. The choice of sonification

We considered the following aspects as added values sonification could bring to the specific case:

- The sonification would keep at the periphery of attention while prompting the retrieval of analytical visual information when needed [10] [19]. Therefore, we won't need to introduce a new visualization system dedicated to cyber-attacks on top of those already in use for routine operations;
- Current alarm systems act on an on/off, 1/0 principle. Acoustic or visual alarms in a control room will either communicate a full alarm or no alarm, which is a radical simplification of continuous data coming from the algorithm. We believe that communicating intermediate status of the system, even if uncertain, via sound [11] might help the operator leverage his/her experience to take further decisions;
- Finally, we hypothesize that the user will, over time, develop a knowledge on "how the system sounds" enabling him/her to make predictions, thus anticipating problems instead of merely reacting to emergencies when these have already occurred. [20]

3. SONIFICATION DESIGN: PROTOTYPING

We designed and implemented two series of prototypes. Description of choices of sounds, mapping strategies and implementation follows.

3.1. Data sets

Given the relevance of water distribution networks for national security, detailed information on cyber-attacks against water utilities are generally not available. Therefore, we adopted simulated, yet realistic, data produced by the numerical simulation software epanetCPA for the case of the C-Town water distribution system, a medium-sized network made of 429 pipes, 388 junctions, 7 storage tanks, 11 pumps and 1 distribution valve, distributed over five demand districts. The employed datasets featured a total of 43 synthetic variables including tank water levels, inlet and outlet pressure for the valve and the pumping stations, as well as their flow and status (on/off). In particular, the datasets were part of the BATtle of the Attack Detection ALgorithms (BATADAL), an international competition on cyber security of water distribution systems. The BATADAL features two training datasets and a test dataset which included a total of 14 different attack scenarios. [2]

Two sonification prototypes (Prototype 1 and 2) were designed based on data from the test data set.

3.2. First Prototype

A first prototype (Prototype 1) was designed based on the concept of embodied metaphors and embodied sonification. [21]

3.2.1. Mapping strategy

An analysis of the datasets led us to consider information related to the specific component of the system as the most relevant. We hypothesized that an early identification of the specific component under attack (i.e., tank, pump, valve) and of the variable under attack (i.e., pressure, status, flow) would allow the operator to more efficiently identify the issue and run further checks. Furthermore, we decided to add information on the geographical location in the network of the component under attack. Consequently, the focus of the first data-to-sound mapping was:

- To represent each network's component with a different sound content. For example, all tanks would have been represented by the same type of sound;
- To link the behavior of each sound to the behavior of each component's variable over time. For example, a specific behavior of the "tank sound" (an increase in volume, or a distortion of the original sound) would represent an anomalous behavior in the tank pressure;
- To play each component in a sequence representing a virtual spatial movement from left to right through the geographical map of the network, to easily locate the component while listening (for example, the first tank sound heard would be the first tank at the extreme left of the network map).

We so obtained a sort of score for a sonification whose duration was determined by user experience criteria. In a first round of sketches, we tried to balance between the need to understand the information with the efficiency of an excessively long duration of the sonification. In its last iteration, the typical duration of sonification for Prototype 1 was of about 2 minutes played every hour.

3.2.2. Sound Design

We assumed with [10] that a successful sound content for peripheral monitoring would be one that the operator would easily relate to real-world experience. Therefore, we designed sounds having in mind how a real tank, pump, or valve would sound. The main choice to represent the anomalous behavior of each component was to apply a distortion parameter to the above-mentioned sounds following experimental results by [11]. The use of other processing such as changing in pitch and volume was also explored. A demo of prototype 1 can be heard following this link.[1]

The prototype has been evaluated through experts' sessions with participants from the field of water management, cyber security, communication design and sound design. Following these sessions, the prototype has been dismissed due to various issues negatively impacting the user experience. To name a few:

- The overall duration of each instance of the sonification (two minutes every one hour) was deemed way too long to be efficiently sustained;
- There was an evident overload in the amount of information conveyed by sound, which included: the type of component; the type of variable for each component; the amount of anomaly for each component and variable; the geographical location of the component;
- Such an amount of information was not only extremely hard to understand in the current situation but made the prototype virtually impossible to scale up to larger networks with more than 7 tanks and 11 pumps.

Nonetheless, feedback from critique sessions helped us radically change strategy for the second series of prototypes. In particular:

- Duration had to radically shrunk in order to limit the impact of sonification with the routine of the control room;
- Information to be conveyed by sound had to be drastically reduced, too. In particular, while geographical information on the specific district under attack seemed very relevant, no added value seemed to resort from information on components and variables;
- In Prototype 1, an anomaly index was artificially introduced for the representation of anomalous behavior in order to normalize data. This index scaled the anomaly level on a 5-steps, scale from no anomaly to extremely serious anomaly. The scale was pre-determined by us, but it is not introduced by the algorithm per se. As we found no clear added value in pre-determining the anomaly level, we decided to leave it to the operator to decide, based on his/her own experience, on the gravity of the anomaly.

3.3. Second Prototype

In a radical pivoting, Prototype 2 followed a strict data – driven approach focusing on the direct communication to the operator of anomalous behavior as it comes from the algorithm. The reason for this shift was mainly a need to go back to a clear formulation of the problem to be solved by sonification in our specific context of application i.e., to allow the operator to quickly identify anomalies due to cyber-attacks for action-taking. As mentioned in 3.1, the numerical value corresponding to the reconstruction error identified by the algorithm, previously hidden to the final user, was introduced into the data set, and it was around this parameter that the second sonification was built.

Following the main feedback emerged from Prototype 1 (see Par. 3.2.2) we decided that only the anomalous behavior of each of the five districts (without reference to components or variables) would be conveyed by sound and that the duration of each sound representing each district would be limited to 3 seconds/sound.

3.3.1. Mapping strategy

In order to subsequently process data in form of sound, we used a Python script to convert our csv database to MIDI. So obtained MIDI files were imported into the commercial software Ableton Live™ for further processing. In particular, we used the open source script miditime 1.1.3[2] to convert data to MIDI format.

Four mapping strategies, later called *Scenarios*, were identified:

1. Scenario 1-Delay: every District is represented by a different sound. The duration of each sound is 3". All sounds (five sounds for five districts) start at

[1] https://sonifying.github.io/UNDERSTANDING-CYBER-ATTACKS-ON-WATER-SUPPLY-SYSTEMS/p1.html

[2] MIDItime: https://pypi.org/project/miditime/

Time 0 and in case of no anomaly, they stop playing after 3". In case of anomaly, the anomalous sound will start delayed by an amount of time directly proportional to the amount of the anomaly as taken from the reconstruction error's data.

2. Scenario 2-Length: starting from the same characteristics as Scenario 1, the duration of each sound increases proportionally to the increase in value of the anomaly level in data.
3. Scenario 3-Repetition: the 3" sounds for each district will cycle over a 10" time span if anomalous. The frequency of cycling is determined by the level of the anomaly coming from data.
4. Scenario 34-Pitch: the 3" sounds representing each district will increase their pitch proportionally to the level of anomaly.

After running a further critique session with experts, we narrowed down the options to the two prototypes which seemed to be more promising. For a series of reasons whose full account exceeds the scope of this paper, we focused our attention on Scenario 2-Length and Scenario 3-Repetition. A demo of all four prototypes as well as a map of the five districts can be found at this link[3].

4. EXPERIMENTAL DESIGN: TESTING WITH REAL USERS

Prototype 2 and 3 (from now on, Prototype 1 and 2) were produced in two versions (A and B, one with tuned one with non-tuned sounds) and were the object of a first experimental phase which included the collection of both quantitative and qualitative information. The experiment involved six expert users in cybersecurity and water management from five different countries (New Zealand, Italy, Singapore, Vietnam, Turkey). Over a two-weeks period, testers were asked to use the prototypes in a real context, during their daily work routine, for eight consecutive hours. One day was dedicated to each scenario/version, for a totality of four full days of testing. To limit biases and expectations we asked testers to keep a few days' break between one prototype and the following, thus obtaining a testing period of two weeks with only four full days dedicated to testing each prototype.

4.1. Experimental Protocol

The experiment included three phases: a preliminary questionnaire, a quantitative test to be completed during the four days of prototype testing, and a final, one hour long semi-structured interview for collection of qualitative feedback. The setup of the various phases was inspired by Research through Design practices [22] and in particular by experimental practices such as Annotated Portfolios [23], Technology Probes [24] and the more recent Design Probes [25].

The preliminary questionnaire gathered self-assessed information on the specific expertise and on the level of music/sound competence of the testers. The quantitative test had to be fulfilled after each sonification (every hour). Despite the very limited number of users not granting statistical relevance, we intend the test as a validation of the efficacy and effectiveness of each prototype in terms of user performance. Specifically, the test has been designed to answer the following questions:

1) can users recognize anomalous behaviors in the system through the sonification?

2) can users attribute a scale of severity to anomalies, identifying corresponding differences in the sound behavior?

3) can users identify in which district of the city is the anomaly occurring, through corresponding differences in sound content?

Interviews, on the other hand, had the goal of gathering a more nuanced series of insights from domain experts with the overall goal of identifying guidelines for the next prototype iteration. Specifically, the interview has been designed to gather feedback on the following:

1) the usage of sonification in a real-world context;

2) the different strategies used for the two scenarios;

3) the different sound content used in the different versions.

5. ANALYSIS OF RESULTS

Quantitative analysis, though not statistically relevant, helped us evaluate the potential of sonification in relationship with the tasks the experts have to carry out in our specific context of usage, i.e., firstly correctly identifying the information coming from the algorithm; secondly attributing a level of gravity to the anomaly and situating the anomaly in the correct district where it is occurring.

We remind the reader that we intend this sonification as an addition and preliminary step to the possibility of finding analytical, more granular information in the visualization system of the control room. We also remind that a pre-determined level of anomaly was not introduced "by design". As such, the interpretation of the level of anomaly in the quantitative testing presents a high degree of subjectivity, as highlighted by some testers. A second series of quantitative testing is planned in order to further investigate the relevance of attributing a level of gravity to the anomaly and of identifying the district, in order to determine their influence on the overall performance of the operator in taking further action. If a strong relevance should emerge, an objective level of anomaly gravity could be embedded in the sonification by design as well as a refined mapping strategy to scale district identification in case of larger networks.

Qualitative analysis helped us identify emerging patterns in the users' relationship with the sonification, in their strategy to learn from it and applying it to their real work context, as well as clustering reactions to the different versions of the prototype and suggestions for further developments.

5.1. Quantitative Analysis

All users showed a high level of performance in the identification of an anomalous status of the system. As Fig.1

[3] https://sonifying.github.io/UNDERSTANDING-CYBER-ATTACKS-ON-WATER-SUPPLY-SYSTEMS/p2.html

illustrates, testers seemed to easily identify anomalous behavior through sound, while attributing an anomaly level and identifying the specific district seemed to be more challenging. In particular, and worth of further investigation, users tended to over-estimate the gravity of the anomaly while under-estimating the number of districts involved. Over-estimation of anomaly level impacts more the sonification under Scenario 2 (and in particular, 2B) which might indicate a certain "anxiety effect" driven by the specific strategy and the specific sound contents used for Scenario 2B. Further investigation is needed to validate this hypothesis. On the other hand, the error in the quantification of the number of district involved in the anomaly is higher for Scenario 1, both A and B (based on length), which could be due to a sonification strategy that tends to superimpose, to an untrained ear, the sound of each district, when more than one district presents anomaly. Both of these interpretations of the result will need further investigation. The majority of users reported improvements over time, from the starting of testing (at 10am) to its conclusion after 8 hours. We take it as a sign that the continuous usage of sonification in daily work operations can lead to a sophisticated capacity of understanding information through sound down to detailed nuances, as it does indeed happen in real life (for example in our relationship with natural soundscapes). Despite some initial mistakes in the understanding of instructions for the quantitative testing by some of the testers, all have been able to develop an individual, self-taught strategy to learn from their mistakes while using the prototypes.

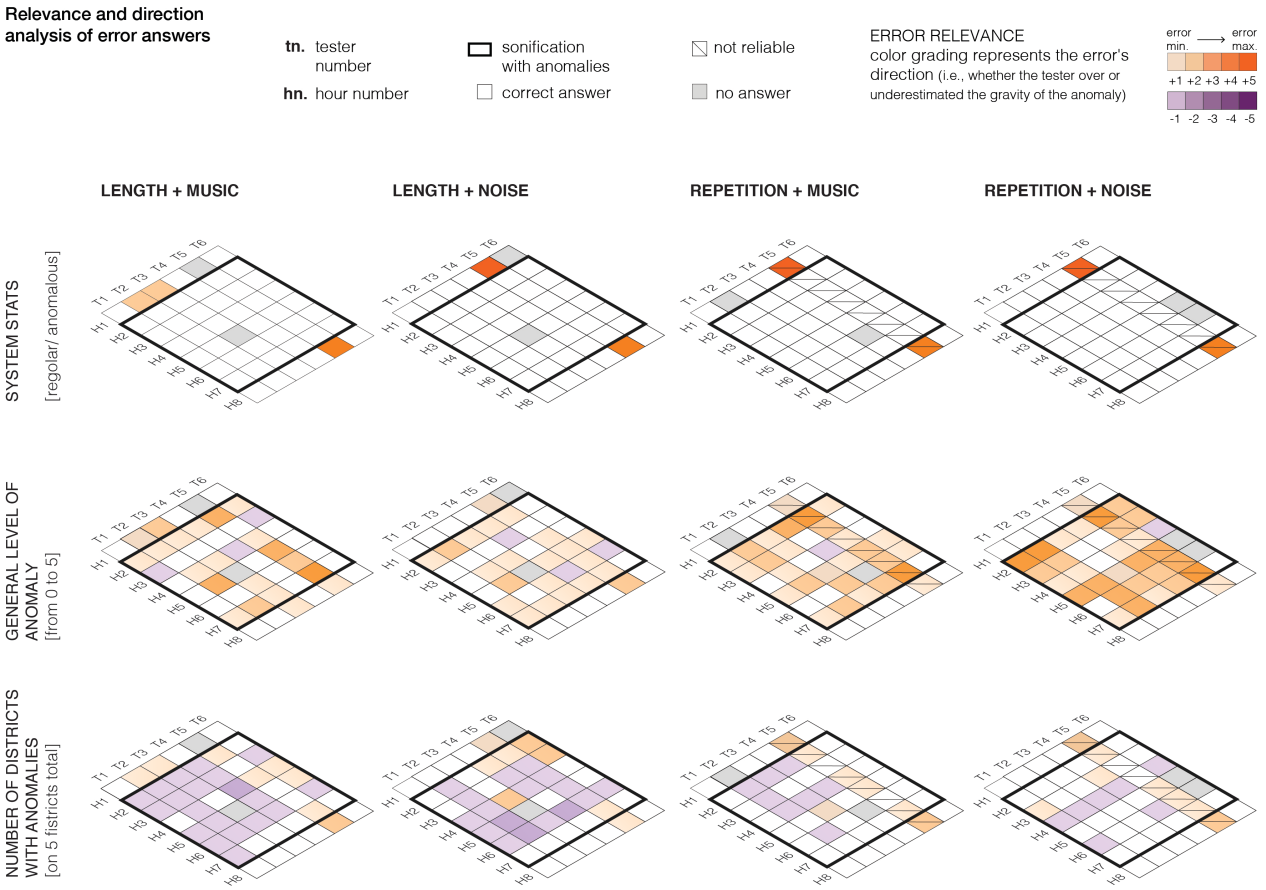


Figure 1. Excerpt from Quantitative Analysis Results.

5.2. Qualitative Analysis

The analysis of interviews highlighted a very positive attitude towards the integration of sonification in control rooms of water plants. Positive comments included: the possibility to hear sound at the periphery of attention while focusing on other (visual) tasks; the smoothness of integrating sonification in the daily routine of the work place; the low cognitive load required by the sonification to gather basic information that might be investigated further on visual tools. Some users highlighted the positive effect on performance of being able to discard an alert as non- dangerous dedicating only 3" of peripheral attention.

All testers strongly highlighted the key role of sound design in their attitude towards the implementation of sonification in real life. One scenario (1B-Length) was clearly rejected as unpleasant, annoying and even scary while the others were all judged as pleasant to various degrees with a slight tendency towards tuned sounds. Some users tended to attribute significance to non -tuned sounds in Scenario 2B (Repetition) using metaphors coming from cartoons, video games or personal experience, opening up room for further investigation on the role of embodied metaphors. An interesting line of discussion emerged on the possibility of using unpleasant sounds for higher anomalies and pleasant sounds for non or little anomalous states, again re-introducing the possibility of embodied metaphors for further design of sounds.

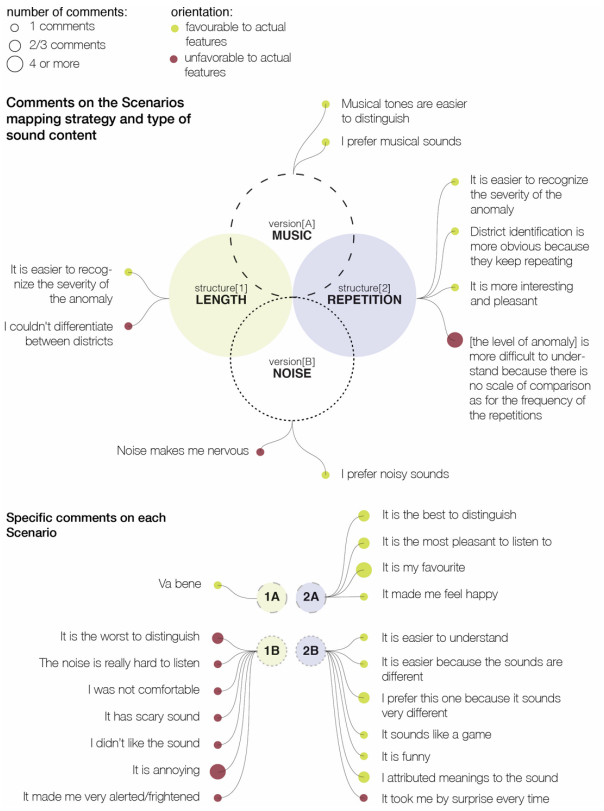


Figure 2. Excerpt from Qualitative Analysis Results.

As Figure 2 illustrates, no definitive judgement can be done on a preferred or "better" version of the sonification: there is no "winning" Scenario. If aesthetic and emotional inclinations seem, for the users, to be an added value for a better comprehension of the sonification itself, they do not seem to entail an improved performance and further testing with a statistically relevant numbers of users will have to be conducted.

6. CONCLUSIONS AND FUTURE DEVELOPMENTS

In general, a very positive, self-reflecting and self-learning attitude was shown by all testers in relation to the introduction of sonification in the daily routine of monitoring cyber-attacks in water plants. None of the testers had reported a particular training or passion for sound/music, leading us to think that, despite the general lack of preparation in reading information through sound, real-world applications can be effectively designed and introduced. All users were vocal in highlighting the need for preliminary training and real-time feedback to self- assess their performance while learning to use the sonification. The setup of the experiment purposely did not provide any such training or feedback. Only a website containing an introduction to sonification and all the instructions was provided and can be found at this link[4].

Based on these encouraging results, a second phase of the project is envisaged. This phase will include testing the sonification along with the corresponding visualization, in a real setting; a new iteration of a single prototype ideally taking into account all the aspects emerged in this first phase, including the possible re-introduction of embodied metaphors; an extended group of testers for a longer period of time in order to evaluate emerging concepts, such as subjectivity/objectivity of the anomaly level and the role of experience in making sense of additional information such as district identification. The final goal is the release of a real-world application, fully integrated with the anomaly detection algorithm which formed the basis of this project.

[4] https://ginevraterenghi.github.io/presentazione-prog/project.html

7. ACKNOWLEDGMENTS

First and foremost, we express our gratitude to the testers who took time to participate to the present research during their working time. We also wish to thank our colleagues at Density Design Lab for the precious advice and help with visualizing the results of the experimental phase. Dr. Taormina and Dr. Galelli were supported in part by the National Research Foundation (NRF), Prime Minister's Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-40) and administered by the National Cybersecurity R&D Directorate.

8. REFERENCES

[1] Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., & Ostfeld, A., "Characterizing cyber-physical attacks on water distribution systems in *Journal of Water Resources Planning and Management*, *143*(5), 04017009, 2017.

[2] Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., Ostfeld, A., Eliades, D. G., ... Ohar, Z. "Battle of the Attack Detection Algorithms: Disclosing cyber-attacks on water distribution networks." in *Journal of Water Resources Planning and Management*, *144*(8), 2018.

[3] Axon L, Creese S, Goldsmith M, Nurse JRC. "Reflecting on the Use of Sonification for Network Monitoring", In: *SECURWARE 2016: The Tenth International Conference on Emerging Security Information, Systems and Technologies*. ; 2016:254-261.

[4] Hermann T, Hunt A, Neuhoff JG, Dombois F, Eckel G. "The Sonification Handbook". In: *The Sonification Handbook*, 2011:301-324.

[5] Ballora M, Pennycook B, Ivanov PC, Glass L, Goldberger AL. "Heart Rate Sonification: A New Approach to Medical Diagnosis", *Leonardo*. 2004; 37(1):41-46

[6] Nesbitt, Keith & Barrass, Stephen. (2004). "Finding Trading Patterns in Stock Market Data", IEEE computer graphics and applications. 24. 45-55. 10.1109/MCG.2004.28.

[7] Hermann T, Hildebrandt T, Langeslag P, Rinderle-ma S, "Optimizing Aesthetics and PRecision in Sonification for Peripheral Process-Monitoring" in *of the 21st International Conference on Auditory Display (ICAD 2015)* 2015:317-318

[8] Vickers, Paul *Sonification for Process Monitoring.* In: The Sonification Handbook. Logos Verlag: Berlin, 2011 pp. 455-492.

[9] Rinderle-Ma S, Hildebrandt T. "Server sounds and network noises", in: *6th IEEE Conference on Cognitive*

Infocommunications, CogInfoCom 2015 - Proceedings. ; 2016:45-50.

[10] Bakker S, van den Hoven E, Eggen B. "Knowing by ear: Leveraging human attention abilities in interaction design" *J Multimodal User Interfaces*. 2011, January 20.

[11] Ballatore A, Gordon D, Boone AP. "Sonifying data uncertainty with sound dimensions". In: *Cartography and Geographic Information Science*, 2018.

[12] Axon L, Nurse JRC, Goldsmith M, Creese S. A Formalised Approach to Designing Sonification Systems for Network-Security Monitoring. In: *International Journal on Advances in Security*. Vol 10. ; 2017:26-47.

[13] Hermann T, Hunt A, Neuhoff JG. (edts.), *Theory of Sonification*, in "The Sonification Handbook", Logos Verlag: Berlin, 2011.

[14] M. Broussard, *Artificial Unintelligence: how computers misunderstand the world*, Cambridge, MA: MIT Press, 2019.

[15] D. A. Norman, *Things that Make Us Smart: Defending Human Attributes in the Age of the Machine*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1993.

[16] Masud L, Valsecchi F, Ciuccarelli P, Ricci D, Caviglia G., "From data to knowledge: Visualizations as transformation processes within the data-information-knowledge continuum." in *Proc Int Conf Inf Vis*. 2010:445-449.

[17] Hildebrandt T, Hermann T, Rinderle-Ma S. "Continuous sonification enhances adequacy of interactions in peripheral process monitoring.", in: *International Journal of Human Computer Studies*. Vol 95. Elsevier; 2016:54-65.

[18] Deb S., Claudio D., "Alarm fatigue and its influence on staff performance", in *IIE Transactions on Healthcare Systems Engineering*, 5:3, pp. 183-196, 2015.

[19] Rönnberg, Niklas & Lundberg, Jonas & Löwgren, Jonas, "Sonifying the periphery: Supporting the formation of Gestalt in air traffic control." in *ISon 2016, 5th Interactive Sonification Workshop,* CITEC, Bielefeld University, Germany, December 2016.

[20] Hildebrandt, T., Hermann, T., Rinderle-Ma, S., "A sonification system for process monitoring as secondary task.", in: 2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom), pp. 191–196Nov. 2014.

[21] Roddy S., Furlong D., "Sonification Listening: an empirical embodied approach", in *Proceedgins of the 21st International Conference on Auditory Display (ICAD 2015).*

[22] Stappers, P. & Giaccardi, E. "Research through Design.", in Soegaard, M. & Friis-Dam, R. (eds.), *The Encyclopedia of Human-Computer Interaction, 2nd edition*, 2017.

[23] Jonas Löwgren. 2013. Annotated porfolios and other forms of intermediate-level knowledge. Interactions 20, 1 (January 2013), 30-34.

[24] Hutchinson H, Mackay W, Westerlund B, et al. Technology probes: Inspiring design for and with families. In: *CHI03 Poceedings of the Conference on Computer-Human Interaction*, 2003:17-24.

[25] Hogan T, Hornecker E. Feel it ! See it ! Hear it ! Probing Tangible Interaction and Data Representational Modality. *Des Res Soc*. 2016:1-13

MIXED SPEECH AND NON-SPEECH AUDITORY DISPLAYS: IMPACTS OF DESIGN, LEARNING, AND INDIVIDUAL DIFFERENCES IN MUSICAL ENGAGEMENT

Grace Li

Georgia Institute of Technology,
648 Cherry St NW
Atlanta, GA 30313, USA
tli.grace@gatech.edu

Bruce N. Walker

Georgia Institute of Technology,
648 Cherry St NW
Atlanta, GA 30313
bruce.walker@psych.gatech.edu

ABSTRACT

Information presented in auditory displays is often spread across multiple streams to make it easier for listeners to distinguish between different sounds and changes in multiple cues. Due to the limited resources of the auditory sense and the fact that they are often untrained compared to the visual senses, studies have tried to determine the limit to which listeners are able to monitor different auditory streams while not compromising performance in using the displays. This study investigates the difference between non-speech auditory displays, speech auditory displays, and mixed displays; and the effects of the different display designs and individual differences on performance and learnability. Results showed that practice with feedback significantly improves performance regardless of the display design and that individual differences such as active engagement in music and motivation can predict how well a listener is able to learn to use these displays. Findings of this study contribute to understanding how musical experience can be linked to usability of auditory displays, as well as the capability of humans to learn to use their auditory senses to overcome visual workload and receive important information.

1. INTRODUCTION

People regularly use visual displays to aid in monitoring data and increasing their situation awareness. With more advanced technology and research, these displays have been beneficial in helping people gather information. However, the amount of information users are able to attend to visually remains limited, even as the demand for more information increases. Researchers have turned to auditory displays as an additional channel, and have studied the benefits of audio versus visual information on peoples' ability to comprehend and retain information presented. This study focuses directly on auditory display design and the impact of listener differences such as musical experience and motivation on usability of auditory displays.

In the example of an anesthesiologist who needs to monitor a patient's vitals during surgery, visual displays can be overwhelming. These displays may prevent anesthesiologists from visually attending to other areas of their workspace. To ease the workload in the visual field, a mix of visual and auditory displays may be used together, where the auditory display would cue the anesthesiologist to look at the information on the visual display. However, in circumstances when the anesthesiologist cannot visually attend to the visual display, auditory displays may prove to be beneficial in informing the anesthesiologist on the status of their patient. Auditory displays prevent overload of information in other daily activities most people encounter, such as listening to the news or weather report in the morning while stuck in bed, or changing music playlists while driving.

The term *sonification* describes a subtype of auditory display that typically uses non-speech audio to present information by translating relationships in data into sounds that human listeners are able to comprehend [1]. Information in auditory displays is mapped to certain sounds that help listeners understand and interpret the information. Bregman and Campbell define auditory streams as a "sequence of auditory events" that are blended together to convey a message or an idea into one single "stream" [2]. These auditory sequences can be different, yet related, in order for them to fit together and present information that makes sense to the listener. These streams of sounds can include manipulations of vario*us* acoustic properties, such as pitch and tempo. Many studies have looked at the use of multi-stream auditory displays in an anesthesiologist's workstation, specifically looking at the effects of mapping multiple pieces of information to fewer auditory streams. Fitch and Kramer mapped eight different health-related variables to two different streams and found that participants improved with practice and were able to manage all the variables [3]. They concluded that auditory systems that simultaneously convey a number of variables can be more effective than visual displays, separating variables into individual pieces of information to perceive one at a time [3]. In a similar study, Loeb and Fitch used actual anesthesiologists to see if they were able to monitor six different variables at once, and found that with little practice, the clinicians were able to identify all the variables in two different streams and simultaneously decipher and respond to critical events [4]. These multi-stream auditory display studies suggest that listeners can accurately monitor up to eight different variables combined into three separate auditory streams within a complex auditory display [3], [5]. Additionally, Schuett found that participants were able to follow about five auditory variables at a time which were blended together to form three more dominant comprehensive streams [6] [7]. Applying the information to a more practical setting, Schuett created auditory displays with three auditory streams using five acoustic parameters, each representing five different variables related to weather or health, and observed participants' ability to interpret information from the auditory display [8]. For instance, one of the health-related variables was Heart Rate, which was mapped to the tempo of one of the streams, while Respiratory Rate was mapped to the frequency of that same stream. Findings suggested that participants were able to learn to comprehend the auditory display and were able to perform better with practice.

https://doi.org/10.21785/icad2019.019

Auditory display research has also looked into the effects of individual differences in listening abilities, familiarity, and practice on the usability of these displays. Watson and Kidd suggest that listeners' perceptual and cognitive abilities play an important role in the systems' usability, while comprehension of these displays may be a result of the listeners' musical ability [9]. They propose that there must be a subjective perceptual difference among the participants when using auditory displays. This points towards musical training and experience as differences that may impact listeners' ability to understand auditory displays. Brochard, Drake, Botte, and McAdams had participants listen to three auditory streams and signal when they found small temporal irregularities within the auditory streams [10]. They found that participants who were musically trained performed better at detecting irregularities than those who were not musically trained. However, there were no significant interactions between musical training and the other variables such as frequency grouping and target location. Lacherez, Seah, and Sanderson concluded that failure in stream segregation was the limiting factor for listeners' perception, even for those who were musically trained [11]. Schuett suggests that although the link between musical experience and stream segregation is unclear, it seems to be that some form of familiarity with the acoustic properties of the auditory display may be helpful in stream segregation [8]. Walker and Nees looked into the role of training and found that practice with feedback led to significantly lower errors in point estimation tasks using sonification more so than no practice, practice only, practice with visual prompts, and conceptual training [12]. Therefore, having knowledge of the results during training can have a positive impact on listeners' improvement and performance.

Because much of the population is not familiar with sonification, there are challenges in incorporating sonification into daily activities. People are becoming more familiar with speech-based auditory displays such as the speech commands in GPSs, Siri and Alexa, which may make speech seem to be a viable alternative to sonification. In some cases, that may be the case; in other cases, not. Nevertheless, when multiple channels of data need to be conveyed, there may well be challenges that would arise with multiple streams of speech. Ericson, Brungart, and Simpson list factors that influence comprehension of speech displays in the context of air force pilots [13]. They determined that the addition of simultaneous voices would decrease the performance of the listener, so keeping the number of speech sounds to a minimum in a display would be best. Differing *characteristics* of the voices can help segregate speech sounds in a display, where pitch/frequency, speaking rate, accents, and intonation might help listeners comprehend the different streams. Finally, *spatially* separating speech sounds in a display can also help listeners comprehend each speech sound, more so than the other techniques, which would only increase intelligibility of one or two speech sounds, at the expense of losing information from the other speech sounds [13]. In a more applied situation, Simpson, Brungart, Dallman, Joffrion, Presnar, and Gilkey tested spatial audio displays in general aviation environments with trained pilots [14]. They found that spatial audio displays effectively improve pilots' situation awareness and safety in general aviation environments when used for both navigation and altitude monitoring. Similarly, Simpson, Brungart, Gilkey, and McKinley found that pilots were very accepting of the spatial audio display and showed low annoyance levels, which suggests that spatial audio displays are important to comprehend spoken information and to prevent overload or annoyance [15].

Ericson et al. have shown that speech displays can make it difficult to monitor different speech streams because speech streams tend to mask each other [13]. Multiple speech streams can be overwhelming and prevent listeners from obtaining adequate information. Methods to keep speech streams separate and intelligible are effective, but there are still limits to how many streams can be followed. Li, Tang, Hickling, Yau, Brecknell, and Sanderson found that speech cues can lead to more accurate responses in identifying information than earcons, however that may be due to the fact that people are more familiar with speech cues than earcons [16].

Even with all the focus on the use of sonification and speech displays to convey data, there has been a gap in knowledge about the effects of speech streams interactions with the sonification on listeners' ability to perceive information. Walker and Nees mention the wealth of knowledge in sonification during concurrent visual and auditory tasks, but a lack in the degree to which non-speech audio interacts with concurrent processing of other sounds such as speech [2]. The purpose of the present study is to combine the benefits of both sonification and speech auditory displays into a mixed auditory display in order to see the effects of the interaction on listeners' comprehension and performance. The mixed displays should minimize the unfamiliarity of sonification, and introduce speech, while also ensuring that there are not too many speech streams to distract or mask the other streams.

2. STUDY OVERVIEW

The study is a continuation and adaptation of Schuett's dissertation [8]. Participants assumed the role of an anesthesiologist and detected trends in body vitals of a virtual patient using an auditory display with five variables combined into three streams. There are a total of four displays variants: one that uses Schuett's [8] "Health" non-speech display; one with all speech; and two with a mixture of speech and non-speech. The speech sounds were added into the display using techniques highlighted by Ericson et al. [13] by separating them spatially, and by frequency. Participants were randomly assigned to one of the displays and took a pretest to see their initial comprehension of the display. Then they completed a practice phase with feedback, and finally, completed a final test to see if they were able to improve their comprehension of the display. The scores were compared across all the three display conditions to see which display had the highest learnability and performance. Additionally, subjective measurements through motivation surveys and musical experiences were used to assess the impact that individual differences may also have on using the auditory displays before and after practice.

3. METHODS

3.1. Participants

Participants in this study were 97 students at a U.S. university between the ages of 17 and 29 (M = 20.0, SD = 1.80), who received extra credit in a college class. There was a total of 32 to 33 participants for each of the three between-subjects conditions tested. Participants all reported normal or corrected-to-normal vision and normal hearing.

Table 1: Display Design

Condition	Context		Basis
Non-Speech	Non-speech: All health variables		Based on the judgments of the sound designers to best-fit health concepts to the acoustic parameters outlined by Schuett (2017)
2 Speech	Speech: Heart rate, Blood pressure	Non-speech: Blood oxygen level, Respiratory rate, Body Temperature	Based on the judgement of which health concepts fit best with speech streams
3 Speech	Speech: Heart rate, Blood pressure, Body Temperature	Non-speech: Blood oxygen level, Respiratory rate	Based on the judgment of which health concepts fit best with speech streams
Speech	Speech: All health variables		Based on the judgment of what sounds the best when all five speech streams are played together.

Table 2: Location and Acoustic Mappings

	Left Ear	Centered	Right Ear
Variable Location	Respiratory Rate	Body Temperature	Blood Oxygen Level
	Heart Rate	--	Blood Pressure
Acoustic Parameters	**Left Pan**	**Centered**	**Right Pan**
	Frequency (Pitch)	Chord (Intensity changes)	Pink Noise (Intensity changes)
	Tremolo (Speed)	--	Filter (filter on pink noise)

Table 3: Acoustic and Speech Parameters

Concept Variable	Acoustic Parameter	Speech Parameter
Respiratory Rate	Frequency	Numeric respiratory rate value Uses lower pitched voice.
Heart Rate	Tremolo	Numeric heart rate value Uses a higher pitched voice.
Body Temperature	Intensity (chord)	Numeric body temperature value Uses a monotone, robotic voice.
Blood Oxygen Level	Pink Noise Intensity	Numeric blood oxygen level value Uses a higher pitched voice.
Blood Pressure	Filter (on pink noise)	Numeric blood pressure value (two numbers) Uses a lower pitched voice.

3.2. Display Design and Mapping

The methods of this experiment followed closely to Schuett's dissertation [8], but, with adjusted sound files and some minor procedural changes. The purpose of the displays is to determine which auditory display mappings (speech, non-speech, and mixed) results in highest performance, by comparing their learnability to one another. There were four display mappings. Table 1 includes all four mappings, using the same health variables. One mapping was identical to the "Health" mapping in Schuett's dissertation [8], which maps five health variables, specifically those used by anesthesiologists. Another mapping used the same health variables in Schuett's dissertation [8] but introduced speech streams based on the study of speech displays by Ericson et al. [13]. The two mixed displays had a combination of speech and non-speech streams. One had two non-speech streams and three speech streams, and the other had three non-speech streams and two speech streams.

The auditory streams were separated in stereo space by panning one into the left ear, one into the right ear, and the third centered. The centered stream was used to only represent one variable, while both the left and right represented two variables combined in one stream. The use of three streams was to segregate the five variables for listeners. Table 2 shows the mapping of the health variables to their respective ears.

Non-Speech mapping. This display is identical to Schuett's "Best-fit Display Mapping (Health)" [8]. Table 2 summarizes the acoustic mapping of the sonifications for the Non-Speech display.

The data trends represented by each of these five parameters could increase, decrease, remain constant, increase-then-decrease, or decrease-then-increase over time. The display was intended to represent informative trends that any of the health parameters could have in a given time frame. The context of the health data was chosen for this condition, which is congruent with past sonification of health related concepts such as Fitch and Kramer [3] and Anderson and Sanderson [5]. Respiratory rate and heart rate were paired together in the left ear because the two are connected conceptually; and similarly, blood oxygen level and blood pressure were also paired due to their connection to one another in the human body. Body temperature is least connected to the other four variables, so it remained in its own stream in the stereo-centered location.

Mixed Displays: 2-Speech mapping and 3-Speech mapping. These displays added speech into certain variables of the display. The mixture of the two stream types incorporated findings from Fitch and Kramer and reflected the optimal design for speech auditory displays as indicated by Ericson et al. [3], [13]. For the 2-Speech display, two of the variables were mapped using speech and three of the variables were mapped using non-speech sonification. For the 3-Speech display, three of the variables were mapped using speech sounds and the other two remained non-speech. The general layout for each display was similar to the Non-Speech display, where each variable remained in its respective ears and followed its respective acoustic parameter. Table 3 includes the acoustic and speech parameters for each variable.

Speech Display. The final display had all five variables represented by five speech sounds. The speech parameters are listed in Table 3. Following pilot testing, this display was not included in the experiment due to the difficulty participants had with it. Even when intentionally listening to the display sounds, it was difficult to concentrate and monitor a single variable, let alone five speech variables.

3.3. Materials

Throughout the duration of the study, participants wore SONY MDR-V150 Headphones, sat in front of a computer in a computer lab, and completed the study via an automated Qualtrics survey. This was a slight procedural differences from Schuett's study in which participants were run one at a time and researchers were heavily involved during each step [8].

Listening Discrimination Task. The point of the Listening Discrimination Task is to see if differences in individual

performance on the task affect performance on the use of the auditory display. Individual differences allow some participants to have a "trained ear", which allows them to be better at discerning smaller differences between acoustic stimuli.

Participants' abilities were assessed separately from the main study. The Listening Discrimination Task after Schuett [8] required participants to listen to one audio track, followed by another, and determine if the first and second track were the same or different. The first and second track were either the same, or differed by one acoustic parameter each time. The task increased in difficulty when the number of acoustic parameters in the tracks increased. When there was only one acoustic parameter, a change across that single parameter was relatively easy for the listener to discern. But when there were multiple acoustic parameters in each track, detecting the presence of a change became increasingly difficult.

For each Listening Discrimination Task trial, participants were presented Track A and then Track B, and given a choice "same" or "different" to choose from. This task consisted of 26 total trials. In half of the trials, Tracks A and B were the same, and in the other half they were different. The trial difficulty was presented in a randomized order for each participant through Qualtrics. The acoustic parameters used for each of the thirteen acoustic groupings are included in Appendix A.

Intrinsic Motivation Inventory. This study also used the same Intrinsic Motivation Inventory (IMI) scale [17], [18], as Schuett's study [8], which participants completed three times throughout the study. The scale measures subjective motivation towards a specific task during the study. The first was administered after the pretest to gauge motivation during the pretest phase. The second occurred at the end of the practice with feedback phase, and the third occurred after the posttest. The purpose of these was to determine if participants got bored or tired throughout the study and if it would have an effect on the participant responses. It was also used to see if their motivation increased between the pretest and posttest. The items in the IMI are listed in Appendix B.

Musical Sophistication Index. Using a shortened version of the Goldsmiths Musical Sophistication Index (Gold-MSI) [19], participants self-reported musical skills and behaviors to assess their history with musical instruments as well as a variety of items that assessed overall level of musical engagement and sophistication. The measure includes four Factors: Factor 1 is related to active engagement in musical activities; Factor 2 is related to perceptual abilities; Factor 3 is related to musical training; and General Factors is a mix of the categories. The MSI items used here are in Appendix C.

3.4. Procedure

Participants were randomly assigned to one of the three display conditions; Non-Speech, 2-Speech, or 3-Speech. The study used a between-subjects design to ensure that participants could focus on becoming familiarized with one display mapping. All sections of the study were presented via Qualtrics, and mp3 files were uploaded and integrated into the survey platform. The first task was the Listening Discrimination Task, followed by an introduction to their assigned display. Then, participants completed the pretest and filled out the first Intrinsic Motivation Inventory. Then they continued to the practice phase, which was on a separate Qualtrics survey. After practice, participants returned to the original Qualtrics survey to fill out the second motivation survey and complete the posttest. Lastly, they filled out the third and final motivation survey and the Musical Sophistication Index.

Listening Discrimination Task. Participants determined if two sound clips were the same or different.

Introduction to the display mapping. The participants were given an introduction to their assigned display. Participants clicked through example sound clips of each of the variables in their display, along with a short explanation of the parameter mapping. Participants were able to listen to the mapping examples and explanations as many times as they liked and were allowed to ask questions.

Pretest. After the participants felt comfortable with their introduction, they were directed to the pretest. The pretest evaluated the listeners' ability to comprehend the data presented within the display initially, without practice, and was used to compare to the posttest results, after practice with feedback. There were a total of 20 questions. Participants listened to a mp3 sound file embedded into the survey that combined all five variables together across the three streams. Then participants were asked to select the trend ("increase", "decrease", "constant", "increase then decrease", and "decrease then increase") of one of the variables from that sound clip. Tracks was presented in a randomized order to each participants.

Practice Phase with Feedback. The practice phase was similar to the pretest phase, but started with a short matching section to review the variable mappings. The survey also allowed participants to go back and replay the sound tracks if needed, and it provided feedback on their answers. The 20 tracks in the practice phase were similar to, but distinct from, the tracks used in the evaluation phase.

Posttest. The posttest phase occurred after practice; it followed the same procedure as the pretest, with the same 20 tracks but in a randomized order.

Motivation Checks. Participants were asked to complete the IMI scale three times: after the pretest, after practice with feedback, and after the posttest.

Musical Sophistication Index. After the participants finished the posttest and the last motivation scale, they completed the abbreviated Goldsmiths MSI.

3.5. Hypotheses

H1. The first hypothesis was (a) that there would be a difference in performance before and after the practice phase, and (b) that participants in the mixed auditory displays would perform differently from participants in the non-speech display.

H2. The second hypothesis was that individual differences such as musical experience and motivation would predict overall listeners' performance on the initial task, and would predict the amount of improvement after practice.

4. RESULTS

There were initially 102 participants in the study. Data from five were removed as statistical outliers in the pretest and posttest score; this left 97 participants for analysis. The data were analyzed with respect to the two primary hypotheses using a split-plot Analysis of Variance (ANOVA) and hierarchical linear regressions.

Table 4: Summary of Test Scores

Evaluation	Mean	Standard Deviation
Pretest	**8.68**	**2.47**
Non-Speech	9.30	2.62
2-Speech	8.66	2.34
3-Speech	8.06	2.36
Posttest	**10.16**	**2.90**
Non-Speech	10.67	3.17
2-Speech	10.00	2.82
3-Speech	9.81	2.71

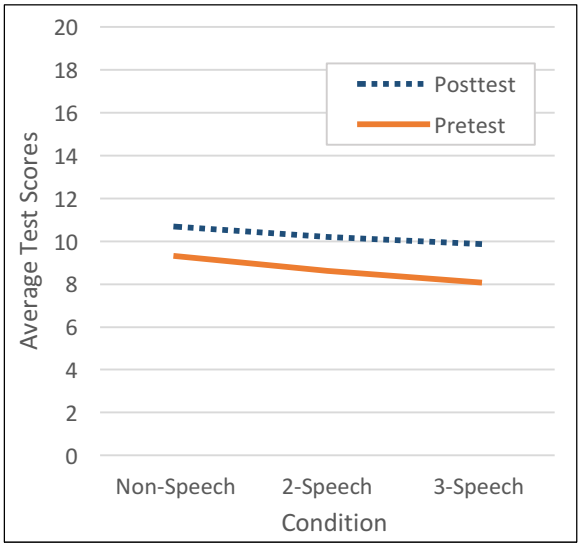


Figure 1. Average pretest and posttest score. This figure highlights the difference in average pretest scores compared to average posttest scores across the three different conditions. The range in scores is from 0-20.

4.1. Hypothesis 1

The first hypothesis was that there would be an improvement in score from pretest to posttest, and a difference in improvement between the three conditions. The results are listed in Table 4. The average pretest score across all conditions was lower than the average posttest score across all conditions. The average pretest score for Non-Speech, was higher than the average pretest score for 2-Speech, which was higher than the average pretest score for 3-Speech. The average posttest score for Non-Speech was also higher than the average posttest score for 2-Speech, which was also higher than the average posttest score for 3-Speech. Results from the split-plot ANOVA showed that there was a significant main effect for test scores $F(1,94) = 28.237$, $p < .001$, but not for condition $F(1,94) = 0.214$, $p = .807$. There was a statistically significant difference between pretest and posttest scores, but no statistically significant difference in improvement among the three conditions. These findings partially support Hypothesis 1, as there was a significant improvement in scores from pretest to posttest. This suggests that practice with feedback affected participants equally regardless of the condition, and that participants were able to improve their scores after the practice phase (Figure 1).

4.2. Hypothesis 2

The second hypothesis was that individual differences in musical experience and motivation would affect performance on the pretest and posttest scores. The results are listed in Table 5. Musical experience is the combination of the Listening Task Score and the four subsections of the Musical Sophistication Index which are Factor 1: Active Engagement, Factor 2: Perceptual Abilities, Factor 3: Musical Training, and General Factors. Motivation was measured three times throughout the study; the first one after pretest, the second after practice with feedback, and the third time after posttest.

Table 5: Summary of Individual Differences

Variable	Mean	Standard Deviation
Listening Task	20.7	6.51
Factor 1: Active Engagement	29.80	10.73
Factor 2: Perceptual Abilities	41.42	6.51
Factor 3: Musical Training	22.47	14.20
General Factors	69.23	22.85
Motivation 1	84.46	15.71
Motivation 2	86.72	16.04
Motivation 3	83.30	17.35

There were a total of fifteen step-wise linear regressions to observe the predictability of musical experience and motivation on pretest scores and on posttest scores. Additionally, pretest scores were also used to determine if they were good predictors of posttest while controlling for musical experience and motivation.

Predicting Pretest scores. For the regressions predicting pretest scores, musical experience and motivation were not significant predictors when all conditions were combined. However, in the Non-speech condition, when controlling for musical experience, motivation accounted for 37% of variance in pretest scores, $\Delta R^2 = 0.374$, $F(8,24) = 3.194$, $p = .013$.

Predicting Posttest scores. For the regressions predicting Posttest scores across conditions, musical experience and motivation were both significant predictors. Musical experience accounted for 13% of the variance in posttest scores, $R^2 = 0.131$, $F(5,91) = 2.753$, $p = .023$. There was a significant contribution of motivation when controlling for musical experience, accounting for 10.8% of variance in posttest scores, $R^2 = 0.239$, $F(8,88) = 3.460$, $p = .002$, $\Delta R^2 = 0.108$, $p = .008$. Motivation was a significant predictor of posttest scores in the 3-Speech condition, accounting for 17.3% of the variance in posttest score when controlling for musical experience, $R^2 = 0.462$, $F(8,23) = 2.465$, $p = .043$, $\Delta R^2 = 0.173$, $p = .088$. For the 2-Speech condition, both musical experience and motivation were significant predictors for posttest scores. Musical experience accounted for 24% of the variance in posttest scores, $R^2 = 0.237$, $F(5,58) = 3.602$, $p = .007$, while motivation accounted for 17% of the variance in posttest scores when controlling for musical experience, $R^2 = 0.407$, $F(8,55) = 4.712$, $p < .001$, $\Delta R^2 = 0.170$, $p = .003$.

Predicting Posttest scores with Pretest scores. The last set of regressions took the pretest score as a final predictor of posttest scores in the step-wise regression. All three predictors (musical experience, motivation, and pretest score) were significant predictors of posttest scores when all three conditions were combined. Musical experience accounted for 13% of the variance in posttest scores, $R^2 = 0.131$, $F(5,91) = 2.753$, $p = .023$, while motivation accounted for 10% of the variance in posttest score when controlling for musical experience, $\Delta R^2 = 0.108$, $F(8,88) = 3.460$, $p = .002$. Additionally, when controlling for both musical experience and motivation, pretest scores accounted for 15% of the variance in posttest scores, $\Delta R^2 = 0.152$, $F(9,87) = 6.205$, $p < .001$. All three predictors were also significant predictors of posttest scores in the two speech conditions combined (2-

Speech and 3-Speech combined). Musical experience accounted for 24% of the variance in posttest scores, R^2 = 0.237, $F(5,58) = 3.602$, $p = .007$, motivation accounted 17% of the variance in posttest score while controlling for musical experience, $\Delta R^2 = 0.170$, $F(8,55) = 4.712$, $p < .001$. Last, when controlling for musical experience and motivation, pretest scores accounted for 7% of the variance in posttest score, $\Delta R^2 = 0.071$, $F(9,54) = 5.480$, $p < .001$.

Musical Experience predicting Posttest scores. Musical experience was a combination of five variables; one listening task and four sections of the Musical Sophistication Index. Each have different standardized coefficients that can be used to determine which one is a better predictor for posttest score. In the conditions where musical experience was a significant predictor of posttest score, the coefficients of Factor 1: Active Engagement was a better predictor than the other three Factors. For instance, when data from 2-Speech and 3-Speech were combined, the first model with just musical experience as a predictor shows that Factor 1, $b = -.440$, $t(64) = -2.424$, $p = .018$ is a better predictor than Factor 2, $b = .170$, $t(64) = 1.139$, $p = .259$, Factor 3, $b = .042$, $t(64) = .211$, $p = .834$, and General Factors, $b = .443$, $t(64) = 1.544$, $p = .128$. The same trend is seen when motivation is added in as a predictor, where Factor 1 is the best predictor of posttest scores, $b = -.499$, $t(64) = -3.015$, $p = .004$, and again, when pretest scores are added as a third predictor, $b = -.474$, $t(64) = -3.018$, $p = .004$. Factor 1 is a better predictor of posttest scores than its counterparts, even when musical experience all together might not be a significant predictor.

5. DISCUSSION

This study was designed to explore listeners' ability to interpret health-related information from auditory displays before and after a practice phase, to see if practice with feedback would help improve performance and comprehension. It also investigated whether or not the display designs would have an impact on listeners' ability to improve their posttest scores after practice. In addition to looking at the effects of practice on the pretest score and posttest score, musical experience and motivation were also measured to see if any of those variables predicted scores.

Overall, practice was helpful and did improve listeners' ability to comprehend information in the auditory displays, however there was no statistically significant difference among the three conditions, Non-Speech, 2-Speech, and 3-Speech. Findings also showed that motivation and musical engagement were significant predictors of posttest scores. The remainder of this section will be split by these two main findings that correspond to each hypothesis.

The first hypothesis was that there would be a difference between pretest scores and posttest scores and that the different display designs may show different effects on that improvement between the pretest and posttest scores. This is based on the evidence that practice with feedback significantly lowers errors while performing sonification tasks [12]. It is also based on the assumption that non-speech and mixed speech auditory displays may have varying difficulty levels, each with specific design factors that can impact performance and usability overall. Findings only partially supported this hypothesis, in that there was a significant difference between pretest and posttest score but no difference among the display designs. These results indicate that regardless of the display design, the participants improved significantly between the pretest and posttest with the help of the practice with feedback. Participants generally started off scoring low during the pretest and were able to improve their score after the practice phase.

Because this task is foreign to participants, they were all starting off on the same level, where their initial performance in the tasks is generally low. However, with practice, as participants became familiar with the sounds and trends of the variables, they were all able to improve roughly the same amount across all conditions. This may also suggest that including speech into a mixed auditory display does not increase familiarity with the display compared to the non-speech display, possibly due to the fact that most participants are not exposed to mixed auditory displays, especially with multiple streams. It would be interesting to compare accuracy in monitoring change in the speech variables versus the non-speech variables to see if familiarity with speech translates to better detection of the speech variables over the non-speech variables. Additionally, workload tasks or measures of usability for each of the display designs may give better insight to how different the displays might have actually been.

The second hypothesis was that individual differences, such as musical experience and motivation, may predict how well individuals perform on the pretest and posttest. Motivation checks were a way to discard data from participants who were not motivated at all, but also because there may be a correlation between motivation scores and test scores. Findings from the hierarchical linear regressions partially supported this hypothesis, where motivation was a significant predictor of posttest scores. The effects were minimal in predicting pretest scores, most likely due to not being bored or tired yet. It serves as a good reminder that motivation plays an important role in participation and obtaining clean, representative data.

Previous research suggests that musical experience such as musical training and expertise may help listeners detect irregularities or changes in auditory streams better than those who do not have musical backgrounds [10]. Though research has not found a clear connection between musical training and stream segregation, there may still be a link that has not been found and is worth looking into [8]. In this study, musical experience included the Listening Task Score and the four sections of the Musical Sophistication Index, and was a significant predictor of posttest score. Factor 1 of the Musical Sophistication Index score is based on active engagement in music and music-related activities. Results show that Factor 1 is usually the best predictor for posttest score compared to the other factors, such as perceptual ability and musical training. This suggests that musical training and expertise is not required for monitoring auditory streams; instead, **active engagement in music** is more likely to impact listeners' ability to monitor auditory streams. In this study, those who scored high on active engagement (Factor 1) may not have had formal musical training, but could still improve significantly on the posttest, compared to someone with years of musical training. Furthermore, participants who scored high on musical training may not have scored high on motivation, while participants who scored high on active engagement may have scored higher on motivation. It would be interesting to see if active engagement correlates with motivation and interest in the study, which can lead to higher posttest scores and a larger improvement. Previous research has reached conflicting conclusions on how musical experiences impacts stream segregation and stream monitoring, but mostly because musical experience has been operationalized in so many different ways [2]. Musical training in an instrument or voice for a certain number of years may not lead to the same level of expertise or ability for each person, so using it as a

measurement may not lead to consistent results. Active engagement in music is more straightforward since it takes into account the amount of time a person spends engaging in music in a given time, while ignoring other factors such as expertise and training. These results indicate that individual differences do not matter when first introduced to an unfamiliar auditory display, but they do matter when predicting how much individuals might improve using the display with practice and feedback. This may be because the unfamiliar auditory display places everyone on the same level, but some individuals improve with practice more than others, due to individual differences.

This study scratches the surface of speech and non-speech mixed auditory display designs, and the effects of active engagement in musical activities on the usability of these displays. It demonstrates that users who actively engage in music are able to learn to use unfamiliar auditory displays better than those who do not engage in music. Continuation of this field of research can lead to better understanding of auditory display designs and training methods for future applications of these displays, such as in an anesthesiologist's workstation, a driver on a long road trip using in-vehicle interfaces, or visually impaired students using STEM education tools. Understanding how to best transform data and information into auditory streams can help reduce the dependence on visual displays and overcome information overload.

6. REFERENCES

[1] Walker, B. N., & Nees, M. A. (2011). Theory of sonification. In Hermann, T., Hunt, A., & Neuhoff, J. G. (Eds.), *The Sonification Handbook*. (9-39). Logos Publishing House, Berlin, Germany.

[2] Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89*(2), 244.

[3] Fitch, W. T., & Kramer, G. (1994). Sonifying the body electric: Superiority of an auditory over visual display in a complex, multivariate system. In G. Kramer (Ed.), *Auditory Display: Sonification, audification, and auditory interfaces.* (307- 325). Reading, MA: Addison-Wesley Publishing Company.

[4] Loeb, R. G., & Fitch, W. T. (2002). A laboratory evaluation of an auditory display designed to enhance intraoperative monitoring. *Anesthesia & Analgesia, 94*(2), 362-368.

[5] Anderson, J., & Sanderson, P. (2004). Designing sonification for effective attentional control in complex work domains. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48*(16), 1818-1822.

[6] Schuett, J. H. (2010). Limits on the number of concurrent auditory streams. Masters Thesis. James Madison University, Harrisonburg, VA.

[7] Schuett, J. H., Winton, R. J., Batterman, J. M., & Walker, B. N. (2014). Auditory weather reports: demonstrating listener comprehension of five concurrent variables. In Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound. ACM. 17.

[8] Schuett, J. H. (2017). Measuring the effect of display design and practice on listener accuracy for auditory displays with multiple streams (Doctoral Dissertation Proposal). Georgia Institute of Technology, Atlanta, Georgia.

[9] Watson, C. S., & Kidd, G.R. (1994). Factors in the design of effective auditory displays. Proceedings of the Second International Conference on Auditory Display ICAD `94, Santa Fe Institute, New Mexico.

[10] Brochard, R., Drake, C., Botte, M. C., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: effect of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1742.

[11] Lacherez, P., Seah, E. L., & Sanderson, P. (2007). Overlapping melodic alarms are almost indiscriminable. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 49*(4), 637-645.

[12] Walker, B. N., & Nees, M. A. (2005). Brief training for performance of a point estimation sonification task. Proceedings of the International Conference on Auditory Display (ICAD2005), Limerick, Ireland.

[13] Ericson, M. A., Brungart, D. S., & Simpson, B. D. (2004). Factors That Influence Intelligibility in Multitalker Speech Displays. *The International Journal Of Aviation Psychology, 14*(3), 313-334.

[14] Simpson, B. D., Brungart, D. S., Dallman, R. C., Joffrion, J., Presnar, M. D., & Gilkey, R. H. (2005). Spatial audio as a navigation aid and altitude indicator. *Human Factors and Ergonomics Society, 49*, 1602-1606.

[15] Simpson, B. D., Brungart, D. S., Gilkey, R. H., & McKinley, R. L. (2005). Spatial audio displays for improving safety and enhancing situation awareness in general aviation environments. *New Directions for Improving Audio Effectiveness, 26*, 1-16.

[16] Li, S. W., Tang, T., Hickling, A., Yau, S., Brecknell, B., & Sanderson, P. M. (2017). Spearcons for patient monitoring: Laboratory investigation comparing earcons and spearcons. *Human Factors, 59*(5), 765-781.

[17] Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450-461.

[18] Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology, 45*, 736-750.

[19] Müllensiefen, D., Gingras, B., Musil, J., & Stewart L. (2014). The Musicality of Non- Musicians: An Index for Assessing Musical Sophistication in the General Population. PLoS ONE, 9(2): e89642

INTERACTIVE AUDITORY NAVIGATION IN MOLECULAR STRUCTURES OF AMINO ACIDS

A CASE STUDY USING MULTIPLE CONCURRENT SOUND SOURCES REPRESENTING NEARBY ATOMS

Danyi Liu

Leiden Institute of Advanced Computer Science
Snellius Gebouw, Niels Bohrweg 1
Leiden, 2333CA, The Netherlands
d.liu.7@liacs.leidenuniv.nl

Edwin van der Heide

Leiden Institute of Advanced Computer Science
Snellius Gebouw, Niels Bohrweg 1
Leiden, 2333CA, The Netherlands
e.f.van.der.heide@liacs.leidenuniv.nl

ABSTRACT

We are interested in sonifying the molecular structures of amino acids. This paper describes the context and the first design choices for our approach. So far, we believe an amino acid molecule is too complex to be perceived at once. Therefore, we have designed an interactive form of sonification in which the listener navigates through the molecule over the network of carbon atoms. We describe our different approaches and discuss the topic of immediacy: the time it takes to recognize the structure surrounding the listener's position while navigating. Furthermore, we touch upon the question how many atoms we can sonify simultaneously and the role auditory masking plays in this context. To overcome auditory masking, we propose to use irregular but easy to recognize sounds. We conclude with an interest in a three-dimensional navigation environment using general molecular structures for further research and development.

1. INTRODUCTION

In our daily lives we are used to navigate through sound environments consisting of multiple sources that not only indicate their positions but also communicate information to us. In laboratory environments, listeners are often presented with rather simple auditory stimuli and listening tasks in order to learn more about our spatial perception. Many studies researched the localization of diverse sound stimuli in the form of single sound sources positioned at various azimuths and elevations [1, 2, 3, 4]. However, relatively few studies focused on our ability to localize two or more concurrent sound resources [5, 6]. In this paper, we illustrate and discuss the approach we have taken to develop an interactive sonification system using multiple sound sources that are spatialized in the horizontal plane around the listener. We are using a simple four-speaker setup in which the positions of the speakers correspond to the directions of the sound sources (see Fig.1). We are currently interested in sonifying the structural formulas of amino acids because of its relatively easy structures. In the future we aim to sonify RNA structures including folding.

Our ability to perceive a sound's direction and estimate the origin of a sound is called sound localization. This works through a process known as binaural hearing. In horizontal plane, our localization relies on a combination of multiple acoustic cues: a) interaural time/phase differences (ITD/IPD), b) interaural intensity differences (IID) and c) the spectral shape [7]. An enormous amount of research has been done on spatial hearing and the ability of a human to localize sound, both using headphones, as well as in free-field setups with loudspeakers. Stevens and Newman conducted experiments in the open air in 1936. Sounds were produced by a speaker which could be moved noiselessly over a circle in the horizontal plane. They concluded that noise was localized more easily than any of the pure tones [1]. Later, Hartmann tested and compared the performance of localizing continuous pure sine tones, broadband noise and complex signals in a room. The result indicated that azimuth judgement became more precise when the spectral density of the sound increased [2]. Lokki et al. did an auditory navigation experiment in 2000 in which the subjects were asked to move in a virtual space with arrow keys of a keyboard and find a point-shaped sound source with a random-position [3]. The sound reproduction equipment was a headphone. They tested three different factors: a) audio stimuli with different spectra including pink noise, artificial flute and recorded anechoic guitar, b) different panning methods for the positioning of the sound, and c) different acoustical conditions: direct sound, combined with early reflections, combined with reverb. The results proved that noise is the easiest stimulus to localize, and reverberation complicates the navigation. Letowski et. al pointed out that sound sources producing impulse sounds (e.g., firearms) are easier to be localized than sources emitting continuous or slowly rising long tones in closed spaces (rooms) [4]. These studies have investigated different aspects that may affect the localization accuracy of single sound sources. On the other hand, Brungart et al. conducted an experiment in which 14 different continuous but independent noise sources were turned on in a sequence within a geodesic sphere consisting of 277 speakers [6]. Each time when a new source was added, the listener was asked to localize it. They found that localization accuracy was modestly better for the sounds with rapid onsets than 1-second ramp onsets. Additionally, accuracy declined as the number of sources increased but was still higher than expected on the basis of chance when all 14 sources were on.

Sound localization is only one possible aspect of sonification. In our study, the sounds represent the type and position of the atoms around us. It is important that the sonification is easy to learn and understand in an intuitive way. In the context of auditory display and sonification, sound has been used to represent complex data, enhance visualizations, as well as support the under-

https://doi.org/10.21785/icad2019.049

standing of subjects in an educational context. Several approaches are distinguished from each other such as the used of earcons, auditory icons, parameter mapping sonification (PMSon) and model-based sonification (MBS) [8]. All of these approaches are based on the human's auditory system, which derives three auditory dimensions that are commonly used in auditory display: loudness, pitch and timbre [9]. With these primary features, humans are able to separate and identify different sound sources, each with their own characteristics. While auditory icons are meant to represent events directly, earcons are synthesized sounds which require a learning process to relate the indirect sound to a specific meaning. When a continuous data stream is involved, it is effective to use PMSon with predetermined relations between the chosen auditory features and the information the data contains. MBS often uses a dynamic model that can include interaction, and utilizes sound to help to analyze a specific data task. Additionally, Carlie showed that the auditory system is sensitive to differences in the duration of a sound larger than 10ms, generally the smallest detectable change increases with the duration of the sounds [10]. This brought us to the idea that duration could also be used as a parameter for identifying different sounds sources. In order to be able to localize and identify the multiple surrounding atoms as fast as possible, our decisions for the sound design were affected by the features mentioned above. We will explain our choices in detail in Section 3.

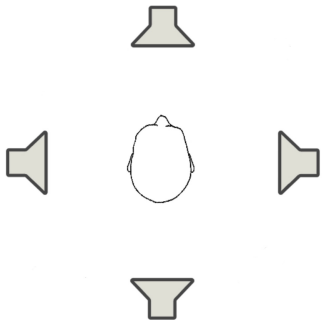


Figure 1: Positions of four speakers setup.

2. INTERACTION DESIGN

The visual field of the human eye has a limited arc while sounds is perceived omnidirectional. Sounds could reveal the existence of something that is difficult to be seen. The three-dimensional structures of proteins attract us, especially the folded parts where amino acids interact with each other. The aim of our research is to sonify multiple surrounding objects simultaneously in the horizontal plane, and to test whether they can be perceived, localized and identified by means of interactive navigation. Due to the complexity and inherent high dimensional order of proteins, we chose to start with exploring the structural formulas of different amino acids in two dimensional schematics. Unlike written chemical formulas, the structural formulas provide a geometric representation of the molecular structure. To simplify the localization task, we have transformed the formulas into flat graphical ones with identical bond angles of either 90 or 180 degrees, and identical bond lengths (see Fig.2). We are aware that this is an extreme simplification of the actual structure but it simplifies the sound spatialization in such a way that the speakers always correspond to the actual directions of the sound sources and we don't need to create phantom source locations in between the speakers. It relates more to how a molecule is drawn on paper than to its spatial shape in three dimensions.

2.1. Speaker Setup

Different from the common quadraphonic speaker setup, we place the four speakers around us to the front, left, back and right (see Fig.1). The physical position of each speaker always corresponds to the position (or direction) of the sonified atoms. We don't need to create phantom source locations in between the speakers and thereby we avoid potential negative effects of spatialization techniques. We sonify the atoms that are connected to a certain carbon atom that forms the (imaginary) center of the speakers and is not audible itself. Detailed sonification and localization implementations will be explained in Section 2.2.1.

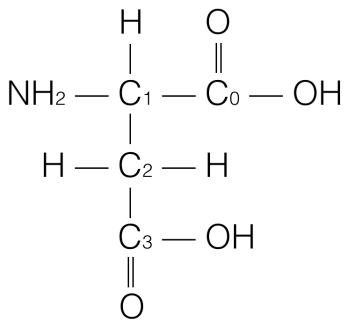


Figure 2: The structural formula of Aspartic acid.

2.2. Interactive Navigation of Structural Formula

In the past decades, structural biology developed into dealing with the molecular structure of biological macromolecules, like proteins, made up of amino acids or nucleic acids. Atoms are organized in a complex ordered 3D manner and form a macromolecule. Grond et al. developed SUMO, an open source software environment to sonify structure data contained in PDB files[1]. They implemented acoustic signatures for each amino acid, where different amino acids had different sounds, and parameterized earcons were used to distinguish pairwise distances and conformation differences of amino acids [11]. SUMO shows how sonification can be complementary to visually displaying macromolecules. Two years later, Grond et al. combined visualization, sonification and interaction in their application to represent the possible secondary structures of an RNA sequence. The application was designed to turn RNA structures into auditory timbre gestalts according to the shape classes they belong to, on the different abstraction levels [12]. Thereby, it became possible for the users to quickly compare structures based on their sonic representation. Additionally, the users were able to learn the meaning of the sound by selecting the visual pieces and playing back the corresponding sound. Compared with sonifying the structures as a whole part in [11], such interactions provide an interesting and effective way for the users to discern the meaning of the sounds.

[1]PDB is a standardized file format saving macromolecular structure data, which contains the positions in x/y/z of all atoms belonging to the corresponded molecule and other relevant information.

In previous studies, sound has been used to enhance the existing structural visualization of static data. Is it conceivable for the listeners to follow the structures when the visuals are removed? What kind of method could help the listeners to learn the meaning of the sounds when there are multiple concurrent sounds? In our design, we would like to only use sound to represent the structural formulas of amino acids. The listeners are able to navigate the structures by moving over the carbon atoms in the molecule with the arrow keys on the keyboard. The navigation task provides opportunities for the listeners to explore the structure and take notice of the surrounding environment on a step by step basis. Meanwhile it allows the listeners to focus on a part of the molecular structure. We assume that such an interactive design would help the listeners to learn the meaning of the sounds and understand the molecular structures.

2.2.1. Navigation Rules

It is necessary to find an accessible way for the listeners to navigate through the structures and not get lost. The 20 natural amino acids contain amine (-NH_2) and carboxyl (-COOH) functional groups, with different R groups (side chains). The common elements are carbon (C), hydrogen (H), oxygen (O), nitrogen (N), while other elements like sulphur (S) and selenium (Se) are found in the R groups of specific amino acids. There is a carbon chain attached to the central carbon atom called C_1 (see Fig.2), which is next to the carboxyl group. Starting from the central carbon, there are several carbon atoms connected and forming the skeleton structure. Therefore, we chose for a navigation method where the user is able to explore the structure by moving from one carbon atom to its neighboring carbon atom(s). The starting point is numbered as C_0, which is the carbon part of the -COOH group and connects directly to the central carbon (see Fig.2). In this case, the user can not move to the right, but only to the left where C_1 is located. If there is an attempt to move into a direction that is not a carbon atom, a short alarm sound will be played as feedback.

2.2.2. Concurrent sound sources implementation

The various elements (atoms) that are connected to the current carbon position will be sonified independently. The -NH_2 and -OH groups are exceptions to this rule and will be sonified as independent groups. In our first stage, only the four atoms/groups connected directly to the current carbon position, will be sonified. For example, when the listener stands on C_0, only -OH, =O and C_1 will be audible (see Fig.2). In this way, the listeners can learn the information conveyed by the sounds and audibly observe the structures by navigating. In our next stage we decided to sonify one more layer of atoms; the atoms connected to the first layer of sonified atoms and in positioned the same direction. In this stage, the groups will be decomposed into single atoms (see Fig.3). Accordingly, N connected to C_1 and H connected to -O are audible (see Fig.3). Thus up to eight atoms will be audible at the same time.

In the future, we would like to sonify even larger areas. For example, all of the atoms in a row of a carbon atom could be sonified simultaneously. When the listener stands on C_1, not only the two layers of atoms connected with it will produce sounds, the O connected with C_3 and the H connected with -O will also be audible (see Fig.3). When the listener moves to C_0, the same atoms in this horizontal row will still be heard but the changes of the surrounding sounds could imply the listener's position changes, and give evidence of how the atoms in this row are positioned. Furthermore, we will consider the use of spatialization techniques to realize phantom sound source locations and work with depth in the sound. For now, we have specifically chosen to make the speaker positions correspond to the location of the intended sound source positions and avoid possible negative side effects that the spatialzation could bring.

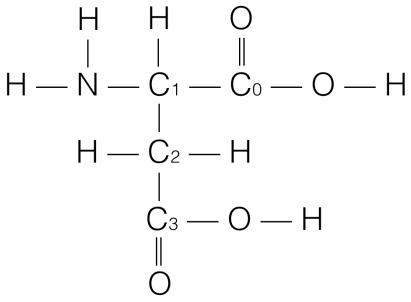


Figure 3: The structural formula of Aspartic acid for the larger area sonification.

3. SOUND DESIGN

In recent decades information sonification in the fields of chemistry and biology, mainly focuses on DNA sequences and macromolecular structures. Many different choices have been made to sonify and represent objects (e.g. amino acids, proteins, nucleotides) and events. For example, a) single note is mapped directly to string data derived from a DNA sequence [13, 14], b) short musical phrases are formed by the Morse code of the amino acids, nucleotides and nucleotide pairs [15, 14], c) parameterized earcons help the users to distinguish similar but different structures such as amino acids. Different parameters in a sound synthesizer can be mapped to the different features of an object or event [11, 12, 16], and d) pre-recorded samples are used as auditory icons to represent events extracted from simulation progress [17]. In these studies, sonification was utilized often to enhance the visual display of complicated structures. However, it remains unclear whether the listeners are able to recognize and comprehend the sounds without the visual input.

For our approach it is essential that the (interacting) listeners can both identify and localize the atoms purely by means of sound. This brings us to the question how the atoms should sound? For atoms there are no metaphorical approaches that are already familiar to us in daily life and therefor auditory icons are not applicable in our context. Therefore we considered earcons as a conceivable way to establish a mapping stratagem between the atoms and their sonic representation. Earcons are defined as short, structured musical messages, where different musical properties of sound are associated with different parameters of the data being communicated [8]. The relations between the earcons and the atoms are supposed to be understood and acquired by the listeners. The goal of our sound design is to be able to easily recognize and distinguish the different sounds from each other, even if they sound simultaneously. We have used Pure Data, a graphical programming language for real-time interactive multimedia processing, for both the interactive navigation and the real-time sound synthesis.

3.1. Sound Synthesis Techniques

We have experimented with different approaches regarding how to sonify the different atoms and how to deal with time (i.e. use rhythmical structures or not). The aims of our sonification are to represent as many surrounding atoms as possible (as many concurrent sounds as possible) and to be able to localize and identify the atoms in as little time as possible. We started with different drum samples because the timbre of different parts from a drum set (e.g. bass drum, snare drum, hi-hat) is easy to be distinguished and such percussion sound is short and easy to localize. In our first early prototype, hydrogens produced closed hi-hat sounds every 400ms, carbons produced snare drum sounds every 1.6s, oxygens and groups generated bass drum sounds every 3.2s. However, the drum samples might be distracting since the listeners can recognize them and may have problems to relate them with chemical elements.

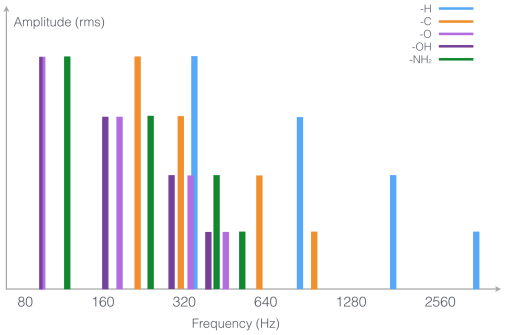


Figure 4: Frequency components for each element.

Then we tried filtered white noise with different amplitude envelopes. The central frequency of the bandpass filter is inversely related to the number of protons in the atom. The fewer protons, the higher filter frequency. This means that the sound that represents hydrogen has the highest frequency setting and the oxygen sound has a lower filter frequency than the carbon sound. The amplitude envelope enables different durations and loudness developments for each of the elements. The oxygen sound is the longest. While the single atoms have a clear and sharp start, the groups have a longer attack time. For example, the frequencies of a single oxygen atom and the -OH group are the same, but -OH has a slower attack time and longer duration at the sustain level. The filtered noise sounds are more abstract than the drum samples. We use pitch as the main feature in this design because the changes are easily perceivable and distinguishable. Hartman examined a tone with a fundamental frequency of 200Hz and 11 harmonics up to 5800Hz and concluded that the mixing of components within a single critical band plays a significant role in localization [2]. Therefore, we decided to add three more bandpass filters for each representation of an atom, resulting in a richer spectrum with four frequency partials, in order to improve the ability to localize the sounds. As shown in Fig.4, the frequency components made up for hydrogen are much higher, which are 352Hz, 877Hz, 1811Hz, 2941.1Hz. As a group, -OH relates to oxygen and the frequency components of -OH are slightly lower than oxygen. Both of them start with 100Hz, then oxygen develops with 201Hz, 350Hz, 461.1Hz and -OH includes 173Hz, 331Hz, 401Hz.

The main problem of this sonification approach is that it is hard to separate the sounds from each other when two or more of the same elements are played together. The similar frequency components produced from identical atoms may cause frequency masking. Also, merging may happen if they are positioned in a row (meaning in the same direction).

3.2. Sound Composition

When there is a complicated sound environment containing multiple concurrent sound sources, Brungart et al. used a sequential localization process to examine localization accuracy in 360 degrees. Each time, the subjects were asked to localize one newly activated sound source, but the previous played sources would remain. The sound sources were physically localized with 277 independently-addressable speakers which formed a geodesic sphere. Furthermore, each source was separated by 45 degrees from all the other sources. Brungart et al. pointed out that this method could avoid that sources originated from same direction, as well as help to reduce proximity-dependent effects of the individual maskers on the target [6]. Our approach does involve multiple sound sources played in parallel. The various frequency components contribute to be able to segregate one object from the others. Nevertheless, there are only four speakers representing four directions in our research, sound sources could be positioned in a row and produced from one same speaker. Later we will discuss other approaches to solve the merging problem when sources are concurrent and even played on one speaker. All of the approaches mentioned below started with the implementation of only sonifying the directly connected atoms and groups of the current carbon position (we call this the first layer). Afterwards we have extended some of the approaches and sonified also the atoms behind the directly connected atoms (we call this the second layer).

3.2.1. Rhythmical Pattern

Several researches have focused on melodic patterns in the field of sonification and auditory display, but there is little relevant research on rhythmical patterns. Rhythmical patterns could be regarded as a sound character to enhance and help the listeners to distinguish and localize multiple sound sources played simultaneously.

Firstly, we divided 4 speakers as 4 beats in a bar, and play a counter-clockwise sequence (front - left - behind - right) with a fixed tempo. This way the sounds will be played sequentially[2]. We implemented the envelope and duration differences mentioned in Section 3.1, combined with the bandpass filter groups. We would like to investigate whether sequenced nature could help the listeners to distinguish the different elements. This approach is a way to solve the problem of the overlapping sounds. However, it takes 2.4 seconds to finish a bar which might be a bit long for the listener to recognize and remember the sounds. It is still possible after several times of repetition but we would like to accelerate the process to achieve a faster and intuitive recognition of the different sounds in a (near) simultaneous way. Therefore, we tried another approach: Besides the envelope and duration differences, we assigned different repetition speeds to different elements. But the position always determines the beat where the sound starts to

[2]A binaural recording example of navigating in the structural formula of Aspartic acid with rhythmical pattern I can be viewed at: https://www.dropbox.com/s/p05lv10fg91equi/Rhythmical_pattern_1_Aspartic_1layer.wav?dl=0

play[3]. For example, when the listener stands on C_1 (see Fig.2), the hydrogen sound repeats at 600 bpm and synchronous to the first beat of the bar. The sound that represents -NH_2 repeats at 45 bpm is synchronous to the second beat in the bar. The carbon sounds repeat at 80 bpm synchronous to both the third and the forth beat. When all four speakers start to play sounds together, it is clear and direct for the listeners to note the similarities and dissimilarities among them. One of the disadvantages of this approach is that each element has an independent and distinct speed that can affect listeners to perceive different tempi at the same time. In addition, the sound results can be chaotic and annoying when there are various elements sonified together.

3.2.2. *Bouncing Pattern*

We also tried loops of a bouncing pattern to create a more interesting pattern for the listeners to identify. Imagine a ball is lifted at a certain height and then released, when it hits a surface it will create a sound, lose some potential energy and bounce into the air again, but lower than the original height. It keeps bouncing until it stops. As for the atoms, they could be balls falling from different height and have various bouncing patterns. Like hydrogen falls at a lower height and produces shorter bounces. Each element has a different bouncing speed and duration. A decay envelope is used to control the decrease in bounce period[4]. The bouncing pattern might be complicated and confusing at some point compared with the previous approaches of rhythmic pattern. The impact sound at the starting point of each loop is always clear, whereas further bounces quickly speed up and become rather intensive. Another problem is that when there are atoms of a same element that generate sounds, the bouncing pattern is also the same. Such bouncing sounds could be mixed up together and challenging for the listeners to separate one from the other, even though they are coming from different speakers. Furthermore, this approach will sound rather confusing when a larger area of the structure is sonified.

3.2.3. *Irregularly Triggered Bandpass Filter Banks*

The bouncing patterns brought us to the idea of a granular structure sound. In order to create a more continuous but irregular pattern, we used colored noise in combination with a comparator with a variable threshold as a way to generate random impulses with random amplitudes. The signal changes vary a lot from white, pink and brown noise. By choosing between different types of noise varying the threshold we can generate different impulse patterns with different desired densities. We chose to give the lighter elements an intensive but (light) pattern and the heavier elements and groups a more extensive pattern with a larger range of amplitude changes. Due to the irregular signal impulses, the all the sounds have their own non repetitive structure. This means that two or more identical atoms still have their own irregular structure. We use the impulse patterns as input signals for banks with four bandpass filters that we used before and mentioned in Section 3.1. Now, even when there are multiple sound sources generated together, the differences are still recognizable[5]. The main difference is that the irregular structure is experienced as a kind of granular-like texture. This makes it easy to recognize the sounds and the listeners are not required to remember the rhythmical patterns and compare them with each other. Now we can play the different sounds in parallel and they can all be identified simultaneously. We have found a way to avoid the merging problem that we had before.

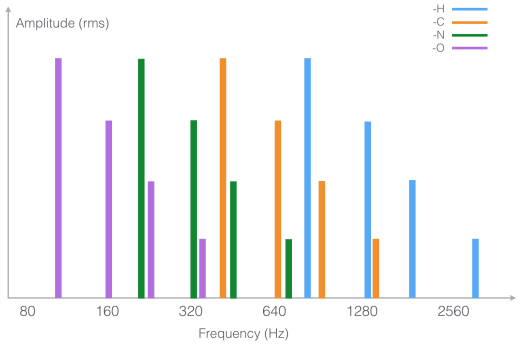


Figure 5: Frequency components for each element.

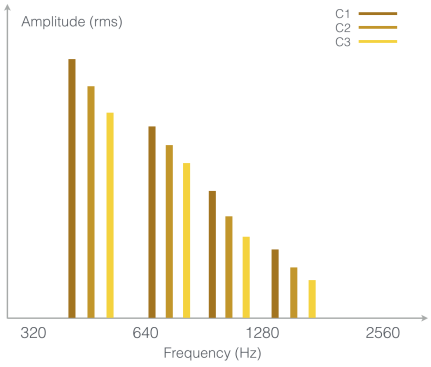


Figure 6: Different frequency components for the same element.

Now that we have achieved this we are curious to know if we can sonify even more atoms in parallel by sonifying the second layer around the carbon atom. Now we don't have to sonify the groups anymore since their individual atoms will both be played. The frequency settings of the filterbanks can be seen in Fig.4. Reverb is employed to enhance the sensation of distance of atoms in the second layer. The amplitude of the direct sound of the atoms from the second layer is one third of the ones from the first layer while the amount of reverb is the same. When the listener stands on C_1, C_2 and C_3 are both sonified (see Fig.3). On one hand, the distance determines the loudness and the sound of C_2 is louder than C_3. On the other hand, the q value of the bandpass filter of C_3 is slightly higher than C_2. The C_3 has more resonance and becomes less sharp and intensive. This is likely to solve the problem that the more intensive sound may mask a less intensive sound. In

[3]A binaural recording example of navigating in the structural formula of Aspartic acid with rhythmical pattern II can be viewed at: https://www.dropbox.com/s/14jkq9urf5l5k83/Rhythmical_pattern_2_Aspartic_1layer.wav?dl=0

[4]A binaural recording example of navigating in the structural formula of Aspartic acid with the bouncing pattern can be viewed at: https://www.dropbox.com/s/bx9nhybgbswoqz4/Bouncing_pattern_Aspartic_1layer.wav?dl=0

[5]A binaural recording example of navigating in the structural formula of Aspartic acid with ITBPFB can be viewed at:https://www.dropbox.com/s/gf7qrcte9z4pwhu/ITBPFB_1_Aspartic_1layer.wav?dl=0

our previous design, some frequencies were too low or too close to each other, which may have had a negative effect on separation and localization when two layers of objects are sonified simultaneously[6]. The frequency components have been adjusted and we have started to use a fixed interval size between the atoms and expanded the used filter frequencies in order to use a wider range (see Fig.5). There is an octave between two elements, for example oxygen is increased to 110Hz, nitrogen starts with 220Hz, carbon has 440Hz and hydrogen gets 880Hz. While oxygen and nitrogen remain with a less dense pattern, the resonance of the bandpass filters for these two elements is higher than for hydrogen and carbon[7]. In order to make the differences easily perceivable when two or more identical elements are positioned in the same direction we have chosen to give the elements of the second order a slightly higher pitch. The difference is small enough so that it is clearly identified as the same atom but larger enough to be able separate the sounds from each other and avoid merging. Fig.6 shows an example of different frequency components of the same carbon elements. There is a fixed ratio between two neighboring atoms. For example, if there are three carbon atoms positioned in a row at the same direction, the closest carbon is made up of 440Hz, 661Hz, 973Hz and 1389Hz and louder than other carbon atoms. The second carbon consists of 484Hz, 727.1Hz, 1072Hz and 1528Hz and the third carbons frequency components increase at the same ration of 1.1. However, it remains unknown what the maximum number of layers is that we can segregate.

4. CONCLUSIONS FUTURE WORKS

In this paper, we have discussed several different approaches to implement the spatial and interactive sonification of amino acids. We have personally evaluated the sound results in a research by design kind of approach. We are aware that part of our work could have been more detailed but have chosen to focus on the experimentation with the different approaches. We started with the concept of earcons in order to achieve the immediacy of sound recognition and localization. Unlike conventional earcons, such as time-based melodies or other sequentially played sound samples, our research focuses on concurrent sounds. We started with using fixed sound samples for the first rhythmical patterns and changed to real-time synthesized sound using banks of bandpass filters. While the repeating rhythmical patterns and bouncing patterns may take a longer learning time, the irregular impulses allow for a faster and simultaneous recognition of the atoms without a separation period. Currently, we combine frequency and irregular density as two main features for our sonification, to help the listeners to identify multiple simultaneous sound sources. By doing this we have expanded our approach that started with earcons toward a model-based sonification. It would be our next step to play an even larger area of concurrently sounding atoms. We already found that making light variations in frequency, density and loudness may (partially) solve the merging problem of multiple identical atoms coming from the same direction. The sound changes are regarded as auditory feedback from the interactive navigation, which may influence the localization accuracy and improve the segregation. In addition, it would be possible to realize a richer spectrum but avoid auditory masking.

Since all of the approaches mentioned above require a learning progress for the listeners to understand the mappings, further experimental investigations are considered to evaluate 1) whether the sounds properly represent the different elements, 2) whether the sounds are intuitive for the listeners to be recognized, and 3) whether the navigation could help to identify and localize multiple concurrent sources. Our main goal is to find out how complex a structure could be while still perceivable and recognizable. We will invite listeners to participate in usability and evaluation tests.

In order to simplify the localization task at present, we are using a particular 4-speaker setup in combination with the flat structural formulas. However, the molecular structures are three-dimensional, and the bond lengths and angles vary from one to another. It would be a logical step to represent the structures in a three- dimensional auditory environment. Setups consisting of more speakers in combination with different spatialization techniques will be considered. Bond lengths and angles could be included in the parameters used for the spatialization. Meanwhile, we are thinking how we can include active head movement in our research, which has proven to reduce front/back confusion and improve localization in elevation [18, 19].

[6] A binaural recording example of navigating in the structural formula of Aspartic acid with ITBPFB can be viewed at:https://www.dropbox.com/s/ylgqw9p3u2nuvwr/ITBPFB_2_Aspartic_2layer.wav?dl=0

[7] A binaural recording example of navigating in the structural formula of Aspartic acid with ITBPFB can be viewed at:https://www.dropbox.com/s/tdxf0949uetdutd/ITBPFB_3_Aspartic_2layer.wav?dl=0

5. REFERENCES

[1] S. S. Stevens and E. B. Newman, "The localization of actual sources of sound." *The American journal of psychology*, 1936.

[2] W. M. Hartmann, "Localization of sound in rooms," *The Journal of the Acoustical Society of America*, vol. 74, no. 5, pp. 1380–1391, 1983.

[3] T. Lokki, M. Grohn, L. Savioja, and T. Takala, "A case study of auditory navigation in virtual acoustic environments." Georgia Institute of Technology, 2000.

[4] T. R. Letowski and S. T. Letowski, "Auditory spatial perception: Auditory localization," Army Research Lab Aberdeen Proving Ground MD, Tech. Rep., 2012.

[5] P. L. Divenyi and S. K. Oliver, "Resolution of steady-state sounds in simulated auditory space," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2042–2052, 1989.

[6] D. S. Brungart, B. D. Simpson, and A. J. Kordik, "Localization in the presence of multiple simultaneous sounds," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 471–479, 2005.

[7] A. N. Popper, R. R. Fay, and A. N. Popper, "Sound source localization," pp. 272–316, 2005.

[8] T. Hermann, A. Hunt, and J. G. Neuhoff, *The sonification handbook*. Logos Verlag Berlin, Germany, 2011.

[9] J. G. Neuhoff, "Perception, cognition and action in auditory displays," *The sonification handbook*, pp. 63–85, 2011.

[10] S. Carlile, "Psychoacoustics," *The sonification handbook*, pp. 41–61, 2011.

[11] F. Grand and F. Dall Antonia, "Sumo. a sonification utility for molecules." International Community for Auditory Display, 2008.

[12] F. Grond, S. Janssen, S. Schirmer, and T. Hermann, "Browsing rna structures by interactive sonification," in *Proceedings of the 3rd Interactive Sonification Workshop*, no. Human Interaction with Auditory Displays, 2010.

[13] N. Munakata and K. Hayashi, "Basically musical," *Nature*, vol. 310, p. 96, 1984.

[14] M. D. Temple, "An auditory display tool for dna sequence analysis," *BMC bioinformatics*, vol. 18, no. 1, p. 221, 2017.

[15] X. Shi, Y. Cai, and C. Chan, "Electronic music for biomolecules using short music phrases," *Leonardo*, vol. 40, no. 2, pp. 137–141, 2007.

[16] A. Tek, M. Chavent, M. Baaden, O. Delalande, P. Bourdot, and N. Ferey, "Advances in human-protein interaction-interactive and immersive molecular simulations," in *Protein-Protein Interactions-Computational and Experimental Tools*. IntechOpen, 2012.

[17] B. Rau, F. Frieß, M. Krone, C. Muller, and T. Ertl, "Enhancing visualization of molecular simulations using sonification," in *2015 IEEE 1st International Workshop on Virtual and Augmented Reality for Molecular Science (VARMS@ IEEEVR)*. IEEE, 2015, pp. 25–30.

[18] W. R. Thurlow and P. S. Runge, "Effect of induced head movements on localization of direction of sounds," *The Journal of the Acoustical Society of America*, vol. 42, no. 2, pp. 480–488, 1967.

[19] M. Kato, H. Uematsu, M. Kashino, and T. Hirahara, "The effect of head motion on the accuracy of sound localization," *Acoustical science and technology*, vol. 24, no. 5, pp. 315–317, 2003.

VISUAL-AUDITORY VOLUME RENDERING OF SCALAR FIELDS

E. Malikova[1]*,V. Adzhiev*[1]*, O. Fryazinov*[1]*, A. Pasko*[2]

[1]Bournemouth University,UK
[2] Skolkovo Institute of Science and Technology, Moscow, Russia
emalikova@bournemouth.ac.uk,vadzhiev@bournemouth.ac.uk,
ofryazinov@bournemouth.ac.uk,apasko@bournemouth.ac.uk

ABSTRACT

This paper describes a novel approach to visual-auditory volume rendering of continuous scalar fields. The proposed method uses well-established similarities in light transfer and sound propagation modelling to extend the visual scalar field data analysis with auditory attributes. We address the visual perception limitations of existing volume rendering techniques and show that they can be handled by auditory analysis. In particular, we describe a practical application to demonstrate how the proposed approach may keep the researcher aware of the visual perception issues in colour mapping and help track and detect geometrical features and symmetry break, issues that are important in the context of interpretation of the physical phenomena.

1. INTRODUCTION

The results of a numerical simulation or experimental measurements are raw and complex scientific data that contains a lot of information. Thus the scientific data can be quite difficult to understand and analyse. One of the examples of such data representation is a scalar field.

Scalar fields are used in many research areas, where computer simulations or experimental studies are involved, such as computational chemistry, medical data analysis and physical phenomena studies. The scalar field can have either discrete or continuous representation. In this work, we consider a more general case of continuous scalar fields. Visualisation of continuous scalar fields, however, is not always straightforward, especially when a complex phenomenon is represented. The examples of such situations are the simultaneous analysis of several scalar fields with different field features and underlying processes; scalar fields after post-processing with image processing applied to 3D textures; application of various optical models in the visualisation pipeline. The main aim of those procedures is to highlight features of interest [1], enhance visual analysis quality [2] and handle image quality issues, arising due to limitations of scanning devices and human perception. Without these techniques, we may get a wrong insight on data, which fails the entire analysis process [3].

Visualisation of scalar fields usually employs Volume Rendering techniques as for the computer systems the scalar fields data is often converted to data volumes stored in the texture memory [4, 5]. In the Volume Rendering the enhancement of optical model is used to address the issues of visual analysis quality improvement [6, 7]. The most recently introduced techniques, such as multidimensional transfer functions, are relatively new and as visualisation tools are the areas of active research [8].

The conventional Volume Rendering techniques can fail as important details and features might be missed, especially when they are relatively small. Moreover, the problem becomes even more apparent for visualisation of the dynamic objects, as it becomes hard to track small, visually hard to distinguish scalar field feature changes. This problem arises in various application domains, where small local changes in the field surfaces should be detected in order to highlight areas of potential physical properties change (e.g., superconductor fields study considered in this work). Tracking of the small changes in the dynamic scalar fields can be solved with additional numerical methods, which makes the whole process even less efficient.

It is well proven that the sensory stimuli operate differently, and thus they can successfully complement each other in the analysis process. Sonification techniques [9] as an approach to data analysis via various sound characteristics proved to be quite effective for multivariate data [10].

A visual system is limited to the perception of a certain amount of colours and shades, can be overloaded and perturbed due to fatigue. The auditory system, on the contrary, can operate in the background mode and act as an early alarm tracking even small changes via sound wave parameters. In this work, we propose a general approach to the visual-auditory analysis of scalar fields by extending the Volume Rendering technique with additional auditory stimuli. The main aim of the approach is to address visual perception issues and enhance analysis quality.

The contributions of the research are:

1. We have proposed the general model for the representation of objects with optic and auditory properties.
2. We have explored similarities between light and sound propagation to suggest a visual-auditory rendering based on the well known ray-marching procedure.
3. We have considered situations when volume rendering is insufficient or does not allow for enhancing a small features analysis. The introduced novel approach can be used to address those problems.
4. To demonstrate how the introduced method works in practice, we have applied it to enhance the quality of visual analysis while detecting small changes in the symmetry of the superconductor field.

https://doi.org/10.21785/icad2019.004

5. We have implemented the basic prototype of the proposed approach to visual-audio rendering. We discuss the application limitations and possibilities of the proposed approach to visual-auditory analysis.

The structure of the paper is as follows. The "Related works" section gives an overview of the scalar fields analysis problem by Volume Rendering technique. The "Visual-auditory volume rendering" section introduces an approach to visual-auditory analysis and specifies the details of visual-auditory rendering and visual-auditory stimuli interpretation. The application case studies are presented in the "Experiments and case study" section. Future directions are presented in the "Conclusions" section.

2. RELATED WORKS

In this work, we concentrate on the Volume Rendering technique for scientific data visualisation. The research in this area emphasises the problems of visual analysis efficiency, perception and quality. A "good" visualisation is the one that leads to the automatic detection and extraction of the requested features and hides unnecessary details[6], [4], [5]. This task is particularly difficult for experimental data as it is obtained with scanning and measuring devices. The device limitations and scans quality issues inevitably arise during the process. The problem is addressed with additional procedures and techniques [11],[12].

Currently, the Direct Volume Rendering technique is used to address issues like enhancement of visual image quality, making images more perceivable and analysis more stable. The relatively new technique is Multidimensional Transfer Functions (TF) [13, 6]. The Multidimensional Transfer Functions design is closely related to techniques and approaches used in image processing [14, 15] in order to introduce a more enhanced optical model to facilitate the visual analysis.

However, the visual analysis has perception limitations that cannot be addressed solely by enhancement of the optical model. The introduction of the other sensory stimuli can significantly enhance the analysis process. The research in this area [10, 16] stresses the visual perception of temporal and spatial resolution limitations. On the contrary, the sound wave most perceptionally efficient parameters are 1000-4000Hz frequency and 0-160Db loudness. The use of sound is a well-known solution to track small changes, operate in background mode as an early alarm system and effectiveness for classification tasks [17].

The use of sound has been widely investigated since early 80-s [18, 17]. The fundamental works were published in 90-s [9] and theoretical research still continues [10]. The technique of data representation using various sound characteristics is called data sonification [19, 10]. The auditory perception brings the unique possibility to distinguish small variations in the parameters of the single sound wave and to compare sound waves. The sound analysis may be efficient for fixing a visual perception [20] or a haptic perception issues [21].

The enhancement of visual analysis of continuous scalar fields with auditory tools is a relatively new area. Some general possible auditory application directions considered in [22]. The problems of visualisation uncertainty [23] and medical data analysis [24] can be identified as areas that can most benefit from sonification. The works [23] and [24] particularly address research areas, where visual analysis of scalar fields may fail. There is an increase of interest to objects sonification and continuous data recently, in particular in augmented reality area [25] that takes advantage of human 3D spatial positioning of sound sources through hearing.

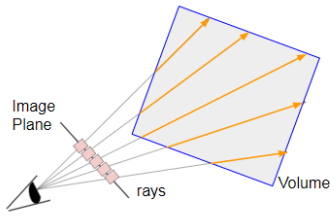


Figure 1: Backward or eye-tracing scheme in computer graphics

In this work, we introduce additional auditory sensory stimuli to address the problems arising in visual analysis. We take advantage of similarities of light and sound propagation to review the Volume Rendering technique and to extend it for a visual-auditory rendering case. We concentrate on the use of ray-tracing procedure application for computation of optical and auditory properties.

3. VISUAL-AUDITORY VOLUME RENDERING

3.1. Approach overview

Scalar field is a function $f(X) : X \to \Re, X \in \Re^n$, which associates any point in space with a scalar value. In computer graphics, scalar fields are often used to represent geometry in an implicit form. In our work, we discuss the generic scalar fields which can represent any type of scientific data. Scalar fields are usually represented in the computer systems as scalar values stored as multi-dimensional data volumes inside the texture memory and are conveniently visualised with Volume Rendering methods.

The core of the conventional Volume Rendering technique is an emission-absorption optical model. The Volume Rendering equation is derived from the rendering equation [26] that is a fundamental concept in computer graphics on how the general optical material properties can be described through physics process of light interaction with the object if the wave nature of light can be neglected [27].The conventional Volume Rendering equation 1 takes advantage of a simplified model of light interaction with an object, that considers only emission and absorption mechanisms [28].

$$I(D) = I_0 e^{-\int_{s_0}^{D} \tau(t)dt} + \int_{s_0}^{D} q(s) e^{-\int_{s}^{D} \tau(t)dt} ds \qquad (1)$$

where I is light traversing from entry point to volume s_0 to exit point towards the camera $s = D$ intensity,

$q(s)$ is a light contribution at point s, it other words term describing emission process,

$\tau(t)$ is used to describe light attenuation, when it reaches point D, in other words, term describing absorption process.

The concept of shooting rays is used to compute the final image that researchers see (see Fig. 1). The ray-casting or ray-marching volume technique [29] solves the equation 1 by approximating the light propagation and interaction with each of volume elements along ray path.

The ray-casting procedure operates a Transfer Function (TF) defining contribution of each volume element (see Fig. 2). In this

work, we consider the Front-to-Back Compositing scheme [30] as a numerical solution of equation 1 for each ray that is shoot from camera (see Fig. 1).

The core idea of our method is that the principles of light and sound propagation are very similar [31]. The acoustic rendering equation [32] is a time-dependent version of the same rendering equation [26]. In this work, we stress the following core similarities and differences in optical and acoustic properties modelling. First, the role of the propagation procedure that allows us to consider the ray-tracing technique as a core component of both models. The successful application of the ray-based model for sound propagation modelling demonstrates that to some extent the wave nature can be neglected. Second, the concept of radiance of the amount of energy transmitted as the result of interaction at a specific point can be used for both optical and acoustic properties modelling[32]. Finally, as follows from the acoustic rendering equation, the main difference between light and sound propagation modelling is that the aural perception of the time/passed distance dependency for the sound [32].

We take advantage of similarities in modelling optic and acoustic properties of objects due to the similar nature of propagation of light and sound and propose an acoustic model that describes the acoustic properties of the object. The model is based on a conventional Volume Rendering optic model and considers time-dependent emission and absorption processes of a sound wave propagation.

The basic concept of the proposed auditory model of sound wave generation as a result of an impulse propagation through the object is shared by both ray-based and wave-based approaches to sound modelling, like digital waveguide and the banded digital waveguide approaches to physically based sound synthesis [33]. These approaches establish relation between sound propagation and modal sound synthesis [33]. The sound wave is modelled as a result of a propagation process, while each activated mode depends on a sound propagation path in the vibrating object and final sound is a contribution of all modes computed for considered rays.

However, in the context of the proposed auditory transfer function, the sound propagation on the basis of the wave model will lead to the possible change of the perceived sound property, namely of the pitch. Thus, the further aural perception and interpretation of such auditory model can be difficult. For this reason the conceptual framework considers wave-based approaches, but in this work does not take direct advantage of them. We will consider the extension of the emission-absorption optical model as the proposed acoustic interaction model and concentrate mainly on the ray-based approach, although, as follows from the above discussions, some parallels to physically based sound synthesis can be made. We will also take advantage of established terms by introducing the "modal areas" that are activated by the travelling ray.

3.2. Object with optical and auditory properties

To keep the auditory model within the concept of scalar fields, we use the idea of a HyperVolume (HV) model [34] described with an equation:

$O = (G, A_o, A_s) : (f(X), S_o(X), S_s(X))$

In the HV model, the scalar field function is augmented with point attribute functions S defining optical A_o and auditory A_s properties. Thus, we effectively define a vector field, or vector-valued function whose first component f is responsible for the object geometry G and maps directly from the input scalar field.

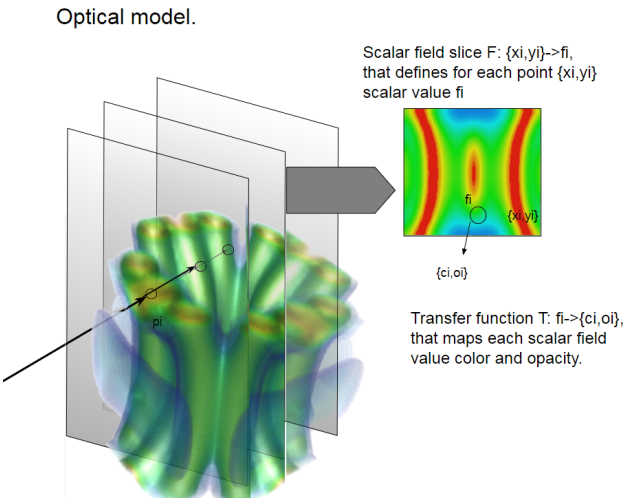


Figure 2: Volume Rendering optical model. Colour at point pi equals to the amount of light that has reached this point multiplied by emission of selected wavelengths perceived as colour. The final pixel colour is a sum of all colours in points pi along ray that is calculated with a ray-marching procedure. $S_o = T(F(x))$, where scalar field $F(X)$, T - transfer function.

Other components serve as the point attributes for visual and auditory properties. Auditory properties in a form of the generated sound wave are defined with $S_s(X)$ and the attribute S_o define the results of mapping to the optical properties such as colour and opacity. Note that in the general case attributes S are not scalars but vectors. For example, the colour information stored in S_t is normally represented as a four-component RGB value and opacity.

The optical properties of the model O can be rendered directly with a Volume Rendering technique, operating a ray-casting procedure. For the model O, the attribute function S_o is a result of the Volume Rendering transfer functions that will operate the scalar field normalisation procedure to perform the mapping. The process is schematically presented in Fig. 2.

We will introduce an auditory model and will describe how the final sound can be obtained with the ray-marching procedure similarly to conventional Volume Rendering technique.

3.3. Auditory model

For the introduced time-dependent auditory model we consider the traditional two parts of sound modelling [31] in terms of listener perception. The first step is considering how a wave propagates through the object represented by the scalar field; the second step is exploration how a resulting sound propagates through an environment and interacts with a listener thus enabling the perception of the sound spatial properties(e.g., sound source position).

For the first part, we define an auditory transfer function to obtain a time-dependent auditory attributes $S_s(X(t))$ of our HV model. The auditory transfer function is designed to be operated by ray-tracing procedure that automatically produces an output allowing for efficient judgement on the considered scalar field properties along the ray. An auditory rendering procedure becomes very much similar to optic rendering procedure in a conventional

Volume Rendering as they both consider physical processes of propagation.

We propose the researcher perceives and analyses the generated auditory properties in terms similar to how the final sound is formed in space as a result of an initial impulse (interaction). Below we describe the proposed auditory transfer function designed to address visual perception limitations as tracking of small changes in the scalar field are required.

For the second part, we consider the use of the pre-computed Head-related transfer function (HRTF) that convolves sounds generated by the object as it propagates to the left and right ears. HRTF is represented with a frequency domain of a head-related impulse response (HRIR). Application of HRTF suggests an extraction of HRIR coefficients and delays to compute sound convolution. A sound source coordinates are matched with HRTF coordinates and HRIR coefficients, and delays are interpolated via bilinear interpolation as the measurements are discrete [35]. The sound pressure for left and right ears H_L and H_R are defined as [35]:

$$H_L = \frac{P_L(r,\theta,\phi,f)}{P_0(r,f)}, H_R = \frac{P_R(r,\theta,\phi,f)}{P_0(r,f)} \quad (2)$$

where the sound source is defined with spherical coordinates (r,θ,ϕ), P_L and P_R are complex sound pressures at the entrance of left and right ears, and P_0 is a complex sound pressure at the centre of listeners head. The process is schematically presented at Fig. 3).

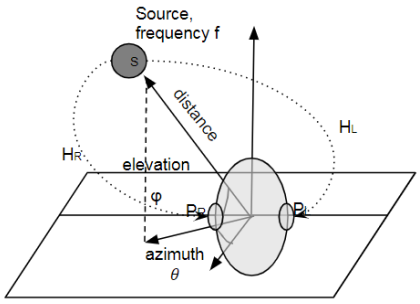


Figure 3: HRTF scheme

3.4. Auditory rendering for tracking small changes in the scalar field

3.4.1. Auditory transfer function

The main requirement to the conventional Volume Rendering transfer function [6] is to highlight features of interest via optical attributes, which normally are colour and opacity. Similarly, the auditory attributes in our model should allow the user to efficiently analyse the scalar field. Traditional auditory properties which allow for doing so are pitch, volume and sound's spatial properties. According to the modal synthesis approach [33], the sound produced by the object can be represented as a sum of the weighted modal modes N' (sound wave components with specified frequencies) extracted with Fourier Transform. However, a normal listener perceives only one or several frequencies as pitches from the entire complex sound and can interpret a sound wave in musical terms as a pitch or chord sequence. All the other components form the sound quality characteristics that allow us to aurally distinguish one musical instrument from another.

Let us consider the mapping $M : f- > w$ of scalar field values f to more pleasant "musical" sounds such as sequences of the specified musical pitches of frequencies w. In our approach, to specify these frequencies we are using MIDI (stands for Musical Instrument Digital Interface). In music, the MIDI format is widely used to formalise the sound representation, and the basic MIDI message tuple $(On/Off, MIDI_{Key}, MIDI_{KeyVelocity})$ can be used to find the wave duration, the frequency and the amplitude directly. Therefore, the auditory properties, which we store in the HV model, are mapped from MIDI message components as $M : f- > MIDI_{Key}$, where field $MIDI_{Key}$ represents frequencies and act like an auditory transfer function.

The $MIDI$ field splits scalar field to separate areas (Fig. 4) that produce a modal vibration, as the sound propagates through them. To roughly define them we introduce a term the separate object "modal area". Thus, to establish the mapping $M : f- > w$, we select a musical scale with degree numbers $0, ..N$ within the specified range N. Whilst small range scales are easier to perceive, a bigger range gives a trained listener more options for judging about small data changes. Our experiments showed that in most cases Cmaj of up to two-octave range is sufficient to auditory highlight areas the visual analysis might miss. The mapping, therefore, is described as follows:

1. To establish the mapping $f- > 0, ..N$, we calculate the scale degree $n_i \in 0, ..N$ for each scalar filed value $f(X)$ within the sub-range as $n_i = \lfloor \frac{f(X)}{\Delta d} \rfloor$, where $\Delta d = \frac{f_{max} - f_{min}}{N}$
2. $0, ..N- > MIDI_n$. The mapping for Cmaj scale of the defined range and the start key can easily be implemented on the basis of knowledge about the major scale structure of a combination of tones (T) and semi-tone (S) intervals between notes (TTSTTTS) [36].
3. The mapping $MIDI_{Key}- > w$ can be obtained with well known MIDI keynote to the frequency conversion equation.

The result of an acoustic transfer function is used by the ray-marching procedure in order to generate the sound, which is perceived by the listener. The idea is to generate the final sound from the initial sound impulse as it propagates along the ray path in a scalar field through time, activating the modes with specified acoustic parameters as described below.

3.4.2. Ray-casting procedure

From a physics point of view, the mapping of scalar field to optical properties defines how the single field point interacts with light[30]. The introduced auditory transfer function describes how a segment of the field of a certain length (a ray-traced modal area) interacts with sound impulse that propagates through the scalar field along a defined path. Thus, the acoustic model follows the same principles and uses a similar to the optical model definition with the discrete number of modal areas and distance-dependent ray-casting procedure employed.

The conventional volume rendering equation is solved with ray-marching procedure by Front-to-Back Compositing scheme

[30]:

$$\begin{aligned} \hat{C}_i &= C_i(1 - \hat{A}_{i-1}) + \hat{C}_{i-1} \\ \hat{A}_i &= A_i(1 - \hat{A}_{i-1})) + \hat{A}_{i-1} \end{aligned} \tag{3}$$

where $\hat{C}_{i-1}$ and $\hat{A}_{i-1}$ are the colour and opacity accumulated on previous step, and C_i and A_i are values that the transfer function returns for the current pixel.

Similarly, as the sound impulse propagates along the ray, the total impulse for the current modal area will be the sum of the accumulated by all previous modes impulse and of the value returned by the auditory transfer function for the current mode. As sound propagation is distance dependent, the modes are activated in time and attenuated depending on the distance passed. We adjust attenuation parameters for each mode in such a way that the produced modal oscillations do not significantly overlap. Thus it is easier to judge on the field values through the frequency of the mode. Taking advantage of a sampled digital wave representation, we represent the entire ray propagation path with a sampled buffer. As the distance for the current mode is computed, we write it into a buffer, regularly updating an output signal, while the ray-casting procedure operates the scalar field acoustic transfer function along a path.

We summarise the process in the following distance dependent equation for a sampled continuous sound wave:

$$I(d) = \sum_{i=0}^{N} A_i * M_i * \begin{cases} 0, & \text{if } D_i - d > 0 \\ e^{dur_i * (D_i - d)}, & \text{otherwise} \end{cases}$$

where d denotes the distance the ray passes, which is proportional to the time parameter of a spreading wave impulse; N is a total number of modes with the times they are activated/intersected as the ray travels; dur_i is a mode duration that can be computed as difference in D_i of the current mode and D_k of the next mode $D_k = \min_{j=0}^{j=n}(D_j) > D_i$. M_i is a mode described in the form of $M_i = sin(w_i * (d - D_i))$, where D_i is a distance along the ray before intersecting a patch area of i mode, w_i is a mode frequency for the MIDI field, which is obtained from the scalar field as described above, and finally A_i denotes the initial amplitude of the mode, which describes the energy the impulse transmits to the mode. As the ray travels, the impulse attenuates due to absorption and is described with equation $A(D_i) = A_{ini} * exp^{-m * D_i}$.

Similarly to the computation of a pixel colour in Volume Rendering, the final acoustic impulse is a weighted sum of all the modes activated along the ray. The ray-marching procedure uses the modal frequency w_i and the initial amplitude A_i in a similar way to the optical model: the modal frequency is mapped from the source scalar field (it acts similarly to the colour attribute in the optical model), and the amplitude acts like opacity in the optical model. Consequently the duration dur_i appears due to the time-dependent nature of sound and is perceived through a delay of activation. The general similarities between optical and auditory models are demonstrated in Fig. 2 and Fig. 4. The similarities of the conventional optical model and the proposed auditory model make the last one convenient to address the issues, where the optical model can fail such as a colour mapping evaluation and tracking small changes in a scalar field.

At the final step, we apply the HRTF convolution to the generated sound wave to allow the listener to track the scanning ray path

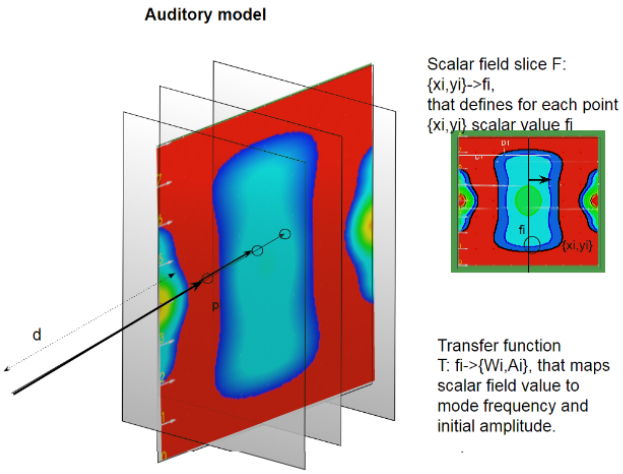


Figure 4: Introduced auditory model. The sound wave at point $p_i = A_i * sin(W_i * d)$ as the impulse has propagated for distance d. W_i controls emission part - frequencies of generated sound; A_i controls absorption part - the amount of initial impulse energy that has reached the point. The final sound wave is a sum of all waves generated along the ray path and is calculated with a ray-marching procedure.

in space by defining the sound source position in the time equal to the current scanning position along the ray.

The introduced scanning ray procedure is a basis for the auditory analysis of the scalar field. As an acoustic impulse propagates, the user evaluates the scalar field features through the time-dependent changes in the pitch. Below we will consider some case studies that demonstrate how the introduced approach to an auditory analysis can enhance the quality of the visual analysis.

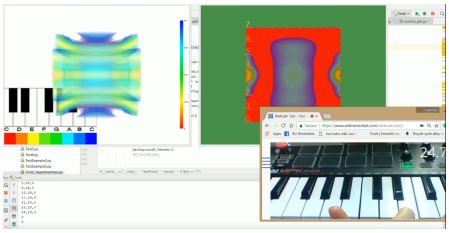
3.4.3. Interactive procedures

Additionally, we describe an interactive data manipulation procedures on the basis of auditory information. The interaction is based on defining "musical queries" that are the sequences of notes. Similar to rendering, the interaction procedures operate with MIDI field. For the musical query simplification, we neglect all the message components except the key number that defines a note.

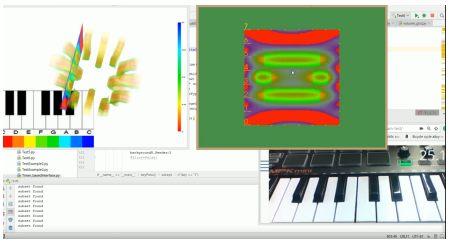
We demonstrate a simple example (Fig. 5 a,b and the accompanying video [37]). We use the MIDI keyboard for a fast extraction of a particular part of a field that demonstrates the musical pattern of interest (the scan is taken along the y-axis). The music pattern is defined with MIDI keyboard (Fig. 5 a). We can search and highlight the scalar field areas demonstrating the same pitch pattern (Fig. 5 b) via the defined field $MIDI_n$ and thus quickly define the area of interest in a scalar field. The technique may be used for the search for a smooth/fast gradient changes detection as well.

4. A CASE STUDY OF SUPERCONDUCTOR'S FIELD

Our approach to visual-auditory analysis has been implemented with C++/Python with using of VTK [38], OpenAL [39] and OpenCV [40] libraries. The interface procedures for the described



(a)



(b)

Figure 5: Interaction via Midi keyboard a) Music pattern definition with Midi Keyboard b) Highlighting of scalar field areas corresponding to defined pitch pattern

example of the physical applications were implemented in Python.

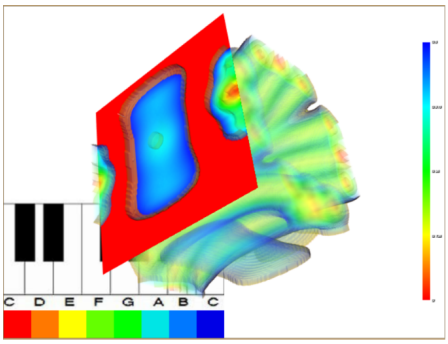
As a case study for our approach, we consider a superconductor field analysis. One of the particular geometric features of this field is so-called Abrikosov vortex [41] (Fig. 6 a), Fig. 7), which represents isosurfaces of the supercurrents as they circulate. The analysis of the vortex arrangements is applied to carry out judgements on the material properties and therefore an extra attention should be paid to the analysis of the superconductor scalar field. Below we discuss how our visual-auditory approach can be applied in this situation.

4.1. Colour mapping quality

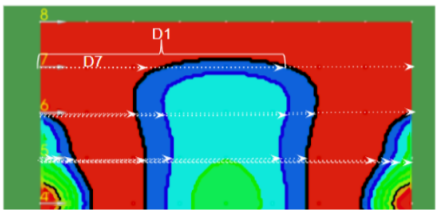
In the general case, the scalar field contains a relatively large range of values. In the visual analysis, those values are interpreted as colours, which are not always easy to distinguish because of human colour perception. However, with a combination of a visual analysis with the auditory analysis we can highlight the regions of interest (ROI) by adjusting the visual mapping parameters (see Fig.6a and the accompanying video [42]).

An auditory approach can be potentially effective for very complex scalar fields with a big range of values as the mapping data one-, two-, or three-octave Cmaj. This range allows us to distinguish the field changes easier than with just a visual analysis and to track the colour mapping quality.

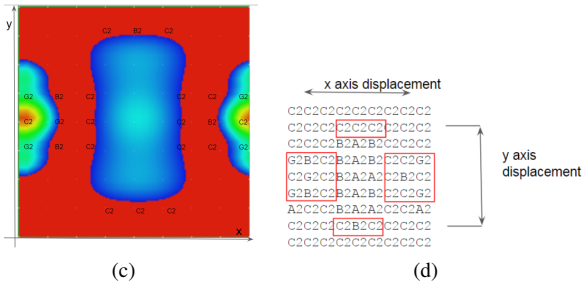
In order to be able to successfully distinguish the musical degrees within the scale, a listener should always keep in mind the first degree sound, which is the tonic, and compare all the other sounds to it. To help with this, the outside domain values are mapped into the tonic. Although such type of analysis is very similar to the procedures that are used by trained musicians and researchers, the entire procedure can be difficult for the untrained researcher.



(a)



(b)



(c) (d)

Figure 6: Examples of visual-auditory exploration of the second type superconductor field: a) The input scalar field with a 2D slice we analyse; the correspondent modal frequencies are denoted by colours and are presented in the form of a piano scale; b) MIDI field tracing with parallel ray casting. The modal areas are represented with the colours as in a); c) The field slice with the corresponding sound matrix representation (d) highlighting the displacements in the field.

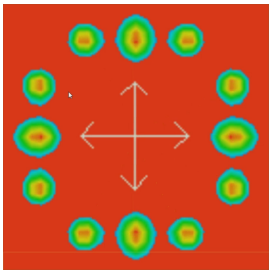


Figure 7: Field 2D slice with the highlighted vortex areas (adjustment of field ROI) and suggested directions for the auditory tracking of the field symmetrical features.

4.2. Symmetry and geometrical features changes tracking

Another important application of our visual-auditory approach in the analysis of scalar fields for superconductors is symmetry and geometrical changes/features detection. For example, researchers in superconductors explore the issues of symmetry breaking and distortion in the vortex lattice or in the shape of the vortex itself [41]. The dynamic changes can be observed and analysed visually with scanning microscopy or with computer simulations. In our method, the small changes in symmetry and vortex geometry can easily be detected with the help of the sound (Fig. 6 c,d). In this example, we shoot two rays to the field areas that are supposed to be equal. The small differences in the field symmetry will sound like a non-perfect unison musical interval that can be aurally distinguished by the untrained researcher.

The main disadvantage of the colour mapping quality evaluation technique that was presented above is a finite sampling resolution of rays casting. The balance should be kept between the details users want to track and the speed of analysis. As a result of a small resolution, the analysis can take a relatively small time, but important features can be missed. One of the possible solutions is to specify ROI for an analysis and in combination with the guidance procedures. Potential ROIs and scanning directions can be identified automatically with the image processing techniques as it is done in the example shown in Fig. 6a and in the accompanying video [43]. For the superconductor analysis, this can be used to track the symmetrical features along the specified ray for a pair of the vortices or examine an area around a vortex to track its distortion.

5. CONCLUSION AND FURTHER RESEARCH

The main result of this research work is a unified theoretical and practical approach to visual-auditory Volume Rendering. The framework takes the scalar field as input and uses the ray-casting procedure that operates on some multisensory transfer function, to render it to the visual-auditory stimuli.

To introduce such an approach, we have taken the following steps. We have considered the colour mapping quality and tracking small scalar field features like symmetry break as two areas concerned with visual perception limitations. We have treated the scalar field and its domain as a representation of an object with optical and auditory properties that should be analysed and have suggested a HyperVolume model for such object representation.

We have discussed the similarities between visual-auditory properties modelling in order to address the problem of an efficient combination of visual-auditory stimuli. On base of those similarities we have proposed the visual-auditory mapping for two problems arising in area of visual analysis: the colour mapping quality and stacking small changes in symmetry of scalar field problems. The proposed mapping allows us to take most of the both sensory stimuli perceptional advantages and balance their disadvantages in the analysis process as they consider the models most close to physically based ones for optical and auditory properties modelling.

The light transfer based optical model or the model of the sound produced as a result of initial impulse propagation can provide not only a high level of realism but an intuitive way of the input parameters control in order to obtain the most realistic, desired result. These models, however, can be computationally expensive. For visualisation and computation purposes certain assumptions and adjustments are made to simplify the modelling in conventional Volume Rendering and the acoustic modelling.

Similarly to light, the sound propagation mechanism defines the perceivable characteristics: opacity and colour vs volume and pitch. Thus, the acoustic properties can also be rendered on the basis of the ray-marching procedure. We have considered the light and sound propagation similarities to introduce an approach to visual-auditory Volume Rendering. We have demonstrated how the auditory representation can be complementary to the visual one in gaining insight into the continuous scalar field in a case study of analysis of the scalar fields of superconductors. The proposed approach to visual-auditory analysis has been applied to some particular case studies. That has allowed us to judge on its advantages, limitations and possible adjustments.

Our experiments show possible limitations of the introduced method. As demonstrated in the colour mapping quality example, it may require some auditory skills from the user. We suggest the use of auditory analysis for the symmetrical data regions that allows for introducing the sound mappings which are easier to interpret.

Another limitation of our method is the sampling resolution. We overcome it through specifying the ROI for analysis and guidance procedures. This can be done automatically with the image processing (as demonstrated in the symmetry tracking example) or machine learning techniques, but currently out of the scope of this research. Finally, in this research, we have considered relatively simple scalar fields. However, the technique can be applied to more complex cases in such research areas as medical image analysis and molecular fields studies. Mapping the field and its additional derived features (gradient, curvature) to sound and an introduction of more complex mappings to music entities such as chords that are based on several rays shooting, are the areas of future research.

6. REFERENCES

[1] Y. Jung, J. Kim, A. Kumar, D. Feng, and M. Fulham, "Feature of interest-based direct volume rendering using contextual saliency-driven ray profile analysis," *Computer Graphics Forum*, vol. 37, no. 6, pp. 5–19, 2018.

[2] S. Djurcilov, K. Kim, P. Lermusiaux, and A. Pang, "Volume rendering data with uncertainty information," in *Data Visualization 2001*, D. S. Ebert, J. M. Favre, and R. Peikert, Eds. Vienna: Springer Vienna, 2001, pp. 243–252.

[3] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual data mining," S. J. Simoff, M. H. Böhlen, and A. Mazeika, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. Visual Analytics: Scope and Challenges, pp. 76–90.

[4] M. Ament, "Thesis: Computational visualization of scalar fields," 2014, accessed on 2017-11-12.

[5] P. Ljung, "Efficient methods for direct volume rendering of large data sets," Ph.D. dissertation, Linkoping University, Visual Information Technology and Applications, 2006.

[6] J. Kniss, G. Kindlmann, and C. Hansen, "Multidimensional transfer functions for interactive volume rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 3, pp. 270–285, July 2002.

[7] C. Johnson, "Top scientific visualization research problems," *IEEE Comput. Graph. Appl.*, vol. 24, no. 4, pp. 13–17, July 2004.

[8] L. Patric, K. Jens, G. Eduard, H. Markus, H. C. D., and Y. Anders, "State of the art in transfer functions for direct volume rendering," *Computer Graphics Forum*, vol. 35, no. 3, pp. 669–691, 2016.

[9] G. Kramer, *Auditory display: sonification, audification, and auditory interfaces*, ser. Proceedings ; vol.18. Reading, Mass Wokingham: Addison-Wesley, 1994.

[10] T. Hermann, A. Hunt, and J. G. Neuhoff, Eds., *The Sonification Handbook*. Berlin, Germany: Logos Publishing House, 2011. [Online]. Available: http://sonification.de/handbook

[11] Y. Allusse, P. Horain, A. Agarwal, and C. Saipriyadarshan, "Gpucv: A gpu-accelerated framework for image processing and computer vision," vol. 5359, 12 2008, pp. 430–439.

[12] M. Haidacher, S. Bruckner, A. Kanitsar, and M. E. Gröller, "Information-based transfer functions for multimodal visualization," in *Proceedings of the First Eurographics Conference on Visual Computing for Biomedicine*, ser. EG VCBM'08. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2008, pp. 101–108.

[13] M. Falk, I. Hotz, P. Ljung, D. Treanor, A. Ynnerman, and C. Lundstrom, "Transfer function design toolbox for full-color volume datasets," in *2017 IEEE Pacific Visualization Symposium (PacificVis)*, April 2017, pp. 171–179.

[14] S. Park and C. L. Bajaj, "Multi-dimensional transfer function design for scientific visualization," in *ICVGIP*, 2004.

[15] S. Fang, T. Biddlecome, and M. Tuceryan, "Image-based transfer function design for data exploration in volume visualization," in *Proceedings Visualization '98 (Cat. No.98CB36276)*, Oct 1998, pp. 319–326.

[16] A. El Saddik, M. Orozco, M. Eid, and J. Cha, *Haptics: General Principles*, 1st ed. Springer Publishing Company, Incorporated, 08 2011, pp. 1–20.

[17] S. Bly, "Presenting information in sound," in *Proceedings of the CHI '82 Conference on Human Factors in Computer Systems*. ACM, 1982, pp. 371–375.

[18] E. Yeung, "Pattern recognition by audio representation of multivariate analytical data," *Analytical Chemistry*, vol. 52, no. 7, pp. 1120–1123, 1980.

[19] H. Kaper, E. Wiebel, and S. Tipei, "Data sonification and sound visualization," in *Computing in Science and Engineering*, vol. 1, no. 4, 1999, pp. 48–58.

[20] L. Gionfrida, A. Roginska, J. Keary, H. Mohanraj, and K. P. Friedman, "The triple tone sonification method to enhance the diagnosis of alzheimer's dementia," in *The 22nd International Conference on Auditory Display (ICAD)*, 2016.

[21] H. Roodaki, N. Navab, A. Eslami, C. Stapleton, and N. Navab, "Sonifeye: Sonification of visual information using physical modeling sound synthesis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 11, pp. 2366–2371, Nov. 2017.

[22] R. Minghim and A. R. Forrest, "An illustrated analysis of sonification for scientific visualisation," in *Proceedings Visualization '95*, Oct 1995, pp. 110–117.

[23] S. K. Lodha, J. Beahan, T. Heppe, A. J. Joseph, and B. Zane-Ulman, "Muse : A musical data sonification toolkit," 1997.

[24] L. Gionfrida and A. Roginska, "A novel sonification approach to support the diagnosis of alzheimer's dementia," *Frontiers in Neurology*, vol. 8, p. 647, 2017.

[25] F. Ribeiro, D. Florłncio, P. A. Chou, and Z. Zhang, "Auditory augmented reality: Object sonification for the visually impaired," in *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, Sept 2012, pp. 319–324.

[26] J. T. Kajiya, "The rendering equation," *SIGGRAPH Comput. Graph.*, vol. 20, no. 4, pp. 143–150, Aug. 1986.

[27] N. Max, "Optical models for direct volume rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 2, pp. 99–108, June 1995.

[28] M. Hadwiger, J. M. Kniss, C. Rezk-salama, D. Weiskopf, and K. Engel, *Real-time Volume Graphics*. Natick, MA, USA: A. K. Peters, Ltd., 2006.

[29] S. Roettger, S. Guthe, D. Weiskopf, T. Ertl, and W. Strasser, "Smart hardware-accelerated volume rendering," in *Proceedings of the Symposium on Data Visualisation 2003*, ser. VISSYM '03. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2003, pp. 231–238.

[30] S. G. Matt Pharr, *Ambient Occlusion*. Pearson Higher Education, 2004.

[31] T. Takala and J. Hahn, "Sound rendering," in *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '92. New York, NY, USA: ACM, 1992, pp. 211–220.

[32] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja, "The room acoustic rendering equation," vol. 122, p. 1624, 10 2007.

[33] P. R. Cook, *Real Sound Synthesis for Interactive Applications*. Natick, MA, USA: A. K. Peters, Ltd., 2002.

[34] A. Pasko, V. Adzhiev, B. Schmitt, and C. Schlick, "Constructive hypervolume modeling," *Graph. Models*, vol. 63, no. 6, pp. 413–442, Nov. 2001.

[35] R. Wu and G. Yu, "Improvements in hrtf dataset of 3d game audio application," in *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, July 2016, pp. 185–190.

[36] M. Hewitt, *Music Theory for Computer Musicians*. Course Technology, CENGAGE Learning, 2008.

[37] "Midi keyboard interaction," 2019, password icad2019. [Online]. Available: https://vimeo.com/323545930

[38] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit–An Object-Oriented Approach To 3D Graphics*, 4th ed. Kitware, Inc., 2006.

[39] "Openal programmers guide," https://www.openal.org/documentation/OpenAL_Programmers_Guide.pdf, 2007.

[40] "Opencv," http://opencv.org.

[41] H. Suderow, I. Guillamon, J. G. Rodrigo, and S. Vieira, "Imaging superconducting vortex cores and lattices with a scanning tunneling microscope," *Superconductor Science and Technology*, vol. 27, no. 6, p. 063001, 2014.

[42] "Colour quality check," 2019, password icad2019. [Online]. Available: https://vimeo.com/323547646

[43] "Scalar field symmetry analysis," 2019, password icad2019. [Online]. Available: https://vimeo.com/323547659

AUDITORY DISPLAYS TO FACILITATE OBJECT TARGETING IN 3D SPACE

Keenan R. May, Briana Sobel, Jeff Wilson, and Bruce N. Walker

Georgia Institute of Technology
Atlanta, 30332
United States

{kmay,bsobel3}@gatech.edu, jeff@imtc.gatech.edu, bruce.walker@psych.gatech.edu

ABSTRACT

In both extreme and everyday situations, humans need to find nearby objects that cannot be located visually. In such situations, auditory display technology could be used to display information supporting object targeting. Unfortunately, spatial audio inadequately conveys sound source elevation, which is crucial for locating objects in 3D space. To address this, three auditory display concepts were developed and evaluated in the context of finding objects within a virtual room, in either low or no visibility conditions: (1) a one-time height-denoting "area cue," (2) ongoing "proximity feedback," or (3) both. All three led to improvements in performance and subjective workload compared to no sound. Displays (2) and (3) led to the largest improvements. This pattern was smaller, but still present, when visibility was low, compared to no visibility. These results indicate that persons who need to locate nearby objects in limited visibility conditions could benefit from the types of auditory displays considered here.

1. INTRODUCTION

There are a variety of situations in which humans need to navigate spaces with limited visual input. Auditory guidance systems, such as purpose-built navigation systems for visually impaired persons [1, 2, 3], or consumer navigation software, have tended to focus on guiding a person to locations of interest on a two-dimensional plane. However, supporting 2D navigation is only part of the solution. Many occupations and everyday tasks involve targeting nearby objects in 3D with limited visual input. For example, in first responder scenarios, personnel may need to quickly locate task-critical objects which are obscured by smoke or debris. Similarly, persons operating underwater or in other unique environments with limited visibility may need to locate tools or machinery. People with visual impairment must solve this problem to carry out everyday tasks. As visual-focused VR/MR (Virtual/Mixed Reality) systems become increasingly common and capable of operation in varied situations, research into the ability of auditory displays to assist with such tasks is needed.

Unassisted, targeting objects can be cumbersome without the use of vision. Searching a 3D space without full quality visual input can take a great deal of time, and be a frustrating experience. This type of task can be divided into two components: determining/recalling the right area to search, and targeting the object itself.

Each of these task components could be supported by different types of information.

First, a person needs to know the general area within which a nearby object is likely to be located. For example, a firefighter might need to locate a control panel, and knows that these are typically mounted roughly at chest height. This component of the task can be considered a knowledge problem as much as a perceptual-motor problem. Information supporting the selection of the correct search area could be retrieved from a person's memory, or an MR system could communicate target information pulled from an object database [4] or inferred via machine vision.

After deciding on the correct search area, a person must then accurately move a hand or tool to their target. If high fidelity visual input is available, a visual search may be conducted to precisely locate the target, followed by a reaching motion that is guided by a visuo-motor feedback loop [5]. However, if sufficient visual input is not available, making precise motor movements to a specific location can be difficult, even if that location is known and serial tactile search is not required. This task component can be considered a perceptual-motor problem as well as a knowledge problem. Interventions might utilize machine vision or wearable sensors to provide precise nonvisual guidance that would assist the user in moving all the way to the target, effectively creating an 'audio-motor' feedback loop.

1.1. The Sound of Space

Sound can be used both to convey 3D location information and to guide movement. However, humans tend to be relatively poor at perceiving the elevation of sound sources. Planar localization can utilize multiple types of information derived from binaural disparities [6], but elevation perception must rely on subtle spectral information derived from the way sounds are occluded by the head, ears, and shoulders, depending their direction of origin.

For virtual sound sources rendered using spatial audio, this inherent difficulty is compounded by the fact that simulating spectral information with high fidelity is more difficult than simulating binaural disparities. Spectral changes can be synthesized using Head-Related Transfer Functions (HRTFs) [7]. HRTFs can be effective if customized to reflect the geometry of an individual's pinnae and head/shoulders, but this is rarely feasible. Generalized HRTFs, while functional, are often not effective enough for a listener to consistently resolve elevation [8]. As such, relying solely on spatial audio effects to represent the position of a target in 3D space is unlikely to be effective.

Some systems have instead utilized Text-To-Speech (TTS) to describe the location of nearby objects. Thakoor et al. [9] tested

https://doi.org/10.21785/icad2019.008

a system that provided TTS denoting the presence of objects recognized by a mobile camera in one of nine areas in front of the user (e.g. "upper right"). May et al. [3] suggested that brief TTS description of object manipulation information (e.g. "trash can, button on lid") could be appropriate in some cases. A system developed by Doush et al. [10] assisted participants with blindness in grasping a specific library book via TTS description.

However, TTS description of object position can be relatively slow and cumbersome. It is also inappropriate in the myriad of situations in which a person's auditory environment is not conducive to TTS comprehension, or, conversely, in those in which a person's capacity to comprehend incoming speech should not be disrupted by TTS. In this study, we instead consider two approaches to utilizing *nonspeech* audio to either (1) quickly convey initial search-limiting location information, or (2) continuously and precisely guide motor movements.

1.1.1. Area Cueing Approach

One form a nonspeech targeting display could take is a discrete, informational *area cue* that informs the user in which area to search for the target object. Such a system could be implemented using information retrieved from a database about expected object locations, or in response to one-time machine-vision recognition of a target object. Systems of this nature could reduce target acquisition times by reducing the space that must be searched. However, they would not assist with the second stage of targeting in which the object must be precisely located and targeted.

Several area cueing systems have been considered in prior work. Chincha and Tian [11] developed a system in which users issued voice commands to initiate machine-vision search for target objects. If the object was in the camera's field of view, a sound confirmed its presence. Schauerte et al. [12] developed a machine-vision-based "lost object finding" system. That system sonified objects within the upper camera-viewable area with higher pitched tones, and lower-area objects with lower pitched tones. Tempo was mapped to object location confidence, and left-right location was represented through sound panning. Users gave the system generally positive ratings.

The effectiveness of an area cue may depend on its ability to swiftly and correctly communicate spatial information, to allow a user to immediately begin moving their extremity toward the target area without waiting to interpret more elaborate TTS or nonspeech displays. As such, choosing sounds that match expectations is crucial to optimizing this information transfer. It has consistently been found that higher pitched sounds tend to be associated with more highly elevated objects, and that lower pitched sounds tend to be associated with lower objects [13,14,15]. This pitch-elevation mapping reflects a statistical regularity of acoustic scenes [16, 17]. Thus, for the area cue sonification evaluated in this study, cue pitch was used to quickly communicate the elevation of the area in which the target resided.

1.1.2. Proximity Feedback Approach

Alternatively, a system could guide the entire process of object targeting by displaying an ongoing sonification of the user's hand position relative to the target. This would allow the user to target the object in 3D space solely through the sonification. While a system of this nature could represent relative position in three dimensions, in this study we considered a simpler, unidimensional display that provided continuous *proximity feedback*. The proximity feedback paradigm is similar to the real-time sonification of human movement, which has been shown to be effective for athletes and others endeavoring to carry out complex, precise movements, even when visual feedback was also available [18]. Unlike area cueing, proximity feedback supports the entire targeting task. However, it could also become distracting in environments with some visibility, and has significant technical requirements such as wearable sensors or cameras.

Displaying proximity feedback entails representing a dynamically changing variable: the current distance from the user's hand to the target. Higher pitch and tempo tend to be conceptually associated with closer proximity, as well as the related property of urgency [19, 20]. As such, in the proximity feedback design tested in this study, pitch and tempo communicated the proximity of the participant's hand to the target as it moved about in 3D space.

1.2. Current Study

The goal of this study was to investigate the effectiveness of area cues and proximity feedback in facilitating object targeting in local space. Participants were asked to walk around a virtual kitchen (Figure 1), and physically reach for target objects, with assistance from either an area cueing display, a proximity feedback display, both displays at once, or without assistance, in either a low visibility or a no visibility environment.

2. METHOD

2.1. Participants

There were 40 participants, with a mean age of 21 (*SD* = 3.23). 27 were male, 10 were female, and 3 elected not to specify. All were undergraduates at a technical university in the southeast United States. Participants reported normal/corrected vision and hearing, and had sufficient mobility/ dexterity to complete study tasks.

2.2. Materials

2.2.1. Virtual Environment

The experiment took place in a virtual environment created in Unity[1]. SteamVR[2] was used to support an HTC Vive VR system. The Unity scene ran on a control computer, which streamed video and audio to the Vive head-mounted display, as well as haptics to a handheld controller. This controller was also used to track the participant's hand position, and accept button-press responses from the participant. Audio was spatialized using the Steam Audio Unity asset, which provides real-time blended HRTF and acoustic simulation effects[3]. The software automatically recorded performance data. The rendered environment consisted of a kitchen-like room approximately 3×3 meters in size (Figure 1). There were two drawer-countertop-cupboard "stacks" along each of the four walls, making for a total of eight possible 2D locations.

Within each of the eight kitchen stacks, a target could exist within three elevation areas: low (in one of the drawers), middle (on the counter), or high (on a shelf within the cabinet, see Figure 2). Each of these areas was populated with 2–5 distractor objects near the target. Distractor objects were plates, coffee mugs, bowls,

[1] https://unity3d.com/

[2] https://steamcommunity.com/steamvr

[3] https://valvesoftware.github.io/steam-audio/

Figure 1: Virtual study environment.

glasses, and white cylinders. At the start of each trial, one of the white cylinders was replaced by a white capsule (Figure 2), which was the target object. Thus, the distractors were all visually similar to the target, to the point where participants in the low visibility conditions would need to move their head close to the objects to tell if the target was present in that area, and/or which object was the target. While participants could complete the task in this way, they could also elect to use the auditory displays to determine the target object's general location or guide targeting.

Figure 2: A kitchen stack, with the target (capsule, center) and distractors (plates, glasses, bowls, mugs, and cylinders).

2.2.2. *Auditory and Haptic Displays*

The 2D navigation beacon was a tone that was spatialized to "point" in the direction of the target stack. Its tempo increased as the participant approached the target stack, similar to [1].

The area cue was a brief sound played just after the participant entered the capture radius of the target stack. One of three variations was played depending on whether the target was in the middle elevation area (countertop), the high elevation area (cupboard) or the low elevation area (drawers).

The area cue was designed to strike a balance between clarity, brevity, and appropriate continuity with the 2D navigation experience. As such, each cue variant was constructed as a composite of several copies of the 2D beacon sound. Some of these copies were pitch shifted up or down, with the original 2D navigation sound always included. This produced a "chord," including the 2D beacon sound as the highest, middle, or lowest comprising note. For the middle elevation area cue, the 2D beacon sound was played alongside components both one octave higher and one octave lower. For the high area cue, components were added that were pitch-shifted upward by up to two octaves. For the low area cue, components were added that were pitch-shifted down by up to two octaves. The higher or lower pitched components faded in gradually over the course of a half second, creating a transition between the 2D beacon and area cue.

The proximity feedback was implemented as a repeating tone whose pitch and tempo changed depending on the proximity of the hand to the target. At maximum range, the tone played approximately once a second, and at minimum range it played approximately 10 times per second, and was one octave higher in pitch. Thus, increasing proximity was displayed via rising pitch and increasing tempo.

In the conditions with both the area cue and the proximity feedback, the area cue played once upon capture radius entry, and then the proximity feedback began playing normally. In order to simulate the ability of a person to search for an object by feeling object contours, the Vive's haptic feedback capabilities were utilized. When the handheld controller (Figure 3) was inside an object, the participant felt a continuous vibration. This vibration was given one of three strengths, depending on the type of virtual object that the participant's hand was inside of.

If the participant's hand was inside of a wall or kitchen structure such as a cabinet or drawer, they felt a weak vibration. If it was inside of a distractor object, they felt a medium vibration. Finally, if the participant's hand was inside of the target object, they felt a strong vibration. Vibration strengths were different enough to be clearly discriminable to a person with typical tactile acuity. In the No Visibility + No Sound condition, participants relied entirely on this haptic information.

The two visibility levels were created using Unity post-processing effects. In conditions with no visibility, post-processing was activated to make the scene completely dark. However, participants were able to see a blue box representing the floor, and a blue wireframe representing the virtual safety boundary. In low visibility conditions, a depth of field effect was applied in order to simulate generic limited visibility conditions. This effect caused objects to appear too blurry for a viewer to resolve precise form at most ranges. From a distance, participants could see the contours of the cabinets, and perceive that objects were present, but could not discriminate between targets and distractors without leaning in closer. Objects only resolved completely when viewed within a distance of approximately 15cm. Instead of leaning closer, which was physically effortful, participants were also able to utilize haptic feedback, or the auditory targeting displays, or could repeatedly guess.

2.3. Procedure

Upon consenting to participate, participants were fitted with the virtual reality headset and instructed in the task. Participants first practiced completing the task in a full visibility training mode. Each condition consisted of a set of object targeting trials. After each condition, participants were given an iPad, which they used to complete the NASA TLX, which assesses subjective workload associated with a task [21]. After completing all eight conditions, participants filled out a demographic questionnaire.

Each trial consisted of two stages. First, the participant used the 2D auditory beacon to walk to the kitchen stack that contained

the target. This procedure was included to increase the validity of the targeting task, and the 'virtual room' paradigm. Upon entering the 0.75-meter capture radius of the target stack, the 2D navigation beacon ceased.

During the second stage, the participant was instructed to find the target object as quickly and accurately as possible, using the different 3D assistance sounds (Figure 3). Doing this required moving the handheld controller, so that it was within the target object, and depressing the trigger on the controller to simulate grasping the target. The three sound types provided different forms of assistance during this second stage of each trial.

Figure 3: Tracked space and participant view during full visibility training.

Typically, humans make goal-directed movements in two parts. First, a large, rapid movement is undertaken that often falls short of the target. Second, after a moment of information uptake, a smaller and slower movement is undertaken to refine the limb position and reach the target [22, 23]. In this study, if the area cue was present, participants could first make a rapid, imprecise movement into the vicinity of the countertop, cupboard, or drawers, as specified by the area cue. Whether or not they heard the area cue, participants ultimately had to determine which of the objects was in fact the target, and guide the controller precisely to it. The proximity feedback assisted with this by providing a continuous sonification of the controller's distance from the target as the controller moved.

In the No Visibility + Area Cue and No Visibility + No Sound conditions, the nature of the targeting task was qualitatively different. Because visibility was zero, participants needed to use the haptic information to determine the layout of the stack and/or to disambiguate targets from distractors. During pilot testing, participants were capable of completing the task in these two conditions, but found it frustrating and time consuming. In response, a 'timeout' procedure was implemented. If a participant took over a minute to complete a trial, the system moved on to the next trial and recorded a 'timeout.' Data were not analyzed for these timed-out trials.

Upon pulling the controller's trigger while it occupied the same virtual space as the target, participants heard a confirmation sound and the next trial began. In the case of a timeout, the next trial began without the confirmation sound.

The target was placed in each of the 8 stacks, 3 times (one each for low, medium or high areas), for a total of 24 trials per condition. The order of trials was randomized. To avoid confusion, participants never had to navigate to the same stack twice in a row, nor to either of the immediately adjacent stacks.

2.4. Experiment Design

There were two independent variables, Sound Type and Visibility Level. Visibility Level could be either no visibility or low visibility (Table 1). Sound Type could be either no sound, area cue, proximity feedback, or the area cue with subsequent proximity feedback (AC+PF). Each participant experienced all of the resulting eight experimental conditions in a single session. The order of conditions was counterbalanced.

		Visibility Level	
		No Visibility	Low Visibility
Sound Type	No Sound	No Sound + No Visibility	No Sound + Low Visibility
	Area Cue	Area Cue + No Visibility	Area Cue + Low Visibility
	Proximity Feedback	Prox. Feed + No Visibility	Prox. Feed + Low Visibility
	Area Cue + Proximity Feedback (AC+PF)	AC+PF + No Visibility	AC+PF + Low Visibility

Table 1: Conditions experienced by each participant.

2.4.1. *Dependent Variables*

Six dependent variables were measured. Task time was measured as the elapsed time from the moment the trial began to the moment the participant found the target. Hand travel distance was measured as the distance the participant's hand traveled from the start of the targeting task, to when it reached the target. A shorter hand travel distance indicated that participants had moved their hand to the target more efficiently. The number of timeouts reflected the number of cases a participant took more than a minute to complete a task, generally reflecting the participant becoming lost or giving up. The number of errors was measured as a tally of instances in which the participant pulled the trigger on the handheld controller without it being within the target.

Although there was always sufficient information to avoid such errors, participants could "guess" by moving the controller and pulling the trigger without waiting to confirm if it was within a target. As such, this error count reflects frustration or impatience more than targeting accuracy. Finally, to assess subjective workload, a NASA TLX composite score was generated.

2.4.2. *Hypotheses and Analyses*

It was hypothesized that the sound types would have different effects depending on the level of visibility.

When no visibility was present, it was expected that the sound types that conveyed the most information about location of the target would perform better, with the AC+PF condition leading to the highest performance, followed by the proximity feedback, area cue, and then no sound conditions.

In the low visibility conditions, it was expected that the area cue would lead to the highest performance, due to the fact that it could provide helpful information without interrupting the task flow of participants who elected to target using the visuals.

Finally, it was hypothesized that all sound types would lead to decreased workload, relative to no sounds, and that these differences would be largest in the no visibility conditions.

For each dependent variable, a two-way Hyunh-Feldt repeated measures ANOVA was conducted, followed, when appropriate, by post-hoc paired Bonferroni t-tests. Post-hoc comparisons between the no visibility and low visibility conditions within each sound type showed significant differences in all cases, and are omitted for brevity. Test statistics for other post-hoc t-tests (represented in results tables) are also omitted.

3. RESULTS

3.1. Visibility Level

Across all dependent variables, participants performed significantly better in the low visibility conditions, compared to the no visibility conditions (Table 2).

	ANOVA Result	No Vis *M, (SD)*	Low Vis M, (SD)
Task Times (seconds)	$F(1, 25) = 280.47$, $p < .001$, $\eta_p^2 = .92$	34.50s (9.15)	8.84s (3.06)
Hand Travel Distance (decimeters)	$F(1, 25) = 150.20$, $p < .001$, $\eta_p^2 = .857$	12.90 dm (5.28)	2.45 dm (0.75)
Number of Timeouts	$F(1, 25) = 66.13$, $p < .001$, $\eta_p^2 = .726$	12.90 (5.27)	2.45 (0.36)
Number of Errors	$F(1, 25) = 91.89$, $p < .001$, $\eta_p^2 = .786$	25.60 (13.89)	1.91 (1.57)
Subjective Workload (0-100 Score)	$F(1, 31) = 91.07$, $p < .001$, $\eta_p^2 = .75$	42.64 (14.60)	23.77 (11.34)

Table 2: Results by Visibility Level.

3.2. Sound Type

Sound Type had an impact on targeting task times, $F(2.56, 63.99) = 70.10, p < .001, \eta_p^2 = .74$. As shown in Table 3, participants were substantially faster with all three types of sounds, compared to when no sounds were present. They took the shortest time when they heard either the proximity feedback or AC+PF. However, task times did not differ between the two conditions with proximity feedback, suggesting that participants did not receive meaningful benefits from the area cue when the proximity feedback was also present.

	No Sound	Area Cue	Prox. Feed.	AC+PF
Mean Time (*SD*)	28.01s (13.17)	25.16s (4.67)	17.06s (5.85)	16.46s (5.30)
Differs from:	Area Cue, Prox. Feed., AC+PF	No Sound, Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue

Table 3: Task time (seconds) by Sound Type.

The distance that participants moved their hand to reach targets was affected by the type of sound that they heard, $F(1, 53.34) = 29.72, p < .001, \eta_p^2 = .535$. Table 4 shows that participants in the two proximity feedback conditions were twice as efficient with their movements toward the target, compared to the area cue and no sound conditions. However, as with other dependent variables, proximity feedback and AC+PF led to equatable performance. Hand travel distance was not different between the area cue and no sound conditions, perhaps reflecting the fact that area cued participants still had to do a significant amount of effortful haptic and/or low visibility visual search to precisely locate the targets.

	No Sound	Area Cue	Prox. Feed.	AC+PF
Mean Distance (*SD*)	11.33 dm (6.01)	9.44 dm (3.79)	4.91 dm (2.19)	5.01 dm (3.60)
Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue

Table 4: Hand travel distance (decimeters) by Sound Type.

The number of times that participants timed out and failed to find the target was affected by the type of sound they heard, $F(2, 50.06) = 51.34, p < .001, \eta_p^2 = .673$. Table 5 shows that the two conditions containing proximity feedback both led to fewer timeouts than the no sound and area cue conditions. However, performance was not different between these two conditions.

	No Sound	Area Cue	Prox. Feed.	AC+PF
Mean Timeouts (*SD*)	6.96 (3.39)	5.89 (3.94)	1.40 (2.63)	1.48 (3.12)
Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue

Table 5: Hand travel distance (decimeters) by Sound Type.

The number of errors made by participants was impacted by the type of sounds they heard, $F(1.78, 44.39) = 51.34, p < .001, \eta_p^2 = .715$. Table 6 shows that, when participants heard either the proximity feedback alone, or AC+PF, they committed fewer errors than when they heard either the area cue or no sound. It was observed that participants tended to "guess" more often in the no sound and area cue conditions, thus increasing error count.

	No Sound	Area Cue	Prox. Feed.	AC+PF
Mean Errors (*SD*)	26.98 (15.44)	21.33 (12.49)	3.03 (10.64)	3.70 (11.66)
Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue

Table 6: Number of errors (count per trial) by Sound Type.

Subjective workload was impacted by the type of sound that participants heard $F(2.23, 69.02) = 19.13, p < .001, \eta_p^2 = .382$. Table 7 shows that, when participants heard either proximity feedback or AC+PF, they reported lower workload, compared to when they heard the area cue or no sound. However, when participants heard the area cue only, they reported the same level of workload as when they heard no sound. This suggests that utilizing the area cue to limit subsequent search area was less impactful on perceived workload compared to the difficulty of carrying out the subsequent targeting movement without assistance from the proximity feedback.

	No Sound	Area Cue	Prox. Feed.	AC+PF
Mean Score (*SD*)	37.80 (15.19)	37.61 (14.04)	28.53 (10.82)	28.88 (12.24)
Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue

Table 7: Subjective workload (NASA TLX, 0-100) by Sound Type.

3.3. Interaction Effects

For all dependent variables, the effect of Sound Type depended on Visibility Level. Overall, Sound Type was more impactful in the no visibility conditions. This was likely because these participants tended to rely on the sounds, in particular the proximity feedback or AC+PF. However, the sounds still led to some performance benefits in the low visibility conditions.

The effect of Sound Type on task times depended on Visibility Level, $F(2.47, 61.63) = 28.18, p < .001, \eta_p^2 = .530$, see Table 8. Sound Type impacted task times in the no visibility conditions, but not in the low visibility conditions.

		No Sound	Area Cue	Prox. Feed.	AC+PF
No Vis	Mean Time (*SD*)	45.85s (8.73)	40.96s (12.40)	25.41s (9.91)	26.26s (11.80)
	Differs from:	Area Cue, Prox. Feed., AC+PF	No Sound, Prox. Feed., AC+PF	No Sound	No Sound

Table 8: Task times (seconds) by Sound Type by Visibility Level.

The effect of Sound Type on hand travel distance depended on Visibility Level, $F(2.25, 56.34) = 21.54, p < .001, \eta_p^2 = .463$, see Table 9.

The effect of Sound Type on timeout count depended on Visibility Level, $F(2.14, 53.57) = 46.61, p < .001, \eta_p^2 = .651$. The number of timeouts differed in the proximity feedback and AC+PF conditions compared to the no sound and area cue conditions when where was no visibility, but when low visibility was present, there were no significant differences (Table 10).

As shown in Table 11, the effect of Sound Type on the number of errors depended on Visibility Level, $F(1.73, 43.16) = 46.78, p < .001, \eta_p^2 = .652$.

		No Sound	Area Cue	Prox. Feed.	AC+PF
No Vis	Mean Distance (*SD*)	18.85 dm (9.81)	16.20 dm (6.26)	7.31 dm (3.48)	7.75 dm (5.97)
	Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound	No Sound
Low Vis	Mean Distance (*SD*)	3.03 dm (1.05)	2.60 dm (1.10)	2.12 dm (0.55)	1.94 dm (0.42)
	Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound	No Sound

Table 9: Travel distance (decimeters) by Sound Type by Visibility Level.

		No Sound	Area Cue	Proximity Feedback	AC+PF
No Vis	Mean Timeouts (*SD*)	13.60 (6.08)	11.37 (7.10)	2.50 (4.50)	3.53 (5.97)
	Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue

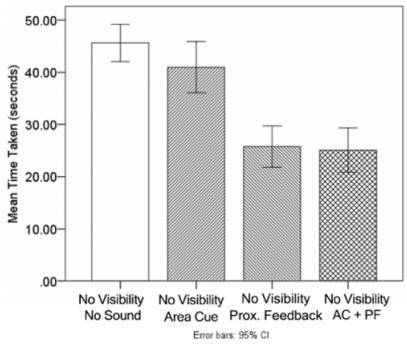
Table 10: Timeouts by Sound Type by Visibility Level.

The effect of Sound Type on subjective workload also depended on Visibility Level, $F(2.54, 78.88) = 4.79, p = .006, \eta_p^2 = .134$. When visibility was low, there were fewer significant differences between conditions (Table 12).

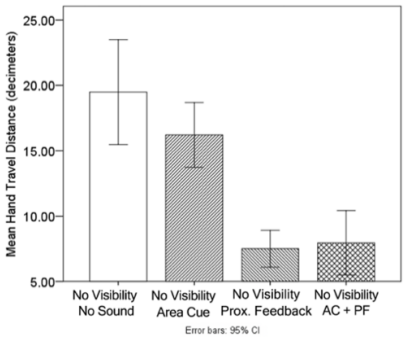
4. DISCUSSION

In this study, three auditory display approaches were evaluated in terms of their ability to assist with finding nearby objects in limited visibility conditions. Using a VR targeting task, the proximity feedback display was found to be most effective at increasing performance and improving the subjective experience of object targeting with limited visibility. The area cue was less effective at achieving these goals, and notably did not lower subjective workload, but did improve performance via several metrics. When both sound types were used in tandem (AC+PF), results were the same as when proximity feedback was used exclusively, indicating that area cue displays may have limited utility when continuous audiomotor feedback can be provided. This pattern of results was similar for both levels of visibility, but less pronounced in the low visibility conditions.

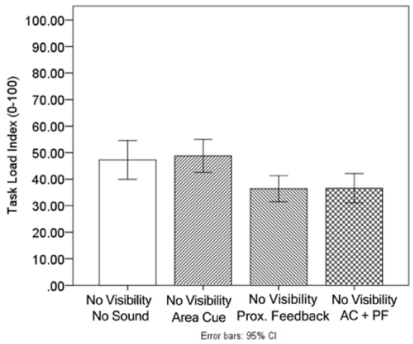
In the no visibility conditions, proximity feedback and AC+PF both led to large improvements across dependent variables (Figures 4a, 4b, and 4c). Notably, the proximity feedback led to a tenfold decrease in errors, indicating that participants were less likely to adopt a "guessing" strategy. Decreases in hand travel distance and task times indicate that, overall, participants were able to utilize the proximity feedback to move more efficiently to the target. The area cue was also effective at increasing targeting performance, but less so than expected, and not via all metrics. Notably, the area cue did not lead to a reduction in workload (Figure



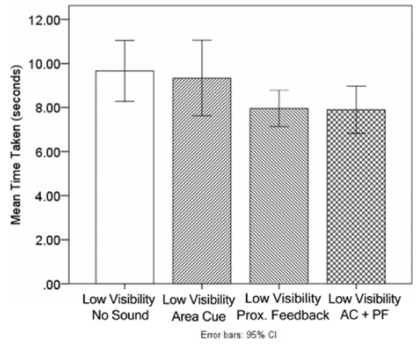
(4a) Mean task times for each sound type, no visibility conditions.



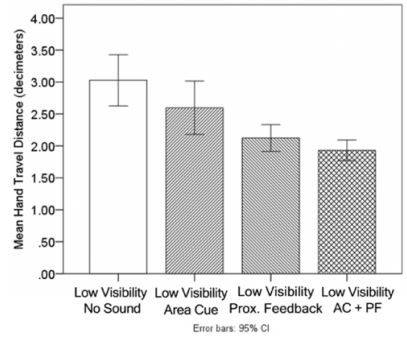
(4b) Mean hand travel distance for each sound type, no visibility conditions.



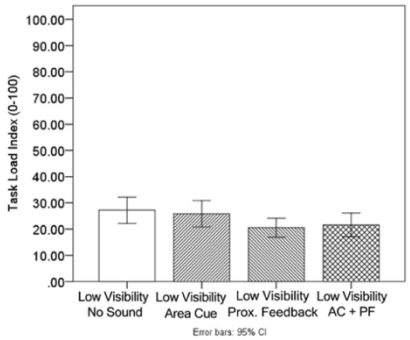
(4c) Subjective workload for each sound type, no visibility conditions.



(5a) Mean task times for each sound type, low visibility conditions.



(5b) Mean hand travel distance for each sound type, low visibility conditions.



(5c) Subjective workload for each sound type, low visibility conditions.

		No Sound	Area Cue	Prox. Feed.	AC+PF
No Vis	Mean Errors (*SD*)	52.24 (28.8)	40.12 (22.56)	5.83 (6.55)	6.39 (7.27)
	Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue
Low Vis	Mean Errors (*SD*)	3.37 (2.16)	2.63 (1.92)	1.02 (1.44)	0.83 (1.92)
	Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue

Table 11: Errors by Sound Type by Visibility Level.

		No Sound	Area Cue	Prox. Feed.	AC+PF
No Vis	Mean Score (*SD*)	47.25 (21.21)	48.78 (18.05)	36.44 (12.25)	36.63 (16.07)
	Differs from:	Prox. Feed., AC+PF	Prox. Feed., AC+PF	No Sound, Area Cue	No Sound, Area Cue
Low Vis	Mean Score (*SD*)	27.26 (14.66)	25.84 (14.62)	20.57 (10.69)	21.60 (13.19)
	Differs from:	Prox. Feed., AC+PF	Prox. Feed.	No Sound, Area Cue	No Sound

Table 12: Workload (NASA TLX, 0-100) by Sound Type by Visibility Level.

4c). While the area cue should have reduced the amount of effort required by a full two thirds, these results suggest that the primary determiner of both subjective workload and task performance was whether or not the participant had to perform the laborious task of object targeting using only tactile information.

In the low visibility conditions, a similar pattern was present: benefits were observed for all sound types compared to no sound. However, the magnitude of the advantage, as well as differences between the displays, was less pronounced compared to when there was no visibility (Figures 5a, 5b, and 5c). This compression of differences suggests that participants utilized visual input when it was available. However, there were still significant performance benefits when the auditory displays were active, as well as a reduction in subjective workload associated with the proximity feedback and AC+PF conditions (Figure 5c). This indicates that persons who are able to complete a targeting task with limited but usable visual input can still be expected to benefit from the pres-

ence of auditory targeting displays, in terms of both performance and workload.

4.1. Conclusion

The three auditory displays evaluated in this study were effective at increasing object targeting performance, and should be incorporated into virtual or mixed reality systems that endeavor to assist humans in limited visibility conditions, depending on the technical abilities of each system and needs of the task. Providing proximity feedback with which motor movements can be guided should be considered when feasible, rather than solely utilizing area cueing displays. Incorporating auditory targeting displays of the types discussed here into future systems could increase the usability of everyday environments without visual input, and support task performance in a variety of low visibility situations.

5. REFERENCES

[1] J. Wilson, B. N. Walker, J. Lindsay, C. Cambias, and F. Dellaert. 2007. "SWAN: System for Wearable Audio Navigation." In *Wearable Computers, 2007 11th IEEE International Symposium on Wearable Computers*, IEEE, Boston, MA. 491–98.

[2] J. Loomis, R. D. Golledge, and R. L. Klatzky. 2001. "GPS-based navigation systems for the visually impaired." In *Barfield W, Caudell T, editors. Fundamentals of wearable computers and augmented reality*. Lawrence Erlbaum. Mahway, NJ. 429–446.

[3] K. R. May, X. Ma, P. Roberts, and B. N. Walker. (Under Review). "Spotlights and Soundscapes: Participatory Design of Mixed Reality Auditory Environments for Persons with Visual Impairment."

[4] R. Yaagoubi, T. Badard, and G. Edwards. 2009. "Standards and Spatial Data Infrastructures to help the navigation of blind pedestrian in urban areas." *Urban and Regional Data Management* (February 2009).

[5] C. Prablanc, J.E. Echallier, M. Jeannerod, and E. Komilis. 1979. "Optimal response of eye and hand motor systems in pointing at a visual target." *Biological Cybernetics* 35, 3 (1979), 183–187.

[6] J. C. Middlebrooks. 1991. "Sound Localization by Human Listeners." *Annual Review of Psychology* 42, 1 (January 1991), 135–159.

[7] D. R. Begault, E. M. Wenzel, and M.R. Anderson. 2001. "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source." *Journal of the Audio Engineering Society*, 49(10), 904–916.

[8] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. 1993. "Localization using nonindividualized head-related transfer functions." *The Journal of the Acoustical Society of America*. 94, 1, 111–123.

[9] K. Thakoor, N. Mante, C. Zhang, C. Siagan, J. Weiland, L. Itti, and G. Medioni. 2015. "A System for Assisting the Visually Impaired in Localization and Grasp of Desired Objects." *Computer Vision—ECCV 2014 Workshops Lecture Notes in Computer Science* (2015), 643–657.

[10] I. A. Doush, S. Alshatnawi, A. Al-Tamimi, B. Alhasan, and S. Hamasha. 2016. "ISAB: Integrated Indoor Navigation System for the Blind." *Interacting with Computers* (2016).

[11] R. Chincha and Y. Tian. 2011. Finding objects for blind people based on SURF features. *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)* (2011).

[12] B. Schauerte, M. Martinez, A. Constantinescu, and R. Stiefelhagen. 2012. "An Assistive Vision System for the Blind That Helps Find Lost Things." *Lecture Notes in Computer Science Computers Helping People with Special Needs* (2012), 566–572.

[13] C.C. Pratt. 1930. "The spatial character of high and low tones." *Journal of Experimental Psychology* 13, 3 (1930), 278–285.

[14] K. Evans and A. Treisman. 2011. "Natural cross-modal mappings between visual and auditory features." *Journal of Vision* 10, 1 (June 2011), 6–6.

[15] K. Pisanski, S. G. Isenstein, K. J. Montano, J.M. O'Connor, and D. R. Feinberg. 2017. "Low is large: spatial location and pitch interact in voice-based body size estimation." *Attention, Perception, & Psychophysics* 79, 4 (2017), 1239–1251.

[16] C. V. Parise and C. Spence. 2009. "'When Birds of a Feather Flock Together': Synesthetic Correspondences Modulate Audiovisual Integration in Non-Synesthetes."

[17] C. V. Cesare V. Parise, K. Knorre, and M. O. Ernst. 2014. "Natural auditory scene statistics shapes human spatial hearing." Proceedings of the National Academy of Sciences 111, 16 (July 2014), 6104–6108.

[18] A. O. Effenberg. 2005. "Movement sonification: Effects on perception and action." *IEEE Multimedia*, 12(2), 53–59.

[19] B. N. Walker. 2007. "Consistency of magnitude estimations with conceptual data dimensions used for sonification." *Applied Cognitive Psych*. 21, 5. 579–599.

[20] J. Edworthy, E. J. Hellier, & R. Hards. 1995. "The semantic associations of acoustic parameters commonly used in the design of auditory information and warning signals." *Ergonomics*, 38(11), 2341–2361.

[21] J. Lyons, S. Hansen, S. Hurding, and D. Elliott. 2006. "Optimizing rapid aiming behaviour: movement kinematics depend on the cost of corrective modifications." *Experimental Brain Research* 174, 1 (2006), 95–100.

[22] M.D. Byrne, M.K. O'malley, M.A. Gallagher, S.N. Purkayastha, N. Howie, and J.C. Huegel. 2010. "A preliminary ACT-R model of a continuous motor task." *PsycEXTRA Dataset* (2010).

[23] S. G. Hart and L. E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." *Advances in Psychology Human Mental Workload* (1988), 139–183.

THE ALCHEMY OF CHAOS: A SOUND ART SONIFICATION OF A YEAR OF TOURETTE'S EPISODES

Thomas J. Mitchell
Creative Technologies Lab
UWE, Bristol, UK
tom.mitchell@uwe.ac.uk

Jess Thom, Matthew Pountney
Touretteshero
London, UK
info@touretteshero.com

Joseph Hyde
Bath Spa Univeristy
Bath, UK
j.hyde@bathspa.ac.uk

ABSTRACT

Touretteshero is the name of a organisation that aims to raise awareness of Touette's syndrome by sharing and celebrating the creativity and humour of the involuntary vocal and movement tics that characterise the condition. This paper documents the development of a Touretteshero project called *The Alchemy of Chaos*, a sound art piece that translates a year of intensive ticcing episodes (or 'ticcing fits') into a six minute sonification. The work emphasises both the faithful representation of data and the aesthetic sound quality, drawing techniques and ideas from sound design for film, which is often used to convey information about a visual scene in ways that can be used for sonfication. Specifically, the work uses Chion's *elements of auditory setting*: short punctual sounds that can express locations with minimal sonic references. Sound parameters are also classified into groups that have 'data significance' and those that do not, with aesthetic interventions limited to those parameters that do not impact on data transparency. The resulting piece was included within a keynote talk at the Royal Albert Hall in the UK and the paper includes a qualitative reflection on the work and the potential value that sound design techniques for film can bring to the auditory display community.

1. INTRODUCTION

Gilles de la Tourette syndrome (Tourette's) is a neurological condition that is characterised by involuntary and uncontrollable movements (motor tics) and vocalisations (vocal tics). Estimates vary, but studies have shown that Tourette's affects between 0.4% and 5.0% of children [1]. The tics associated with Tourette's can vary in severity, from subtle muscle contractions to sustained and intense spasms that can appear like seizures [2].

This paper documents the sonification of a year of intense ticcing episodes recorded by an individual with Tourette's between 2011 and 2012. The records were kept initially to enable longitudinal analysis by medical practitioners but in this work the data are used to define the structure of a sound art piece, to reveal and share the human experiences of Tourette's. The paper begins with an introduction to the Touretteshero project before introducing the salient literature relevant to sonification and sound art. The methods and processes adopted for this work are presented and the final piece is then discussed. The paper concludes with remarks on the importance of aesthetic considerations in sonification design and, in particular, how sound design techniques from film sound have the capacity to enhance and humanise information when expressed as sound.

2. TOURETTESHERO

Individuals with Tourette's can often experience discriminatory behaviour in public spaces and often withdraw from social activities to avoid confrontation [3]. This lack of public understanding and the resulting social isolation can have a negative impact on the lived expedience and quality of life of people with Tourette's. 'Touretteshero' is both the alias of Jess Thom and the name of an organisation that was set up in 2010 to raise awareness of Tourette's and its challenges.

Jess was diagnosed with Tourette's in her early twenties and at the time of writing, her tics are frequent and varied. Her vocal tics produce combinations of sounds and words with the occasional appearance of offensive language, referred to as Coprolalia, which affects 10% - 15% of individuals with Tourette's [4]. However, more frequently Jess's vocalisations produce highly creative and humorous phrases that originally inspired Touretteshero. Jess's movement tics include, blinking, shrugging, jumping, head jerking and leg bending, which when combined make walking difficult, so Jess uses a wheelchair. Sometimes, these movements intensify into 'ticcing fits': distinct periods of overpowering and constant motor movements that typically last between 10 minutes and an hour (although sometimes considerably longer).

A hallmark of Touretteshero is the celebration of the humour and creativity of Jess's tics, embracing the utterances and movements as a source of inspiration for a range of artistic outputs. These outputs typically involve collaborations with artists to create imagery inspired by Jess's vocal tics, see Figure 1(a) and (b), the publication of biographical writings [5] and events that include performers with Tourette's, Figure 1(c). This theme of creativity motivated this sonification work: the desire to translate a medical record of ticcing fits, that represent an immense amount of discomfort into something engaging and beautiful.

3. SONIFICATION, SOUND DESIGN AND MUSIC

The relationship between sonification and music has formed the basis of much discussion and disagreement in the auditory display community. For example, Vickers and Hogg go some way to reduce the differences between sonification and music to the perceptions of the listener [6, 7], Scaletti argues that both endeavours

https://doi.org/10.21785/icad2019.040

(a)

(b)

(c)

Figure 1: a) *'God's Moving to Watford on Sunday'* by Jess Thom, b) *'Postman Pat Gave Me Herpes'* by Luke Turner, c) Photograph from a performance of Samuel Beckett's *'Not I'* by Touretteshero at Battersea Arts Centre. Courtesy of: James Lyndsay

would benefit from being considered entirely distinct [8]. Other researchers describe a tension between musicality and utility, where sonifications become decreasingly useful and transparent as they become increasingly musical and satisfying [9, 10], although Vickers has recently brought this dualism into question [11]. A consensus does form, however, around the use of aesthetics to enhance the communicative and expressive qualities of an auditory display and to reduce fatigue [12, 13]. This observation has led to numerous calls for interdisciplinary collaboration to encourage the integration of artistry and craft into sonification research [14, 15, 16].

Sonification and music can be considered on a continuum between representation and abstraction, or *infomatica* and *musica* as described by Vickers and Hogg [7]. Representation tends to emphasise an information theoretic approach, viewing the auditory system as a communications channel and prioritising the faithful expression of information with little regard for aesthetics. Conversely, abstraction is concerned with aesthetics over representation, as the underlying information may be symbolic or serve only as inspiration. The extremes are easily identified, for example, representation is a priority for sensory substitution devices that translate image into sound (e.g. 'The vOICe' [17]), and abstraction is emphasised in data inspired music (e.g. Alvin Lucier's 'Music for Solo Performer' [18]) where sonification manifests as an artistic device. Common to all sonifications is the intention of the designer/artist to convey information with sound, combined with the delegation of some aspect of the aural fabric to a data source [19]. If the mapping is intended to have scientific utility, the data points of interest must be rendered faithfully such that they may be inferred by listeners.

The aim of this work is to explore a mid-point on the continuum from representation to abstraction, prioritising sonification aesthetics without compromising the accuracy and legibility of the underlying data with an artificial 'musical' framework. We believe this approach has the potential to increase the agency of sonification in terms of access, usability and ergonomics through careful design choices rather than exclusively artistic activity. With this in mind, inspiration is taken from the field of sound design, with elements from the (somewhat related) field of acousmatic music. An outline of this approach is laid out in Section 4.2 below.

The remainder of this paper describes the creation of a data inspired composition that translates a year of Jess's ticcing fits into a sound piece. Many aspects of the data within the records are sonified to preserve and convey the frequency and relentlessness of these episodes, with aesthetic and sound design choices that help to bring this data into the human realm, inviting listeners to contemplate the lived experience of Tourette's.

4. THE ALCHEMY OF CHAOS

The Alchemy of Chaos was prompted when Jess Thom (Touretteshero) was invited to deliver a keynote at the Royal Albert Hall in the UK at the TEDxAlbertopolis event. It had been a year since Jess had started to experience intensive ticcing fits and she wanted to share this aspect of her condition at the event.

At the onset of these episodes, Jess began keeping detailed records to look for trends, patterns and possible causes with her medical consultants. During a ticcing episode, any part of Jess's body may move, shake, contort or lock into painful positions, which can resemble an epileptic seizure but with major differences: Jess remains fully conscious and aware of her surroundings and although unable to speak she is normally able to communicate

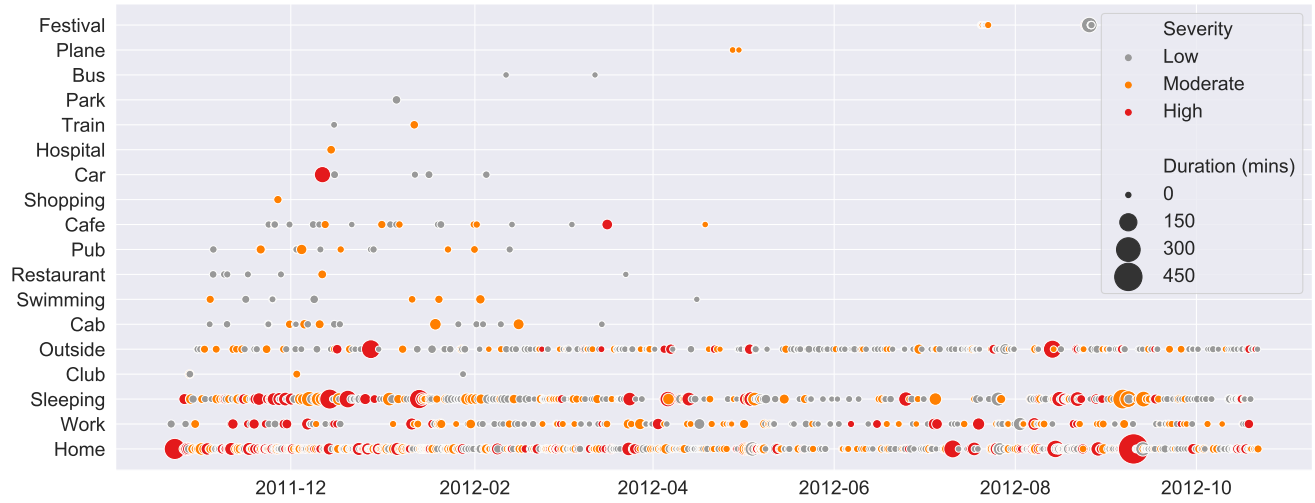


Figure 2: Visualisation of the ticcing fit data showing the location, intensity, duration and time of each episode

by blinking. Ticcing fits can occur at any time and in any location, they vary in severity, duration and frequency and are likely to result in injury without the help of a support worker. While these episodes can be physically exhausting, as soon as they calm, Jess is immediately able to resume the activity she was attending to previously.

Unlike Jess's vocal tics that inspire the majority of the Touretteshero artwork, humour is notably absent from this aspect of her symptoms. However, the desire to create an artform from the records was shared with the co-authors of this paper and subsequent discussions resulted in plans for a data inspired sound piece. The aim was to translate the alphanumeric information of the records into music that preserves, humanises and coveys this aspect of Jess's everyday lived experience.

4.1. Data

Following each episode, Jess or her support worker would note the start date, time, duration, severity (low, moderate or high) and location along with some additional notes (i.e. body parts affected and whether speech was lost). Records were transcribed and shared with the project team in spreadsheet format. The complete data set contained records for 2011 episodes that took place between 21st October 2011 and the same day one year later.

The data was initially examined, converted to CSV (comma separated values) and preprocessed to identify and correct any transcription errors and inconsistencies that were found. For example, typos, wording and capitalisation variations, and ordering errors (owing to mixed 24/12 hour clock entries). Occasionally, entries were recorded with two locations because the episode may have started in one location and ended in another, e.g. 'Outside/Cab'. Alternatively, these items were a result of the emergent nomenclature mixed with more specific labels, e.g., 'Outside/Park'. Instances of the former were replaced with the start location of the episode and instances of the latter were replaced with the more specific location. A plot visualising the data is provided in Figure 2.

4.2. Sound Design

Several suggestions and guidelines that define sonification have been proposed previously, for example, Hermann's definitions [20] set out criteria that promote objectivity and reproducibility. Other authors have provided less prescriptive definitions that highlight the importance of *authenticity* and the preservation of relative time structures [10].

In this work, sound design took a compositional approach, drawing inspiration and working practices from acousmatic music, acoustic ecology and sound design. Precise timing was considered a priority, with the aim of preserving and representing the relative timing of events within the piece. Consequently, playback time for the piece was considered in units of *days per minute* (*dpm*) with the sound events for each episode scheduled in relation to its timestamp within the data.

The general approach for generating the sound content took a Foley approach, using primarily short, clearly identifiable, close-miked sounds to construct the dominant audio texture. For each location, a representative symbolic sound (or auditory icon [21]) was selected. Particular inspiration was taken from the pioneering sound design work of Walter Murch on several seminal films made in the 1970s by directors such as Francis Ford Coppola and George Lucas [22]. The soundtracks to these films blur the boundaries between the traditionally-demarked areas of sound design and music, with one often taking on the role of the other. For example, in the opening sequence of Lucas's *TXH1138*, a sequence of controls on a console light up and the corresponding 'beep' tones form a melodic contour that functions as part of the music, offering a segue into the following scene where the same tones become the bell of an elevator. In the 'Tiger Scene' in Coppola's *Apocalypse Now*, high-pitched insect sounds fulfil the tension-building function more commonly facilitated with high 'Psycho' inspired strings.

In 'The Alchemy of Chaos' recognisable sounds are used to represent features of the data being sonified. Harmonic principles were employed in selecting these sounds, and in some cases manipulation (re-tuning), such that they collectively form harmonic, melodic patterns. This is a good example of sound design fulfill-

ing the 'function' of music - beyond this choice, no other musical logic was imposed upon the data, but this design choice enables the result to still sound 'musical'. Note that the pitch of the sounds has no significance to the sonification, allowing this creative 'liberty' to be taken without compromising the transparency of the sonification. Where parameters do have data significance, no such interventions were made.

Another of Murch's sound design principles is also adopted here, defined by Chion as *elements of auditory setting* (EAS), that is:

> 'a punctual source... which help to define a film's space by means of specific, distinct small touches. Typical sounds of the auditory setting are the far-away barking of a dog, or the ringing of a phone in the office next door, or a police car siren' - [23], p54

EAS allows locations to be defined with minimal sonic references, an efficient approach allowing much flexibility and space for other elements of the sound design/music. In this sonification, EAS are used to define the location of each ticcing fit, whilst allowing other sound parameterisations, timing, amplitude and reverberation to define the start time, intensity and duration of each episode respectively.

These short sounds, arranged into distinct categories according to the location of the tics, resemble Pierre Schaeffer's *l'objets sonores*. Our intention here is somewhat at odds with Schaeffer's (a founding principle of *musique concrete*). His idea, rooted in the phenomenology of Husserl, was that sounds might be divorced from their source and any *a priori* meaning, and treated as abstract plastic entities. In this sonification, the sounds are explicitly intended to have 'meaning' (as representing each location). This corresponds more with 'causal listening' as defined by Shaeffer rather than 'reduced listening'. However, the parameterisation of timing, amplitude and reverb is informed by Schaeffer's approach, as subsequently expanded by Denis Smalley in his writings on Spectromorphology*s* [24]. Certainly, the approach to sonification taken here has more in common with *musique concrete* than it does with more traditional forms of music.

For each location in the data set, a symbolic sound was chosen. In most cases, these were recorded by Jess in the locations where many of these episodes may have taken place, establishing a link to the acoustic ecology of Jess's day-to-day environment [25]. In some instances this was impractical ('club') or inappropriate ('hospital'), in which case, suitable substitutions were found. A complete list of locations, their incidence and the representative sounds is provided in Table 1.

4.3. System Overview

The sound piece was made possible with the development of two applications, a *Data Player* and a *Sampler*. Both applications ran on a single machine and communicated by Open Sound Control (OSC) [26] over a network datagram socket. The system architecture is shown in Figure 3.

The Data Player was a simple application developed in C++ using the Juce library [27] that opened, parsed and 'played' the data in a variety of modes. Accepting as input a CSV file containing the episode data, the application produced as output UDP packets containing an OSC message for each episode with the address pattern `/episode` and the arguments shown in Table 2.

The data playback rate was in units of *dpm* as described in section 4.2. For the final audio piece, this was set to 60 *dpm* where each day can be imagined as a beat at a tempo of 60 *bpm* and the resulting year of ticcing episodes translates to 6:05 seconds. To ensure that messages were precisely timed, all event messages were scheduled and sent on a high priority timing thread.

Location	Incidence	Sound
Bus	2	bus - interior and exterior
Cab	20	cab - exterior and exterior
Café	21	cups and saucers
Car	8	ignition and engine
Club	7	indicative music clip
Festival	37	crowd
Home	967	'soundmarks' from Jess's home
Hospital	1	hospital machine (from library)
Outside	304	birdsong
Park	1	children playing
Plane	2	plane
Pub	16	crowd - interior
Restaurant	7	knives and forks
Shopping	1	musak clip
Sleeping	398	Jess's alarm clock
Swimming	10	swimming pool
Train	2	station announcement
Work	207	typing

Table 1: List of episode locations and their sound representations

Argument	Type	Description
counter	int32	index of episode in the dataset (0 - 2010)
duration	int32	rounded to nearest minute
severity	int32	where 0 = low, 1 = moderate, 2 = high
location	int32	0 - 17 see Figure 2

Table 2: OSC message format for each episode

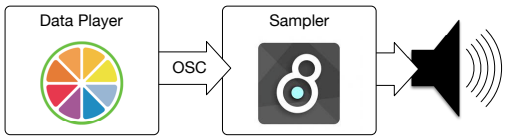


Figure 3: System block diagram showing *Data Player* and *Sampler* applications

The application GUI provided controls for loading a data file, toggling playback positioning/resetting the playhead and playback rate. Three timing modes were also available and explored: *Proportional*, where the timing structure and intervals are maintained precisely, as described above; *Day Beat*, where all events in a single day are played simultaneously on the onset of their respective *day beat*; and *Day Beat Quantize*, where events are quantized to the nearest semiquaver of their respective *day beat*. After some initial experimentation only the *Proportional* method was used as relative timing was considered more important than the aesthetic rhythmic qualities that the other methods introduced.

The Sampler was a simple audio file playback system created in Max/MSP that received and decoded the OSC messages sent from the Data Player and triggered audio files in response. The sampler was loaded with the Foley samples representing each location (so, for example 'café' might be represented by the clink of a cup on a saucer). Rather than use a single sample for each

Figure 4: Still taken from the video *'Tourette's syndrome – why it doesnt define me — TEDxAlbertopolis'*, showing Touretteshero addressing the audience at the Royal Albert Hall

location, a small library of Foley sounds (between 4 and 20) were loaded into a sound bank and could be loaded and triggered in a randomised order, a technique used in game audio to avoid fatiguing and artificial repetition.

The sample amplitude is then scaled to reflect the intensity of the episode before being summed into a simple reverberator with the diffusion time scaled to reflect the duration of each episode (which was deemed aesthetically preferable to actually lengthening the sounds using pitch shift or timestretch).

5. QUALITATIVE REFLECTION

An excerpt of the resulting sonification[1] was played to an audience of 5000 delegates at Royal Albert Hall as part of the TEDxAlbertopolis[2] event, see Figure 4.

Listening to the piece, it is possible to identify the rhythm of passing days from the alarm clock sound, signifying the 'Sleeping' location. With episodes occurring every four hours, on average, most days included an entry taken at night and this regular pulse persists throughout the sound piece, contextualising the time structure and conveying the frequency and disruptive nature of these episodes.

As commented by Jess in her writings [28], seasonal changes are also recognisable as the piece progresses, with the number of 'Outside' sounds increasing in frequency from the midway mark as dates pass into spring and summer. The piece is noticeably calmer and sparse during this period, as can be observed in Figure 2 where, in general, episodes tended to occur less frequently, were less severe and shorter in duration. An observation again confirmed by Jess, who notes that ticcing fits can intensify at times of stress and anxiety, symptoms that are both relieved by the longer days and warmth of the summer season.

The sound palette can also be heard evolving throughout the piece with the full range of locations and chaotic sound textures present in the first half, settling into a consistent sound texture from the mid-point onward. This evolution in sound ecology reflects Jess adaption as these ticcing fits became a permanent and uncontrollable aspect of everyday life. As episodes can occur at any time and in any location, Jess's behaviour change is signified by a stabilising sound palette as certain locations within the data set (e.g. swimming, eating out, clubbing, etc.) disappear. In her writings about the sound piece, Jess comments on this change and notes that the ticcing fits have led to a reduced sense of independence and freedom [28]. While the tics themselves cannot be controlled, it is possible to control the factors that make them more manageable or less likely to occur.

6. CONCLUSION

This paper documents a translation of a year of intensive ticcing episodes (or 'ticcing fits') into a 6:05 minute sound art piece that was played to an audience of 5000 people at the Royal Albert Hall in the UK. The data included the time, location, duration and intensity of 2011 ticcing fits and were recorded by Jess Thom, who

[1] Which can be accessed here:
https://soundcloud.com/josephhyde/touretteshero-whole-year

[2] Which is available here:
https://www.youtube.com/watch?v=_jmTlQld2Z8

spreads awareness of Tourette's and its challenges by sharing the challenges and creativity of her tics through the Touretteshero organisation. The aim of this work was to convert a data set representing a great deal of discomfort, into something creative and engaging, while preserving and conveying the relentlessness and lived experience of this aspect of Jess's condition. The collaborative team used sonification techniques to represent the recorded data variables as sound, drawing inspiration and techniques from sound design for film to enhance the aesthetics of the piece without compromising the data representation. In particular, Chion's *elements of auditory setting* are used to rapidly convey a sense of location for each ticcing fit through short, punctual sound icons. The paper also introduces the process of grouping sonification parameters into those that have elements of 'data significance' and those that do not, with aesthetic and creative interventions limited to sound parameters that do not impact on the data transparency of the sonification.

7. ACKNOWLEDGMENT

TJM and JH acknowledge support from Leverhulme Trust, Royal Society, the British Academy and the Royal Academy of Engineering under grant APX\R1\180118. The authors would like to thank Dr David Glowacki, Alex Jones and the members of the Glowacki Group and CT Lab.

8. REFERENCES

[1] R. H. Bitsko, J. R. Holbrook, S. N. Visser, J. W. Mink, S. H. Zinner, R. M. Ghandour, and S. J. Blumberg, "A national profile of tourette syndrome, 2011-2012," *Journal of developmental and behavioral pediatrics*, vol. 35, no. 5, pp. 317–22, 2014.

[2] J. S. Stern, "Tourette's syndrome and its borderland," *Practical neurology*, vol. 18, no. 4, pp. 262–270, 2018.

[3] H. Smith, J. Fox, and P. Bunton, "The lived experiences of individuals with tourette syndrome or tic disorders: A metasynthesis of qualitative studies," *British Journal of Psychology*, vol. 106, no. 4, 2014.

[4] M. M. Robertson, "Tourette syndrome," *Psychiatry*, vol. 4, no. 8, 2005.

[5] J. Thom, *Welcome to Biscuit Land: A Year in the Life of Touretteshero*. Souvenir Press, 2012.

[6] P. Vickers, "Ars informatica – ars electronica: Improving sonification aesthetics," in *Understanding and Designing for Aesthetic Experience (workshop at The 19th British HCI Group Annual Conference)*, 2005.

[7] P. Vickers and B. Hogg, "Sonification abstraite / sonification concrète: An 'aesthetic perspective space' for classifying auditory displays in the ars musica domain," in *Proceedings of the 12th International Conference on Auditory Display*, 2006.

[8] C. Scaletti, "Sonification $\neq$ music," in *The Oxford Handbook of Algorithmic Music*. Oxford Univeristy Press, 2018.

[9] T. Bovermann, J. Rohruber, and A. de Campo, "Laboratory methods for experimental sonification," in *The Sonification Handbook*. Logos Verlag, 2011.

[10] S. Gresham-Lancaster, "Relationships of sonification to music and sound art," *AI and Society*, vol. 27, no. 2, 2012.

[11] P. Vickers, "Sonification and music, music and sonification," in *The Routledge Companion to Sounding Art*. Routledge, 2016.

[12] G. Kramer, "An introduction to auditory display," in *Auditory Display: Sonification, Audification and Auditory Interfaces*. Addison-Wesley, 1994.

[13] P. Vickers, "Lemma 4: Haptic input + auditory display = musical instrument?" in *Proceedings of the First International Conference on Haptic and Audio Interaction Design*, 2006.

[14] S. Barrass and G. Kramer, "Using sonification," *Multimedia Systems*, vol. 7, no. 1, 1999.

[15] M. Barra, T. Cillo, A. De Santis, U. F. Petrillo, A. Negro, and V. Scarano, "Multimodal monitoring of web servers," *IEEE MultiMedia*, vol. 9, no. 3, 2002.

[16] S. Barrass, "Sonifications for concert and live performance," *AI and Society*, vol. 27, no. 2, 2012.

[17] P. B. L. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

[18] V. Straebel and W. Thoben, "Alvin lucier's music for solo performer: Experimental music beyond sonification," *Organised Sound*, vol. 19, no. 1, pp. 17–29, 2014.

[19] P. Sinclair, "Sonification: what where how why artistic practice relating sonification to environments," *AI and Society*, vol. 27, no. 2, 2012.

[20] T. Hermann, "Taxonomy and definitions for sonification and auditory display," in *Proceedings of the 14th International Conference on Auditory Display*, 2008.

[21] W. W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human–Computer Interaction*, vol. 2, no. 2, 1986.

[22] M. Ondaatje and W. Murch, *The conversations: Walter Murch and the art of editing film*. A&C Black, 2002.

[23] M. Chion, *Audio-Vision: Sound on Screen*. Columbia University Press, 1994.

[24] D. Smalley, "Spectromorphology: Explaining soundshapes," *Organised sound*, vol. 2, no. 2, 1997.

[25] A. Polli, "Soundscape, sonification, and sound activism," *AI and Society*, vol. 27, no. 2, 2012.

[26] M. Wright and A. Freed, "Open sound control: A new protocol for communicating with sound synthesizers," in *Proceedings of the International Computer Music Conference*, 1997.

[27] R. Ltd. The juce library. [Online]. Available: https://juce.com

[28] J. Thom. Rah. [Online]. Available: https://www.touretteshero.com/2013/09/23/rah/

MULTILAYERED NARRATION IN ELECTROACOUSTIC MUSIC COMPOSITION USING NUCLEAR MAGNETIC RESONANCE DATA SONIFICATION AND ACOUSMATIC STORYTELLING

Falk Morawitz

University of Manchester,
Oxford Road,
Manchester, M13 9PL, United Kingdom
falk.morawitz@manchester.ac.uk

ABSTRACT

Nuclear magnetic resonance (NMR) spectroscopy is an analytical tool to determine the structure of chemical compounds. Unlike other spectroscopic methods, signals recorded using NMR spectrometers are frequently in a range of zero to 20000 Hz, making direct playback possible. As each type of molecule has, based on its structural features, distinct and predictable features in its NMR spectra, NMR data sonification can be used to create auditory 'fingerprints' of molecules. This paper describes the methodology of NMR data sonification of the nuclei nitrogen, phosphorous, and oxygen and analyses the sonification products of DNA and protein NMR data. The paper introduces *On the Extinction of a Species*, an acousmatic music composition combining NMR data sonification and voice narration. Ideas developed in electroacoustic composition, such as acousmatic storytelling and sound-based narration are presented and investigated for their use in sonification-based creative works.

1. INTRODUCTION

Spectroscopy is a field of research that analyses the interactions of electromagnetic waves and physical matter [1]. It is used to examine the molecular structure, bonding strengths, energy level distribution, compound weight, and many other factors of chemical compounds. Sonification of spectroscopic data has been used in scientific research to determine quantum coupling in oscillating atoms [2] or to perceptualize the properties of subatomic particles [3]. In a musical context, infra-red spectroscopy data have been used to create microtonal musical scales [4] and have been sonified for the exploration of new sonorities [5][6].
The focus of this paper is the sonification of NMR data, with sections 1.1 and 1.2 detailing the basis of NMR spectroscopy and the use of NMR data sonification in science and art, respectively. The second part of the paper is concerned with the presentation and the aesthetics of the sounds created. This paper explores the use of narration as key to support sonification and examines strategies developed in acousmatic music compositions, including acousmatic storytelling and sound-texture-based narration for their use in sonification-based compositions. The acousmatic work *On the Extinction of a Species* is discussed as a case study in chapter 3.2.

1.1. An introduction to NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy examines the resonance frequencies of magnetic nuclei observable under strong magnetic fields. Subjected to a magnetic field, each type of magnetic nucleus resonates at a characteristic frequency, its resonance frequency. Nuclei will derivate from this resonance frequency, depending on the other nuclei it is bound to. This derivation is known as chemical shift and is typically recorded relative to the resonance frequency in parts per million (ppm). Depending on the chemical compound's structure, the resonance of a nucleus can split further into sets of resonances (figure 1), so-called splitting patterns. The resonance frequencies of a chemical compound, their chemical shift, and its resonance splitting patterns are highly indicative of a compound's chemical structure, making NMR spectroscopy one of the most valuable and most used tools in organic chemistry structure elucidation and validation.

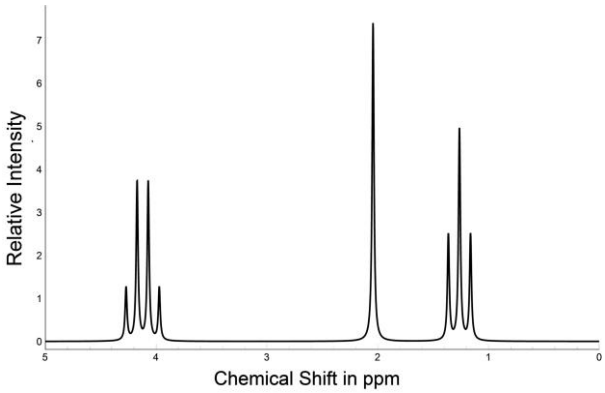


Figure 1: The hydrogen-1 NMR spectrum of ethyl acetate, with the signals at 4.1 ppm and 1.2 ppm split into quadruplet and triplet patterns, respectively.

As a simple analogy, a chemical compound can be thought of as a guitar string, which, by itself, is silent. Only when this string is fixed onto a guitar and plucked (analogous to subjecting a chemical to a strong magnetic field and exciting it with a radio pulse) a signal is emitted. In the case of the guitar string, this signal is a mechanical wave. In NMR spectroscopy an electromagnetic signal, also known as free induction decay, or FID, is recorded (figure 2). The FID is Fourier transformed and plotted as an NMR spectrum.

https://doi.org/10.21785/icad2019.052

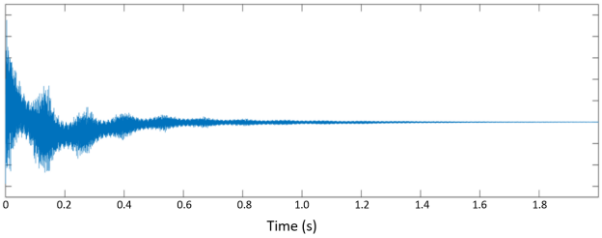


Figure 2: Free induction decay of ethylmalonic acid. Displayed using DOSYToolbox [7]. Data taken from [8].

1.2. NMR sonification in science and art

The chemical shifts in an NMR spectrum are conventionally displayed in parts per million but can be converted to frequencies values using (1), where f is the chemical shift in hertz, δ the chemical shift in parts per million, γ is the gyromagnetic ratio of the examined nuclei and B_0 is the strength of the external magnetic field.

$$f = \delta \gamma B_0 \quad (1)$$

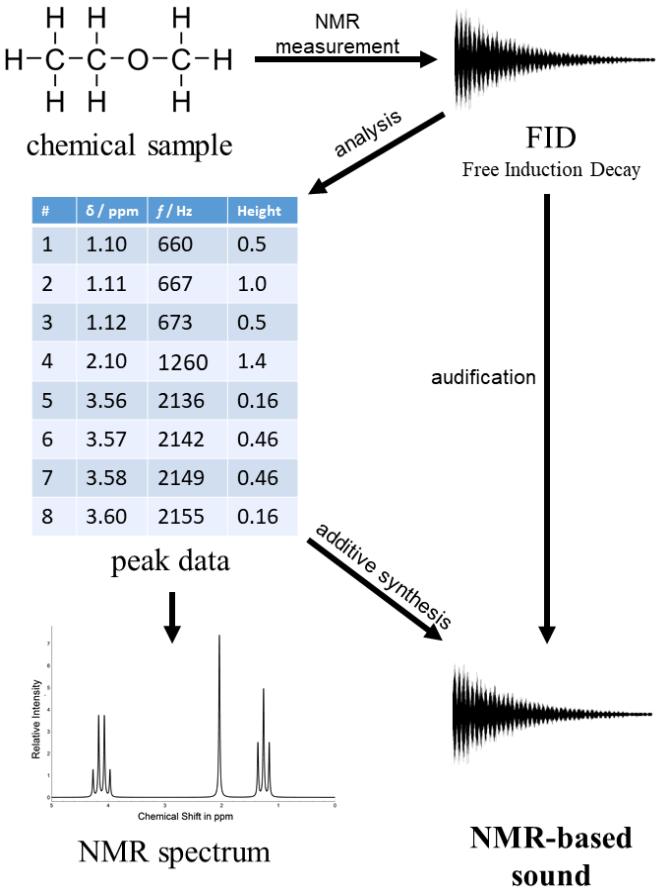
A feature of NMR spectra is that chemical shifts of a great number of different types of nuclei tend to lie in the range of zero to 20000 hertz, making NMR data an ideal candidate for audification. The audification of NMR data was sporadically used in analytical laboratories in the 1970s, as listening to the reference sample's pitch before an experiment was a fast way to check that the NMR machine was calibrated correctly [9]. As computing power increased, this tuning process was relegated to computer algorithms and the sound modules were eventually removed. In 2015, a program that sonified NMR data for analysis purposes was released [10], and its efficacy evaluated using simple NMR spectra.

The use of NMR data in an art-science context has only recently been explored, with the incorporation of audified NMR data into electroacoustic music compositions [11][12]. It has been shown that NMR sonification is not only a source of new musical timbres, but it can be used as an auditory indicator of inter-molecular interactions in virtual reality art installations [13]. The use of hydrogen-1 and carbon-13 NMR data for sonification has been examined elsewhere [11] and is well suited to display structural features of small molecules as sound. It is less suited for the display of structural features of molecules with only a small number of carbon or hydrogen atoms present, such as amino acids, organophosphates or nucleotides, and their macromolecular counterparts, such as proteins and DNA strands.

This paper expands the usage of NMR sonification by examining the sonification products of the NMR-active nuclei of three atoms commonly found in these organic compounds: nitrogen, phosphorous, and oxygen.

2. NMR DATA SONIFICATION

NMR data can be sonified using a variety of methodologies, including audification, additive synthesis, FM synthesis, or model-based synthesis. The sonification methodology employed here is based on the additive synthesis of NMR data (see figure 3).



#	δ / ppm	f / Hz	Height
1	1.10	660	0.5
2	1.11	667	1.0
3	1.12	673	0.5
4	2.10	1260	1.4
5	3.56	2136	0.16
6	3.57	2142	0.46
7	3.58	2149	0.46
8	3.60	2155	0.16

Figure 3: Methodology for the sonification of NMR data. Modified, based on [11].

2.1. 15N and 14N NMR

Nitrogen, together with carbon, hydrogen, oxygen, and phosphorous, is one of the most commonly found elements in organic matter. Nitrogen is a core component of nucleic acids and amino acids and macromolecular structures such as DNA, RNA, and proteins. Two different isotopes, nitrogen-14, and nitrogen-15, are used in NMR spectroscopy. Unlike nitrogen-15, nitrogen-14 is a quadrupolar nucleus. Subjected to a magnetic field, a quadrupolar nuclei does not split into two magnetic orientations, but more. A nitrogen-14 nucleus has three possible orientations in a magnetic field. This means that instead of a single signal peak, nitrogen-14 will split into at least two signals. In 15N and 14N NMR, most signals lie in the range of zero to eight kHz for amines and amides and 10 – 25 kHz for aromatic rings. The decay time of each nucleus ranges from 0.1 seconds to 3 seconds depending on its bonding partners, with tertiary nitrogen (that is, nitrogen bound to only carbon atoms) generally having longer decay times [14]. Nitrogen atoms tend to be less common in organic molecules compared to hydrogen or carbon. This means that 15N NMR spectra of small molecules, such as amino acids, often only contain one or two nitrogen signals (figure 4). As the nitrogen quantity for small molecules is low, 15N NMR can be used for the analysis and structure validation of macromolecules, such as DNA or proteins (figure 5), by comparing measured 15N NMR spectra to computer predictions.

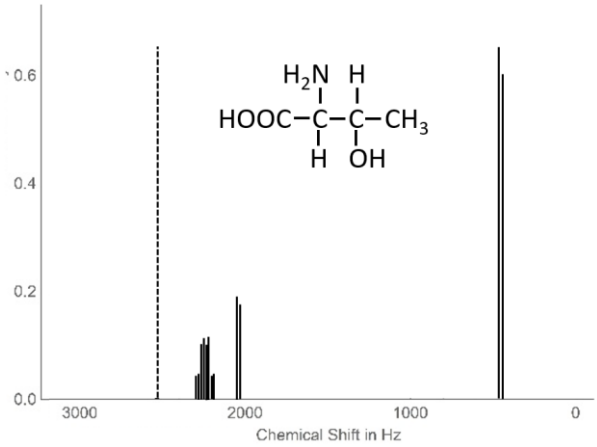


Figure 4: Comparison of NMR spectra of hydrogen-1 (continuous lines) and nitrogen-15 nuclei (dashed line) of L-allothreonine. The 1H NMR contains 3 frequency clusters with a total of 12 signals, whereas the 15N NMR spectrum contains only one signal. Data taken from [15].

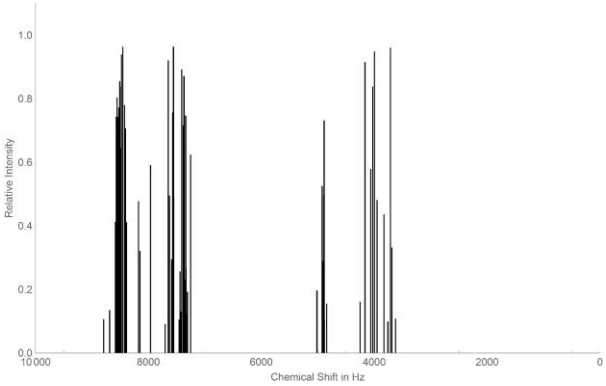


Figure 5: 15N NMR of an RNA segment of a lead-dependent ribozyme. Data taken from [16]. Sound example accessible via [17].

Sonification results of DNA bases contain a mix of high and mid-frequency content. Adenine, Guanine, and Cytosine contain an amine group resulting in a sharp triplet peak around 3 to 4 kHz when measured under a magnetic field of 11.7 Tesla, a magnetic field strength commonly used for NMR analysis. All DNA bases exhibit signals around 7 to 10 kHz due to their pyrimidine-type structural features, with Adenine and Guanine exhibiting additional resonances around 6 kHz. Sonification results of DNA bases contain a mix of high and mid-frequency content. Adenine, Guanine, and Cytosine contain an amine group resulting in a sharp triplet peak around 3 to 4 kHz when measured under a magnetic field of 11.7 Tesla, a magnetic field strength commonly used for NMR analysis. All DNA bases exhibit signals around 7 to 10 kHz due to their pyrimidine-type structural features, with Adenine and Guanine exhibiting additional resonances around 6 kHz. Due to chemical factors, not all resonance can be resolved as sharp peaks, leading some peaks to sound closer to band-pass filtered noise. The final timbre of 15N NMR sonification of DNA nucleotides is the sum of these noise bands, sharp peaks, and oscillating triplet patterns. DNA and RNA, being a sequence of a high number of DNA bases, exhibit similar resonance patterns as singular DNA nucleotides, however, each DNA nucleotide in

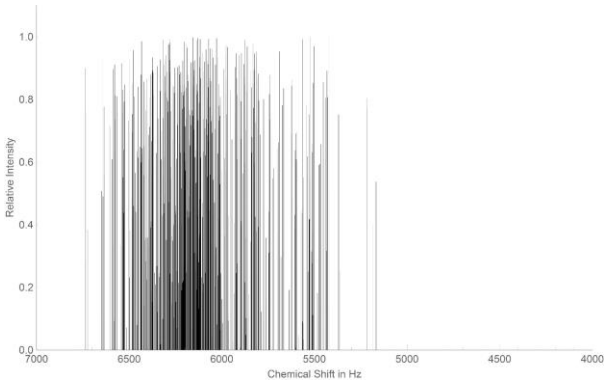


Figure 6: 15N NMR of an alkaline phosphatase protein. Data taken from [16]. Sound example accessible via [17].

a DNA sequence will have slight derivations of its resonance frequencies due to its chemical environment, leading to highly dense peak clusters (figure 5). 15N NMR spectra of proteins can be sonified, as well. In a protein, most nitrogen will be bound in the form of amides resonating in dense frequency clusters in a range of 5.5 to 6.5 kHz (figure 6).

2.2. 31P NMR

The NMR active isotope of phosphorous is phosphorous-31. 31P NMR typically returns frequency peaks in the range of up to 1 kHz for phosphates and up to 10 kHz for phosphines and phosphine oxide with decay times of up to 230 milliseconds (figure 7). Phosphorous is incorporated into a small percentage of organic structures of which ADP, ATP and the sugar backbone of DNA and RNA are the most common. Due to their inharmonic low-frequency resonances, sonification products of 31P NMR resemble the sound of metallic bells. Unlike 15N or 1H NMR spectra, features such as frequency clusters or sound wave pulsing of split peaks is less common in 31P NMR. Peak patterns of 31P NMR resemble more closely 13C NMR peak patterns, albeit at lower frequencies.

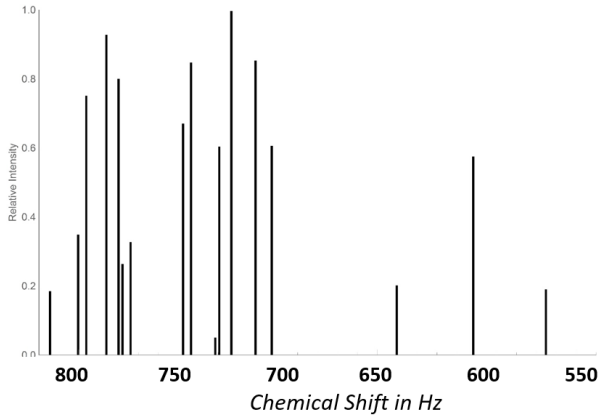


Figure 7: 31P NMR spectrum of an RNA strand of Bacillus subtilis**.** Data taken from [16]. Sound example accessible via [17].

2.3. 17O NMR

Like nitrogen-14, oxygen-17 is another quadrupolar nucleus with multiple resonance frequencies. Due to its nuclear composition, under a magnetic field, oxygen-17's spin states will occupy 6 distinct energy levels, leading to five

resonances when examined via NMR spectroscopy. 17O NMR is used comparatively little due to the low natural abundance of oxygen-17 and its fast FID decay time of only 20 milliseconds, leading to broad NMR peaks. It is, however, useful for examining biochemical systems as well for the investigation of structural and dynamic features of organic and metal-organic compounds [18]. Sonification of 17O NMR data predominantly yields noise-based textures with broad peaks occupying frequency ranges of zero to 4 kHz for esters and hydroxyl groups and between 13 and 15 kHz for carboxylic functional groups.

2.4. Data Sources

The Biological Magnetic Resonance Data Bank [16] contains more than 9000 freely accessible 15N spectra of proteins, peptides, DNA and RNA, as well as 125 phosphorous-31 datasets of DNA and RNA structures. The web database nmrshiftdb2 contains approximately 100 nitrogen-15 and 50 phosphorous-31 spectra of small molecules [19]. The Wiley NMR collection contains 7500 spectra for nitrogen-15, 21000 datasets for phosphorous-31 and more than 5500 entries for oxygen-17 [20], however, these data are not freely available. No free prediction or simulation software for 14N, 15N, 31P or 17O spectra exist, however, the necessary functionality is included in analysis software, such as ACD/Spectrus Processor [21] or NMRPredict [22], both offering free trial periods. 15N NMR spectra for proteins can be predicted via [23].

3. NMR SONIFICATION IN ACOUSMATIC MUSIC COMPOSITION

Acousmatic music is a form of electroacoustic music that is presented using loudspeakers only. Acousmatic composition focuses on the gestural, textural, and spatial development of sound material [24], exploring sound creation and structuring processes beyond harmony, pitch, or meter [25][26].

A number of acousmatic compositions based on NMR data have been composed [12]. In these compositions, NMR data have been used for purely aesthetic purposes, to create and experiment with these data-based timbres and their impact on compositional procedures. NMR data have also been implemented as auditory information on molecular states for 3D molecular representations in virtual reality [13]. One of the biggest challenges for the deployment of NMR data sonification in music and sound art is the abstract nature of NMR data [13], making it hard for the audience to link the sounds heard to the underlying data. However, this link between data and sonification-based sound material is important for art and music-based works, as it has been noted for various art installations that the audience gains a deeper appreciation of the artwork when the link between data and audio-visual spectacle is clear [27]. In those cases, the audience can perceive science-based installations as “science and not art” [28] resulting in a shift in interaction and appreciation depending on the individual’s preconceived notions of science [29]. Various pieces based on NMR-data have tried to supply the missing information in a variety of formats, including program notes, videos, workshops, talks, or interactive environments [12]. Another possibility, discussed in the following section, is the supply of necessary information via a hybrid drama of oral narration and acousmatic composition.

3.1. Acousmatic storytelling, narration, and sonification

Acousmatic storytelling is defined by Amelides as a combination of an acousmatic sound world with a voice narration [30]. The aim of music pieces based on acousmatic storytelling is not the exploration of sound transformations in space and time, as it is usual for acousmatic compositions, but the creation of a narrated sound world ‘closer to human experience’ [30], to contextualize and present cultural information and human experiences [30]. Examples of acousmatic storytelling are, H. Westernkamp’s *Kit’s Beach Soundwalk* or L. Ferrari’s *Far-West News*. By combining abstract and referential sounds with a spoken narrative, Amelides argues, acousmatic storytelling can be utilized as a vehicle for historical presentation [30], as a way to transform private meaning into public meaning [31] or to present a personalised story.

The narrator in an acousmatic piece can take many forms, from passive omniscient observers to real or fictional characters, recounting (or trying to remember) first-hand experiences [32]. Voice narration arcs are often created from assembling pieces of recorded interviews and can combine multiple different points of view (e.g. the same moments recounted by a son and his mother) to form a full narrative. The reoccurrence of narrators can, in those cases, act as leitmotifs [33]. Other sound materials used in acousmatic storytelling are predominately field recordings, archival sound material and cultural sound icons (e.g. musical quotations of national anthems or the sound of church bells) [30], each being able to create their own independent non-vocal narratives [34].

The combination of voice narration, field recordings, abstract sound and their subsequent transformation in time creates a multilayered story. It falls to the listener to combine the parallel streams of narration to create a multifaceted representation of events described in the acousmatic piece. A process that engages the listener to interact with the acousmatic piece more closely. Amelides argues that as acousmatic storytelling is less abstract than pure acousmatic composition, a wider audience is reached [30].

Principles of acousmatic composition have been proposed as guidance to increase the communicative and aesthetic properties of sonification work [35][36]. Acousmatic storytelling, with its focus on the contextualisation of information, the creation of shared stories and the use of acousmatic sound material to support information delivery, can be a valuable contribution to the discourse on sonification aesthetics. *On the Extinction of a Species*, described in the following section, is an exploration of such a combination of acousmatic storytelling and sonification.

3.2. On the Extinction of a Species

On the Extinction of a Species [37] is a 23-minute, 7.1 channel, acousmatic composition that combines voice narration, field recordings, analog synthesis and sound material created from the sonification of NMR data. The structure of the piece follows events surrounding the demise and potential resurrection of the passenger pigeon (table 1 and figure 8).

The first section of the piece depicts a time where billions of pigeons roamed North America, using the sound of pigeons taking flight and pigeon calls as cultural sound icons complimented by a stable harmonic sound texture.

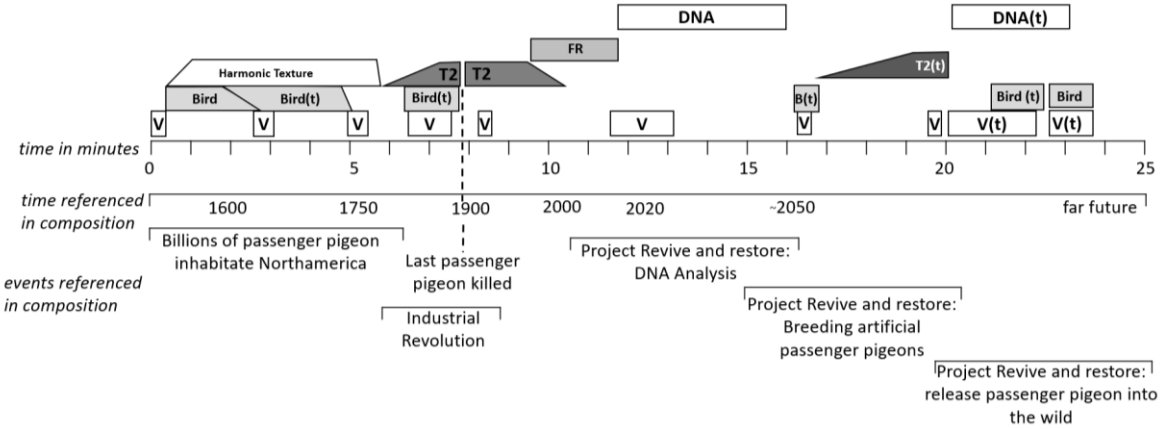


Figure 8: Structure of *On the Extinction of a Species* and its relation to referenced events. Sound material is abbreviated as follows: V – voice narration, Bird – sound of pigeon calls and birds taking flight, Harmonic Texture – Tonal music texture employed in section 1, T2 – noise-based sound texture, FR – field recording, DNA – sound material based on the sonification of DNA sequences of the passenger pigeon. Sound materials with the index (t) are sounds that were transformed based on the sound source indicated, e.g. V(t) is sound material created by the transformation of sound type V. Graphic adopted from [12].

Section	**1**	**2**	**3**	**4**	**5**
Time in piece	0 – 6 min	6 – 9 min	9 – 12 min	12 – 16 min	16 – 23 min
Time referenced	1000 – 1800 AD	1800 – 1950 AD	1950 AD – now	near future	far future
Events referenced	Billions of pigeons roaming free in North America	The industrial revolution and the extermination of the passenger pigeon	Ongoing industrialisation in a world without the passenger pigeon	Project 'Revive and Restore': analysis of the pigeon's genome	Digital resurrection and digital preservation of the passenger pigeon
Narration content	AI introduces itself and its purpose	AI describes the extermination of the passenger pigeon	-	AI proclaims its desires to resurrect the pigeon, by starting to sequence the pigeon's genome	AI allocates more and more resources to the resurrection of the pigeon, until its main functions, including its voice interface, break down.
Field recordings	-	-	City recordings	-	-
Cultural sound icons	Pigeon calls, sounds of birds taking flight, sounds of songbirds, with more obvious sound transformations towards the end of the section	Processed bird calls with strong glitch-type elements	Rain hitting a concrete floor as well as the sound of cars and trains	-	Transformations of the AI's voice to bird calls and reverse transformations.
Abstract sound material	Harmonic sound textures	Noise-based sound texture derived from trains and other heavy machinery	Heavily processed bird calls	**NMR data sonification of DNA sequences**	Transformed sound material based on DNA sequences

Table 1: Structure of *On the Extinction of a Species*, including the main sound elements used to create multilayered narrations.

Section 2 guides the listener through the industrial revolution and the extinction of the species in the late 18th century, using abstract noise-based sound textures reminiscent of an industrial landscape. Section 3 represents the 'now time', a bleak field recording of a noisy city soundscape. The fourth section of the piece envisions a future for the pigeon where its DNA is restored from preserved samples using genome sequencing, inspired by the real-world endeavors of 'Revive and Restore' [38]. The fifth section sees the story to its conclusion re-introducing and transforming sound material from the previous sections.

The piece is narrated by a fictional character, an artificial intelligence which is tasked with the retrieval, storage, and presentation of all data relating to extinct species. It addresses the audience directly and recounts the life of the passenger pigeon accompanied by the sound material previously described. In section 4, a key section of the piece, the AI decides to sequence the genome of the passenger pigeon from incomplete DNA data to create a virtual

passenger pigeon, alive in digital eternity. Inspired by the project 'Revive and Restore', the AI musically 'processes' sequences of DNA strands to complete the analysis of the passenger pigeon's genome. The section is an homage to Hayashi and Munakata's sonification experiments in which DNA bases were assigned midi notes to find patterns in DNA sequences [39]. However, in the piece, DNA bases are not assigned to musical notes but are represented by the sound based on the sonification of NMR data of DNA nucleotides. The voice narration links the sounds heard to their data origin, by calling out the name of each DNA nucleotide when its NMR-based sound is heard for the first time. Using voice narration, this section introduces data as part of its narrative arc. Section 5 continues to present DNA-based sound material and voice narration as main driving forces of the composition, however, both voice and sonification-based sound materials are being heavily transformed approaching the conclusion of the piece. The voice narration becomes erratic and the voice itself sporadically transforms into bird song. The sonified DNA sequences jump erratically in pitch, playback speed, and continuity. Using these sound transformations, this section aims to break down the clear distinction of voice narration and acousmatic sound world, merging both to symbolize the breakdown of the AI narrator.

During the composition of the piece, a compositional conflict arose from the divergence of the story told by the narrator and the "data story" of NMR sonification, the data story being the influence of the nucleotide's structure on the sound characteristics of the NMR-based sound. As NMR sonification is not explained as part of the narration, the audience has no way of knowing if and what sonification methodologies have been employed to produce these sounds, reducing the merit of NMR-based sounds in this context to their aesthetic features. With *On the Extinction of a Species'* focus on telling the story of the demise of passenger pigeon, the data story inherent to NMR sonification had to be omitted.

4. CONCLUSION

A methodology for the audification of 14N, 15N, 17O and 31P NMR data is presented and possible sonification products of DNA, RNA and protein sequences are characterized. The combined efficacy of acousmatic storytelling and NMR sonification for the presentation of data as part of a fictional story was explored using the acousmatic piece *On the Extinction of a Species*. By means of a spoken narrative, field recordings, cultural sounds, abstract and sonification-based sounds, a multilayered narrative was created.

On the Extinction of a Species is only an initial exploration in the combination of sonification and acousmatic storytelling and proposed future work includes the investigation of acousmatic storytelling in a more sonification-focused context, using data stories as a focal point for voice narration and the design of the acousmatic sound world.

5. REFERENCES

[1] R. Herrmann, and C. Onkelinx, "Quantities And Units In Clinical Chemistry: Nebulizer And Flame Properties In Flame Emission And Absorption Spectrometry (Recommendations 1986)", *Pure And Applied Chemistry*, 58.12, 1986, pp. 1737-1742.

[2] S. V. Pereverzev, A. Loshak, S. Backhaus, J. C. Davis, and R. E. Packard, "Quantum Oscillations Between Two Weakly Coupled Reservoirs Of Superfluid 3He", *Nature*, 388.6641, 1997, pp. 449-451.

[3] E. Hill, J. Cherson, S. Goldfarb, and J. A. Paradiso, "Atlas data sonification: a new interface for musical expression and public interaction", *38th International Conference on High Energy Physics*, 2016, pp. 1–3.

[4] S. Alexjander and D. Deamer, "The Infrared Frequencies Of DNA Bases: Science And Art", *IEEE Engineering In Medicine And Biology Magazine*, 18.2, 1999, pp. 74-79.

[5] T. Delatour, "Molecular Music: The Acoustic Conversion Of Molecular Vibrational Spectra", *Computer Music Journal*, 24, 2000, pp. 48-68.

[6] T. Delatour, "Molecular Songs", in *Molecular Aesthetics*, 1st edn (Karlsruhe: MIT - Press, 2013), pp. 293 – 311.

[7] M. Nilsson, "The DOSY Toolbox: A new tool for processing PFG NMR diffusion data," *Journal of Magnetic Resonance*, 200, pp. 26-302, 2009.

[8] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, and C. Knox, "HMDB 3.0 - The Human Metabolome Database in 2013," *Nucleic Acids Res*., vol. 1, p. 41, 2013.

[9] http://www.chemie.uni-erlangen.de/bauer/music4.html [Accessed 12 August 2018].

[10] J. W. Newbold, A. Hunt, and J. Brereton, "Chemical spectral analysis through sonification", *The 21th International Conference on Auditory Display*, 2015, pp. 329 – 330.

[11] F. Morawitz, "Molecular Sonification Of Nuclear Magnetic Resonance Data As A Novel Tool For Sound Creation", *Proceedings Of The International Computer Music Conference 2016*, 2016, pp. 6–11.

[12] F. Morawitz, "Portfolio of Compositions", *PhD Thesis*, University of Manchester, 2019.

[13] F. Morawitz, "An Art-Science Case Study On Sonification And Sound Design In Virtual Reality", *2018 IEEE 4th VR Workshop On Sonic Interactions For Virtual Environments (SIVE)*, 1, 2018.

[14] A. Wei, M. K. Raymond, and John D. Roberts, "5N Nuclear Magnetic Resonance Spectroscopy. Changes in Nuclear Overhauser Effects and T1 with Viscosity", *J. Am. Chem. Soc.*, 119, 1997, pp. 2915-2920.

[15] https://www.chemicalbook.com/SpectrumEN_144-98-9_1HNMR.htm [Accessed 29 March 2019].

[16] http://www.bmrb.wisc.edu/search/query_grid/NMR_param_grid.html [Accessed 29 March 2019].

[17] Sound material can be accessed for review purposes on https://www.dropbox.com/sh/ogr30l9e3a42e5s/AACCvy4ORcFWVFHR-ZLxr511a?dl=0

[18] I. P. Gerothanassis "Oxygen-17 NMR spectroscopy: Basic principles and applications", *Progress in Nuclear Magnetic Resonance Spectroscopy*", 56.2, 2010, pp. 95-197.

[19] http://nmrshiftdb.nmr.uni-koeln.de [Accessed 29 March 2019].

[20] https://sciencesolutions.wiley.com/solutions/wiley-spectra-lab/nmr [Accessed 29 March 2019].

[21] https://webstore.acdlabs.com/software-solutions/acd-spectrus-processor [Accessed 29 March 2019].

[22] http://mestrelab.com/software/mnova/nmr-predict/ [Accessed 29 March 2019].

[23] http://www.shiftx2.ca [Accessed 29 March 2019].

[24] D. Smalley, "Spectromorphology: Explaining Sound-Shapes", *Organised Sound*, 2, 1997, pp. 107-126.

[25] E. Varese and C. Wen-chung, "The Liberation of Sound", *Perspectives of New Music*, 5, 1, 1966, pp. 11-19.

[26] T. Wishart, *On Sonic Arts*, Routledge, 1996.

[27] D. Glowacki, P. Tew, J. Hyde, L. Kriefman, T. Mitchell, J. Price, and S. McIntosh-Smith, "Using Human Energy Fields to Sculpt Real-Time Molecular Dynamics", *Molecular Aesthetics*, 1.4, 2013, pp. 246 – 257; here, 249

[28] A. Vandso, "Listening To The World", *Soundeffects*, 1.1 2011, pp. 67-81.

[29] S. Emmerson, *Living Electronic Music*, Aldershot: Ashgate, 2007, pp. 35-57; here, 39.

[30] P. Amelides, "Acousmatic Storytelling", *Organised Sound*, 21.3, 2016, pp. 213-221.

[31] M. Jackson, *The Politics of Storytelling: Violence, Transgression and Intersubjectivity.* Tusculanum Press, 2002.

[32] W. C. Booth, *The Rhetoric of Fiction*, University of Chicago Press, 1961.

[33] J. Young, "Figures of Speech Oral History as an Agent of Form in Electroacoustic Music", *Leonard music journal*, 28, 2018, pp. 88-94.

[34] J. Andean, "Narrative modes in acousmatic music" *Organised Sound*, 21.3, 2016, pp. 192–203.

[35] P. Vickers, B. Hogg, "Sonification Abstraite/Sonificiation Concrete: An aesthetic perspective space for classifying auditory displays in the ars musica domain", *Proceedings of the 12th International Conference on Auditory Display*, 2006.

[36] F. Grond and J. Berger, "Parameter Mapping Sonification" in *The Sonification Handbook*, Logis Verlag, 2011.

[37] https://soundcloud.com/falk-morawitz/on-the-extinction-of-a-species-stereo-version [Accessed 29 March 2019].

[38] "About The Passenger Pigeon", Revive & Restore <http://reviverestore.org/about-the-passenger-pigeon/> [Accessed 15 January 2018]

[39] H. Kenshi, and N. Munakata, "Basically Musical", *Nature*, 310, 1984, p. 96.

EIGHT COMPONENTS OF A DESIGN THEORY OF SONIFICATION

Michael A. Nees

Lafayette College
Oechsle Hall
Easton, PA, 18040 USA
neesm@lafayette.edu

ABSTRACT

Despite over 25 years of intensive work in the field, sonification research and practice continue to be hindered by a lack of theory. In part, sonification theory has languished, because the requirements of a theory of sonification have not been clearly articulated. As a design science, sonification deals with artifacts—artificially created sounds and the tools for creating the sounds. Design fields require theoretical approaches that are different from theory-building in natural sciences. Gregor and Jones [1] described eight general components of design theories: (1) purposes and scope; (2) constructs; (3) principles of form and function; (4) artifact mutability; (5) testable propositions; (6) justificatory knowledge; (7) principles of implementation; and (8) expository instantiations. In this position paper, I examine these components as they relate to the field of sonification and use these components to clarify requirements for a theory of sonification. The current status of theory in sonification is assessed as it relates to each component, and, where possible, recommendations are offered for practices that can advance theory and theoretically-motivated research and practice in the field of sonification.

1. INTRODUCTION

In 1997, The Sonification Report [2] identified the lack of a theory of sonification as a major barrier to advancement of the field. In 2011, Walker and Nees's Theory of Sonification chapter [3] reiterated these concerns while pointing to incremental progress toward theory as a reason for optimism. Yet that incremental progress seems to have stalled, and the same dilemma remains with little evident momentum toward a resolution (see [4]). Although the reasons for the lack of sustained, intensive efforts toward theory-building in sonification are unclear, two potential explanations are disciplinary differences regarding the definition, role, and value of theory, and the fledgling nature of the field. Interdisciplinarity can be viewed as a strength of the auditory display community, but different disciplinary understandings of the forms and roles of theory might impede theory development [3]. Further, systematic progress in the field only began around 30 years ago [5].

Regardless of the reasons, sonification theory remains so underdeveloped that even the path to advance theory-building for sonification remains unclear. Recently, however, sonification researchers have begun to consider how lessons learned from broader areas of inquiry in design research might be translated to the study of sonification (see [6]). Design research has developed approaches for dealing with barriers similar to those facing sonification theory. This position paper draws connections between design theory and sonification theory in an attempt to identify paths toward advancing sonification theory. Regarding scope, design theory is most relevant to sonification for the purposes of conveying information in human-machine interfaces, and that is the focus of this paper. Although some of the discussion presented here incidentally might be applicable to sonification as art or composition, I have not attempted to examine or elaborate those connections.

2. STATUS OF SONIFICATION THEORY AND PRACTICE

Vickers recently said, "I think our knowledge of sonification design and theory is still fairly primitive" (as quoted in Quinton and colleagues [4]), and this sentiment seems to be widely held among sonification experts. The sonification literature, however, has featured various attempts at *theorizing*—what Weick [7] described as "activities like abstracting, generalizing, relating, selecting, explaining, synthesizing, and idealizing" (pp. 389) that result in pseudo-theory before fully- developed theory emerges. The sonification literature has produced scholarship with long lists of references cataloging variables and constructs [3], taxonomies [8], [9], design space maps [10], conceptual models [11], design guidelines [12], and frameworks for capturing design patterns [13], yet none of these are theories of sonification (see [14], [15]).

In some applied fields, a wealth of knowledge resides in practices that have not yet been codified formally as theory. Much has been written about gaps between theory and practice in design fields (e.g., [16]). Theoretical research—characteristic of academic approaches and whose purpose is to discover generalizable knowledge—has been criticized for being too abstract or removed to guide specific applications of knowledge in practice. Practice in design fields, on the other hand, is devoted to solving particular instances of immediate real-world problems and, as such, may result in one-off solutions that are not broadly shared and/or offer little contribution to re-usable knowledge. This creates a dilemma such that research discoveries may not be translated into practice (i.e., the knowledge is unknown, unused, or unusable for the practitioner), while designs used in practice may be produced on an ad hoc basis each time a problem is encountered with little awareness by the designer of why the resulting artifact was effective (or ineffective) and little concern for preserving the solution for future use by others.

In general, however, the field of sonification has been dominated by academic research. In fields characterized by theory-practice gaps, practitioner-designers solve problems in systems that are deployed or imminently will be deployed. For example, auditory alarms have been widely-used in applications for some time, and auditory alarms arguably have enjoyed the benefits of symbiotic exchanges in knowledge between research on auditory alarm design and

https://doi.org/10.21785/icad2019.048

information gleaned from analyzing the outcomes associated with auditory alarms as they have been used in widespread deployment (see, e.g., [17]). This example illustrates the theory-practice gap in the more traditional sense. There are few if any other examples, however, of ubiquitous deployment of sonification in practice (for a recent discussion, see [18]). Thus, for sonification, the theory-practice gap is different from the gap in other domains for which robust academic research and widespread practical applications co-exist.

The theory-practice gap in sonification as it stands currently seems to be more of a chasm between (1) academic research on potential sonification solutions to information display; (2) a (nearly complete) dearth of actual use of sonification in practice. Sonification as a field appears to be characterized to a nontrivial extent by the on-going development of sonification techniques in the absence of both generalizable theory and widespread (or any) use of sonification in practice. This type of approach—which I describe as *audio for the sake of audio*–produces novel sonification techniques, often without evaluation, as proofs-of-concept that audio artifacts can be produced using particular processes. This work generally appears to be accomplished by academics, yet it is largely atheoretical (in that in produces one-off concept-designs rather than programmatic, generalizable knowledge) and also does not appear to be driven by need or demand for an immediate audio solution to any practical problem, even if the design space does include legitimate practical problems that could be addressed using audio. Although proof-of-concept research can offer scholarly contributions to a field, it is representative of pre-theoretical stages of inquiry [19].

3. SONIFICATION THEORY AS DESIGN THEORY

In this pre-theory stage, a specification of the requirements of a theory of sonification could help to provide a framework in which progress toward a theory of sonification could proceed. The formulation of a theory of sonification currently appears to be an exercise in examining potentialities rather than extant real-world conditions, which complicates our ability to begin to articulate what a theory of sonification should accomplish. To some extent, design research already has grappled with this dilemma. In trying to make the case for design as science, Simon [20] said, "The natural sciences are concerned with how things are…Design, on the other hand, is concerned with how things ought to be…" (pp. 69). Design research would seem to offer a useful launch point for specifying the requirements of a theory of sonification [6].

Design fields deal with artifacts—artificial human creations in the form of technology, so theory-building occurs in a way that is different from the way theory develops in natural sciences (e.g., [15], [20]). Design theory must explain phenomena related to the form of the artifacts themselves, the creation of artifacts, and the use of artifacts in practice. Design theory helps to ensure that research contributes to programmatic accumulation of knowledge such that: (1) research findings can be integrated into a general framework of understanding; (2) successes and best practices are carried forward and expanded upon; and (3) mistakes are not repeated.

A focus on theory-building would have several benefits across the spectrum of research and practice in the field of sonification. Venable, for example, [21] placed theory-building as the center hub of a trio of other design science activities, including (1) inventing/creating the technology; (2) defining the problem space of the technology; and (3) evaluating the technology. The sonification literature to date, has emphasized the creation of sonification as a scholarly activity, with some (but perhaps less) attention paid to defining the problem spaces for sonification and evaluating sonification's ability to meet goals within a problem space—steps that will be imperative for sonification to be effective in practice (for a discussion, see [22]). To explain, the *audio for the sake of audio* approach has undertaken the creation of audio solutions under the assumption that an audio solution is necessary for some problem space—as assumption that may or may not hold across many potential applications (see, e.g., [23]). Further, only a fraction of the novel sonification approaches that have been presented have been subjected to rigorous evaluation. Sonification theory, as a central hub of activities related to inventing sonification, defining the problem space of sonification, and evaluating sonification, could help to provide the crucial link between existing activities in the field—particularly the pursuit of sonification methods and approaches—and other important but relatively under-developed activities related to evaluating the usefulness of sonification in real problem spaces for which audio may offer viable solutions.

4. GREGOR AND JONES' ANATOMY OF A DESIGN THEORY APPLIED TO SONIFICATION

In a highly-cited work on the requirements of a design theory, Gregor and Jones [1] synthesized multiple perspectives to derive eight essential components of design theories: (1) purposes and scope; (2) constructs; (3) principles of form and function; (4) artifact mutability; (5) testable propositions; (6) justificatory knowledge; (7) principles of implementation; and (8) expository instantiations. These components offer a relatively complete account of the meta-requirements for theory in design fields that emphasizes the unique challenges of formulating design theory. This section examines sonification theory and assesses the current completeness of sonification theory with respect to the components.

4.1. Purpose and Scope

Gregor and Jones [1] defined the purpose and scope of design theory as "the set of meta-requirements or goals that specifies the type of system to which the theory applies and in conjunction also defines the scope, or boundaries, of the theory" (pp. 325). The requirements enumerated in a statement of purpose and scope are "meta" in that they should generalize to all (or at least a class of) sonification artifacts rather than a particular instance.

As a useful starting point for considering the purpose and scope of a theory of sonification, The Sonification Report [2] stated, "Sonification is defined as *the use of nonspeech audio to convey information*. More specifically, *sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation"* (pp. 4, italics retained from original). Hermann [9] parsed this definition in a manner that is helpful for establishing the boundaries of a sonification theory. The set of artifacts to which a sonification theory applies are specified as *nonspeech audio* and implicitly the tools used to create the audio. This immediately excludes speech sounds from the scope of the theory. *Nonspeech audio* could include naturally occurring environmental sounds, music, etc., though the second statement further clarifies that sonification begins with *data*

relations that are *transformed* (presumably deliberately) into *perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation.* This further refines the scope to include only those sounds that have been deliberately created to represent relations in data for the purpose of understanding the data. Hermann further elaborated a set of four conditions that represent meta-requirements for an audio artifact to fall under the purview of a theory of sonification: (1) the sound must represent "objective properties or relations in the input data"; (2) the transformation from data to sound must be systematic such that "there is a precise definition provided of how the data…caused the sound to change"; (3) the sound must be reproducible; and (4) the sonification system must be reusable with the same or different data (for a detailed discussion of these conditions, see [9]).

As such, the purpose and scope of a theory of sonification were apparent in the early definitions of sonification (e.g., [2]). Further, the field has examined and debated the boundaries of sonification (e.g., [9]). In this regard, sonification theory has achieved a degree of maturity that offers a solid grounding regarding its purpose and scope. A theory of sonification explains how, when, and why to use nonspeech sounds to convey information in systems using audio artifacts that are objective, systematic, reproducible, and reusable. To interpret further, this scope includes both audio-only and multimodal use of nonspeech sounds and excludes speech sounds, music, and incidentally occurring environmental sounds except to the extent that a consideration of these excluded factors might impact the use of nonspeech sounds to convey information. Edge cases (e.g. spearcons, see [24]) may challenge our understanding of the boundaries of a theory, and there is some ambiguity in the field about what it means to "convey information" (see section 6.5 below). Also, it is not clear if sonification could be captured in a single grand design theory, or if many related theories will be required for different uses of sonification. Thus, further refinement of the purpose and scope of sonification theory may occur in the future. Yet the purpose and scope of a theory of sonification appear to be articulated in a manner that is clear enough for mature theory to develop.

4.2. Constructs

Gregor and Jones [1] defined constructs as "representations of the entities of interest in the theory…these entities could be physical phenomena or abstract theoretical terms" (pp 325). Constructs in a design theory must entail a broad conceptualization of representations to capture the entities of interest. Constructs in a theory of sonification must include terms used to describe the audio artifact, terms used to describe the perception of the artifact by a listener, and terms used to describe the tasks to be undertaken by a listener. For example, a theory might explain how to use earcons (the audio artifact construct) to capture attention (a psychological construct) during monitoring (a task construct). Each construct would in turn need to be operationalized with a formal way of quantifying or identifying the construct. One could arguably extend the entities of interest in a theory of sonification to include terms used to describe the data from which the sonification is derived, etc., but those are discussed here under 6.6 below.

Even before the first ICAD conference, researchers had begun to operationalize sonification constructs such as auditory icons [25] and earcons [26]. More recently, Nees and Walker (e.g., [3], [11], [27]) have presented overviews of auditory display that taxonomize types of auditory displays, tasks to-be-accomplished with auditory displays, and listener variables to consider when designing auditory displays. de Campo's Sonification Design Space Map [10] offered a framework to define and relate the types of audio artifacts produced by sonification to one another. Early work by Barrass [28] and recent work by Verona and Peres [22] emphasized the critical role of task demands in the design of auditory displays and offered examples of how to use task analysis to precisely hone in on task constructs. Perceptual research in psychology has produced decades of literature on constructs relevant to auditory perception (see, e.g., [29], [30]). Although refinement of constructs to resolve confusions represents an on-going process in the development of a theory of sonification (see, e.g., [9], [31]) the constructs of sonification appear to be articulated in a manner that is clear enough for mature theory to develop.

4.3. Principles of Form and Function

Gregor and Jones [1] defined this component as "the principles that define the structure, organization, and functioning of the design product or design method…this component gives an abstract 'blueprint' or architecture for the construction of an…artifact" (pp. 326-327). In the sonification literature, Barrass described several general principles of design [32]. Specific guidelines have been provided for designing auditory alarms [33], and an international standard exists for medical device alarms [34]. A sustained critical examination of these guidelines has occurred (see [35], [36]). Guidelines exists for auditory graphs and tables ([12], [37]), earcons [38], model-based sonification [39], and general use of nonverbal sounds in interfaces [40].

Still, the available principles tend to be articulated in broad terms, and most represent an initial or preliminary attempt to codify the blueprints for sonification. For example, the standards for medical device alarms—one of the more formal and specific statements of principles of auditory design available—have been legitimately criticized for producing poor designs (e.g., [41], [42]). A lack of usable guidance is a contributor to the theory-practice gap in human-computer interaction in general [16] and in sonification specifically [43]. As such, principles and guidelines for designing sonification, though present, remain incomplete. Improved and expanded principles will be required as sonification theory develops.

4.4. Artifact Mutability

Simon [20] said "…a science of artificial phenomena is always in imminent danger of dissolving and vanishing" (pp. 68). Since sonification and its related artifacts depend upon technology, the artifacts explained in a theory of sonification have the potential to exist in a tentative state that, in some cases at least, is subject to extinction from unanticipated changes that can arrive capriciously. For example, since sonification tools generally have been created independently from mass-marketed software and hardware, updates to the infrastructure supporting the tools can render tools unusable until the developer of the tool—often one researcher or lab—dedicates time to updating the tool. To sustain sonification tools requires a commitment from a researcher or lab to devote resources more or less continuously toward addressing difficulties that arise from software and hardware changes over which the tool developer often has little or no control.

This is the work required to simply keep the tools usable before any resources are devoted to substantive improvements or modifications to the tools.

As a result of these challenges, the field seems to be characterized by a proliferation of one-off, novel tools and techniques whose usable lifespan is fleeting. In fact, many of the sonification tools described in ICAD proceedings are never publicly released for use by other researchers or practitioners, much less supported and updated over time. Tools (and in some cases their associated artifacts) effectively become extinct when their developer no longer has the interest in supporting and/or resources to support the tool for other users, so designers new to sonification face considerable technical obstacles to using sound in applications (see [18], [43]). Sonification might enjoy more widespread use and deployment, which in turn would broaden the base of knowledge and feed back into the development of theory, if more general audiences (e.g., in user interface design, user experience, etc.) had access to sonification tools with sustained technical support. A consideration of the mutability of artifacts seems to be a particularly underdeveloped component of a theory of sonification.

4.5. Testable Propositions

A theory should create new, testable predictions. Gregor and Jones [1] argued that the most general predictions of a design theory are that the goals and purpose (see section 4.1 above) will be met when the design principles of the theory (see section 4.3 above) are applied correctly. The specificity of predictions can vary considerably across different applications of a theory, but a theory should be capable of providing a framework for guiding action and a set of criteria against which the success of that action can be judged. A mature field of inquiry will focus its scholarship efforts toward examining the testable propositions of theory to refine and qualify the theory, resolve contradictions, etc.

In the sonification literature, this component is closely related to discussions regarding how to evaluate sonifications. Bonebright and colleagues, in particular, have presented practical overviews of methods for evaluating sonifications (see [44]), and evaluation has been recognized as a critical activity for the effective design of auditory displays (e.g., [27]). Yet the issue of evaluation holds a somewhat contentious place in the field. Supper [45] has documented an epistemological rift in the auditory display community between advocates of systematic user evaluation and those who believe formal evaluation is unnecessary. Effectively the difference lies in empirical versus heuristic approaches to evaluation. Testing advocates value evidence from a representative sample of users, whereas their detractors believe that an "expert" or "trained" listener can use her knowledge as a heuristic substitute for objective evidence from formal evaluations. In general, the former perspective is more characteristic of theory-building; for example, Supper [45] identified "theoretical contextualization" as a quality desired by proponents of user testing. Heuristic evaluation can be important for the design evaluation process and can provide information that is different from formal user testing (e.g., [46]). Yet it is not clear how a field in a pre-theoretical stage could formulate broadly successful heuristics in the absence of broadly successful theory. Critics of user evaluations take the position that the intended information is obviously available to the listener in the audio artifacts they produce. Currently, the heuristic evidence that an otherwise unevaluated sonification conveyed information seems to be that the creator of the sonification believes as much, which ignores the possibility that the positive evaluation could result from well-documented threats to validity [47]. For the foreseeable future, theory-driven approaches likely will require formal, rigorous evaluation, though a standardization of heuristic principles of evaluation for sonification could be useful.

As Gregor and Jones explained, testable propositions "can take the general form: 'If a system or method that follows certain principles is instantiated then it will work, or it will be better in some way than other systems or methods'" (pp. 327). A fair critique of sonification research is that it runs the risk at times of becoming an industry of designs that compare audio artifacts to other audio artifacts (or nothing at all) under the assumption that an audio approach is inherently valuable, regardless of the value added as defined by task- and goal-specific criteria (for a discussion, see [22]). Novel sonification approaches should be met with scrutiny until evidence is provided that such approaches have value for meeting the goals of sonification for a particular task (see [23]).

The act of formally testing propositions alone will not necessarily produce an adequate knowledge base for a theory of sonification, because the quality of the evidence produced by testing propositions is affected by the quality of the research undertaken. There is reason to be concerned about the quality standards of user testing in the current sonification literature. Related domains of study have recently experienced a reckoning of sorts regarding the reproducibility and replicability of their findings. The "replication crisis" in psychology has revealed methodological and statistical shortcomings that have called into question a surprisingly high amount of empirical evidence in the field (see [48]). Subdisciplines in psychology (e.g., cognitive psychology) that are somewhat aligned with sonification research (with respect to both content and typical methodologies) generally have fared better under replication scrutiny than other sub-disciplines, such as social psychology (see [48], [49]). But data from studies in psychology—a field that explicitly trains students in statistics and research methods and generally requires empirical evidence (the sonification equivalent of user testing) to warrant publishable contributions —appear to be unreliable at unacceptable (or at least previously underestimated) levels.

There is evidence to suggest that interdisciplinary fields like sonification also should be concerned about research quality. As an illustrative snapshot, of the 29 papers (excluding the editor's introduction) currently archived from the 2018 ICAD conference[1], roughly half (n = 15) presented a formal user evaluation. Of note, five papers purported to introduce a new or novel sonification approach or technique, with just two of those papers providing a formal evaluation of the new approach. In the papers reporting evaluations, the median sample size was N = 17 (ranging from 1 to 24). Although adequate sample size depends on a number of factors, it appears that research reporting evaluations at ICAD tends to be underpowered. This is problematic not only in that null results are ambiguous (i.e., they could result from lack of effects or lack of power), but also because positive findings in underpowered research can be more likely to represent Type I (false positive) statistical errors [50].

Sample size is an imperfect surrogate for overall research quality, but as one indicator, the tendency for

[1] https://smartech.gatech.edu/handle/1853/60062

sonification studies to be underpowered suggests there is reason for concern regarding the quality of research findings in the sonification community. Sonification researchers have yet to apply the scrutiny to their own body of evidence that is currently being applied to the base of evidence in other fields such as psychology. Given the relatively lax research standards in sonification research (e.g., empirical testing of designs is viewed as optional and small sample sizes are typical), however, it seems difficult to imagine that replication and reproducibility of findings in sonification research would fare better than psychology, and it is easy to imagine that sonification research would fare worse.

In summary, the testable propositions of a theory of sonification extend readily from the definition of the term *sonification* (see section 5.1). There appears to be disagreement about the value of testing, however, which has resulted in disparate evaluation approaches in the field. Given recent replicability issues in related fields such as psychology, there also is reason to be concerned about the existing knowledge base for sonification.

4.6. Justificatory Knowledge

Design theories draw upon existing disciplinary bases of knowledge to inform and explain design decisions. Sonification's interdisciplinary roots require a theory of sonification to draw upon relevant theories in auditory perception and cognition, music, computer science, acoustics, data science, etc. This justificatory knowledge should support a theory of sonification not only by providing guidance on *how* to design and implement sonification, but also by explaining *why* those design and implementation strategies will satisfy the goals of the theory (see [1]). To some extent, then, the adequacy of a theory of sonification will be contingent upon the adequacy of its supporting justificatory knowledge from theories in related disciplines—what Walls and colleagues [51] described as "…*kernel theories* from natural or social sciences which govern design requirements" (pp 42; italics retained from original).

Although a complete review of the types of justificatory knowledge that could support a sonification theory is beyond the scope of this paper, several overviews have provided markers (e.g., [2], [3]). Presumably, a theory of sonification will draw connections with related work in all three elements of the auditory display system (information, display, and listener, see [52]), and representative examples of each approach can be found in the literature. In one of the earliest examples of auditory display research, Pollack [53] applied principles of information theory to benchmark performance with auditory displays. McGookin and Brewster [54] used Bregman's Auditory Scene Analysis [55] theory to improve the recognizability of co-occurring earcons. Walker and Kramer [56] provided explicit linkages between the knowledge base of traditional psychoacoustics and auditory display. In general, a rich base of justificatory knowledge is available to support the design and implementation of sonification, but translational work remains needed to elicit relevant and useful connections with related areas of inquiry.

4.7. Principles of Implementation

Gregor and Jones [1] defined this component as "the means by which the design is brought into being─a process involving agents and actions" which could include "…an abstract, generic design method or development approach" (pp. 328). This is different from the component outlined in section 4.3, which described the principles for creating specific types of sonifications. For sonification theory, principles of implementation entail both (1) generic principles to guide the design cycle for sonifications; and (2) generic principles for the deployment of sonifications. There are several good examples of the former in the sonification literature, but there are few if any examples of the latter.

General descriptions of sonification design cycles exist. Barrass's [57] sonification design patterns approach provided a narrative framework for the sonification design process. Johannsen [58] described a "life cycle development of auditory displays." Anderson [59] described a decision-making process for designing sonification. Watson and Sanderson [60] detailed how the process of ecological interface design could be applied to the development of sonification for monitoring patients under anesthesia. Nees and Walker [27] described a process for designing auditory displays for in-vehicle technologies. Each of these approaches offered generic guidance for designing sonifications.

Guidance on how to implement sonification within existing sociotechnical ecosystems is less readily available, perhaps because there are few examples of deployments of sonification at scale. Some general implementation advice (e.g., regarding strengths and limitations of audio) was offered by Kramer [52]. Edworthy [35] has discussed the implementation of auditory alarms from a holistic, systems-thinking perspective (e.g., by considering the potential negative consequences of the proliferation of alarms across devices in real world implementation, also see [36]). Tomlinson and colleagues [61] reported on lessons learned during a two-year deployment of auditory graphs in classrooms for students with visual impairments (also see [62]). Previously, the SonEnvir project also reported lessons learned from an attempt to integrate sonification broadly into work in multiple disciplines [63]. Despite the ambitious nature of these projects, there is not currently enough evidence available to formulate generic advice on how to deploy sonifications in sociotechnical systems—particularly from a macro-ergonomics perspective that addresses social, organizational, and technical challenges in less than ideal implementation circumstances. Such advice does exist in other domains (e.g., [64]) and could serve as a model for how sonification theory might develop in this regard.

4.8. Expository Instantiation

Gregor and Jones [1] stated, "A realistic implementation contributes to the identification of potential problems in a theorized design and in demonstrating that the design is worth considering" (pp. 329). Their conceptualization of this component included mock-ups, prototypes, and simulations—examples of the artifacts described and explained by the theory that help to illustrate the principles of the theory. In this regard, sonification research has produced numerous instantiations of sonifications, and this activity has been particularly valued by the sonification community. As Gregor and Jones point out, however, "If the instantiation or artifact is all that there is, rather than a theory of design…the level of knowledge is that of a craft-based discipline" (pp. 329). As sonification moves from a pre-theoretical stage to more developed theoretical positions, presumably the instantiations of sonification will be adapted to align with theoretical principles. As described above, sonification research has resulted in a proliferation of sonification examples and prototypes, so the on-going development of expository

instantiations should remain a strength of sonification research into the future.

5. CURRENT STRENGTHS AND WEAKNESSES OF SONIFICATION THEORIZING

Considering sonification theory as design theory under the rubric developed by Gregor and Jones [1], some areas of strength emerge regarding the current state of sonification theory. In general, sonification research appears to have adequately articulated purposes and scope, and a shared understanding of constructs has emerged. Sufficient justificatory knowledge exists to advance sonification theory, and sonification research has produced a proliferation of potential expository instantiations. These four areas represent relative strengths for theory-building.

Several of the components appear to be relatively underdeveloped at this time. Although principles of form and function have been proposed in the sonification literature, these principles have not been widely tested and refined. Further, existing principles may be articulated at a level that is too general for designing sonifications for many practical applications (see [4], [43]). Similarly, the principles of implementation in the sonification literature have been expressed in general terms (e.g., by specifying circumstances when audio is an appropriate design choice). The lack of deployment at scale of most types of sonification has left large gaps in knowledge regarding *how* to implement sonification in practice, particularly with respect to organizational, social, and technical challenges that may arise. Thus, principles of form and function and principles of implementation currently have achieved a preliminary status that will need further refinement and development to advance a theory of sonification.

Our current understanding and practices appear to be especially weak for at least two of the components. Although current theorizing in the sonification literature does produce testable propositions, current research practices often leave testing and evaluation of theoretical claims optional. Further, sonification researchers have not begun to consider the reproducibility and replicability of their base of knowledge, so the quality of evaluations to date may be suspect. Related fields (e.g. psychology) have had empirical findings called into question, and the psychology literature has emphasized rigorous experimental methods and quantitative analysis moreso than the sonification literature. There is reason to be concerned that replication problems also affect the sonification literature. Finally, considerations of artifact mutability have been almost entirely absent from the sonification literature. As a design field that relies on technology in the production and delivery of its artifacts, sonification theory will need to seriously grapple with solving problems related to supporting and sustaining sonification and its tools in the face of rapidly-changing technological landscapes. Currently, many sonification tools never become available to other researchers and practitioners, and one-off tools are prone to quickly become inviable. A full consideration of the lifecycle of sonification artifacts and tools must consider design, deployment, mutability, and eventual degradation of the sounds and the tools that make them.

6. RECOMMENDATIONS FOR THEORY-BUILDING

A number of potential recommendations for theory-building in sonification can be gleaned from a consideration of sonification theory in the context of Gregor and Jones' [1] anatomy of design theories. Explicit consideration of each component at the outset of projects could help ensure that research advances theory.

Regarding purposes and scope, before design begins the criteria for success (i.e., the information to-be-conveyed by a sonification) should be defined, and these criteria should be linked to task- and goal-specific outcomes. This process likely will involve the specification of relevant constructs. Justificatory knowledge also should be made as explicit as possible at this stage of research.

Where possible, the principles of form and function that were used in the design of a sonification should be made explicit, and successes or failures of principles should be noted explicitly. Where appropriate, new principles and suggested refinements of old principles should be offered.

Robustness against changing circumstances—especially those related to software—appears to be a particular vulnerability of sonification. A deeper consideration of artifact mutability likely would involve stronger commitments to making sonification tools and examples (including design patterns) openly available. Repositories (e.g., Github, Open Science Framework) are a superior option to personal webpages, which often become defunct despite the best intentions of researchers at the time of creation and publication. Sustaining tools, examples, and design patterns over time likely will require a concerted effort involving collaborations across the sonification community. General or multi-purpose sonification toolkits possibly could generate broader interest (e.g., from HCI/UX professionals) than one-off, specific tools. That interest, in turn, might increase the collective motivation and commitment of the sonification community to sustaining and regularly updating such toolkits.

To advance theory, sonification research must formally test the extent to which a sonification tool or audio artifact meets the stated purposes of sonification. To the extent that the purpose of sonification is to convey information to listeners, it is incumbent upon researchers to provide evidence that the intended information has, in fact, been conveyed. Where possible, evaluation criteria should be linked to objective real-world outcomes (clinical outcomes, benchmarking against current best practice, etc.). The specific criteria that must be met in the evaluation phase will vary across use scenarios and stages of research (early/exploratory versus advanced/confirmatory, etc.). If a particular application domain is, for example, dominated by visual displays, it seems of little use to compare one sonification prototype to another unless both are also referenced to the level of performance achieved using existing approaches or the required level of performance for a particular task while using the display. One sonification could be statistically superior to another, with both falling short of criteria related to real-world usefulness.

Sonification as a field also likely would benefit from an examination of the reproducibility of its research findings. This might include the development of formal statements regarding best practices in research methods and statistical analyses. For example, psychology has seen a push toward pre-registration of research studies, open sharing of

research data, and reforms of statistical practices[2]. Further, some have begun to advocate for (and coordinate) replication studies of important findings by students as part of training[3], which partially addresses the problem of lack of incentives for researchers to invest resources in replication studies. It would be in the interest of sonification researchers to follow these developments closely and adopt practices that improve research quality.

Beyond user testing—and taking into account the resistance to user testing in some quarters of the field—the development of formal heuristic forms of evaluation could potentially be of value for sonification. Useful heuristics may be difficult to derive until other areas described in this paper are developed more completely. At some point in the future when theoretical evidence has accumulated, however, a formal heuristic checklist for sonification design (like those in HCI/UX[4]) could be useful.

Sonification remains mostly unexamined at any scale of implementation in practice, because significant barriers exist to implementing sonification in design [18]. In perhaps the only systematic attempt to understand how audio is viewed in design practice, Frauenberger, Stockman, and Bourguet [43] conducted a survey regarding the use of audio in interface design. Barriers included the lack of standards, lack of successful design patterns, and lack of guidance for using audio, and lack of appropriate tools for design. Research to follow-up and expand upon the questions posed by Frauenberger et al. [43] seems warranted. Ultimately, a great deal more information is needed to understand how to support the delivery of sonification across organizational, social, and technical contexts, because so little information is available about actual implementation of sonification beyond lab studies. To address this gap in knowledge likely will require sustained, coordinated efforts across multiple research labs. Indeed, overcoming many of the obstacles to the development of sonification theory likely will require intensive collaboration. From the perspective of theory development, efforts to thoroughly evaluate and technically support select promising sonifications through a deployment life cycle of actual use would seem to be more valuable than the one-off, proof-of-concept projects that have characterized a considerable proportion of research in the field to date.

7. REFERENCES

[1] S. Gregor and D. Jones, "The Anatomy of a Design Theory," *J. Assoc. Inf. Syst.*, vol. 8, no. 5, pp. 313–335, May 2007.

[2] G. Kramer *et al.*, "The Sonification Report: Status of the Field and Research Agenda. Report prepared for the National Science Foundation by members of the International Community for Auditory Display," 1999.

[3] B. N. Walker and M. A. Nees, "Theory of sonification," in *Principles of Sonification: An Introduction to Auditory Display*, T. Hermann, A. Hunt, and J. Neuhoff, Eds. Berlin, Germany: Logos Publishing House, 2011, pp. 9–39.

[4] M. Quinton, I. McGregor, and D. Benyon, "Investigating effective methods of designing sonifications," in *Proceedings of the 24th International Conference on Auditory Display (ICAD 2018)*, Michigan Technological University, 2018.

[5] S. P. Frysinger, "A brief history of auditory data representation to the 1980s," in *Proceedings of the 11 International Conference on Auditory Display (ICAD 2005)*, Limerick, Ireland, 2005, pp. 410–413.

[6] M. Jeon, B. N. Walker, and S. Barrass, "Introduction to the Special Issue on Sonic Information Design: Theory, Methods, and Practice, Part 1," *Ergon. Des.*, vol. 26, no. 4, pp. 3–3, Oct. 2018.

[7] K. E. Weick, "What Theory Is Not, Theorizing Is," *Adm. Sci. Q.*, vol. 40, no. 3, pp. 385–390, Sep. 1995.

[8] T. Letowski *et al.*, "Human factors military lexicon: Auditory displays," Army Research Laboratory Technical Report, 2001.

[9] T. Hermann, "Taxonomy and Definitions for Sonification and Auditory Display," in *Proceedings of the 14th International Conference on Auditory Display,* Paris, France, 2008.

[10] A. de Campo, "Toward a sonification design space map," in *Proceedings of the 13th International Conference on Auditory Display (ICAD 2007)*, Montreal, Canada, 2007, pp. 342–347.

[11] M. A. Nees and B. N. Walker, "Listener, task, and auditory graph: Toward a conceptual model of auditory graph comprehension," in *Proceedings of the 13th International Conference on Auditory Display (ICAD 2007)*, Montreal, Canada, 2007, pp. 266–273.

[12] L. M. Brown, S. A. Brewster, S. A. Ramloll, R. Burton, and B. Riedel, "Design guidelines for audio presentation of graphs and tables," in *Proceedings of the 2003 International Conference on Auditory Display*, Boston, MA, 2003.

[13] C. Frauenberger, T. Stockman, and M. L. Bourguet, "Pattern Design in the Context Space A Methodological Framework for Auditory Display Design," in *Proceedings of the 13th International Conference on Auditory Display (ICAD 2007)*, Montreal, Canada, 2007.

[14] R. Sutton and B. Staw, "What Theory is Not," *Adm. Sci. Q.*, vol. 40, no. 3, pp. 371–384, 1995.

[15] J. Hooker, "Is Design Theory Possible?," *J. Inf. Technol. Theory Appl. JITTA*, vol. 6, no. 2, Jul. 2004.

[16] L. Colusso, C. L. Bennett, G. Hsieh, and S. A. Munson, "Translational Resources: Reducing the Gap Between Academic Research and HCI Practice," in *Proceedings of the 2017 Conference on Designing Interactive Systems*, New York, NY, USA, 2017, pp. 957–968.

[17] J. Edworthy, "Medical audible alarms: A review," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 3, pp. 584–589, May 2013.

[18] M. A. Nees, "Auditory Graphs Are Not the 'Killer App' of Sonification, But They Work," *Ergon. Des.*, vol. 26, no. 4, pp. 25–28, Oct. 2018.

[19] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Q.*, pp. 337–355, 2013.

[20] H. A. Simon, "The Science of Design: Creating the Artificial," *Des. Issues*, vol. 4, no. 1/2, pp. 67–82, 1988.

[21] J. Venable, "The role of theory and theorising in design science research," in *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology (DESRIST 2006)*, 2006, pp. 1–18.

[22] D. Verona and S. C. Peres, "A Comparison between the Efficacy of Task-Based Vs. Data-Based sEMG Sonification Designs," in *The 23rd International Conference on Auditory Display (ICAD 2017)*, Pennsylvania State University, 2017.

[23] J. Edworthy, "Does sound help us to work better with machines? A commentary on Rautenberg's paper 'About the importance of auditory alarms during the operation of a plant simulator,'" *Interact. Comput.*, vol. 10, pp. 401–409, 1998.

[2] https://improvingpsych.org/

[3] https://osf.io/wfc6u/

[4] https://www.nngroup.com/articles/ten-usability-heuristics/

[24] B. N. Walker *et al.*, "Spearcons (Speech-Based Earcons) Improve Navigation Performance in Advanced Auditory Menus," *Hum. Factors*, vol. 55, no. 1, pp. 157–182, 2013.
[25] W. W. Gaver, "Auditory Icons: Using Sound in Computer Interfaces," *Human–Computer Interact.*, vol. 2, no. 2, pp. 167–177, Jun. 1986.
[26] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and Icons: Their Structure and Common Design Principles," *Human–Computer Interact.*, vol. 4, no. 1, pp. 11–44, Mar. 1989.
[27] M. A. Nees and B. N. Walker, "Auditory displays for in-vehicle technologies," in *Reviews of Human Factors and Ergonomics*, P. Delucia, Ed. Thousand Oaks, CA: Sage Publishing/Human Factors and Ergonomics Society, 2011, pp. 58–99.
[28] S. Barrass, "TaDa! demonstrations of auditory information design," presented at the Proceedings of the 1996 International Conference on Auditory Display, 1996.
[29] J. G. Neuhoff, *Ecological Psychoacoustics*. New York: Academic Press, 2004.
[30] C. L. Baldwin, *Auditory Cognition and Human Performance: Research and Applications*. Boca Raton, FL: CRC Press, 2012.
[31] M. A. Nees, "Have we forgotten auditory sensory memory? Retention intervals in studies of nonverbal auditory working memory," *Front. Psychol.*, vol. 7, 2016.
[32] S. Barrass, "Some golden rules for designing auditory displays," in *Csound Textbook*, Cambridge, MA: MIT Press, 1998.
[33] R. D. Patterson, "Guidelines for auditory warning systems on Civil Aircraft," 1982.
[34] I. E. Commission and others, *IEC 60601-1-8: Medical electrical equipment–General requirements, tests and guidance for alarm systems in medical electrical equipment and medical electrical systems*. Geneva, Switzerland: Author, 2005.
[35] J. Edworthy, "The design and implementation of non-verbal auditory warnings," *Appl. Ergon.*, vol. 25, no. 4, pp. 202–210, Aug. 1994.
[36] J. Edworthy and E. Hellier, "Fewer but better auditory alarms will improve patient safety," *Br. Med. J.*, vol. 14, no. 3, pp. 212–215, 2005.
[37] J. H. Flowers, "Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions," in *Proceedings of the 11 International Conference on Auditory Display (ICAD 2005)*, Limerick, Ireland, 2005.
[38] S. A. Brewster, P. C. Wright, and A. D. Edwards, "Experimentally derived guidelines for the creation of earcons," in *Adjunct Proceedings of HCI*, 1995, vol. 95, pp. 155–159.
[39] T. Hermann, "Model-based sonification," in *The Sonification Handbook*, 2011, pp. 399–427.
[40] J. Hereford and W. Winn, "Non-speech sound in human-computer interaction: A review and design guidelines," *J. Educ. Comput. Res.*, vol. 11, pp. 211–233, 1994.
[41] A. N. Wee and P. M. Sanderson, "Are melodic medical equipment alarms easily learned?," *Anesth. Analg.*, vol. 106, no. 2, pp. 501–507, 2008.
[42] P. Lacherez, E. L. Seah, and P. M. Sanderson, "Overlapping melodic alarms are almost indiscriminable," *Hum. Factors*, vol. 49, no. 4, pp. 637–645, 2007.
[43] C. Frauenberger, T. Stockman, and M.-L. Bourguet, "A survey on common practice in designing audio user interface," in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers*, 187-194, 2007.
[44] T. L. Bonebright and J. H. Flowers, "Evaluation of auditory displays," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Logos Verlag Berlin, Germany, 2011.
[45] A. Supper, "'Trained ears' and 'correlation coefficients': A social science perspective on sonification," in *Proceedings of the 18th International Conference on Auditory Display (ICAD 2012)*, Atlanta, GA, 2012.
[46] W. Tan, D. Liu, and R. Bishu, "Web evaluation: heuristic evaluation vs. user testing.," *Int. J. Ind. Ergon.*, vol. 39, no. 4, pp. 621–627, 2009.
[47] D. T. Campbell and J. C. Stanley, *Experimental and quasi-experimental designs for research*. Ravenio Books, 2015.
[48] O. S. Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, Aug. 2015.
[49] G. Mitchell, "Revisiting Truth or Triviality: The External Validity of Research in the Psychological Laboratory," *Perspect. Psychol. Sci.*, vol. 7, no. 2, pp. 109–117, Mar. 2012.
[50] K. S. Button *et al.*, "Power failure: Why small sample size undermines the reliability of neuroscience," *Nat. Rev. Neurosci.*, vol. 14, no. 5, pp. 365–376, May 2013.
[51] J. G. Walls, G. R. Widmeyer, and O. A. El Sawy, "Building an information system design theory for vigilant EIS," *Inf. Syst. Res.*, vol. 3, no. 1, pp. 36–59, 1992.
[52] G. Kramer, "An introduction to auditory display," in *Auditory Display: Sonification, Audification and Auditory Interfaces, SFI Studies in the Sciences of Complexity, Proceedings*, 1994, pp. 1–77.
[53] I. Pollack, "The information of elementary auditory displays," *J. Acoust. Soc. Am.*, vol. 24, no. 6, pp. 745–749, 1952.
[54] D. K. McGookin and S. A. Brewster, "Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition," *ACM Trans. Appl. Percept. TAP*, vol. 1, no. 2, pp. 130–155, 2004.
[55] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press, 1990.
[56] B. N. Walker and G. Kramer, "Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making," in *Ecological psychoacoustics*, J. Neuhoff, Ed. New York: Academic Press, 2004, pp. 150–175.
[57] S. Barrass, "Sonification design patterns," Jul. 2003.
[58] G. Johannsen, "Auditory displays in human-machine interfaces," *Proc. IEEE*, vol. 92, no. 4, pp. 742–758, 2004.
[59] J. Anderson, "Creating an empirical framework for sonification design," Jul. 2005.
[60] M. Watson and P. M. Sanderson, "Designing for Attention With Sound: Challenges and Extensions to Ecological Interface Design," *Hum. Factors*, vol. 49, no. 2, pp. 331–346, Apr. 2007.
[61] B. J. Tomlinson, J. Batterman, Y. C. Chew, A. Henry, and B. N. Walker, "Exploring Auditory Graphing Software in the Classroom: The Effect of Auditory Graphs on the Classroom Environment," *ACM Trans Access Comput*, vol. 9, no. 1, p. 3:1–3:27, Nov. 2016.
[62] S. M. Hetzler and R. M. Tardiff, "The three 'R's: Real students in real time doing real work learning calculus," Montreal, Canada, 2007.
[63] A. de Campo, C. Frauenberger, K. Vogt, A. Wallisch, and C. Daye, "Sonification as an interdisciplinary working process," Jun. 2006.
[64] C. Shupe and R. Behling, "Developing and implementing a strategy for technology deployment," *Inf. Manage.*, vol. 40, no. 4, p. 52, 2006.

SONIFICATION WORKSTATION

Sean Phillips

Media Arts and Technology Department
University of California Santa Barbara
Goleta, CA USA
seanphillips@ucsb.edu

Andrés Cabrera

Media Arts and Technology Department
University of California Santa Barbara
Goleta, CA USA
andres@mat.ucsb.edu

ABSTRACT

Sonification Workstation is an open-source application for general sonification tasks, designed with ease-of-use and wide applicability in mind. Intended to foster adoption of sonification across disciplines, and increase experimentation with sonification by non-specialists, *Sonification Workstation* distills tasks useful in sonification and encapsulates them in a single software environment. The novel interface combines familiar modes of navigation from Digital Audio Workstations, with a highly simplified patcher interface for creating the sonification scheme. Further, the software associates methods of sonification with the data they sonify, in session files, which will make sharing and reproducing sonifications easier. It is posited that facilitating experimentation by non-specialists will increase the potential growth of sonification into fresh territory, encourage discussion of sonification techniques and uses, and create a larger pool of ideas to draw from in advancing the field of sonification. Source code is available at **https://github.com/Cherdyakov/sonification-workstation**. Binaries for macOS and Windows, as well as sample content, are available at **http://sonificationworkstation.org**.

1. INTRODUCTION

When referring to sonification applications we mean finished software programs targeting end-users, with a focus on creating sonifications. A broad definition of sonification is best for our purposes, and "the technique of rendering sound in response to data and interactions," which is found in section 1.1 of *The Sonification Handbook* is suitable [1]. This includes methods which convert data samples directly into amplitudes, known as *audification* and sometimes treated separately.

Comprehensive figures on the software used in sonification research are not readily available, but in 2012 Bearman and Brown reviewed 51 articles on sonification and found domain-specific programming languages to be the most common tools [2]. Their survey found *Supercollider* [3] and *Pure Data* [4] to be especially popular, with *Supercollider* leading the way amongst research published in The Proceedings of the International Conference on Auditory Display (ICAD) [5]. This is despite the existence of multiple sonification applications.

Sonification tools can generally be placed on a spectrum from most to least flexible, which usually correlates with the degree of complexity and therefore has an inverse relationship with accessibility to novices. On one extreme of this spectrum are the general-purpose programming languages. Domain-specific programming languages fall along the middle of this spectrum, while at the other extreme are found end-user applications dedicated to creating data sonifications.

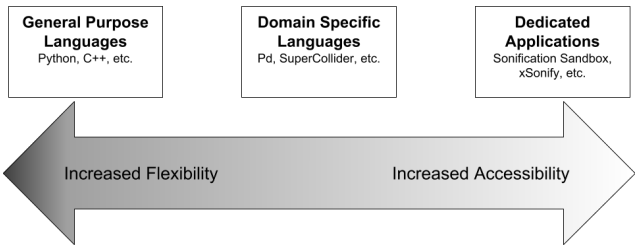


Figure 1: Sonification tools on a conceptual spectrum, illustrating a common trade-off between flexibility and ease-of-use.

End-user applications have been designed to sonify specific data types [6], but we are concerned here with applications designed for general sonification tasks. This implies applications which can load datasets of assorted sizes, sonify them in multiple, user-selected ways, and which place few restrictions on what the dataset represents. The given criteria still allow for a wide range of software types and a handful have been tested over the years, though few appear to be actively under development.

2. RELATED WORK

This section briefly describes some of the more significant, dedicated sonification applications that have been developed.

2.1. Sonification Sandbox (2003)

Sonification Sandbox was "motivated by the need for a multi-platform, multi-purpose toolkit for sonifying data" [7]. The program generates MIDI output, rather than audio. The graphical interface provides tabs for viewing the data, altering the parameter mappings, and adding context. In a sonification, context is non-signal information added to the output to help the listener interpret what they hear, analogous to the axes and trend lines on a visual graph [1]. In the *Sonification Sandbox* these context cues can include reference pitches for comparing to data values and click tracks to assist in interpreting time. The software is in beta and hasn't been updated for newer versions of Java, but it is still

https://doi.org/10.21785/icad2019.056

available for download at the Georgia Tech School of Psychology website [8].

2.2. SonART (2003)

The *SonART* toolkit used *The Synthesis ToolKit* (STK) [9] for synthesis and audio output, but added a scheduler and "parameter matrix engine," in an effort to provide cross-platform GUI tools for auditory display [10]. Akin to an audio matrix router in operation, the parameter matrix arrayed data parameters along the top of a 2D matrix and synthesis parameters down the right-hand side. This matrix arrangement allowed for arbitrary mapping of data to control parameters. The original paper proposes an ambitious long-term plan, with stated goals of "laying the foundation for an ongoing open-source collaborative effort," and "establishing and maintaining a well-documented and publicly accessible repository of sonification development tools." However, looking at internet archives, it appears the only download link for *SonART* dates to 2004 or earlier, not long after publication. Unfortunately, the downloads are no longer available [11]. The paper describes *SonART* as cross-platform and seems to contain screenshots from a Windows build, but it was not ultimately released for that platform.[1] *SonART* emphasized image sonification over general sonification tasks in the final release.

2.3. xSonify (2006)

NASA's *xSonify* was developed to sonify one-dimensional space physics data [12]. This makes it narrower in scope than other programs under consideration, but it still targets many datasets and has a history of practical application that makes it interesting. *xSonify* offers pitch, loudness, and rhythm modes, and some data pre-processing. *xSonify* includes text-to-speech facilities for menu navigation and a strong focus of the project has been accessibility for the visually-impaired. Co-author of the original *xSonify* paper Wanda L. Diaz Merced is blind and has used sonification in her physics research for many years [13, 14]. Merced also used a prototype of *xSonify* with visually-impaired students at the University of Puerto Rico [12]. *xSonify* is available at the Sonification Research page of NASA's website [15].

2.4. Sonifyer (2008)

Sonifyer is meant to be an easy-to-use sonification program, accessible even to amateurs. The authors became interested in such a project while sonifying EEG data with the Max/MSP [16] framework, writing that their Max sonification system became increasingly difficult to teach newcomers as it grew in complexity [17]. They also noted the steep learning curve of *Supercollider*, which they acknowledged as a popular sonification tool. *Sonifyer* is an effort to bring the user-friendliness of consumer software to the sonification space, including easy availability and installation, citing iTunes as a benchmark example. The original paper on *Sonifyer* also stressed the need for a more active community and easy sharing of sonifications. To address such needs the authors introduced a companion website alongside *Sonifyer*, which they hoped would provide a place to share audio samples and community knowledge. As of this writing the *Sonifyer* website appears to have very little content posted after 2009, and no samples posted after 2011 [18]. Curiously for a project aimed at wide adoption, *Sonifyer* will not function without obtaining a license from the makers via e-mail, and is available only for macOS. *Sonifyer* provides audification (which appears to be a strong suit) and limited FM parameter-mapping sonification.

2.5. Rotator (2016)

Rotator was created at MIT by Juliana Cherston [19]. It is a client-side web application, written in JavaScript, React, and Flux. It has a novel interface, which allows for visualization and sonification of multiple data streams at the same time. The software is aimed at "diversifying the way that users distribute data across their senses" [19, 20]. *Rotator* assumes the data has a geometric relationship and a user-provided schematic of the data origin-points is the key UI component. Users place bounding boxes around clusters of data streams on the schematic; one bounding box dictates the streams currently being visualized, another dictates the streams being sonified. The two boxes are fully independent and can overlap or delineate exclusive areas of the schematic. There are six synthesis possibilities, including audification. The *Rotator* project was largely for experimentation and is not under active development at the time of writing.[2]

3. SONIFICATION WORKSTATION

3.1. Motivation

The preceding overview of existing sonification software should help clarify motivations for the *Sonification Workstation* project. The state of the field suggests an opening for current work on sonification applications. Domain-specific programming languages such as *Supercollider* and *Pure Data* have contributed to numerous publications, and exhibit ongoing development [21, 22]. This contrasts with the more experimental nature and limited life-span of dedicated software, and invites additional efforts in the application space.

Sonification Workstation is an attempt to capture some of the utility of the domain-specific language solutions, while providing the simplified access to established sonification techniques and processes sought by prior dedicated software. Additionally, technical decisions were made to ease ongoing development and hopefully increase the project's longevity (see 4.2).

3.2. Application Overview

The *Sonification Workstation* interface consists of a single window, divided into two parts; the *data view* and the *patcher view*. The data view is the main user-interface for controlling playback of the data being sonified, and is analogous to the waveform view in a Digital Audio Workstation (DAW), such as *Pro Tools* or *Reaper*. The patcher view is where the user creates the synthesis tree that will determine the character of the sound. These two interfaces work together to allow data playback, parameter mapping, and synthesis design, without interrupting the flow of listening and evaluating.

3.2.1. Data View

The data view is populated whenever a new dataset is loaded via the *File* menu. Currently, data can be imported from CSV files. CSV columns are converted to tracks and plotted. Columns are more

[1] J. Berger, personal communication, September, 2017

[2] J. Cherston, personal communication, March, 2019

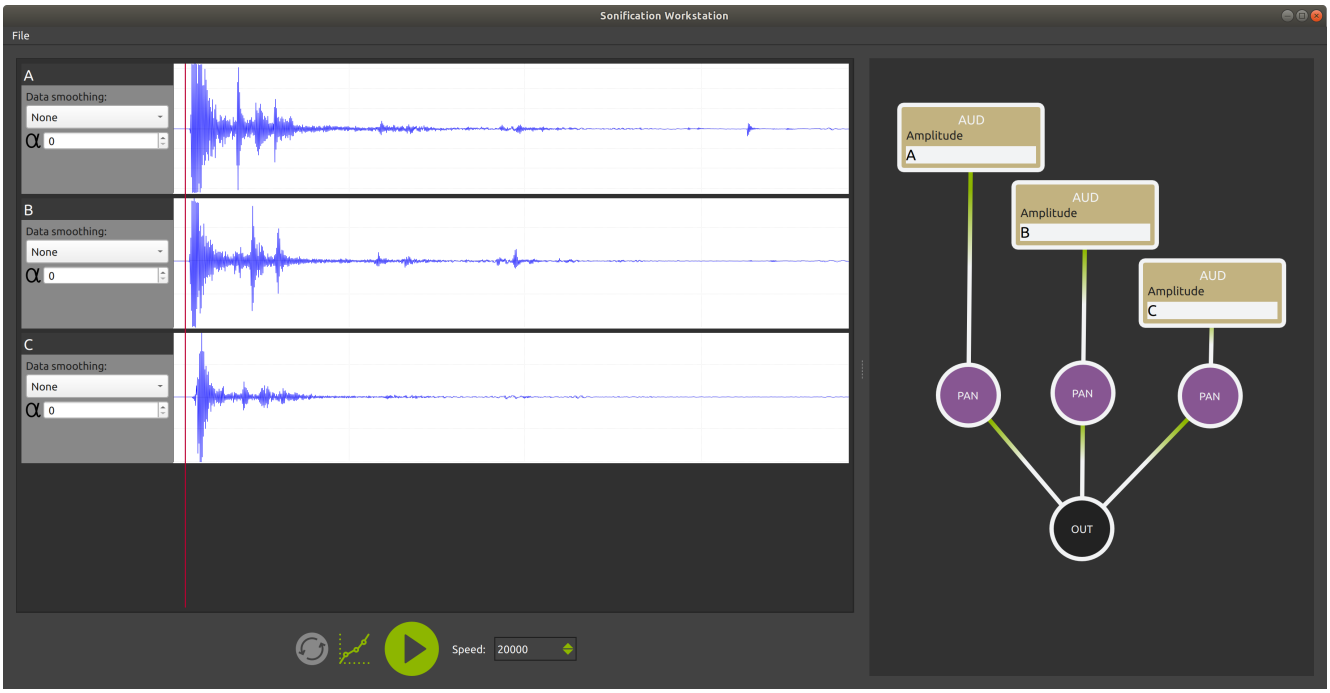


Figure 2: *Sonification Workstation* application window, configured for audification of a three channel seismic dataset at a playback rate of 20,000 data samples per second and using interpolation.

suited to represent individual data dimensions than rows, since common CSV editors can access only a limited number of columns in a single CSV file. Such programs will truncate a long series of values if it is entered in a row. Once loaded, data tracks are assigned variable names, for use in parameter mapping expressions (see 3.4).

Data is treated as a signal source, and the time domain can be quickly navigated by mouse. Clicking anywhere along the plotted data will move the playback cursor, right-clicking and dragging will create a bounded area for looping playback. Transport controls, seen directly below the data tracks, provide controls for play/pause, setting the playback speed, enabling looped playback, and enabling real-time interpolation between data points during playback. Playback speed is equivalent to the number of data points read every second and scales from zero to audio rate (48kHz). Playback is synchronized across all tracks, so that each sample of data playback represents a single point in the dataset, across all dimensions.

3.2.2. *Patcher View*

One of the important contributions of *Sonification Workstation* is providing a flexible way to construct a sonification, without the need for domain-specific knowledge, by consolidating the main techniques into a very simple patching interface. With a small set of synthesis components and the available data mapping and scaling features (see following sections), *Sonification Workstation* offers tools for additive synthesis, subtractive synthesis, frequency and amplitude (AM and FM) modulation, and audification. Context for values can be created via fixed frequency oscillators or noise beds. Context for time can be added with short AD envelopes, triggered at data playback rate, or synchronizing modulation and playback rates.

Figure 2 shows a three-track seismic dataset. The patcher interface has been populated with synthesis components to audify and pan the three channels. The Audification (AUD) synthesis components have been maximized (see section 3.4), showing tracks A, B, and C, have each been mapped to a single AUD component.

3.3. High-Level Synthesis Components

The high-level synthesis components in *Sonification Workstation* encapsulate the data mapping and audio settings. The patcher interface is inspired by domain-specific patcher languages such as Pure Data and Max, but is quite simple in comparison. There are fewer than a dozen instantiable types and no differentiation between mono, stereo, or control signals. A brief description of the existing high-level synthesis components follows.

The OSC Component

A sinusoidal oscillator. Accepts arbitrary functions mapped to frequency and optionally scales the frequency value within a user-selected range.

The AM Component

An amplitude modulator. Will modulate the amplitude of any parent synthesis component. The frequency accepts arbitrary mappings and the value can be scaled.

The FM Component

A frequency modulator. Will modulate the frequency of parent OSC, AM, and FM components. Will also modulate the pan position of the PAN object. Frequency and depth parameters accept arbitrary mappings and can be scaled.

The AUD Component

Turns the results of data mappings directly into amplitudes

for audification. Values are always scaled within the range [-1.0,1.0] to prevent clipping and maximize gain.

The PAN Component

Pans the output of connected components in the stereo field. Pan position accepts arbitrary mappings and can be scaled, but values are clipped to the range [-1.0,1.0].

The ENV Component

Applies an AD envelope to connected components. Attack and Decay values accept arbitrary mappings and the values can be scaled.

The VOL Component

Scales the output of connected components. Accepts arbitrary mappings and can be scaled. The gain value will be clipped to the range [-1.0,1.0]. Negative values allow VOL to be used for phase inversion.

The NSE Component

Generates white, pink, or Brownian noise.

The EQ Component

Biquad filter with mappable resonance and frequency. Switchable high-pass, low-pass, band-pass, or notch. Combines with the NSE Component for subtractive synthesis.

The OUT Component

The root of the synthesis tree, connecting to the OUT Component will pass a component's signal to the audio callback for output to the computer sound card.

3.4. Parameter Mapping

Sonification Workstation synthesis components accept parameter mappings in the form of a mathematical expression. Data tracks are assigned names on import, these names can be used in expressions as variables and multiple data tracks can be included in the same expression. Valid mappings are constants (e.g. setting an oscillator to a fixed frequency of 440Hz), data tracks, or an arbitrary expression including both. Expressions are evaluated in real-time.

In their minimized state, each high-level synthesis component is a colored circle with a text identifier. Double-clicking maximizes the component, revealing controls for mapping and audio settings. Figure 3 shows four examples of maximized synthesis components and their settings, clock-wise from top-left these are:

1. A noise generator (NSE), set to generate white noise.
2. An amplitude modulator (AM), with frequency mapped to the square root of data track A minus data track B. It is shown modulating the amplitude of the connected oscillator directly below.
3. An oscillator (OSC), with frequency mapped to the values of data track C. The oscillator also has scaling enabled, which will scale the incoming values to fit, in this case, between 100Hz and 800Hz.
4. A pan control (PAN), set to full left.

While not implemented in the current build, the authors are also interested in adding parameter mapping to the transport. An earlier version of *Sonification Workstation* allowed parameters to control the rate of playback, essentially making time a mappable parameter.

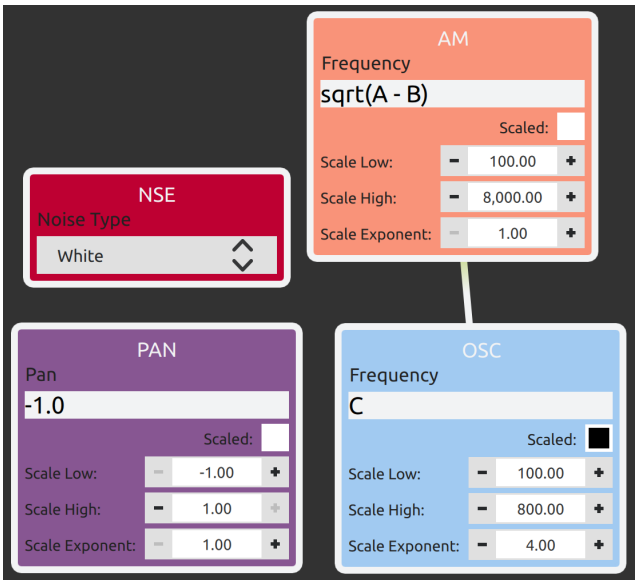


Figure 3: Maximized synthesis components, showing their data mapping and audio settings.

3.5. Data Scaling

Synthesis components provide for the scaling of data mappings to usable values. In Figure 3 the scaling controls can be seen at the bottom of the maximized AM, OSC, and NSE interfaces. Parameter scaling is controlled with the following values:

- Scaled: enables scaling. If disabled, the data values are used as-is. There is no protection against aliasing.
- Scale Low: the lowest output value desired, corresponding to the lowest data value in an assigned row, or the lowest value of the assigned expression, given the current dataset.
- Scale High: the highest output value desired, corresponding to the highest data value in an assigned row, or the highest value of the assigned expression, given the current dataset.
- Scale Exponent: controls the shape of the curve the data values are mapped to.

The formula for scaling is taken from the *scale* object in the Max programming language [23]. Inverted mappings are also possible, by setting *Scale Low* to the top of the desired scale range and *Scale High* to the bottom of the range. Inverted mappings have been identified as useful in cases where increasing data values have an intuitively inverse relationship to the data, e.g. when size is mapped to pitch [1].

3.6. Session Files

The state of a session, including the synthesis tree and the path to the currently loaded dataset, can be saved to a session file. Session files are .json files, with the dataset path and objects representing the state of each synthesis component stored in human readable form. Sharing a session file, together with the dataset, affords complete reproduction of the sonification and enables collaboration, modification and experimentation.

3.7. Color Scheme

The color scheme for UI elements in *Sonification Workstation* draws heavily from the so-called "Kelly colors." This is a set of 22 colors, meant to provide maximum contrast for color coding tasks, published by Kenneth L. Kelly for the Inter-Society Color Council in 1965 [24]. According to Green-Armytage, the ISCC has worked recently to bring Kelly's list up to date, but improving on it was difficult [25]. Kelly's list is designed so that the second color provides maximum contrast with the first, the third color will contrast maximally with colors one and two, etc. Kelly chose the first nine colors for their differentiability by individuals with red-green color blindness.

Major UI elements of *Sonification Workstation* utilize the first thirteen colors of the list. While using *Sonification Workstation* is not a color coding task per se, the data track view requires color coding and differentiable colors were desired for all of the synthesis components. It was not possible to keep to the first nine Kelly colors, while providing unique colors for each of the synthesis components. Additionally, multiple neutral shades are used to provide some organization to the main application window, where Kelly only provides values for white, black and a single grey. To further aid users less-sensitive to color differences, all synthesis components are surrounded by a white ring, providing very high contrast against the dark background of the patcher view. They are also labelled with a text-based code that indicates their function.

4. TECHNICAL NOTES

4.1. Processing Rates

The software operates at three different processing rates.

4.1.1. Audio Rate

Audio rate processing happens at the sample rate of the audio system, which is set to 48,000Hz. This is the rate at which all synthesis classes generate audio, regardless of the playback rate or the rate of change in any associated data parameters. If mapped data parameters are changing more slowly than the audio rate (a typical case), the synthesis components will generate multiple audio samples from a single data point.

4.1.2. Command Processing Rate

User commands to sonification components and the transport are buffered in a lock-free ring buffer, for consumption by the audio callback, which happens at *command process rate*. This allows the user to issue changes to synthesis parameters and data playback settings without interrupting audio processing. The command processing rate is set to the audio block rate, consequently user commands are processed once for every time the audio buffer is filled.

4.1.3. Step Rate

Step rate is the rate at which values from the dataset are read and passed to all synthesis blocks in the patcher view. This is dependent upon the data sample playback rate, which is set in samples per second. Therefore, at the default speed of "1", the step function is called on every synthesis component once every second. The step function provides a way for synthesis blocks to take special actions when new data points are reached. The ENV class, for example, can re-trigger envelope generation at the step rate.

4.2. Qt Framework

The *Sonification Workstation* is written in C++, QML, and JavaScript, using the Qt framework. The graphical front end is where the QML and JavaScript code resides, while core classes for synthesis and playback are written in C++. The current build targets Qt LTS version 5.12 and C++17. Source code is available at the project's GitHub repository (source: https://github.com/Cherdyakov/sonification-workstation). It is hoped that building with the Qt framework will help ongoing support and development of *Sonification Workstation*, since the framework itself is well-established and receives regular updates. Additionally, a number of existing sonification applications are only available for a single platform (see section 2. Related Work), limiting their potential audience. Targeting the Qt framework and using cross-platform libraries for synthesis and audio i/o means that *Sonification Workstation* can be built for macOS, Linux, and Windows.

4.3. Gamma Synthesis Library

Most of the high-level patcher objects contain lower-level synthesis classes, commonly referred to as *unit generators*. OSC contains an Oscillator unit generator, ENV contains an envelope unit generator, and so on. Many unit generators contained in the *Sonfication Workstation* synth objects are from the Gamma C++ synthesis library (source: https://github.com/LancePutnam/Gamma), written by Lance Putnam [26].

4.4. Mathematical Expression Toolkit

Evaluation of parameter mapping expressions is handled by the C++ Mathematical Expression Toolkit, written by Arash Partow (source: https://github.com/ArashPartow/exprtk).

5. TAXONOMIC EVALUATION AND FUTURE WORK

In examining the fitness of the completed project for its purpose, some criteria must be chosen. Several experts have attempted to generate sonification taxonomies. It stands to reason that a successful general purpose sonification application would address one or more well-considered taxonomies. This section evaluates the merits and shortcomings of the *Sonification Workstation* through the lens of example taxonomies, then notes significant improvements indicated by this evaluation.

5.1. Functional Taxonomy

Summarizing available research, the Sonification Handbook [1] describes the following functional categories:

1. Alarms, alerts, and warnings
2. Status, process, and monitoring messages
3. Data exploration
4. Art, entertainment, sports, and exercise

At this point in time, the software addresses category three most fully, with additional application to the artistic element of category four.

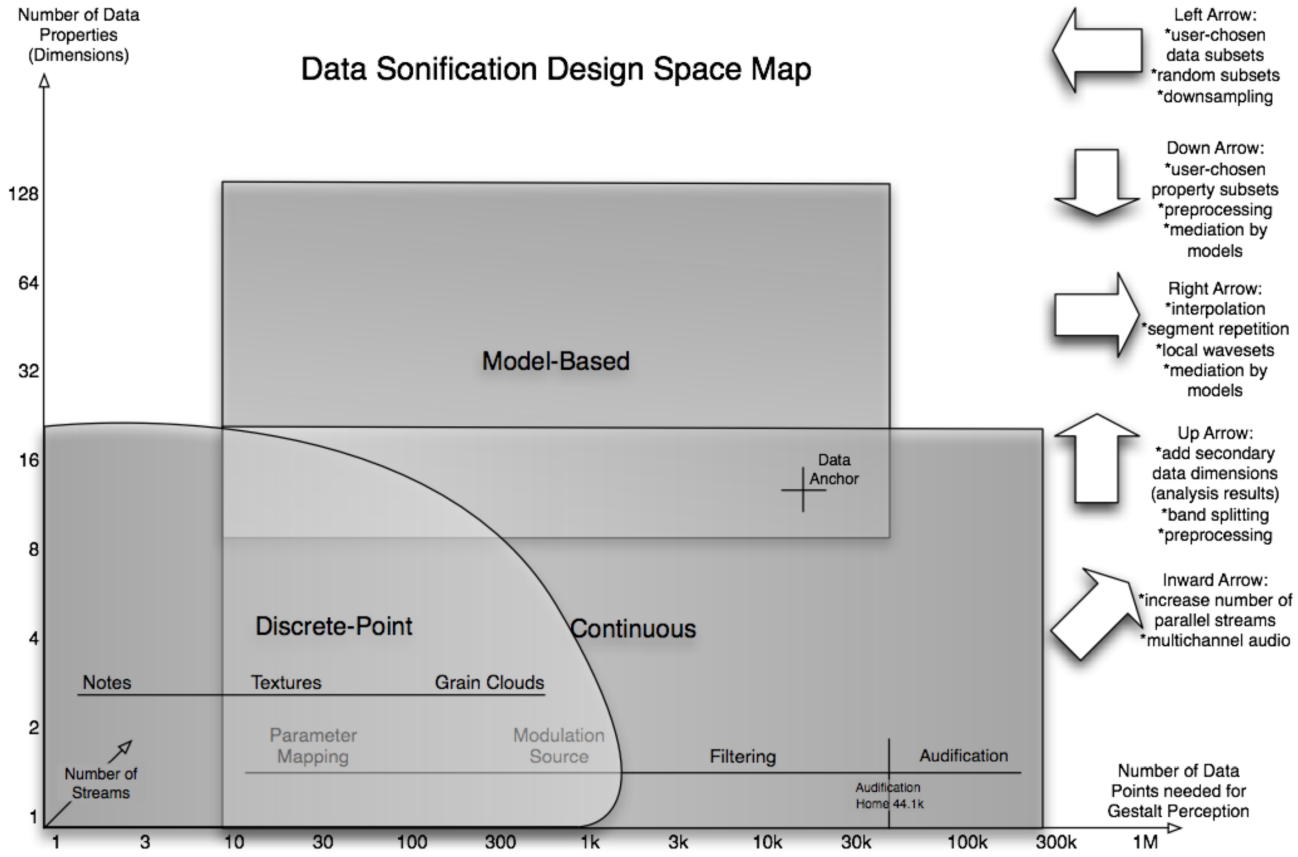


Figure 4: De Campo's Data Sonification Design Space Map, taken from the original paper. *"The overlapping zones are fuzzy areas where different sonification approaches apply; the arrows on the right refer to movements on the map, which correspond to design iterations."* — de Campo

Category one, *alarms, alerts, and warnings*, refers to sounds which "indicate something has occurred, or is about to occur" or "require immediate response or attention." Category two likewise refers to monitoring things "current or ongoing." These can be prototyped in *Sonification Workstation* using recorded data, but live streaming of input data is not currently implemented. Real-time input would be highly desirable for users working on these sorts of tasks, and sports and exercise sonification could also benefit from real-time sonification. The project should be expanded in the future, to incorporate real-time process monitoring and event based sonification via network messages.

5.2. De Campo's Sonification Design Space Map

In a 2007 paper [27], de Campo began by classifying sonification strategies into three categories:

1. Continuous Data Representation
2. Discrete Point Data Representation
3. Model-Based Data Representation

De Campo employed these categories in creating the *Data Sonification Design Space Map*. The map is intended as a guide for choosing sonification strategies, based upon the number of data dimensions, the number of simultaneous streams, and the number of data points required to comprise a single "gestalt." Briefly, number of streams refers to the number of data dimensions that are sonified in parallel, either through spatialization or using sonically distinct frequency ranges per stream, etc., while a gestalt is a perceivable pattern or recognizable audio structure that is of interest.

De Campo describes the use-case for the Design Space Map this way: "The Design Space Map enables a designer or researcher to engage in systematic reasoning about applying different sonification strategies to his/her task or problem, based on data dimensionality and perceptual concepts."

De Campo's map posits a space in which the sonification designer can move freely between many different techniques and dataset types. This view of sonification design is highly compatible with the idea of a generalized sonification application. De Campo has perhaps provided a map not only for sonification designers, but for the developers of sonification software as well. The prospect of a design aid, such as the Design Space Map, combined with a software tool meant to realize the methods it describes, is a potentially exciting development in sonification software design.

Sonification Workstation already incorporates the ability to transition along the various axes of the Design Space Map. This is not a coincidence, de Campo's paper had a strong influence on the current project. Realizing the concept of freedom of movement along these axes is only a starting point however, and specific areas of the

map are covered thinly or left unaddressed. Currently, *Sonification Workstation* offers parameter mapping, modulation source, filtering, and audification. Incorporating physical modeling is an intriguing possibility and it would help cover a large area on the design map. The challenge would be in incorporating physical modeling that is applicable to general sonification tasks. Many model-based sonifications reflect the real properties of an object or physical system under study and can be difficult to generalize. Some general model-based sonification strategies have been offered; Lee, Sell, and Berger proposed methods based on digital waveguide meshes [28], while Hermann and Ritter describe a system based on a crystal growth model [29]. Such a method is an excellent candidate for inclusion in *Sonification Workstation*, even if experimentally. *Sonification Workstation* was designed to be extensible through the addition of new synthesis components such as these.

6. SUMMARY

The authors have presented the *Sonification Workstation*, software uniquely suited to generalized sonification work, with a low barrier to entry. Prior work in this space was presented for context and the present work was examined in relation to existing sonification taxonomies, illuminating the strengths and weaknesses of the approach, and providing future direction for the project.

7. REFERENCES

[1] Various, *The Sonification Handbook*, J. G. N. Thomas Hermann, Andy Hunt, Ed. Berlin, DE: Logos Verlag Berlin, 2011.

[2] N. Bearman and E. Brown, "Who's sonifying data and how are they doing it? a comparison of icad and other venues since 2009," in *Proc. of the 18th Int. Conf. on Auditory Display*, 2012, pp. 231–232.

[3] J. McCartney, "Supercollider: A new real time synthesis language," in *Proc.of the Int. Computer Music Conference*, 1996, pp. 257–258.

[4] M. S. Puckette, "Pure data," in *Proc.of the Int. Computer Music Conference*, 1996, pp. 224–227.

[5] http://www.icad.org, accessed: 2017-10-17.

[6] R. A. Khan, R. K. Avvari, K. Wiykovics, P. Ranay, and M. Jeon, "Lifemusic: Reflection of life memories by data sonification," in *Proc. of the 22nd Int. Conf. on Auditory Display*, 2016, pp. 90–92.

[7] B. Walker and J. Cothran, "Sonification sandbox: A graphical toolkit for auditory graphs," in *Proc. of the 9th Int. Conf. on Auditory Display*, 2003, pp. 231–232.

[8] http://sonify.psych.gatech.edu/research/sonification_sandbox/, accessed: 2019-03-29.

[9] P. R. Cook and G. P. Scavone, "The Synthesis ToolKit (STK)," in *Proc. of the Int. Computer Music Conference*, 1999, pp. 164–166.

[10] ——, "Sonart: The sonification application research toolbox," in *Proc. of the 8th Int. Conf. on Auditory Display*, 2002.

[11] https://ccrma.stanford.edu/~woony/software/sonart/, accessed: 2019-03-29.

[12] R. M. Candy, A. M. Schertenleib, and W. L. D. Merced, "xSonify sonification tool for space physics," in *Proc. of the 12th Int. Conf. on Auditory Display*, 2006.

[13] http://www.npr.org/2017/01/20/510612425/how-can-we-hear-the-stars, accessed: 2019-03-29.

[14] https://www.cfa.harvard.edu/sed/projects/star_songs/pages/xraytosound.html, accessed: 2017-09-02.

[15] https://spdf.sci.gsfc.nasa.gov/research/sonification/, accessed: 2019-03-29.

[16] M. S. Puckette, "The patcher," in *Proc.of the Int. Computer Music Conference*, 1988, pp. 420–429.

[17] F. Dombois, "Sonifyer: A concept, a software, a platform," in *Proc. of the 14th Int. Conf. on Auditory Display*, 2008, pp. 1–4.

[18] http://www.sonifyer.org/?lang=e, accessed: 2019-03-29.

[19] J. M. Cherston, "Auditory display for maximizing engagement and attentive capacity," MIT, 2016.

[20] J. Cherston and J. A. Paradiso, "Rotator: Flexible distribution of data across sensory channels," in *Proc. of the 23rd Int. Conf. on Auditory Display*, 2017, pp. 86–93.

[21] https://puredata.info/downloads/pure-data, accessed: 2019-05-27.

[22] https://supercollider.github.io/archive, accessed: 2019-05-27.

[23] https://docs.cycling74.com/max7/maxobject/scale, accessed: 2019-03-29.

[24] K. L. Kelly, "Twenty-two colors of maximum contrast," *Color Engineering*, vol. 110, no. 3, pp. 26–27, 1965.

[25] P. Green-Armytag, "A colour alphabet and the limits of colour coding," *Colour: Design & Creativity*, vol. 5, no. 5, pp. 1–23, 2010.

[26] L. Putnam, "Gamma: A C++ sound synthesis library further abstracting the unit generator," in *Proc.of the Int. Computer Music Conference*, 2014, pp. 1382–1388.

[27] A. de Campo, "Toward a data sonification design space map," in *Proc. of the 13th Int. Conf. on Auditory Display*, 2007, pp. 342–347.

[28] G. S. K. Lee and J. Berger, "Sonification using digital waveguides and 2- and 3-dimensional digital waveguide mesh," in *Proc. of the 11th Int. Conf. on Auditory Display*, 2005, pp. 140–145.

[29] T. Hermann and H. J. Ritter, "Crystallization sonification of high-dimensional datasets," in *Proc. of the 8th Int. Conf. on Auditory Display*, 2002, pp. 1–6.

TESTING SPATIAL ASPECTS OF AUDITORY SALIENCE

Zuzanna Podwinska, Bruno M Fazenda, William J Davies

University of Salford
School of Computing, Science & Engineering
43 Crescent, Salford, M5 4WT, UK
z.podwinska@edu.salford.ac.uk

ABSTRACT

Auditory salience describes the extent to which sounds attract the listener's attention. So far, there have not been any published studies testing if the location of sound relative to the listener influences its salience. In fact, not many experiments in general test auditory attention in a fully spatialised setting, with sounds in front and behind the listener. We modified two experimental methods from the literature so that they can be used to test spatial salience - one based on oddball detection and artificially created sounds, the other based on self-reported attention tracking in a more ecologically valid scenario. Each of these methods has its advantages and each presents different challenges. However, they both seem to indicate that high frequency sounds arriving from the back are slightly less salient. We believe this result could likely be explained by loudness differences.

1. INTRODUCTION

1.1. Motivation

Certain sounds in the environment involuntarily attract attention. This happens outside of the listener's control, and depends on the properties of the sound itself. It is also task-independent: even if the listener is consciously paying attention to a radio programme or a piece of music, her attention will be drawn to a new *salient* sound in the environment. Salience can be defined as the property of sound which makes it stand out among other sounds [1].

Although there have been studies on salience of acoustic features such as loudness, brightness or tempo [2, 9, 10], no studies so far have shown how salience might be related to the location of the sound. To date, spatial attention studies have focused on target-distractor separation and relied on focused top-down attention (e.g. [3, 4]). But do different locations of sound around the listener have inherently different salience, regardless of what the person is focused on? It is not unreasonable to suspect that it might be the case. For example, one could argue that there would be an evolutionary advantage to humans being more alert to sounds arriving from behind them, where vision provides little useful information. The difficulty in studying this question lies mainly in determining where a person's attention was directed. Unlike in vision, where eye-tracking is often used, humans do not have auditory organs which would indicate the direction of the attentional 'spotlight'.

In this work, we propose two methods of testing spatial auditory salience, which are extensions of previously published salience experiments.

1.2. Measuring auditory salience

There is not one widely agreed upon way of behaviourally measuring auditory salience. Perhaps the most straightforward way of testing whether a sound is salient is asking human subjects directly. For example, in an annotation task [5], participants were asked to manually mark 'interesting' sounds in a recording of a scene. Another type of experiment which involves human judgement is a comparison of two sounds (or scenes) in terms of their salience or 'interestingness' [6, 7, 8]. This type of experiment has the advantage of being able to sort test sounds from least to most salient. The downside is the subjectivity of the word 'salient' or 'interesting', which can have different meanings to different people (especially since there is no single, universally accepted definition of auditory salience). Some researchers [9] avoid this issue by asking participants to indicate where their *attention* is, and to do so in real time. This is somewhat analogous to gaze tracking in visual attention, but a less direct representation of the phenomenon, as it also involves conscious tracking of one's attention.

Another way of testing salience is through sound detection - for example, of sound in noise [6]. This is more objective but seems more removed from the notion of salience. It assumes that more salient sounds will be easier to detect, which might not be strictly true. Another task involved detection of a salient event in a scene [10], which still might confound salience and energetic masking issues. A different paradigm is based on oddball detection - detecting a stimulus which is different from a series of standard, regular ones, often in the presence of competing streams. Response time and detection rate are indicators of stimulus salience (e.g. [2, 11]).

Finally, some experiments use task interference paradigms, where participants are asked to perform a task while unrelated distracting stimuli are played to them. Sound salience is assumed to be directly related to the amount of distraction caused, so changes in response time and error rate are an indication of stimulus salience.

In the following section, we present two methods which are spatial extensions of two of the published salience measuring paradigms discussed above [2, 9].

https://doi.org/10.21785/icad2019.042

2. EXPERIMENT 1

2.1. Methods

The first experiment is based on the Segregation of Asynchronous Patterns (SOAP) paradigm [12]. It is based on the idea that two perceived auditory streams compete for attentional resources, and as a result one of them becomes *foreground*, and the other will be *background*. If no arbitrary top-down effects are in place, a more salient stream will win the competition and become the foreground. The main assumption here is that it will be easier to detect changes in the foreground (more salient) stream.

In the original SOAP experiment, two sound patterns were presented dichotically through headphones. Both patterns consisted of short birdsong excerpts separated by constant interstimulus interval (ISI). A crucial part of the design is to make sure that the two patterns are asynchronous, to avoid creating a rhythm which could be morphed into a single auditory object. The participants' task was to detect a change in ISI in one of the streams. No instructions were given about which stream should be attended to. According to the SOAP framework, listeners should be statistically more likely to attend to, and detect changes in, the more salient stream.

In order for the SOAP framework to account for spatial effects, we modified it so that sound patterns arrive at the listener from 2 out of 6 locations around them, rather than just left and right. The participant was seated in an acoustically treated listening room, surrounded by loudspeakers as in Figure 1. In [12], participants were asked to choose between the left or right stream. However, in this experiment we wanted to avoid requiring participants to localise sounds, as we were not interested in their localisation ability as such. Therefore, we decided to use two distinctively different stimulus types: short noise bursts, either high- or low-pass filtered at 2 kHz. Each pattern contained only one type of stimulus, and participants were asked to detect a shortened ISI and indicate whether it occurred in the high or low frequency pattern. The sounds were designed so that there was no overlapping spectral content, to ensure that it was easy to segregate and follow one of the streams without too much interference from the other. To ensure asynchrony, one of the two patterns always included shorter stimuli than the other (200 versus 150 ms). This resulted in one pattern sounding faster than the other (a property which is referred to here as tempo). Independent variables were then: sound location (1 to 6, as shown on figure 1), frequency (high and low), and tempo (fast and slow). Each participant was exposed to all conditions.

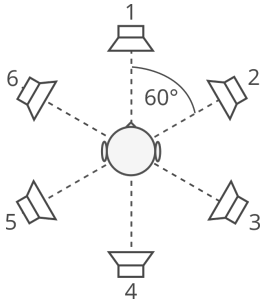


Figure 1: Loudspeaker set-up in the listening room.

Before the main experiment, participants completed a short training session and a baseline test, where only one pattern was present at a time. 19 volunteers took part in the experiment, all with self-reported normal hearing, average age 30.4, 4 female, 18 right-handed.

2.2. Results

Time elapsed from the end of the shortened ISI to the button press was recorded as response time (RT). Only correct responses were taken into account. The data was analysed with a Generalised Linear Mixed Model (lme4 package in R [13]) with an inverse Gaussian distribution and an identity link function, to account for a non-normal distribution of response times. Fixed effects were location, frequency, and tempo, and random effects were participant and background sound location. A model including frequency-tempo and frequency-location interactions was used as it gave the best fit (based on the Akaike information criterion).

Table 1: GLMM results on response time data. Significant predictors are in bold.

Fixed effects	Coeff.	SE	Z	p-value
(Intercept)	**0.843**	**0.031**	**27.13**	**< 0.0001**
Location 2	0.019	0.026	0.74	0.458
Location 3	-0.007	0.025	-0.29	0.773
Location 4	-0.038	0.024	-1.56	0.119
Location 5	-0.026	0.024	-1.09	0.276
Location 6	-0.029	0.024	-1.18	0.238
Frequency (high)	-0.005	0.026	-0.21	0.835
Tempo (fast)	0.003	0.014	0.21	0.831
Frequency:Tempo	**-0.070**	**0.020**	**-3.47**	**0.0005**
Location2:Frequency	-0.032	0.034	-0.92	0.356
Location3:Frequency	0.013	0.034	0.38	0.707
Location4:Frequency	**0.138**	**0.035**	**3.89**	**< 0.0001**
Location5:Frequency	0.061	0.034	1.79	0.074
Location6:Frequency	0.059	0.034	1.73	0.084
Random effects	Standard deviation			
Participant	0.103			
Background location	0.016			

The results, shown in Table 1, indicate that there are significant interactions: frequency-tempo and frequency-location. A post-hoc analysis of contrasts shows that, for low frequency stimuli, there are no significant differences between locations. However, for high frequency stimuli, there are significant differences between front and back locations ($p = 0.0018$), back and right-front ($p = 0.0002$), and back and right-back ($p = 0.005$). Figure 2 shows estimated mean response times and confidence intervals for the two interactions.

2.3. Discussion

There was no difference between participants' responses to different locations and tempo when the stimuli were low frequency noise. However, for high frequency stimuli, responses were on average 67ms faster for fast compared to slow patterns. Additionally, for high frequency stimuli, responses were significantly slower (about 100 ms) if target sound was behind the listener, than if it was in front of them.

The results show an interaction between tempo and frequency: slow patterns were more salient if they were low frequency, and fast patterns were more salient if they were high frequency. Interestingly, the study this experiment was based on [2] also found

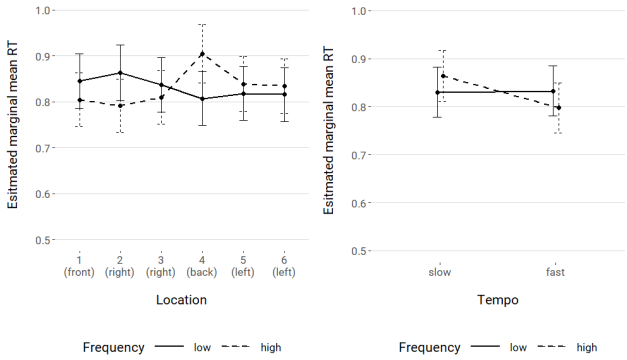


Figure 2: Estimated marginal means with 95% confidence intervals show interactions between spectrum of the noise bursts and and location (left panel) and tempo (right panel).

an interaction between these variables, but in the opposite direction: "the sounds with higher salience [...] are those with faster tempo and lower spectral centroid". They also concluded that in general, sounds with a lower spectral centroid were more salient, which was not found here. This last result is also in contrast to some other studies, such as [9], which found a significant increase in brightness in salient events. Our experiment did not find a significant effect of spectral centroid on salience.

3. EXPERIMENT 2

3.1. Methods

One of the shortcomings of the first experiment was that the stimuli were simple, synthetic sounds. Although this allowed for straightforward manipulation of the sound and minimised effects of context or semantic meaning, it could be argued that the perception and responses to those stimuli does not accurately represent everyday listening situations.

The goal of the second experiment was to test spatial salience in a more ecologically valid scenario. The experimental procedure was inspired by [9], who tested salience of sound events in two competing scenes. The participants heard one scene in each ear, and were asked to continuously indicate which one they were focusing on. For that, a mouse and a visual interface were used.

A similar procedure was used here, but with stimuli arriving from different locations all around the listener instead of just left and right. Additionally, it can be argued that the situation would be more realistic if competition for attention was between sound *events*, rather than full scenes, presented dichotically. Therefore, different locations in this experiment did not correspond to different scenes, but rather to events. Similarly to [9], the participants were asked to indicate, in real time, to which location in the scene their attention was directed. To do that, they used a joystick, and no visual display was provided, partly to avoid forcing participants to focus their attention on a display in front of them. Participants were allowed to move their heads slightly, but were reminded to indicate the location of the sound in relation to the room, rather the direction they were facing.

The experiment by [9] used recordings of different types of existing sound scenes. However, using recordings of full scenes would make manipulation of experimental variables difficult, so here, the scenes were designed from individual sounds instead. They consisted of a steady background and two types of events: distractors and targets. The experiment checked how often participants paid attention to targets, while responses to distractors were not analysed (they were effectively treated as part of the background). Position in time of distractors was randomised but the same for all participants. Position of targets was randomised for each participant separately, in an attempt to average out any interactions between specific distractors and targets.

The experiment was a full-factorial repeated-measures design with the following independent variables:

- target loudness (2 levels)
- target spectral centroid (2 levels)
- target location (4 levels)
- target semantic category (3 levels)
- background type (2 levels)

This results in 96 different conditions. Because habituation to a particular sound might make it less salient (as it is less surprising), it was crucial not to use the same stimulus more than once. For this reason, 96 different sound events were used as targets.

Because this design relies on accurate localisation of targets, a baseline experiment was conducted directly after the main experiment, with the same target stimuli and the same reproduction method, but with no background or distractors. The participants were asked to indicate which direction each target was coming from, as soon as they heard it, and to return to the centre after the sound was over. This allowed collection of baseline data which indicated individual localisation accuracy.

3.1.1. Target sounds

Targets were short clips from recordings of real-world sounds (from [14], [15] and ([16]), on average 3 seconds long. Time spacing between consecutive stimuli varied randomly from 2 to 4 s. The stimuli belonged to three different semantic categories, which were determined based on the soundscape taxonomy established in a sorting experiment by [17]. The categories were: Nature (subcategory: Animals), People (subcategory: Voices), Manmade (subcategory: Industrial).

Spectral centroid represented an objective measure of the perceived brightness of the sound, and was calculated as:

$$SC = \frac{\sum_{n=1}^{N} f(n)Y(n)}{\sum_{n=1}^{N} Y(n)}$$

where $Y(n)$ is the amplitude of the nth bin of the spectrum, and $f(n)$ is the centre frequency of that bin. To avoid any artefacts that come with filtering, and the risks of sounding unnatural, sound spectra were not manipulated. Instead, events were chosen so that their spectral centroid falls within one of two groups: 1000-2500 Hz or 4000-5500 Hz.

Short-term loudness of sound was calculated using the Dynamic Loudness Model [18] available through the PsySound3 toolbox in Matlab [19]. As an indication of loudness of each sound, the maximum of time-smoothed short-term loudness was used (STL window = 2 ms, smoothing window = 100 ms). Sound level was manipulated to create two levels with loudness means 8.4 and 14.4 sones, and standard deviation of 0.2 sones. These two levels correspond to the loudness of a 1kHz tone at about 70 and 78 dB SPL.

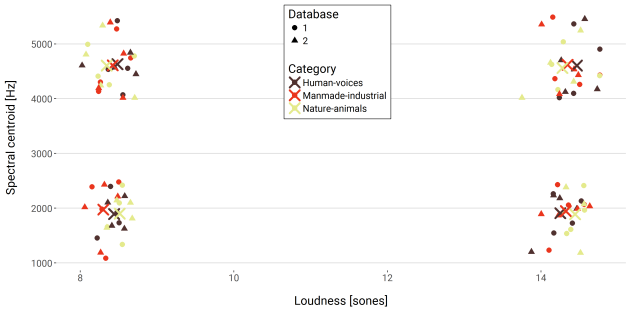
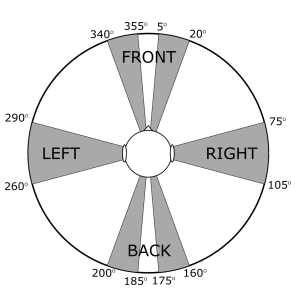


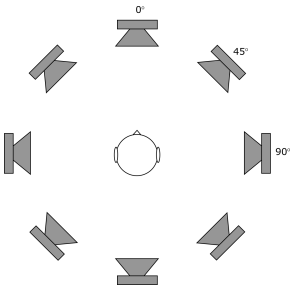
Figure 3: Stimuli used in the experiment. Database corresponds to two stimuli groups, used with different backgrounds. Colours indicate one of the three semantic categories. Recordings were chosen to fall within the two spectral centroid levels, and then their loudness was manipulated, while keeping the pairs of brightness groups as similar as possible.

Each sound was assigned to either one of the two levels in a way that minimised mean and variance differences between brightness levels. Figure 3 shows all targets on the loudness-brightness spectrum.

Targets were placed at one of four 30° areas (cones in Figure 4a) around the listener: front, back, right and left. The exact location of stimuli varied randomly within these areas. The choice of cone width was guided by a trade-off: on one hand, it would be best to avoid the borders between areas (e.g. 45° front/right border), where small localisation errors would be more problematic. On the other hand, from the perspective of scene realism, the cones should be wide enough so that the targets do not always appear at the exact same location. Additionally, 10° cones around the front and back locations were excluded in order to minimise front-back confusion effects (see Figure 4a). The location of each target was determined randomly for each participant, while keeping the number of targets in each area equal. Elevation was always the same, at approximately ear level.



(a) Grey areas show cones where target sounds were localised



(b) Loudspeaker positions; distractors and background were played from all loudspeakers.

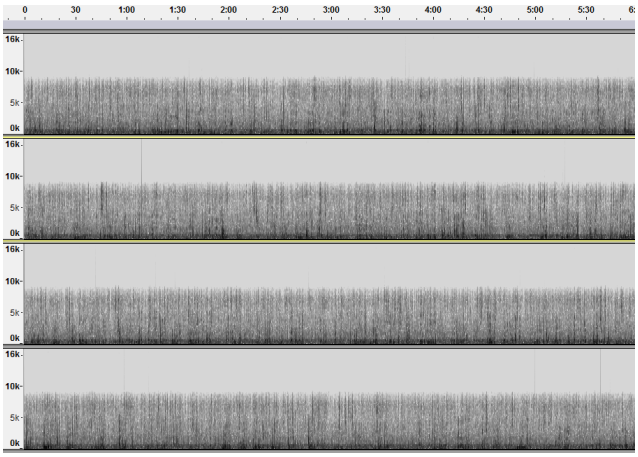
Figure 4: Target locations and experimental setup.

3.1.2. Scenes

These targets were used in two different sound scenes, each about 5 minutes long, each with different background sound and distracting events. Targets were divided into 2 balanced groups (this is represented by different shapes in Figure 3) and each group was played over one of the backgrounds. The 2 targets/backgrounds combinations, as well as the order of the scenes, were randomised between participants.

In the first scene (*speech*), the background was steady babble noise with distracting louder speech excerpts (from [20]). Most of the time, there was more than one talker present at the same time, but never in the same channel. The speech was in 9 different languages and participants were asked about their knowledge of these languages in a questionnaire after the test, and no one reported knowing any of the languages well enough to understand any of the sentences. The speech was originally recorded at 16000 Hz sampling frequency. Spectrogram of the *speech* background is shown in figure 5.

Figure 5: Spectrogram of the *speech* background. Each row represents one channel. For clarity, only 4 of the 8 channels are shown.



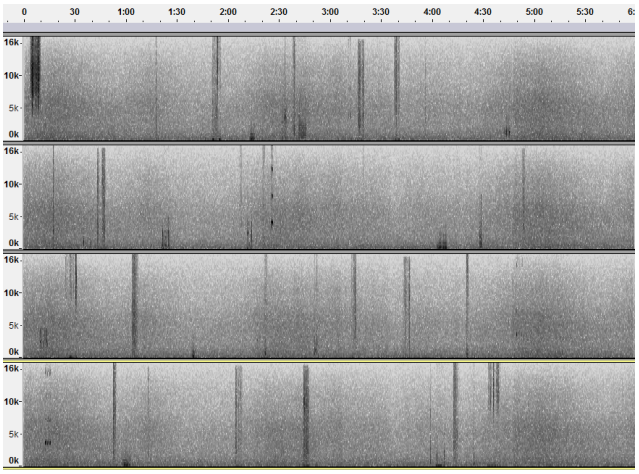
The second scene (*nature*) had a steady wind sound as background, and distracting sound events from the semantic category Nature, but different subcategories than the targets: 48 were sounds of insects, 32 of leaves and branches, and 16 of water, all positioned evenly across all 8 channels. These distractors were distributed over the background in a similar manner as target events, with one or two distractors present at any given time, and 2-4 s gaps in-between. Some distractors overlapped with targets, but because of the randomisation of target positions and timings, this overlap was different for each participant. Average background loudness was 4.3 sones, and average distractor loudness: 11 sones. Spectral centroid of distractors ranged between 780 Hz and 13600 Hz. Figure 6 shows a spectrogram of this background.

3.1.3. Reproduction system and participants

The target stimuli were reproduced over a 2nd order ambisonic system, using the Higher Order Ambisonic Library Matlab toolbox [21]. The reproduction system was 8 loudspeakers placed on an octagon, at ear-level (see Figure 4b). Background was not ambisonic but rather an 8-channel signal sent directly to the loudspeakers. All sounds were reproduced with a 44100 Hz sampling frequency.

15 volunteers took part in the experiment, 8 male and 7 female, mean age = 28.3, 13 right-handed and 2 left-handed.

Figure 6: Spectrogram of the *nature* background. Each row represents one channel. For clarity, only 4 of the 8 channels are shown.



3.2. Results

3.2.1. Data preprocessing

Figure 7 shows an example of raw data collected from the joystick movements of one of the participants in the baseline experiment.

A target event was considered attended to (a "hit") if, within a certain time window (acceptance window), the joystick was in the quadrant of the event. Thus, two things needed to be decided: limits of the acceptance window and the size of each quadrant. Both were determined from the baseline experiment.

No participants responded within the first 400 ms of any event, so this value was chosen as the lower limit of the acceptance window. We assume this to be the minimum time required for the cognitive and motor functions necessary to give a response in this setting. The upper limit of the window was set to 2 s, with which all participants were very close to their best localisation performance. A longer window could overlap with subsequent targets, and a shorter one would miss correct responses, unnecessarily reducing participants' performance.

The joystick area was divided into quadrants, each including one of the areas where targets were present, and also allowing for localisation errors around these areas (analysis quadrants were 90° wide, while target areas - only 30°). Because participants were instructed to keep the joystick in the centre if they were unsure what they were listening to, this area had to be removed from analysis. Analysis of joystick movements in the baseline experiment showed that the result is not very sensitive to the size of the central area (until it becomes close to the size of the whole joystick area). Figure 7 shows the chosen centre area and response quadrants.

3.2.2. Localisation errors

Average localisation accuracy in the baseline experiment varied from 68% to 100% between participants, indicating that, despite removing direct front and back locations from playback, localisation errors were still an issue. This accuracy was different for different sound locations, on average: 79% for the front, 81% for the back, and 99% for left and right. As expected, the main difficulty lied in localising sounds positioned in the front and back, while sounds on the left and right were localised almost perfectly.

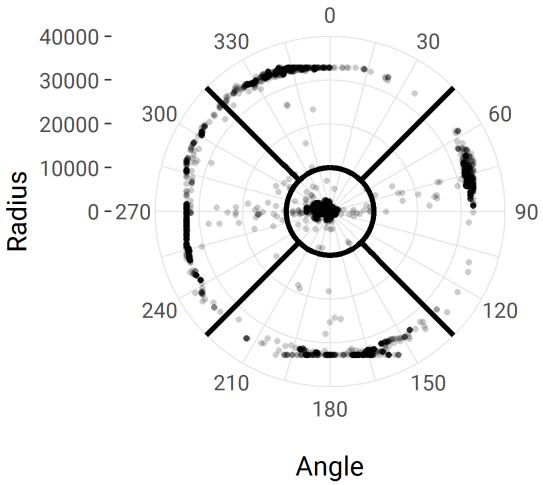


Figure 7: Raw joystick movement data for one of the baseline experiment participants. Dots are joystick positions sampled at regular time intervals. The darker the region, the more data points there are. Solid black lines show how the space was divided into quadrants and the centre area.

A GLMM model confirmed that none of the other factors (loudness, brightness, category) had an effect on localisation accuracy, nor were there any significant interactions between them.

These localisation errors will likely influence main experiment responses as well. The following section discusses how these these errors could be disentangled from effects of attention and distraction.

3.2.3. Main experiment

The total percentage of target sounds attended varied among participants, with an average of 64% and a standard deviation of 10%.

To study the effects of experimental variables on the hit/miss responses, data from the baseline and main experiments was pooled together, forming a new variable in the analysis - experiment type. By looking at interactions between 'Experiment' and other variables, we can see if adding distracting sounds - in other words, introducing attentional effects - had an effect on any of these variables.

A Generalised Linear Mixed Model (logit link, binomial distribution) was fitted with Participant as a random effect, and 2-way interactions between the Experiment type and the other independent variables (loudness, brightness, location, category and background type). The results are shown in Table 2. Wald tests indicate significant interaction effects between experiment type and loudness, and between experiment type and location.

Analysis of contrasts confirms that participants were 1.7 times more likely to attend to loud than to quiet targets in the main experiment ($p<0.0001$), while no effect is observed in the baseline. This is to be expected, as louder sounds will be more salient, and loudness should not affect localisation. However, there is also a possibility that some of this effect is due to different levels of energetic masking.

Comparison of contrasts between different locations shows the same significant differences for main and baseline experiments: front/right, front/left, back/right, back/left. These differences seem to be mainly due to localisation errors. All of these effects, however, are smaller for the main experiment than the baseline. The effect of experiment type on responses to different locations can be seen on Figure 8. Clearly, the 'hit rate' in the main experiment is generally lower than in the baseline, because in the former, participants were not asked to attend to target sounds and there were distractors. The general trend looks similar in both experiments, with more 'hits' to the sounds on the right and left, and fewer for front and back.

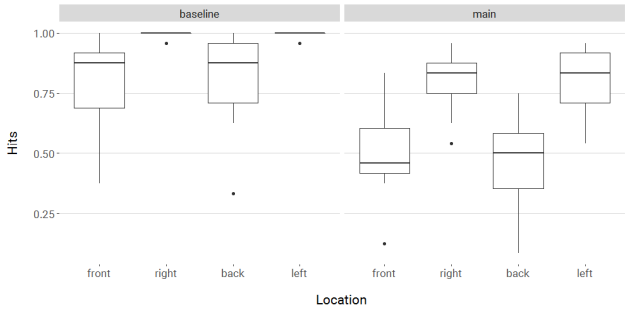
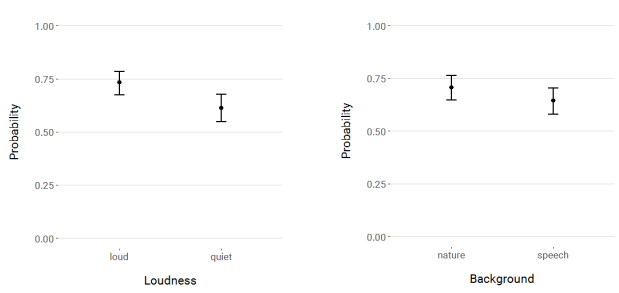


Figure 8: Responses to sounds in different positions for the baseline localisation experiment (left panel) and the main experiment (right panel). Boxplots show hit scores calculated for a particular condition and for each participant.



(a) Effect of loudness (b) Effect of background type

Figure 9: Probability of attending to target sounds in the main experiment, based on model in Table 3. Error bars show 95% confidence intervals.

To see if there were any interactions between independent variables, we analysed the main experiment data separately from the baseline data. A GLMM model with the best fit based on AIC included one interaction: location/brightness (see Table 3). The model indicates that brightness significantly changes responses to front and back locations. Analysis of contrasts shows that in the main experiment, although no significant differences were found for low brightness targets in front and back, there is a significant difference between high brightness targets presented in front and back locations, with sounds in front being more salient - see Figure 10.

The model also confirms a significant main effect of loudness, and suggests that there is a significant effect of background type,

Table 2: Coefficient estimates of the interactions in the fitted model, their standard errors, Z statistics and p-values. Note that we are mostly interested in how the main experiment interacted with other variables, not in the main effects. Predictors in bold are statistically significant.

Fixed effects	Coeff.	SE	Z	p-value
(Intercept)	**1.54**	**0.29**	**5.38**	**< 0.0001**
Channel - right	**3.91**	**0.72**	**5.42**	**< 0.0001**
Channel - back	0.15	0.19	0.77	0.444
Channel - left	**4.61**	**1.01**	**4.56**	**< 0.0001**
Loudness - loud	-0.12	0.19	-0.66	0.509
Brightness - high	0.02	0.19	0.09	0.925
Category - manmade	-0.22	0.23	-0.93	0.351
Category - nature	-0.22	0.23	-0.93	0.351
Background - nature	0.12	0.19	0.66	0.509
Experiment - main	**-2.18**	**0.31**	**-7.07**	**< 0.0001**
Experiment:Background	0.17	0.22	0.74	0.458
Category-manmade:Experiment	0.53	0.28	1.93	0.054
Category-nature:Experiment	0.43	0.28	1.57	0.116
Brightness:Experiment	0.11	0.22	0.51	0.613
Loudness:Experiment	**0.68**	**0.22**	**3.01**	**0.003**
Location-right:Experiment	**-2.43**	**0.74**	**-3.27**	**0.001**
Location-back:Experiment	-0.25	0.25	-1.03	0.302
Location-left:Experiment	**-3.11**	**1.03**	**-3.03**	**0.002**
Random effect	Standard deviation			
Participant	0.51			

with higher probability of attending to targets in the *nature* background. This is not surprising, as compared to *speech*, the *nature* background was less busy. Figure 9 shows both of these effects.

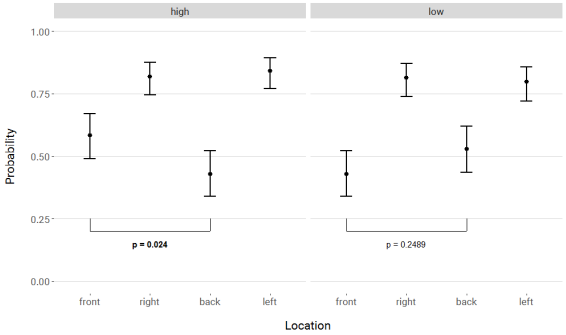


Figure 10: Probability of attending to sounds in different positions in the main experiment, split by brightness of the sound. Error bars show 95% confidence intervals. Based on model in Table 3.

3.3. Discussion

As expected, participants paid attention to louder sounds more often, which is in agreement with other studies on salience of loudness [10, 9]. The results also suggest an interaction between brightness and location of sound - there is a small decline in salience of sounds arriving from behind the listener, but only for high brightness sounds.

There are significant differences between sound categories. This could point to an influence of semantic meaning on salience. However, it is worth keeping in mind that, while the targets were balanced on the loudness and brightness scales, there might be

Table 3: Results of the GLMM model fitted with main experiment data. Significant predictors in bold.

Fixed effects	Coeff.	SE	Z	p-value
(Intercept)	**-0.89**	**0.23**	**-3.89**	**<0.0001**
Location - right	**1.76**	**0.25**	**7.12**	**<0.0001**
Location - back	0.41	0.22	1.85	0.064
Location - left	**1.66**	**0.24**	**6.81**	**<0.0001**
Brightness - high	**0.62**	**0.22**	**2.83**	**0.005**
Loudness - loud	**0.55**	**0.12**	**4.54**	**<0.0001**
Category - manmade	**0.32**	**0.15**	**2.14**	**0.033**
Category - nature	0.22	0.15	1.47	0.142
Background - nature	**0.29**	**0.12**	**2.41**	**0.016**
Location-right:Brightness	-0.59	0.35	-1.68	0.094
Location-back:Brightness	**-1.03**	**0.31**	**-3.31**	**0.001**
Location-left:Brightness	-0.32	0.35	-0.92	0.357
Random effect	Standard deviation			
Participant	0.43			

other properties of the sounds (e.g. impulsiveness) which vary between the categories. A more thorough analysis of the acoustic properties of sounds in different categories could be useful.

Because natural sounds were used as targets, other factors not taken into account in the design could influence the results, especially participant-specific subjective effects, such as personal experience or emotional reaction to a sound. With enough data points, these effects should average out, leaving the effects of the target sounds themselves. These effects will be the focus of a further study.

4. GENERAL DISCUSSION

4.1. Comparison of methods

Each of the two experiments used a different method to study the effect of sound location on auditory salience. There are a few important differences between them. Firstly, the tasks used in the two experiments were very different. It is reasonable to assume that tracking one's attention - Experiment 2 - is a more complex task, more prone to errors than the oddball detection task used in Experiment 1.

Secondly, unlike Experiment 1, Experiment 2 used a method which relied to some extent on sound localisation, which is not always perfect, and might add additional errors. This introduced the need for a way to separate localisation errors from attentional effects.

Thirdly, although no instructions about what to listen to were given in Experiment 2, it allowed for possible effects of top-down attention and personal preference for a specific sound or sound category. This makes Experiment 2 more sensitive to subjectivity and processes beyond bottom-up attention.

Finally, the experiments used very different sounds as stimuli. The advantage of Experiment 2 was its use of real-world sounds and a more ecologically valid listening environment.

4.2. Comparison of results

Despite the differences in the two methods, both experiments were designed to test auditory salience in a spatial setting. Both experiments seem to show a small decline in saliency of sounds arriving from behind the listener, but only for higher frequency sounds. This effect could potentially be explained by loudness differences

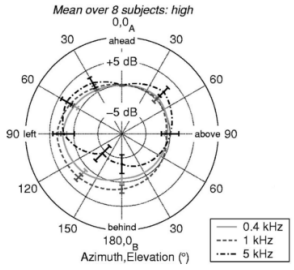


Figure 11: Directional loudness sensitivities at 65 dB SPL, reproduced from [22]

caused by pinna shadowing. [22] measured loudness for different locations around the listener (only on the left, however, as they assumed symmetry). Their results, shown in Figure 11, suggest lower sensitivity from the back for 5 kHz sounds, and almost no difference for 400 Hz and 1000 Hz sounds (third-octave noise bands), consistent with the results of the two experiments. No other effects of spatial salience were found.

Neither of the experiments showed a clear main effect of spectral content of the sound on salience, in contrast to the original studies the experiments were based on. However, it is worth pointing out that the two original studies provide contradicting results. While [2] report that lower sound patters were more salient, in [9], an increase in brightness causes an increase in salience. Both studies use the spectral centroid as a representation of brightness, however [2] only found the significant effect when the spectral centroid was calculated on sounds previously weighted with equal-loudness contours. It might be that the relationship between spectral content of sound and its salience is more complex and needs further research.

5. CONCLUSIONS AND FUTURE WORK

We have designed and tested two methods for testing spatial aspects of auditory salience. Having these methods not only lets us study the effect of location of sound on salience, but also allows conducting salience experiments in a more ecologically valid listening situation. Method used in Experiment 1 gives results which are easier to interpret, however it is difficult to use more natural sounds as stimuli. On the other hand, the method used in Experiment 2 is more prone to errors and effects of top-down attention, but allows a more natural listening environment, with real-life sounds.

The results suggest that high frequency sounds arriving from behind the listener are less salient, but the effect is not large and could probably be explained by loudness differences. If this indeed is the case, it confirms the usefulness of sound for interfaces, where an auditory alert can be placed anywhere around the person and still effectively attract their attention.

Because of the possible effects of subjectivity and top-down attention in Experiment 2, more participants will be invited to participate in it in the future. With more data points, we will be more confident that the errors caused by subjective effects average out, leaving only the effect of the sound itself. Additionally, it might be interesting to confirm this result in a distraction-type experiment, as well as to more carefully account for differences in loudness of sounds in different locations.

6. REFERENCES

[1] F. Tordini, A. S. Bregman, and J. R. Cooperstock, "Prioritizing foreground selection of natural chirp sounds by tempo and spectral centroid," *Journal on Multimodal User Interfaces*, vol. 10, no. 3, pp. 221–234, Sep 2016.

[2] ——, "The loud bird doesn't (always) get the worm: Why computational salience also needs brightness and tempo," in *21st International Conference on Auditory Display (ICAD2015), July 6-10, 2015, Graz, Styria, Austria*, 2015, pp. 236–243. [Online]. Available: https://smartech.gatech.edu/handle/1853/54145

[3] C. J. Spence and J. Driver, "Covert Spatial Orienting in Audition: Exogenous and Endogenous Mechanisms," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 20, no. 3, pp. 555–574, 1994.

[4] V. Best, F. J. Gallun, A. Ihlefeld, and B. G. Shinn-Cunningham, "The influence of spatial separation on divided listening," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1506–1516, 2006.

[5] K. Kim, K. H. Lin, D. B. Walther, M. A. Hasegawa-Johnson, and T. S. Huang, "Automatic detection of auditory salience with optimized linear filters derived from human annotation," *Pattern Recognition Letters*, vol. 38, no. 1, pp. 78–85, 2014.

[6] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[7] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," *European Signal Processing Conference*, no. Eusipco, pp. 1206–1210, 2007.

[8] T. Tsuchida and G. W. Cottrell, "Auditory saliency using natural statistics," *Proc. Annual Meeting of the Cognitive Science*, pp. 1048–1053, 2012.

[9] N. Huang and M. Elhilali, "Auditory salience using natural soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2163–2176, 2017.

[10] E. M. Kaya and M. Elhilali, "Investigating bottom-up auditory attention," *Frontiers in human neuroscience*, vol. 8, no. May, p. 327, 2014.

[11] R. Southwell, A. Baumann, C. Gal, N. Barascud, K. J. Friston, and M. Chait, "Is predictability salient? A study of attentional capture by auditory patterns," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160105, feb 2017.

[12] F. Tordini, A. S. Bregman, and J. R. Cooperstock, "Toward an improved model of auditory saliency," in *ICAD*, 2013, pp. 189–196.

[13] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[14] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia (MM'13)*, ACM. Barcelona, Spain: ACM, 21/10/2013 2013, pp. 411–412.

[15] Bbc sound effects library. [Online]. Available: https://www.sound-ideas.com/Product/152/BBC-Sound-Effects-Library-Original-Series

[16] Xeno-canto, https://www.xeno-canto.org, [Accessed: 17/07/2018]. [Online]. Available: https://www.xeno-canto.org

[17] O. B. Bones, T. J. Cox, and W. J. Davies, "Sound categories: category formation and evidence-based taxonomies," *Frontiers in Psychology*, vol. 9, p. 1277, 2018.

[18] J. Chalupper and H. Fastl, "Dynamic loudness model (dlm) for normal and hearing-impaired listeners," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 378–386, 2002.

[19] D. Cabrera, S. Ferguson, and E. Schubert, "'psysound3': Software for acoustical and psychoacoustical analysis of sound recordings." Georgia Institute of Technology, 2007. [Online]. Available: http://www.psysound.org

[20] A. Al Noori, P. Duncan, and F. Li, "Training "on the fly" to improve the performance of speaker recognition in noisy environments," in *Audio Engineering Society Conference: 2017 AES International Conference on Audio Forensics*, Jun 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18744

[21] A. Politis, "Microphone array processing for parametric spatial audio techniques," 2016. [Online]. Available: https://uk.mathworks.com/matlabcentral/fileexchange/54833-higher-order-ambisonics-hoa-library

[22] V. P. Sivonen and W. Ellermeier, "Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2965–2980, 2006.

TRACES OF MODAL SYNERGY: STUDYING INTERACTIVE MUSICAL SONIFICATION OF IMAGES IN GENERAL-AUDIENCE USE

Niklas Rönnberg

Linköping University
Media and Information Technology
SE-581 83 Linköping, Sweden
niklas.ronnberg@liu.se

Jonas Löwgren

Linköping University
Media and Information Technology
SE-581 83 Linköping, Sweden
jonas.lowgren@liu.se

ABSTRACT

Photone is an interactive installation combining color images with musical sonification. The musical expression is generated based on the syntactic (as opposed to semantic) features of an image as it is explored by the user's pointing device, intending to catalyze a holistic user experience we refer to as modal synergy where visual and auditory modalities multiply rather than add. We collected and analyzed two months' worth of data from visitors' interactions with Photone in a public exhibition at a science center. Our results show that a small proportion of visitors engaged in sustained interaction with Photone, as indicated by session times. Among the most deeply engaged visitors, a majority of the interaction was devoted to visually salient objects, i.e., semantic features of the images. However, the data also contains instances of interactive behavior that are best explained by exploration of the syntactic features of an image, and thus may suggest the emergence of modal synergy.

1. INTRODUCTION

Photone is an interactive installation combining photographic images and musical sonification [1]. In Photone, an image is displayed and a dynamically changing musical score is generated based on the overall color properties of the image and the color value of the pixel under the touch-point on the touch-screen. Consequently, the music changes as the user moves the finger exploring the image and simultaneously using the image to explore the music.

When we developed the first version of Photone, we found the quality of *modal synergy* to be potentially relevant when designing multimodal interaction, such as interactive sonification. Modal synergy refers to how two or more modalities fuse in interaction to create a user experience that goes beyond the simple sum of the parts, forming expressions that are not easily predictable, and thus stimulates engagement driven by ludic motivation and the curiosity of exploration. Our work is specifically oriented towards the visual and auditory modalities in the forms of images and musical sonification, even though we feel that the concept of modal synergy might be generative also in designing for other multimodal combinations.

In our previous work, the extent to which modal synergy actually manifests itself in the general use of Photone, and the occurrence and nature of actual explorative interaction, were left as more or less open questions. Our purpose here is to start addressing these questions by *studying actual use of Photone by a general audience in the context of a public exhibition*. We hope that this represents a worthwhile contribution to the sonification research community by providing design ideas on the interactive generation of musical elements based on syntactic image elements, as well as insights into the experiential qualities of multimodal interaction in which sonification plays a constituent role.

2. USING PHOTONE

To make the arguments and discussions presented in this paper easier to understand, let us try to convey a sense of the synergistic interaction experience we are talking about. A short video demonstration of Photone to complement the following vignette can be found here: https://vimeo.com/322740494.

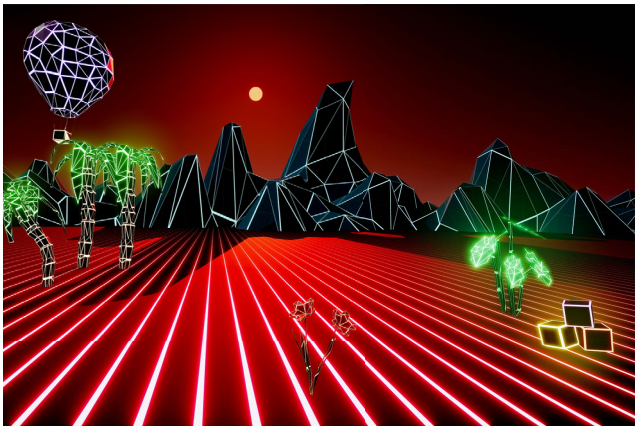


Figure 1: One of the images used in Photone. Image courtesy Norrköping Visualization center C.

Consider the image from Neonland Experience (Figure 1), entering the image with the finger on the touch-screen in the top right corner. The music is a bit muffled, quite low in its amplitude and with the high frequencies attenuated. The impression is that it is dark, both the music and the image, and the music is experienced as somewhat anticipatory with the ominous red sky and futuristic

https://doi.org/10.21785/icad2019.010

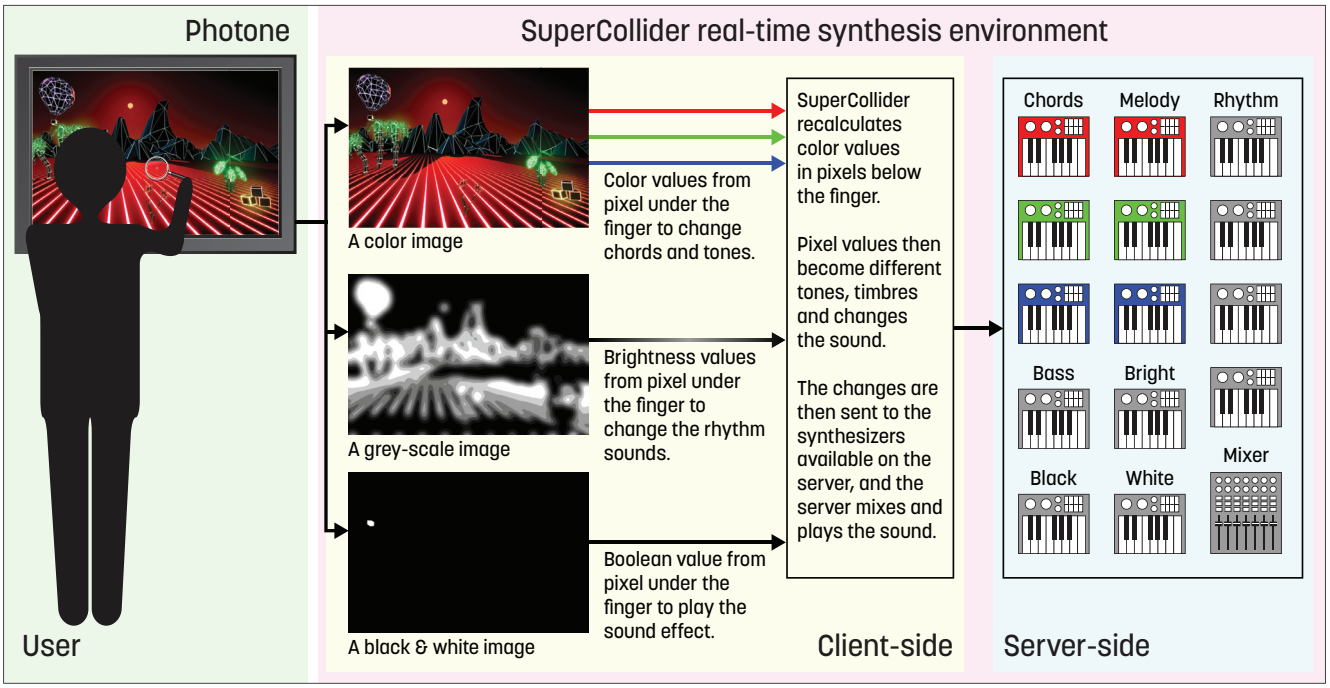


Figure 2: Schematic of the structure in Photone showing the user, the client-side, and the server-side. The interactive sonification is created in SuperCollider by three images, and the color values in these images are mapped to different musical parameters. The server runs the synth definitions for the harmonic ambience (Chords), the melodic components (Melody), the low bass tones (Bass), the high light intensity chord (Bright), the bell-like sound (White), the low frequency sweep (Black), the rhythmic instruments (Rhythm), and the output mixer with reverberation and low-pass filter (Mixer).

digital landscape. As the finger moves towards the brighter area close to the mountain range, the amplitude of the music increases and rising melody keeps the explorer company. If you go back towards the darker area, the melody changes direction and goes downwards in pitch while the background harmonics become more muffled again. When you come close to the mountains, a rhythmic beat accompanies the melody and the harmonic background. As you enter the mountain area most of the musical elements die away, apart from the rhythmic instruments that play along emphasizing the contrasts, the white lines on the black background of the mountains. These bright lines in the mountains now feel like the keys on a piano keyboard, creating outbursts of high pitch tones when you cross over each one of them. As you continue down from the mountains and enter the area of the bright green plant, new harmonic content comes to the fore with new melodic tones that mix with the harmony and tones from the red. The bright lines in the mostly red gradient to the left of the green plant also act like piano keys, playing rising and falling melodic movements as the red color changes from darker red in the middle between the bright lines to almost white in the middle of each line.

With this example we hope to give a passing acquaintance with the interaction and user experience in Photone. We believe that it is rather fruitless to discuss whether one modality augments or supplements the other in Photone. The interaction with image and music has holistic qualities that combine into what we call modal synergy, creating an experience that is larger than its individual components. The example is also meant to show that the image is considered a collection of pixels with specific color values, and that the temporal trajectories in the music are formed by spatial movement across the surface of the image.

3. DESIGN OF PHOTONE

In Photone, color values in the image at the specific pixel under the touch-point on the touch-screen are read and mapped to different musical elements in the sonification.

3.1. Musical elements

The composition in Photone consists of seven musical elements (see Figure 2). These elements are 1) the overall harmonic ambience, 2) melodic components, 3) two low bass tones, 4) a high light intensity chord, 5) a bell-like sound for pure white, 6) a low frequency sweep for pure black, and 7) rhythmic instruments to highlight areas in the image that are rich in contrasts. The synthesis method used for the musical elements 1 to 6 are described in [1], and the following text will describe the composition of Photone used in the current study.

Similar to the previous version of Photone, the overall harmonic ambience and the melodic components are composed with the three color channels of red, green, and blue (RGB) in mind. The harmonic ambience has been slightly simplified compared to the previous version, and now consists of two-tone intervals multiplied over five octaves, creating a harmonic ambience with ten tones for each color channel. Depending on the pixel value (i.e. the color) at the touch-point under the finger the harmonic ambience varies from a two-tone interval (when information in only

one-color channel is present) to a complex chord (when information in all three color channels are present). The light intensity of the individual color channel determines the amplitude of the components in the harmonic ambience, reflecting how the perception of loudness is closely linked to the perception of brightness [2]. The light intensity is also mapped to the cut-off frequency of a second order band-pass filter between 100 and 4000 Hz for each channel. This makes the harmonic ambience louder and with more high frequency content in bright areas in the image compared to darker areas.

The number of tones for the melodic components is in the current version increased to ten tones for each color channel which are played one tone at a time. The intensity level in each color channel is divided into ten steps and one of the tones is used accordingly. This creates an upwards going melodic movement when intensity in that specific color channel increases, and a downwards going melody when intensity decreases. There is an association between pitch of tones and colors where, for example, higher pitched tones are associated with lighter and brighter colors (see for example [3, 4, 5]). Similar to the harmonic ambience, the melodic components also vary in amplitude and in band-pass filter cut-off frequency according to the intensity level in the pixel value at the touch-point under the finger.

As in the previous version of Photone the two low bass tones are only present when the overall intensity level is low, to emphasize the impression of darker colors. The high light intensity chord is composed with three tones and is only present when the overall light level is high to create an airy and high-intensity feeling. The short bell-like sound is used to further accentuate the dazzling intensity of white, and the downwards sweeping low frequency sound is used to emphasize the change in intensity from different shades of color to darkness.

3.2. Rhythmic instruments

In the current version of Photone rhythmic instruments are added. These rhythmic instruments are synthesized to mimic congas, triangle, and hi-hat sounds, and are used to rhythmically emphasize the amount of contrasts in the images. A script in Matlab divides each image used in Photone into 8 levels of contrast (see Figure 2), where contrast level 0 has no rhythmic instruments but higher levels of contrasts are sonified with increased level of rhythmic sounds. The levels of contrast are chosen as discrete values, so there are clearly defined levels of rhythmic components (see Table 1).

Table 1: The amplification levels of the four rhythmic instruments in the eight levels of contrast.

	0	1	2	3	4	5	6	7
Instrument 1	0	0.5	1	1	1	1	1	1
Instrument 2	0	0	0	0.5	1	1	1	1
Instrument 3	0	0	0	0	0	0.5	1	1
Instrument 4	0	0	0	0	0	0	0.5	1

3.3. Overall image color

Similar to the previous version of Photone each image is determined to have an overall color that affects the composition and musical expression. However, in the current version of Photone the color of an image is determined by a human rather than by weighted means in the RGB color channels. As the RGB color model is not well adapted to the human color perception [7], this method of choosing the overall color in an image should better correspond to a users impression of an image. The colors available to choose from are yellow, orange, red, white, purple, green, and blue (see Table 2). Based on psychology of colors [6] the composition, as well as the synthesis of the sounds, are adapted to better fit, not to mimic but rather to complement, the impression of the overall color in the image. The reason for adjusting the composition according to the overall color is to attract the user to continued exploration of Photone and to vary the musical expression between images.

A number of musical elements are adjusted according to the selected overall color (see Table 2). The harmony of the harmonic ambience is changed due to the color, where major chords are used for the warmer colors while minor chords are used for the colder colors. Colors with more positive impressions are thus accompanied by chords in major, which in turn might be experienced as more positive [8]. Furthermore, the complexity of the chords is chosen to correspond, at least to some degree, to the energy and the complexity in the colors. The complexity of the chords is connected to the experience of the musical sounds, as a more complex harmonic sound is more captivating for a listener compared to a simpler harmonic sound [9]. However, it is important to keep in mind that the impression of the musical sonification is created by the combination of musical elements, and not by the selection of chord alone.

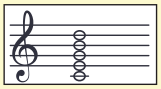
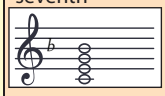
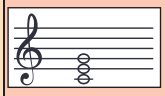
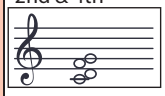
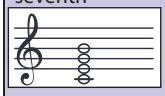
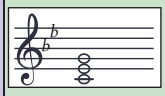
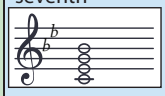
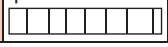
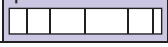
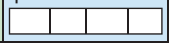
In Photone the dissonance of each tone, i.e. the spread in frequency of the pitches creating each tone, used in the harmonic ambience varies in relation to the impression of the colors. Colors with more energy have a greater dissonance, creating tones with more energy, while colors with less energy have more unison and relaxing tones. The timbre of the harmonic ambience and the melodic components is changed by altering the pulse-width of the square wave forms, where a deviation from 50% pulse-width creates more harmonics compared to 50%. Colors associated as more positive with more energy have pulse-widths creating more harmonics. A softer timbre is experienced as more negative compared to a brighter timbre [10]. By changing the pulse-width the sound changes from rich and prominent sound at 80% pulse-width to a simpler and more hollow sound at 50% pulse-width. The musical sonification is output through a low-pass filter and the cut-off frequency is adjusted according to the overall color, where the sonification for the more positive colors has more high frequency content while the less positive colors have their high frequencies attenuated. The tempo of the rhythmic composition is also generally faster and the rhythm is more complex for the colors with more energy. These musical elements together create a difference in the musical expression between the different colors.

4. IMPLEMENTATION

Photone (see Figure 2) is implemented in SuperCollider 3.10, which is a real-time audio synthesis programming environment [11, 12].

At the request of the science center where Photone is part of the public exhibition, an Easter egg is implemented in each image. This Easter egg is a sound sample that is played back, and mixed with the musical sonification, if the user happens to explore a certain small area of the image (see Figure 2). These sound samples are sound effects that are connected to motif of the image, for

Table 2: Each image used in Photone is determined to have an overall color that affects the composition and musical expression. The table shows the colors used in the current version of Photone, the impression of the color according to Cleary [6], and the musical elements affected of these colors.

	Yellow	Orange	Red	White	Purple	Green	Blue
Color impression	Happiness & vitality	Energy & joyfullness	Passion & intensity	Purity & cleanliness	Creativity & loveliness	Nature & serenity	Calm & sadness
Harmony	Major, ninth 	Major, minor seventh 	Major 	Suspended 2nd & 4th 	Major, major seventh 	Minor 	Minor, minor seventh 
Dissonance	+/- 30 cents	+/- 22.5 cents	+/- 15 cents	+/- 1 cent	+/- 10 cents	+/- 10 cents	+/- 2.5 cents
Timbre (PW)	80% 	70% 	65% 	50% 	60% 	60% 	55% 
LPF cutoff	14kHz 	12kHz 	10kHz 	8kHz 	6kHz 	4kHz 	3kHz 
Tempo (BPM)	98 	96 	94 	96 	94 	92 	90 
Rhythm	Most complex and dense rhythmic pattern. 	Slightly less complex and dense rhythmic pattern. 	Less complex and dense rhythmic pattern. 	Less complex and dense rhythmic pattern. 	Not that complex and dense rhythmic pattern. 	Not complex and sparse rhythmic pattern. 	Least complex and dense rhythmic pattern. 

example the imaginary sound of a science fiction hot air balloon (see Figure 1). The science center argued that searching for these Easter eggs would engage and motivate the user to further explore Photone.

The previous version of Photone was explored with the mouse cursor, however the present version is implemented in a touch-screen environment. To avoid covering the area of interest with the finger, an overlay of an image showing a magnifying glass is implemented with a position offset compared to the actual touch-point (see Figure 2 and Figure 3). The image of the magnifying glass is chosen since the normal use of a magnifying glass is to look into the glass while holding the magnifying glass with an offset to the area of interest.

5. SONIFICATION, MUSIC, AND IMAGES

Even if we claim that the scope of Photone is something new, the combination of images and music is nothing new in itself. The history is full of interesting examples of composed music for images and motion pictures. In film, music is used to help the audience to realize the meaning of the film, to guide the audience to understand the films dramatic and emotional value [13]. Music is also used to create a convincing atmosphere of time and place [14]. The list of the roles of music as a creator of emotions and continuity can be made very long [15], but the use of musical elements in film music differs from Photone. In most cases where music is used as a complement to an image, the music is composed to images based on their denotative and connotative meaning. For example, a sad image is mirrored by sad sounding music, maybe with minor chords and a slow tempo, while a happy image might be reinforced by positive rising melodies, major chords, and a more forward-going tempo. The aesthetics of the music and of the intended meaning is then a psychological analysis, an interpretation and explanation of the experience of the music for the composer as well as the listener [16]. Our artistic intention in Photone is another: By building the sonification upon pixel values of hue and brightness, that is, syntactic rather than semantic properties of an image, we aim to cut through conventional ways of seeing to a more foundational level. Instead of adding sad music to a sad image, Photone elaborates the small elements that create an image, where dark areas in the image are more attenuated and have a more dull timbre compared to brighter areas, where a gradient creates rising and falling melodies, where clear and sharp contrasts between different hues create rhythmic patterns, and where the finger on the touch-screen becomes the conductor's baton.

As motion pictures, film music, and technology evolved, artists discovered the new tool of optical sound for music production and aesthetic sound synthesis [17]. The artists manipulated the image input to the photocell of the image-to-sound converter [18], unlike film music using the image in syntactic rather than denotative or connotative ways. However, these early artistic explorations and expressions lacked the interaction that unites the visual and auditory modalities. With the development of computers and the use of these as a dynamic medium the artistic emphasis shifted towards the interactive experience of audiovisuality [17]. In an aesthetic experience, in a fully interactive installation, both sound

and image are the means through which the user interacts, and the products of interaction [19]. In Photone the visual expression on the display is not affected by the interaction, and thus the outcome of the interaction, the modal synergy, is formed by nonlinear exploration of the static image and the variable musical sonification. Superficially, the concept of Photone might be described as musicalization of a visual image, as the interaction incorporates musical aspects into the image making it possible to listen to the music of the colors [20]. However, in Photone image, sound, and interaction are tightly integrated, and we argue that the three elements are aspects of the same emergent experience in use [1].

Tanaka [21] converted photographic images to sound. The idea was to create a musical work that replaced the image evoking the same emotional response as the image. This was done by converting the image data, scanning pixels in the image, to sound in different ways, for example brighter values of gray became sound samples of higher amplitudes. But, even if this approach was similar to Photone in that it sonified pixel values rather than the motif in the image, it was not interactive. For sonification to be useful for data exploration, and we would like to argue that sonification of image elements to some extent could be seen as data exploration, dynamic human interaction is necessary [22, 23]. There have been different approaches towards interactive sonification for images apart from Photone (see some examples, discussions, and variations in [24, 25, 26, 27]). O'Neill and Ng [28] presented an interactive sonification to provide feedback in exploration of structure in images. In the user testing O'Neill and Ng found that the sonification supported the foundation of a higher level of understanding of the image structure. Similarly to Photone, O'Neill and Ng used different parameters in the sonification, for example modulation of timbre and pitch. In Photone the sonification is adjusted to every individual pixel in the image, while O'Neill and Ng rather used sonification to differentiate between segments in the image. Heath and Gordon [29] suggested a "primitive" sonification of an image, where an input audio signal was transformed according to statistical interpretation of the average intensity levels in the RGB color channels of an image, in regard to semitone, scale, and chord. This sonification approach has similarities to Photone, even though the proposed sonification was not reported with a user study exploring the user experience or the sonification. Furthermore, it is not clear what the aim was with the suggested sonification: to support visual perception, to evaluate a multi-modal user interaction, or to provide an artistic experience. There has also been examples of sonification of images for the visually impaired (see examples in [30, 31, 32, 33]), even if these show interesting sonification approaches they aim to achieve something slightly different than Photone. These examples try to define the environment for a visually impaired individual, for example in object or obstacle detection, rather than providing a musically interesting multi-modal user experience that aims for modal synergy.

6. METHOD

Photone is exhibited in the science center at Campus Norrköping, Linköping University. It is part of the exhibition Decode the code that aims to explain computer graphics and visualization techniques to the general public (see Figure 3). The science center has about 100,000 visitors per year of all ages, even though school classes of 10- to 14-year-old kids are particularly frequent visitors. In Photone, the user can select from twelve different images to explore and experience. These images come from the different installations in the science center. All images in the exhibition are synthetic and used to explain volumes, voxels, triangles, shaders and similar computer and visualization concepts for the general public. For the use in Photone, the images were somewhat enhanced in terms of richness of colors, and in range between darker and brighter areas in the image.

Figure 3: Photone being explored by a visitor in the science center, with the ultrasonic distance sensor visible in the dark space between the purple and blue area below the touch screen.

The data from the interactions were collected for two months where roughly 8,000 persons visited the center. The data saved consists of the image that is shown on the display, the position of every pixel explored and the timestamp in milliseconds for each pixel. The update frequency of the data recording is 120 Hz, and the data is saved as a log file named with the date and time for the interaction.

An ultrasonic distance sensor is used to determine if a user is present in front of Photone. When the detected distance exceeds 1 meter the system interprets this to mean that the user has left Photone and saves the data. There are multiple challenges in determining if a user is in front of Photone, for example, if a user leans too far to one side this might be interpreted as the user has left Photone and the interaction will end up in two shorter log files. Moreover, if a user replaces another user too closely, this might go undetected with one log file ending up containing two actual sessions. These challenges could have been overcome by video recording of Photone and the environment around it. However, the distance sensor was deemed to be the best tradeoff between accuracy and privacy.

The data has been analyzed in terms of the total amount of interaction time for each log file. Log files shorter than 30 seconds or longer than 50 minutes have been discarded as either being too short to be interesting for the analyzes or too long to be considered interactions from only one user. The remaining interaction files (n = 233) have been analyzed and visualized with a histogram.

The longest 10% of all interactions (n = 23) have been further analyzed to study the interaction behavior for users who interacted a long time with Photone. The analysis of these has explored the interaction patterns within each image. The interaction patterns for each image have been analyzed using a heat map where the number of visits to any pixel in the image gets a higher value in the heat map. Based on the heat maps the images have been studied to

identify the image features that seem to attract the users.

7. RESULTS

The level of engagement, as measured in interaction session times, is a long-tail distribution. Photone is not engaging for everyone, and it is clear that most users only interact with Photone for 5 minutes or less, and that the most frequent time interval is 0.5 to 1.5 minutes (see Figure 4). We will focus on the long sessions here, assuming they represent more engaged use.

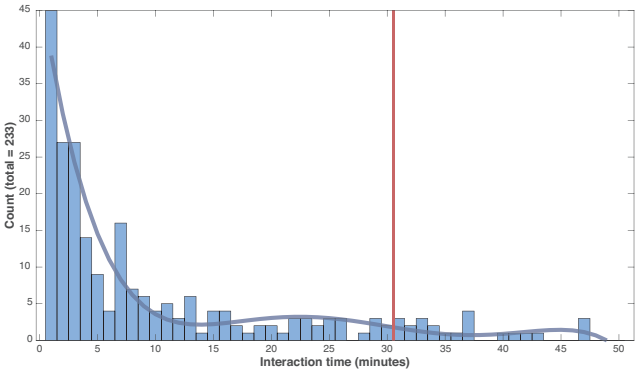


Figure 4: Histogram displaying the interaction time for all interactions (n = 233), with a trend line in dark blue. User interactions to the right of red vertical line (n = 23) are used in the detailed level of the analysis.

The analysis of the 10% longest interactions (n = 23) was done with heat maps displaying the most visited areas in each image. Most of the engaged use seems to follow (semantic) visual salience, showing a pattern very similar to what conventional eye-tracking data of image viewing would look like (see, for example, [34, 35]). Figures 5 and 6 illustrate this type of distribution.

However, the data also reveals some indications of interaction engagement that are not as clearly similar to visual salience effects in regular image viewing. Some examples are shown in Figures 7 and 8.

8. DISCUSSION

Already in our previous study [1], our conjecture was that only a limited proportion of a general audience would find interacting with Photone engaging beyond the cursory examination and superficial poking. The distribution of the log data in our present work (Figure 4) confirms this conjecture.

What we did not expect was that among the engaged users (longest 10% of recorded sessions), a majority of the interaction engagement follows a pattern that closely resembles the results we would expect from a conventional eye-tracking study of image viewing. In other words, it would seem like the visually salient objects in an image guide the attention as well as the interactive exploration of the image to a great extent. These areas in the images generally contain more clearly defined visual contrasts resulting in more rapid changing melodies and differences in tonal qualities, as well as more rhythmic instrumentation compared to the background. A plausible scenario could be that a user's attention is first attracted by the salient visual object, and then the user stays in that area of the image for further exploration of dramatic musical effects and satisfying rhythm.

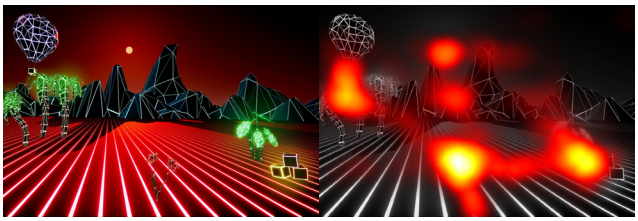


Figure 5: The heat map (to the right) clearly shows that the users mainly explored visual objects in the image, such as the flowers in the foreground, the bush to the right, the palm trees and the hot air balloon to the left, and the sun in the middle.

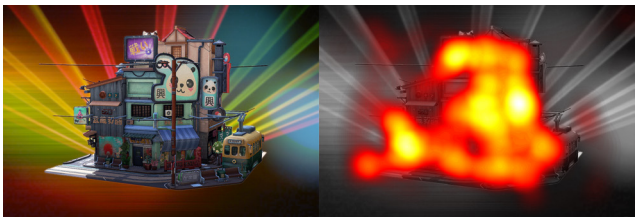


Figure 6: The heat map (to the right) suggests that the block of houses in the middle of the image got much more attention from the user than the background and the color bars emitting from behind the block.

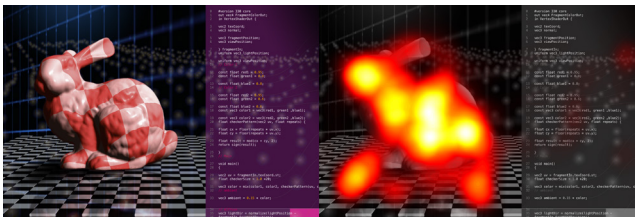


Figure 7: The heat map (to the right) shows that the rabbit was getting more explored than the areas around it. However, for this image the users also explored the shadow beneath the rabbit as well as the reflection in the checkered marble floor.

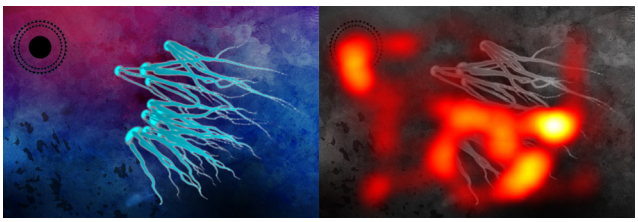


Figure 8: The heat map (to the right) suggests an interaction pattern where the visually salient objects as well as parts of the non-depictive background were explored.

However, our data does contain examples of interaction engagement also in image areas that are not visually salient in the conventional sense of the word. Looking more closely at those areas, we find that they contain textures and gradients that may yield

engaging visual-auditory synergistic effects in the current design of Photone. For example, in the shadows in front of the rabbit (see Figure 7), the checkers pattern creates clear differences between darker, more sparse musical expression and brighter, more energetic music. The colored reflections from the rabbit also add to the timbre and color the tonal content, as the area is rich in contrasts which creates a quite complex rhythmic instrumentation. A similar experience can be found in the long tails of the cyan shapes (see Figure 8), where the bright tails and darker background create similar experiences as the checkers in the previous example. Also here the area is quite rich in contrasts creating an interesting rhythmic instrumentation.

These are examples of exploration of image areas with high syntactic complexity, from the point of view of the sonification algorithm, but without strong semantic salience in the conventional image-viewing sense. As such, these examples do suggest that there is some amount of modal synergy at work in the experience guiding the user's exploration, albeit less than we had expected when starting our data collection.

One potential reason for the discrepancy is that the first version of Photone, forming the basis for the concept of modal synergy, used proper photographic images. The current version, for which the general audience data were collected, contains synthetic images such a 3D-renderings and even a few flat 2D vector graphics with solid colors as the images used are drawn from an exhibition about computer graphics and visualization. There are significant visual differences between these two versions, in that the photographs have much more nuance, texture and detail than the synthetic images. The sonification mechanism of Photone was originally designed to reward exploration of high-spatial-frequency and visually intricate image areas, which are less prevalent in synthetic images. We suspect that testing a version of Photone with proper photographs would yield slightly different data in general-audience use, possibly showing a more even distribution between visually salient semantic objects and visually-auditively interesting syntactic features.

Another option could be to use less depictive images, where the lack of clearly identifiable visual objects might mitigate the power of visual salience. It would be tempting to test a version of Photone with less-figurative paintings, chosen from e.g. expressionism or pointillism, where a relative lack of visually salient semantic objects is combined with visually complex and engaging syntactic features resulting from the artists craft skills and personal techniques in using brushes and paints with different properties on a textured canvas.

9. CONCLUSION AND FUTURE WORK

Photone does not attract everyone in a public exhibition space, but some of the visitors engage more deeply in the interaction. In general, visually salient objects in an image guide the attention of an engaged user, and exploring them may or may not represent an experience of modal synergy for the user. However, our data also shows interaction engagement in image areas which do not contain visually salient objects in the conventional, semantic sense but rather syntactic features that may yield engaging visual-auditory synergistic effects. Consequently, we claim that there may be traces of modal synergy at work in the general-audience use of Photone. We find these results encouraging enough to warrant further exploration.

A necessary first step would be to validate the method used in the present study by combining log data with qualitative data to start unpacking the nature of the user experience while interacting with Photone, and specifically how subjectively perceived modal synergy relates to log data as represented in our heat maps above. Self-reflection through prompted recall would be a suitable approach to collect qualitative experiential data than can be correlated with previously collected logs, and think-aloud protocols during use can also be considered even though there is always the danger that verbalization during the interaction changes the nature of the experience significantly.

Such validation can be expected to provide guidelines for more robust interpretation of future interaction logs. Next, a more systematic quantitative study of different types of images, as discussed in section 8 Discussion, all using the same sonification algorithm would be most enlightening. More generally, these considerations form the basis for a slightly more systematic approach to visual image space where candidate images for Photone sonification can be placed along two dimensions: level of complexity in syntactic visual features, and level of semantic figurativeness. Empirical testing with the same sonification algorithm applied to images from all four quadrants of this space could yield two major insights: (1) a better understanding of which kinds of images work best with the current sonification algorithm to evoke modal synergy experiences, and (2) inspiration for re-designing sonification algorithms for each of the four quadrants towards more modal synergy. Such a qualitative study would also answer questions about the experience of the sonification in itself to provide insights in, for example, the quality of the synthesis or the use of musical elements.

A third step in this progression, assuming that we find differences between image types, would be to re-design sonification algorithms specifically to engender modal synergy for each of the main types of images, and to assess them through further rounds of quantitative evaluation.

10. REFERENCES

[1] N. Rönnberg and J. Löwgren, "Photone: Exploring modal synergy in photographic images and music," in *Proc. International Conference on Auditory Display (ICAD 2018)*, 2018, pp. 73–79.

[2] R. W. Pridmore, "Music and color: Relations in the psychophysical perspective," *Color Research & Application*, vol. 17, pp. 57–61, 1992.

[3] W. G. Collier and T. L. Hubbard, "Musical scales and brightness evaluations: Effects of pitch, direction, and scale mode," *Musicae Scientiae*, vol. 8, pp. 151–173, 2004.

[4] L. E. Marks, "On cross-modal similarity: Auditoryvisual interactions in speeded discrimination," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 13, pp. 384–394, 1987.

[5] J. Ward, B. Huckstep, and E. Tsakanikos, "Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all?" *Cortex*, vol. 42, pp. 264–280, 2006.

[6] S. P. Cleary, "Using the psychology of color schemes to create an appreciative advising environment," *Journal of Appreciative Education*, vol. 2, pp. 24–38, 2015.

[7] C. Ware, *Information Visualization: Perception for Design*, 3rd ed. San Francisco: Morgan Kaufmann Publishers Inc., 2013.

[8] E. G. S. Patrick G. Hunter and U. Schimmack, "Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 4, pp. 47–56, 2010.

[9] S. A. Iakovides, V. T. Iliadou, V. T. Bizeli, S. G. Kaprinis, K. N. Fountoulakis, and G. S. Kaprinis, "Psychophysiology and psychoacoustics of music: Perception of complex sound in normal subjects and psychiatric patients," *Annals of General Hospital Psychiatry*, vol. 3, pp. 1–4, 2004.

[10] P. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of New Music Research*, vol. 33, pp. 217–238, 2004.

[11] J. McCartney, "Supercollider: A new real-time synthesis language," in *Proc. International Computer Music Conference (ICMC)*, 1996, pp. 257–258.

[12] ——, "Rethinking the computer music language: Supercollider," *IEEE Computer Graphics & Applications*, vol. 26, pp. 61–68, 2002.

[13] R. M. Prendergast, *Film Music: A Neglected Art*. W.W. Norton, 1992.

[14] C. Gorbman, *Unheard melodies: narrative film music*. London: Bloomington: BFI Pub.; Indiana University Press, 1987.

[15] B. Langkjær, *Filmlyd & filmmusik: fra klassisk til moderne film*. Museum Tusculanum Press, 1997.

[16] C. E. Seashore, *Psychology of Music*. New York, US: Dover, 1967.

[17] L. Ribas, "Sound and image relations: a history of convergence and divergence," *Divergence Press*, 2016.

[18] T. Y. Levin, "Tones from out of nowhere: Rudolph Pfenninger and the archaeology of synthetic sound," *Green Room*, vol. 12, pp. 32–79, 2003.

[19] G. Levin, "Audiovisual software art," in *Audiovisuology: Compendium*, D. Daniels and S. Naumann, Eds. Köln: Verlag Walther König, 2010, pp. 271–283.

[20] S. Naumann, "The expanded image: On the musicalization of the visual arts in the twentieth century," in *Audiovisuology, A Reader, Vol. 1: Compendium, Vol. 2: Essays*, D. Daniels and S. Naumann, Eds. Köln: Verlag Walther König, 2015, pp. 504–233.

[21] A. Tanaka, "The sound of photographic image," *AI & Society: Knowledge, Culture and Communication*, vol. 27, pp. 315–318, 2012.

[22] T. Hermann and A. Hunt, "The discipline of interactive sonification," in *Proc. of the Int. Workshop on Interactive Sonification Workshop (ISON-2004)*. Germany: Bielefeld University, 2004, pp. 1–9.

[23] A. Hunt and T. Hermann, "The importance of interaction in sonification," in *Proc. of the 10th Meeting of the International Conference on Auditory Display (ICAD 2004)*, Sydney, Australia, 2004, pp. ICAD04–1–ICAD04–8.

[24] R. Sarkar, S. Bakshi, and P. K. Sa, "Review on image sonification: A non-visual scene representation," in *Proc. 1st International Conference on Recent Advances in Information Technology (RAIT)*, Dhanbad, India, 2012, pp. 86–90.

[25] T. Hermann, A. Hunt, and J. G. Neuhoff, *The Sonification Handbook*, 1st ed. Berlin, Germany: Logos Publishing House, 2011.

[26] T. Pinch and K. Bijsterveld, *The Oxford Handbook of Sound Studies*. Oxford University Press, 2012.

[27] K. Franinovic and S. Serafin, *Sonic Interaction Design*. MIT Press, 2013.

[28] C. O'Neill and K. Ng, "Hearing images: Interactive sonification interface for images," in *Proc. of the International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution*. Florence, Italy: IEEE Computer Society, 2008, pp. 25–31.

[29] M. D. Heath and G. Hunter, "Listen to the picture! the statson sound sonification system, using vst and dsp," in *Proc. of the Acoustics 2012 Nantes Conference*. Nantes, France: Société Française d'Acoustique, 2012, pp. 3861–3866.

[30] M. Banf and V. Blanz, "A modular computer vision sonification model for the visually impaired," in *Proc. of the 18th International Conference on Auditory Display (ICAD 2012*, Atlanta, GA, USA, 2012, pp. 121–128.

[31] ——, "Sonification of images for the visually impaired using a multi-level approach," in *Proc. of the 4th Augmented Human International Conference (AH13)*, Stuttgart, Germany, 2013, pp. 162–169.

[32] S. Cavacoa, J. T. Henriquesb, M. Menguccia, and F. M. Nuno Correiaa, "Color sonification for the visually impaired," *Procedia Technology*, vol. 9, pp. 1048–1057, 2013.

[33] P. Skulimowski, M. Owczarek, A. Radecki, M. Bujacz, and P. Strumiłło, "Interactive sonification of the u-disparity maps of 3d scenes," in *Proc. 5th Interactive Sonification Workshop (ISON-2016)*. Germany: CITEC, Bielefeld University, 2016, pp. 18–22.

[34] D. Massaro, F. Savazzi, C. D. Dio, D. Freedberg, V. Gallese, G. Gilli, and A. Marchetti, "When art moves the eyes: A behavioral and eye-tracking study," *PLoS ONE*, vol. 7, pp. 1–16, 2012.

[35] D. Villani, F. Morganti, P. Cipresso, S. Ruggi, G. Riva, and G. Gilli, "Visual exploration patterns of human figures in action: an eye tracker study with art paintings," *Frontiers in Psychology*, vol. 6, pp. 1636 1–10, 2015.

SOCCER SONIFICATION: ENHANCING VIEWER EXPERIENCE

Richard Savery and Madhukesh Ayyagari

Georgia Tech Center for Music Technology
Georgia Institute of Technology
Atlanta, USA
rsavery3@gatech.edu
madhukesh.ayyagari@gmail.com

Keenan May and Bruce Walker

Sonification Lab
Georgia Institute of Technology
Atlanta, USA
bruce.walker@psych.gatech.edu
kmay@gatech.edu

ABSTRACT

We present multiple approaches to soccer sonification, focusing on enhancing the experience for a general audience. For this work, we developed our own soccer data set through computer vision analysis of footage from a tactical overhead camera. This data-set included X, Y, coordinates for the ball and players throughout, as well as passes, steals and goals. After a divergent creation process, we developed four main methods of sports sonification for entertainment. For the Tempo Variation and Pitch Variation methods, tempo or pitch is operationalized to demonstrate ball and player movement data. The Key Moments method features only pass, steal and goal data, while the Musical Moments method takes existing music and attempts to align the track with important data points. Evaluation was done using a combination of qualitative focus groups and quantitative surveys, with 36 participants completing hour long sessions. Results indicated an overall preference for the Pitch Variation and Musical Moments methods, and revealed a robust trade-off between usability and enjoyability.

1. INTRODUCTION

Sports generate a wealth of data, including long-term statistics across games, seasons and careers, as well short-term analysis of player and ball movement during games. A large collection of research has developed across the last four decades focusing on using this data to improve physiology, psychology, and biomechanics[1]. In this paper we present and evaluate multiple approaches to soccer sonification, specifically geared towards entertainment for a general audience. Our goal is to use data to create an enhanced experience through increased perception of key events and complementary music.

2. RELATED WORK

Soccer data tracking and analysis is ubiquitous in professional soccer. These data are analyzed to help manage player fatigue [2], manage and identify long term trends such as the increased distance covered by players [3] and, crucially, to discover how these factors contribute to winning games[4].

Sonification for sports and physical activity has been explored in many different research projects, although, to our knowledge, not for pure entertainment enhancement. Barrass et al. [5] compared six approaches to sonifying accelerometer data for non-specific exercise, focusing on user enjoyment during exercise. Amongst these approaches, which were algorithmic music, sonification, weather metaphor, formants, musicification and stream-based, the algorithmic music approach was shown to be the most popular, with participants noting it had a large amount of variety and was sensitive to their actions. Specific movements have been sonified to assist with physical activity such as squats [6], or to help predict future movements in sport [7], or to guide tactics [8]. Using sonification to optimize and improve athletic performance has been studied in the context of specific sports (elite sport rowing techniques [9]) as well as general techniques for real-time heart rate monitoring geared towards athletes [10]. Schaffert [11] presented results from a workshop on the use of real-time sonification to increase performance by athletes. Sports have also been sonified to allow visually impaired users to participate, such as sonified aerobics [12]. Conversely, many audio sports have been created, including an interactive soccer environment using only audio [13].

3. SYSTEM OVERVIEW

Our system is divided into three distinct components. First, the data are created through computer vision. They are then processed and mapped in MaxMSP, which in turn sends MIDI messages out to sample libraries and controls other parameters (such as tempo) in Logic. Figure 1 displays the system overview.

3.1. Data Creation

In our original sound design, creations we were able to use a data set collected directly by an established football club. However, for the purposes of evaluation and public sharing, we were required to use an external data set due to player data privacy laws. Some data sets exist containing soccer movements [14] although we were unable to find a data set that contained player movement and professional quality video. Professional clubs are understandably guarded about player and team data as it may lead to a competitive advantage. To create data we collected video from the tactical camera view of all rounds from the 2018 FIFA World Cup (see Figure 2). This footage allowed us to implement some relatively straightforward computer vision techniques to extract player and ball X,Y positions throughout the game. We started by analyzing the slight camera panning and movement using previously

https://doi.org/10.21785/icad2019.037

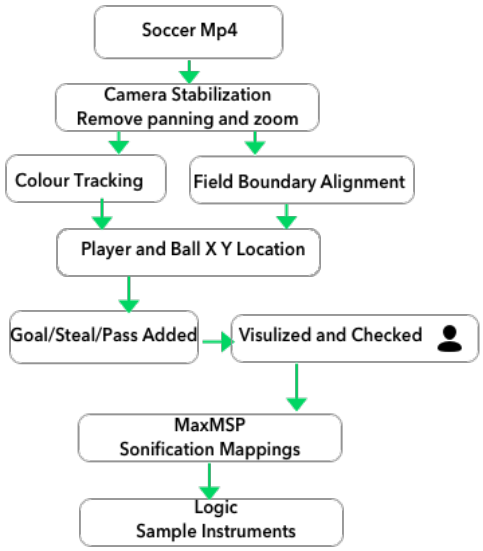


Figure 1: System Overview

created video analysis[15]. We then applied colour on the players and the ball, with the colours set by a human user before running the analysis. We were able to then bind the positions of the player by the boundaries of the field. These data were validated and, when necessary, corrected with the help of a Matlab visualization.

With X and Y data created for players and the ball, key soccer features were then generated, including possession, passes, steals, and goals. Possession was set by whichever player is closest to the ball. Goals were set whenever the ball passed through a set threshold. Passes were created when a ball travels a certain distance from a player, while steals were labelled when possession changed from one team to another. These higher level metrics are certainly not perfect, however, and required confirmation by a human.

Figure 2: Tactical Camera View

3.2. MaxMSP and Logic X

After the data are created, they are loaded into MaxMSP[1] where mapping and processing occurs. In MaxMSP all calculations are done such as tempo mapping (linear or exponential), acceleration of the ball and players, and the distance between objects. From MaxMSP, MIDI pitches and control channels are sent to Logic where many sample instruments are controlled. Sample instruments include built-in Logic libraries, as well as libraries from 8dio, EastWest and Native Instruments.

[1]https://cycling74.com/

4. SONIFICATION MAPPING

From the outset designing for entertainment and working with a specific football club guided us to certain grounding decisions. All sonifications were created from the viewpoint of one team, assuming the audience were supporting a set team. In early iterations we designed around a club's branding and nationality, however for evaluation and demos we moved to a generalized team sound. We also assume sounds would be used for clips no longer than three minutes long, allowing us to worry less about listening fatigue that could take place across a 90-minute game. Initial tests included possible representations of change throughout a season that would have allowed a players' sonic world to develop between games. The design went through an iterative process, creating many different approaches to the sound creation.

While we never aimed to directly replicate events shown through crowd noise we found this naturally occurred as plays built towards goals, or when possession changed. The design went through an iterative process creating many different approaches to the sound creation. After many divergent creations, we placed sonifications into four broad categories: pitch data mapping, tempo data mapping, musical moment alignment, and key moments.

4.1. Game Clips

To evaluate our different approaches, we used four clips from Belgium against Tunisia in Group G of the Fifa World Cup. We chose this game due to the variety of available plays and goals, with the final score 5 - 2, to Belgium. We used Belgium as the supported team. Clip 1 (Goal 1) begins with the ball in Tunisia's possession, before a breakaway steal leads to a Belgium goal. Clip 2 (Goal 2) features a Tunisia goalie dropkick, followed by multiple Belgium passes eventually leading to a goal. Clip 3 (Penalty goal) is the shortest clip and features a goal scored after a penalty. Clip 4 (No goal) displays two shots on goal by Belgium with both blocked by the goalie, before the play disperses to midfield.

4.2. Data Mapping

Direct data mapping of ball and player distances from the goal, or to each other, became a key element of a two subgroups of our sonifications. These subgroups focused on mapping distances to either tempo, or pitch and the many possibilities that arise from this linkage. The following section describes guiding principles for each subgroup, followed by the evaluation where specific implementations are described.

4.2.1. Tempo Mapping

Through early internal testing we found operationalizing tempo as a measure of excitement was an effective technique. We created multiple demos featuring drum tracks generated using the author's previous system[16, 17]. This generative drum system allows control of rhythmic density of each cymbal, drum or percussive element individually. This category primarily focused on mapping the ball's distance from the goal to the tempo of the piece.

Figure 3 displays the tempo curve (top line) followed by three lines of dynamic display, representing key player's distance from the ball. In this example each player had their own loop with the volume increasing as they became more or less involved in the play. In short term clips we found mapping dynamics ineffective as listeners were unable to quickly associate a sound with a player and therefore could not determine which dynamics corresponded to which player.

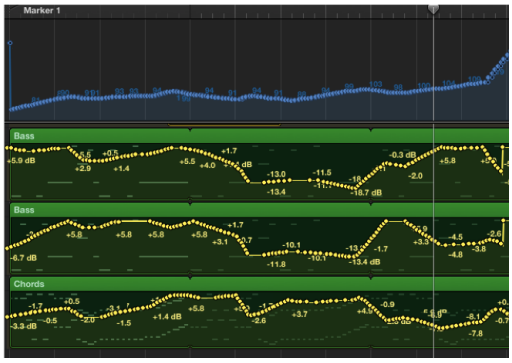


Figure 3: Dynamics Mapping

4.2.2. Pitch Mapping

Pitch variation utilized the data from the match in an almost identical way to tempo variation, however instead operationalized pitch as the driver for excitement. As in tempo mapping, certain elements of pitch were mapped to the distance of the ball from the goal.

4.3. Key Moments

For the key moment approach, we focused on only significant events and not player or ball positions on the field, including passes, steals, and goals. While these key events were also included in all the other sonifications this subgroup emphasized a deliberate focus on these interactions. They additionally allowed for enhanced listener focus on elements of these comments, such as the sonification of the ball speed during a pass.

4.4. Musical Moment Alignment

This category used existing pieces of music and aligned the most important moments in the soccer plays with the musically important features. This was done through intelligent identification of the important points in the data, and then working back through what was possible to align either through slight tempo variations, or cutting and reordering small sections of the composition. In general, the most important part of the clip was the goal, followed by a steal, and then other passes were aligned. This category is by far the least scalable as it requires human mapping, whereas all the other categories are automatically created by the pipeline described in Section 3.

5. EVALUATION AND QUALITATIVE RESULTS

5.1. Process and Stimuli

We ran evaluations with 36 participants, across 10 focus groups with 3 to 4 participants each. Participants were students in undergraduate classes, and were given partial class credit for their participation. Focus group sessions involved listening/viewing, plus discussion mixed with filling out an online survey form (using Qualtrics). Each session consisted of the following structure, and took an hour to complete.

Note that all stimuli are available at: http://richardsavery.com/soccersonification

1. Basic background questions about soccer playing and watching experience and regularity
2. Introduction to all four soccer clips used for evaluation without sound
3. A description of our goals in the project and an introduction to sonification
4. A tutorial track (tempo mapping for Goal 1) to demonstrate how sonification could be applied
5. For each of the subcategories (i.e. Key Moments):
   - Viewing of each videos with discussion after
   - Completing a Buzz Audio UX Scale [18]
   - Writing out individual thoughts on the form
6. Final discussion and closing thoughts including consideration of all examples in comparison to each other

In the focus groups we aimed to ask about and analyze high level variables, such as the application of each category and understanding what did and did not work with each sonification approach. In general we aimed not to focus on low level parameters, such as the specific instrument sounds, as we were looking to develop a broader understanding of how sonification can enhance soccer and not focus too deeply on our own implementations of these sonification methods.

5.2. Tempo Mapping

5.2.1. Stimuli

For evaluation, we created four clips using tempo mapping. For Goal 1 we emphasized the tempo using a drum kit playing a groove throughout, with tempo linearly mapped between 80 beats-per-minute (BPM) and 260 BPM to the ball's distance from the goal. An electric bass was mapped to the passes for the opposing team. The speed of the ball dictated whether 1, 2, or 3 notes were played by the bass. For the supported team, an electric guitar was mapped to passes, steals, and goals. This clip was designed as a general tutorial for the participants in the evaluations described later.

For Goal 2, we used a 2 bar drum loop with a repeating bass line to demonstrate the tempo, in this case mapped exponentially between 40 BPM and 260 BPM to the distance from the goal. Passes, steals and goals were mapped with a chime synth, with pass length tied to note length. The penalty goal used only the drum kit, to demonstrate the contrast between relatively fixed positions, as the clip begins with the ball stationary. The penalty goal had a small range of x,y positions, with the associated tempo ranging between 80BPM and 180 BPM. The non-goal used a new drum

groove underneath, again exponentially mapped between 80 BPM and 260 BPM, combined with a different pass sound.

5.2.2. Feedback

In general, respondents found that having tempo mapped underneath added extra excitement and increased tension to the play. There was a consensus throughout all groups that the sounds amplified what is happening on field, with one noting that it was like listening to a more detailed version of the crowd. As to be expected, participants varied in their style preferences for the underlying groove. However, there was agreement that grooves featuring not just drums conveyed a clearer sense of ball position. The variation in contrast between exponential and linear mapping was noticed, although preferences were split between each category.

For the penalty goal only using drums, we heard repeatedly that drums do not convey much information, and that the contrast from a slow tempo to fast tempo with only drums wasn't particularly clear. Many participants noted that *Tempo gives the idea of how fast the players are moving*, with most arguing this as a positive; however two participants felt the slower sections made the play feel slower and less interesting. In addition some argued that these contrasts didn't always replicate the real *situational intensity* of the play. For the clips overall there was disagreement about whether the intensity level was correct or not, with some describing the music as *intense for what would happen during the game*, while others thought the music wasn't intense enough.

5.3. Pitch Mapping

5.3.1. Stimuli

Three evaluation tracks were created for pitch mapping. The first clip (sonifiying Goal 2) used an electric bass playing repeating eighth notes at 120 BPM. The pitches were then exponentially mapped to the ball's distance from the goal, the closer the ball to the goal the higher the pitch. Underneath the bass a generated drum track was created, with variations in density also mapped to this distance, although divided into 7 density levels. Passes, goals and steals were mapped using the same guitar sound used for goal 2 tempo mapping. The second pitch mapping example was created for the penalty goal and uses a bassoon and flute. Both instruments play eighth notes at 120 BPM. The bassoon's pitch is mapped to the ball's distance from goal. The flute's pitch is mapped to the shooting player's distance from the ball: as he approaches the ball the pitch rises, and then falls again as the ball travels away. The third pitch mapping track was created for the non goal clip. In this clip pitch changes were quantized per measure of music, with the average distance over that measure used to set the pitch.

5.3.2. Feedback

Overall, Pitch Mapping and Key Moments received the most positive qualitative feedback from participants. The clip created for Goal 2 was many participants' favorite track, with some labelling it like a song. Others described the track made them feel a *good nervous*, as it was *like a car chase*. They also noted that while tempo was good for excitement, pitch was generally easier to understand, and its mapping conveyed the ball's position in a clearer manner. Ultimately for the Goal 2 clip many participants agreed it was *All encompassing of what you look for in the game*. The second pitch clip created for the penalty goal was in general well received for its information content. For the pitch clip for the non goal all participants found the pitch hard to understand, due to only shifting pitch per measure. We believe this was all due to the lack of a tonic creating a guide for the pitch, so adjustments were very hard to distinguish for a general audience.

5.4. Key Moments

5.4.1. Stimuli

The first key moments example was for Goal 2, and used only a drum kit. This track featured changes in density, volume, and cymbals/parts of the drum kit to demonstrate when a key moment occurred. The second example was created for the penalty goal, with just the shot and goal sonified, through a change of musical tone. This was done by moving from a V chord to the I after the goal, with a change of groove. The final clip created for the non goal used solely piano, with each key moment sonified through mapping of pitch, volume, and note length. Actions between teams were differentiated by the octave of the piano, with a lower octave given to the opposing team. The note length was mapped to the duration of each pass. Passes from each team were sonified with a piano tremolo: over a minor chord for the opponent team, and over a major chord for the supported team.

5.4.2. Feedback

The clip for Goal 2 was generally disliked by participants, with many describing only drums as confusing, and *Not expressing any information by itself, but supporting the story*. The penalty goal was described as matching the joy of scoring a goal, with the music accurately capturing the mood and the *euphoria of scoring a goal*. The final key moments clip created for the non goal received many positive comments. Participants noted that the use of piano changed the perception of the ball, with it at times seeming *lighter* than in other clips. The use of silence in this clip received different interpretations, with some describing it as helpful to only emphasize important parts while others described the silence as distracting. Many participants said this clip was the clearest to follow, with significant differences between each teams' actions and a good portrayal of what was happening.

5.5. Musical Moment Alignment

5.5.1. Stimuli

Goal 1 used a carefully chosen collection of rock loops. The ending of the song lined up with the goal, while the bridge was able to line up with the steal in the play. The bridge also featured a crescendo, and rise in pitch that loosely corresponded to the ball's distance from the goal. Using slight variations in tempo several passes were also aligned with the pulse of the piece. For the second goal we used loops of a funk soul groove, with the ending of the piece synchronized to the ball entering the goal. Other alignments included saxophone and trumpet layers coming in and out at important moments of the play. The third musical moment alignment clip was created for the penalty goal. This clip lines up just when the goal is scored, combining suspended trumpet rubato line pre-goal, followed by a mariachi inspired groove after the goal.

5.5.2. Feedback

Musical Moment Alignment was unsurprisingly described as the most musical, due to the fact standard pieces of music were used for the creations. As expected, participants also described that these examples were the least helpful in understanding the game; however, many noted it did support the play. One participant noted that the way the piece lined up showed that soccer itself is musical. Some participants described Goal 1 as making it harder to focus on the fine points of the play, even though they noted the sounds made the track more enjoyable. No participants noticed the horn lines syncing with passes for Goal 2, likely due to an inconsistent mapping. Overall, almost none of the participants found the approach to Goal 2 effective. The penalty goal for this category was by far the most polarizing clip used in the evaluation. Responses ranged from describing it as perfectly matching the tension of a penalty goal, followed by the joy of scoring, while others thought it drastically overplayed the clip. Participants did unanimously enjoy this section of clips, although commonly noted it made the clip feel like a highlight reel, and not like they were involved in the play itself.

6. QUANTITATIVE RESULTS

For quantitative analysis we used the BUZZ Audio User Experience Scale [18]. The BUZZ scale is comprised of eleven questions about the usability,usefulness and aesthetics of the sounds used in auditory displays and user interfaces. The BUZZ scale was designed to be applicable to a variety of different systems, as well as generalizable, allowing comparisons to be made across different systems. It is typically analyzed both by combining all eleven questions into one composite score, and by decomposition via factor analysis.

6.1. BUZZ Composite Scores

A Hyunh-Feldt repeated-measured ANOVA indicated that there was a significant effect of Sonification Method on BUZZ composite scores, $F(2.856,97.114) = 5.344$, $p = .002$, $\eta_p^2 = .136$. As shown in Figure 4 and Table 1, participants rated the Musical Moment and Pitch Mapping conditions more highly than the Key Moment condition. Additionally, a linear regression model revealed that an unweighted composite of the soccer experience and preference questions was not a significant predictor of BUZZ composite scores, nor did this item interact with Sonification Method.

6.2. BUZZ Subscale Scores

To identify factors within the BUZZ results, a principle factor analysis with Varimax rotation and Kaiser Normalization was conducted, as recommended by [18]. Items 2,3,8, and 9 loaded on a factor reflecting the enjoyment and appeal of the sounds, and items 1,4,5,7,10, and 11 loaded on a factor reflecting ease of use. Those items were combined into an unweighted sum, to produce scores for those two factors. Analyses of those factors are recounted below.

6.2.1. Enjoyment and Appeal

A Hyunh-Feldt repeated-measured ANOVA indicated that there was a significant effect of Sonification Method on BUZZ Enjoyment and Appeal scores, $F(2.771,94.208) = 21.474$, $p < .001$, $\eta_p^2 = .387$. Table 2 and Figure 5 show that participants rated the Pitch Variation condition more highly than the other three conditions. Within those three conditions, the Key Moment condition was rated lower than the other two. This indicates that participants found the Pitch Variation Sonification Method most enjoyable and appealing to listen to.

6.2.2. Ease of Use

A Hyunh-Feldt repeated-measured ANOVA indicated that there was a significant effect of Sonification Method on BUZZ Ease of Use scores, $F(2.661,90.466) = 10.232$, $p < .001$, $\eta_p^2 = .231$. Table 3 and Figure 6 show that participants rated the Musical Moment and Pitch Sonification conditions more highly in terms of ease of use. Notably, the Pitch Sonification was rated lower in terms of ease of use compared to Musical Moment.

7. DISCUSSION

Through evaluation we developed multiple takeaways. These focused on what worked as we expected, what surprised us, and what we could use for future developments.

7.1. Point of Focus

Different sonification methods significantly shifted the way participants watched the game and their point of focus. Depending on the sonification tactic employed, participants would focus on different aspects of the play. This included a change between micro and macro level aspects, with some sonification methods drawing listeners to focus less on the broader game and more on the movement. There is no clear answer to what is best to draw attention to, and particularly for entertainment this will vary between viewers.

7.2. Managing Interpretation of Events

The sonification of key moments drew several comments about the placement of the sound. For passes some participants noted that the event happened before the sound, while others commented the sound was too early. Attempting to sonify a sporting event ultimately forces many of these decisions to be made by the creators, with just soccer passes open to many interpretations and sonification methods.

7.3. Supporting a Team

Our approach to always have a single supported team was well received in these evaluations. With a data-driven approach this seems a logical choice to sustain, since either team could be automatically sonified. In general, though, participants also believed that many of our sonification methods could be used for either team, with approaches-to-goal being conveyed as suspenseful for both the defending and attacking team.

7.4. The Use of Drums

Overall the use of drums as a guiding feature in the sonifications was not effective. There was a general attitude that subtleties could not be distinguished as relating to moments in the soccer game.

	Tempo	Key	Musical	Pitch
Mean	33.31	30.60	35.94	37.54
SE	51.36	55.38	54.12	62.76
Differs From:	–	music,pitch	key	key

Table 1: BUZZ Composite Scores (out of 77) by Sonification Method

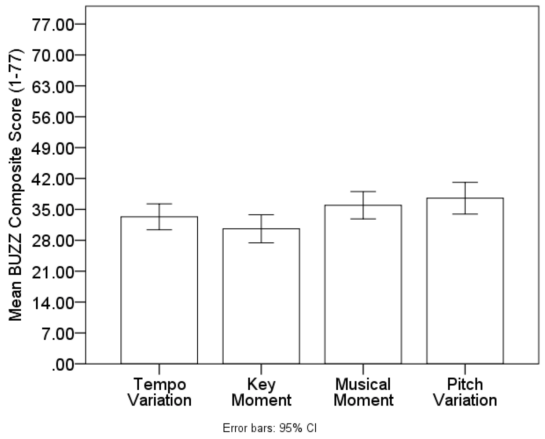


Figure 4: BUZZ Composite Scores (out of 77) by Sonification Method

	Tempo	Key	Musical	Pitch
Mean	12.54	10.74	8.71	15.57
SE	20.16	24.60	21.24	31.62
Differs From:	music,pitch	pitch	pitch tempo	tempo,key music

Table 2: BUZZ Enjoyment and Appeal Scores (out of 21) by Sonification Method

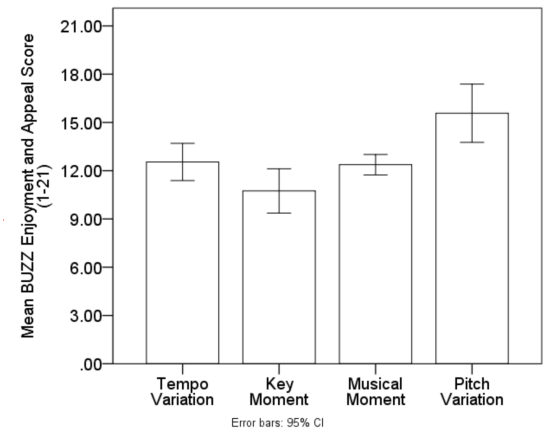


Figure 5: BUZZ Enjoyment and Appeal Scores (out of 21) by Sonification Method

	Tempo	Key	Musical	Pitch
Mean	21.60	19.86	27.23	21.97
SE	19.38	39.72	47.58	44.10
Differs From:	music	music	tempo,key, pitch	music

Table 3: BUZZ Ease of Use Scores (out of 56) by Sonification Method

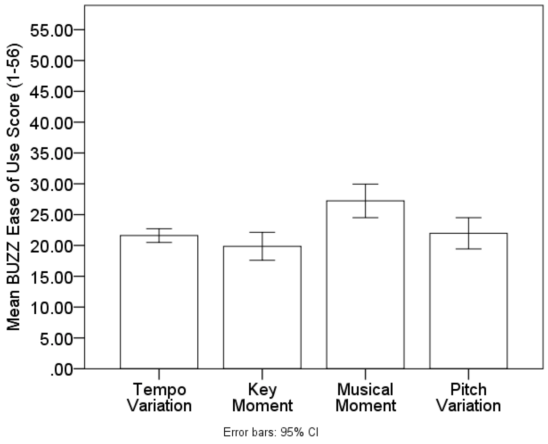


Figure 6: BUZZ Ease of Use Scores (out of 56) by Sonification Method

7.5. Sonification Methods for Mood

Each sonification method we employed received different positives, and many commented on the mood created by each method. Musical moment alignment was commonly described as good for a highlight reel, but made the play feel as though it was not happening in real-time. Conversely, pitch mapping, tempo mapping and key moments made viewers feel much more involved in the play and as if the play was in real time.

7.6. Creating for Entertainment

There are many unique challenges when creating sonifications for entertainment and sport. The balance between entertainment and analysis is significant and emerged in both the quantitative and qualitative evaluations. The importance of each factor can be expected to vary between each viewer; levels of sonic information that could be distracting to some viewers might be considered insightful by others. There was a general consensus that the more information that was conveyed through sound, the less like music the sonification sounded. For some this meant a less enjoyable experience, while for others this enhanced their understanding of the play and their overall enjoyment in watching the play. In addition to participant discussions, the presence of this type of enjoyability-usability trade-off was also shown through the BUZZ scores. Although the Pitch Sonification condition exhibited the highest overall BUZZ scores, analyses of BUZZ sub-scales revealed that this advantage came from the fact that participants found this version to be the most enjoyable, even though they rated it as less usable than the Musical Moment condition. This indicates the presence of a trade-off between enjoyability and usability in these two higher-performing auditory display approaches.

8. CONCLUSION

Through a divergent creation process we established four methods of sonification that can be used for soccer. After evaluation, each method showed varied strengths and weaknesses, however we had the most positive overall response to Key Moments and Pitch Mapping, with the former being more usable and the latter being more enjoyable. While we developed multiple strategies for sonification and lessons learned from evaluation, there are still many open questions and new potential strategies applied. Going forward, a key step will be to evaluate interactive user control over sonification choices and how this impacts user experience. Ultimately, we have demonstrated the potential for improved viewer experience through soccer sonification.

9. ACKNOWLEDGEMENTS

We are grateful to Steve Gera and Dave Anderson from the GAINS Group for discussions and suggestions during the beginning stages of this project, and for encouraging us to keep moving ahead on this project to make sports more accessible through the use of technology.

10. REFERENCES

[1] C. Carling, J. Bloomfield, L. Nelsen, and T. Reilly, "The role of motion analysis in elite soccer," *Sports medicine*, vol. 38, no. 10, pp. 839–862, 2008.

[2] E. Rampinini, F. M. Impellizzeri, C. Castagna, A. J. Coutts, and U. Wisløff, "Technical performance during soccer matches of the italian serie a league: Effect of fatigue and competitive level," *Journal of science and medicine in sport*, vol. 12, no. 1, pp. 227–233, 2009.

[3] T. Strudwick and T. Reilly, "Work-rate profiles of elite premier league football players," *Insight*, vol. 2, no. 2, pp. 28–29, 2001.

[4] G. Vigne, A. Dellal, C. Gaudino, K. Chamari, I. Rogowski, G. Alloatti, P. Del Wong, A. Owen, and C. Hautier, "Physical outcome in a successful italian serie a soccer team over three consecutive seasons," *The Journal of Strength & Conditioning Research*, vol. 27, no. 5, pp. 1400–1406, 2013.

[5] S. Barrass, N. Schaffert, and T. Barrass, "Probing preferences between six designs of interactive sonifications for recreational sports, health and fitness," *Proceedings of ISon*, pp. 23–29, 2010.

[6] J. W. Newbold, N. Bianchi-Berthouze, and N. E. Gold, "Musical expectancy in squat sonification for people who struggle with physical activity." Georgia Institute of Technology, 2017.

[7] G. Schmitz and A. O. Effenberg, "Perceptual effects of auditory information about own and other movements." Georgia Institute of Technology, 2012.

[8] O. Höner, T. Hermann, and C. Grunow, "Sonification of group behavior for analysis and training of sports tactics," in *in Proc. of the International Workshop on Interactive Sonification, Bielefeld*, 2004.

[9] N. Schaffert, K. Mattes, and A. O. Effenberg, "A sound design for the purposes of movement optimisation in elite sport (using the example of rowing)." Georgia Institute of Technology, 2009.

[10] B. Stahl and B. Thoshkahna, "Real-time heart rate sonification for athletes." Georgia Institute of Technology, 2015.

[11] N. Schaffert, K. Mattes, S. Barrass, and A. O. Effenberg, "Exploring function and aesthetics in sonifications for elite sports," in *Proceedings of the 2nd international conference on music communication science (ICoMCS2)*, vol. 83. HCSNet, 2009, p. 86.

[12] T. Hermann and S. Zehe, "Sonified aerobics-interactive sonification of coordinated body movements." International Community for Auditory Display, 2011.

[13] T. Stockman, N. Rajgor, O. Metatla, and L. Harrar, "The design of interactive audio soccer." Georgia Institute of Technology, 2007.

[14] S. A. Pettersen, D. Johansen, H. D. Johansen, V. Berg-Johansen, V. Reddy, A. Mortensen, R. Langseth, C. Griwodz, H. K. Stensland, and P. Halvorsen, "Soccer video and player position dataset," in *MMSys*, 2014.

[15] R. Savery and G. Weinberg, "Shimon the robot film composer and deepscore," in *Computer Simulation of Musical Creativity*, 2018.

[16] R. Savery, "An interactive algorithmic music system for edm," *Dancecult: Journal of Electronic Dance Music Culture*, vol. 10, no. 1, 2018.

[17] R. J. Savery, "Algorithmic improvisers," Master's thesis, 2015.

[18] B. J. Tomlinson, B. E. Noah, and B. N. Walker, "Buzz: An auditory interface user experience scale," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '18. New York, NY, USA: ACM, 2018, pp. LBW096:1–LBW096:6. [Online]. Available: http://doi.acm.org/10.1145/3170427.3188659

A PSYCHOACOUSTIC SOUND DESIGN FOR PULSE OXIMETRY

Sebastian Schwarz

University of Hamburg
Institute of Systematic Musicology
Neue Rabenstr. 13, 20354 Hamburg, Germany
bav7222@studium.uni-hamburg.de

Tim Ziemer

University of Hamburg
Institute of Systematic Musicology
Neue Rabenstr. 13, 20354 Hamburg, Germany
tim.ziemer@uni-hamburg.de

ABSTRACT

Oxygen saturation monitoring of neonates is a demanding task, as oxygen saturation (SpO_2) has to be maintained in a particular range. However, auditory displays of conventional pulse oximeters are not suitable for informing a clinician about deviations from a target range. A psychoacoustic sonification for neonatal oxygen saturation monitoring is presented. It consists of a continuous Shepard tone at its core. In a laboratory study it was tested if participants (N = 6) could differentiate between seven ranges of oxygen saturation using the proposed sonification. On average participants could identify in 84% of all cases the correct SpO_2 range. Moreover, detection rates differed significantly between the seven ranges and as a function of the magnitude of SpO_2 change between two consecutive values. Possible explanations for these findings are discussed and implications for further improvements of the presented sonification are proposed.

1. INTRODUCTION

In a clinical environment auditory displays can be very beneficial for patient monitoring, especially when visual attention is committed with another task [1]. The translation of input data to sound is called sonification, which is considered as the central element of an auditory display [2]. As sound is a temporal medium, process monitoring seems to be a very promising candidate for sonifications [3]. In a monitoring situation temporally-related data has to be observed and it is important to recognize changes in the current state of the process to be able to intervene appropriately in time [3]. In a clinical context auditory displays are already very common. For example there exists a huge variety of different alarms for patient monitoring. However, there seem to be drawbacks using them [4]. Apart from auditory alarms, auditory displays have the potential to inform the listener continuously about the current state of a patient, rather than putting him in a sudden state of alert [5]. This way the issue about when information is presented can be avoided and moreover the sonification also informs about normal states of the process, while attention is not attracted in an inappropriate way [6]. For example in the case of pulse oximetry, auditory displays seem to be of great use for patient monitoring, as they can shorten reaction times [5] and improve performances in time-shared tasks [5], [7].

Pulse oximeters are used to monitor oxygen saturation (SpO_2) and to prevent unwanted deviations [8]. The realization of a high level of SpO_2 was often supported by the aim to avoid negative consequences of hypoxemia and tissue hypoxia [9]. However, optimal oxygen saturation differs significantly across ages [1], [10]. Mainly patients at the extremes of age are at high risk of potential detriments of hyperoxia [10], [11]. In a meta-analysis the effect of functional oxygen saturation targets in premature infants was examined, which revealed an increased relative risk for mortality and necrotizing enterocolitis and a reduced relative risk of severe retinopathy of prematurity for a low compared to a high oxygen saturation target [12]. According to these results, the functional SpO_2 should lie between 90- and 95% in case of a gestational age under 28 weeks until 36 weeks postmenstrual age [12]. It is therefore of high importance to keep the oxygen saturation level in newborns in a particular range [1]. However, the maintenance of SpO_2 in a particular range using a pulse oximeter seems to be difficult, as could be shown in the case of preterm infants [13], [14]. In a conventional pulse oximeter a tone can be heard on each heartbeat and the pitch of the tone is varying with the oxygen saturation [15]. With the oxygen saturation rising or falling, the pitch is accordingly going up or down. Although most manufacturers include a variable pitch tone in their pulse oximeters, the acoustic properties of this tone are not standardized [16], which can lead to confusion interpreting the sonification [17]. For example the mapping between SpO_2 and frequency can be linear or logarithmic, whereby pitch perception is logarithmic rather than linear in nature [18]. Accordant to that, anaesthetists could estimate absolute oxygenation values as well as the size of oxygenation level differences significantly more accurate with a logarithmic pitch scale than with a linear scale [18]. Nonetheless, considering the specific demands on oxygen supply for neonates, a clinician would need more direct information, if and to what extent the SpO_2 level is moving out of a target range, unless he regularly checks the SpO_2 level on a visual monitor [1].

In a recent study a novel pulse oximetry sonification for neonatal oxygen saturation monitoring was proposed [1]. In two experiments it was tested, if nonclinician's ability to identify a target range of SpO_2 (90-95%) would improve with a modified version of a conventional pulse oximeter with a logarithmic mapping between SpO_2 and pitch. Two different redesigns of the conventional sonification were compared to the control condition. For the first sonification the pitch differences became very small in the target zone and increasingly large outside the target zone. This design didn't improve range identification accuracy compared to the control condition. In a second redesign [1] a fixed-pitch reference tone was included, when SpO_2 was outside of the target range. The pitch of this reference tone corresponded to the pitch at a SpO_2 level of 93% and it preceded every fourth pulse. This sonification significantly improved the accuracy of SpO_2 range identification in comparison to the control condition (85% vs. 60%). Consequently a modified sonfication seems to be beneficial for the listeners ability to detect a specific range of SpO_2. In a subse-

https://doi.org/10.21785/icad2019.024

quent study different levels of tremolo were added to a conventional pulse oximeter to test, if this would help listeners to identify SpO_2 ranges, direction of change and target transitions [19]. SpO_2 ranges were subdivided into five ranges, a target range and two ranges below and above the target range. In the target range no tremolo was used, whereby three cycles of tremolo were added each time a SpO_2 range was reached, that deviated further form the target range. SpO_2 ranges and transitions into and out of the target range were identified more accurately with the advanced sonification, than with the conventional sonification of a pulse oximeter. According to this, adding tremolo to a conventional pulse oximeter seems to be beneficial for identifying SpO_2 ranges and might even be more effective than the use of a reference tone [19]. Similarly in another study, tremolo and brightness were used to differentiate three SpO_2 ranges [20]. Participants of this study could successfully identify SpO_2 ranges (*Mdn* = 100 %), as well as transitions into and out of the target range (*Mdn* = 100 %).

This work proposes a novel sonification for pulse oximetry to convey information about current SpO_2 of neonates receiving supplemental oxygen. Unlike the examples discussed above, this design deviates further from the auditory display of a conventional pulse oximeter, as a Shepard tone [21] forms the basis of the sonification. Among other things, this approach is motivated by the aim to differentiate a larger number of SpO_2 ranges. In a listening test the effectiveness of the proposed sonification for identifying seven different SpO_2 ranges was tested. On the basis of the results of the listening test, further adjustments of the sonification are discussed.

2. THE SONIFICATION

The sonification is derived from the psychoacoustic sonification for navigation that has been introduced in [22] and discussed in a clinical context in [23]. The technical implementation is explained in [24]. The central element of the sonification is a continuous Shepard tone. In a preliminary study the Shepard tone has proven to be helpful in finding a target region [22]. As it might be important for a clinician to be able to estimate the distance of current SpO_2 from a predefined target range, a Shepard tone was used instead of the conventional mapping of SpO_2 to pitch. The Shepard tone contains the carrier frequencies

$$f_n = f_0 2^n \text{Hz}, \tag{1}$$

whereby $n = 0, \ldots, N-1$. In total the Shepard tone contains six carrier frequencies with $f_0 = 100$ Hz. If SpO_2 values are above or below the center of the target range, the Shepard tone is rising or falling in frequency respectively. This way the information about SpO_2 being below or above the center of the target range is conveyed by a simple binary coding. All carrier frequencies are rising or falling with the function

$$f(\phi) = f_0 2^{\frac{\phi N}{2\pi}}. \tag{2}$$

This way neighboring carrier frequencies are always one octave apart. In Eq. (2) ϕ is the phase of one cycle, such that the frequency rises from f_0 to f_N. The phase ϕ is defined as

$$\phi(\theta, t) = \arg[\sin(2\pi\theta t)], \tag{3}$$

whereby θ is a function of the distance to the center of the SpO_2 target range. This way the speed of the Shepard tone (rising or falling) is dependent on the distance to the center of the SpO_2 target range (90-95%), such that the speed increases the further SpO_2 deviates from the center. The amplitude of one frequency is weighted by a simple bell shaped curve. Consequently the amplitude of partials close to f_0 and at f_N are gradually reaching 0. A temporal envelope curve is used to create a pulse like sound, as the Shepard tone is supposed to get integrated in the sound design of conventional pulse oximeters. The frequency interval every pulse goes through, is increasing or decreasing with the Shepard tone gaining or losing speed respectively. This way a continuous mapping for the distance of current SpO_2 from the center of the target range is provided. A logarithmic mapping from distance to speed is used, such that a 1% change of SpO_2 would result in an approximately equal change of the perceived frequency interval. As the partials of the Shepard tone are continuously rising or falling, it is likely to happen, that the phase is varying between different pulses. Therefore, it is important that the Shepard tone is reseted to the starting point of its period T with every pulse of the oximeter. This means that the point of origin is held constant for every pulse, avoiding possible confusion, as the period of the Shepard tone contains no additional information.

The aim of this sonification was to enable the listener to differentiate between seven different ranges of SpO_2 illustrated in Figure 1. This is achieved by subdividing the target range (90-95%) into five ranges, consisting of a center range (92-93%) and two ranges below (90-91% and 91-92%) and above (93-94% and 94-95%) the center range. The remaining two SpO_2 ranges are defined as below ($< 90\%$) or above ($> 95\%$) the target range. SpO_2 ranges are numerated starting with range 1 at the top (see Figure 1). Pink noise is used to indicate that SpO_2 is within the target range. It provides a continuous background sound, such that it does not only occur within the time window of every pulse. Pink noise is used, as it is considered to be more pleasing to hear than white noise. This way the most critical information about the current SpO_2 is provided by placing only a minimum of cognitive workload on the clinician. Further information about the position of SpO_2 can be inferred by the direction and the speed of the Shepard tone. Within the center range (92-93%) the speed of the Shepard tone is set to 0, resulting in a pulse with a constant pitch. Deviations below or above the center range result in an increasingly falling or rising speed respectively. Thus, by identifying a rising or falling motion of the Shepard tone, a clinician should be able to locate current SpO_2 below or above the center range. To further differentiate between the remaining two ranges below (90-91% and 91-92%) and above (93-94% and 94-95%) the center range, the listener has to rely on the size of the interval the particular pulse goes through. The speed of the Shepard tone reaches its maximum at 90% and 95% of SpO_2 respectively, such that further deviations of SpO_2 do not result in an additional increase of speed. SpO_2 values outside the target range (90-95%) are made audible by the vanishing of the pink noise, whereby the direction of the Shepard tone still indicates, if current SpO_2 is below or above the center range. Nonetheless, a redundant coding is chosen, to make the ranges below ($< 90\%$) and above ($> 95\%$) the target range more distinguishable. A redundant coding by a second parameter can increase the robustness of the auditory display, as it may reinforce the representation parameter [25]. For SpO_2 values below the target range, frequency modulation is used to increase the perceived roughness of the Shepard tone, whereas for SpO_2 values above the target range the sound is not further manipulated. By using FM-synthesis to create roughness the perceived inharmonicity, roughness and noisiness increases with an increasing modu-

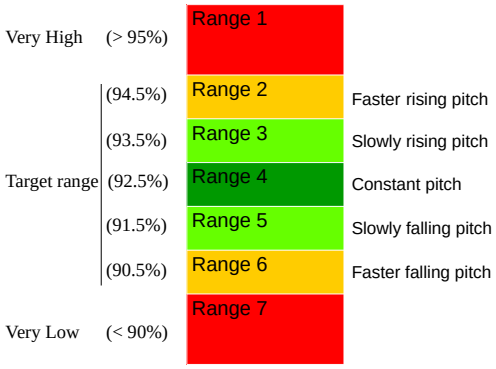


Figure 1: Subdivision of SpO_2 ranges. The target range (90-95%) is further subdivided into five SpO_2 ranges.

lation depth, such that the sound is perceived as more urgent [2]. Consequently the proposed sonification suggests a higher need for action in the case of hypoxia than in the case of hyperoxia. A demo video can be found on the second authors Youtube channel (`https://youtu.be/5kwzCunbLrA`).

3. METHOD

A convenience sample was recruited, consisting of students (N = 5) and staff (N = 1) of the Institute of Systematic Musicology at the University of Hamburg. In total 6 participants (1 female and 5 male) with an average age of 27.6 years (age range: 22-32 years) took part in the listening test. With only 6 participants the sample was rather small and not very representative, which should be kept in mind, while interpreting the results. All participants were non clinicians and except for one participant had no or little experience with sonifications. Participants were seated around two broadband loudspeakers, approximately 2-3 meters away. Due to economic reasons, all participants were tested simultaneously and were therefore instructed not to communicate with each other during the listening test, to prevent potential bias in the individual performances. The primary outcome variable was the detection rate calculated as the percentage of correct identified SpO_2 ranges. As described earlier, the principle of the proposed sonification is continuous between 90 and 95% of oxygen saturation. More precisely the frequency interval the Shepard tone went through got continuously bigger between 93- and 95% and 92- and 90% of SpO_2. As the participants had to discriminate between two different SpO_2 ranges in each of these cases, they could solely rely on the magnitude of the corresponding interval to do so. In the proximity of the transition from one range to the other it would be almost impossible to identify the correct range by hearing alone. Therefore, all values in the range of 93- to 95% and 90- to 92% were replaced by the mean of the corresponding range. The value of 90.2% was for example replaced by the corresponding value of 90.5%, which is the mean of 90 and 91%.

At first the sonification was explained, in particular the theoretical background and the applied mapping of data and sound, which was supported by auditory examples. After that the participants took part in a training session, which lasted about 5 minutes. In this session, participants had to listen to the modified pulse oximeter, which produced a pulse-like sound with a heart frequency of 60 Hz. Since it was assumed that the identification of the correct SpO_2 range each second would be too demanding for an untrained person, the value of the oxygen saturation was changed every two pulses. This way participants had two seconds for every SpO_2 value to identify the correct range. SpO_2 values were chosen arbitrarily, to cover all relevant ranges in a relatively short amount of time. Altogether, the training session consisted of four blocks, whereby in each block participants had to identify the correct SpO_2 range of five consecutive SpO_2 values. For each SpO_2 value the participants had to tick the correct box in a 7x5 table, whereby each row corresponded to one of the seven oxygen ranges and each column to one of the five SpO_2 values. After each part of the training session a feedback in terms of the correct answers was provided and a short break of approximately 30 seconds was taken. To indicate the start of a sequence, two pulses with the corresponding sound of 92.5% of oxygen saturation were always played at the beginning.

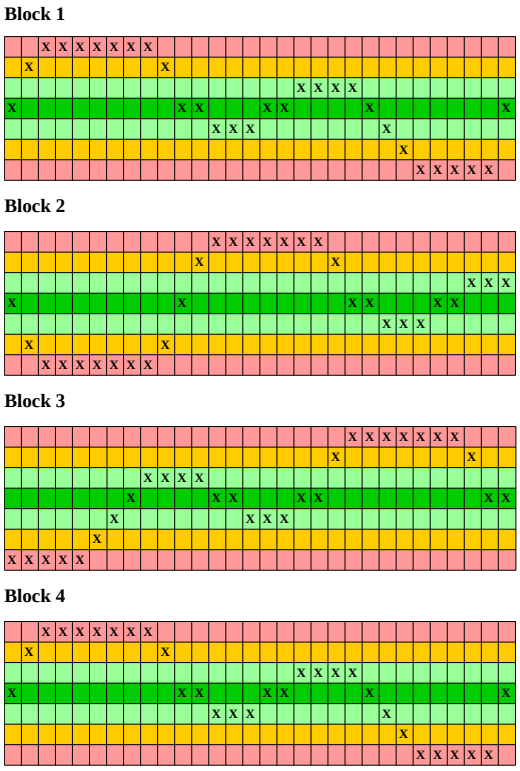


Figure 2: In each of the four blocks participants had to identify 30 SpO_2 values by ticking the correct box in a 7x30 table. The 7 rows correspond to the 7 SpO_2 ranges and each column to one SpO_2 value, which changed for every second pulse with a frequency of 30 Hz. The sample solution for each block is depicted above.

After the training was completed, the actual experimental task was performed, which lasted for approximately 10 minutes. In

Table 1
Median (Mdn), upper (HQ) and lower (LQ) quartiles for the detection rate of each SpO_2 range

	Mdn	*LQ*	*HQ*
Range 1	89%	60%	96%
Range 2	18%	12%	25%
Range 3	100%	78%	100%
Range 4	100%	100%	100%
Range 5	100%	87%	100%
Range 6	87%	78%	87%
Range 7	100%	100%	100%

Note. Detection rates were calculated as percentage of correct SpO_2 range identifications.

Table 2
Effect sizes (r) for multiple post hoc comparisons

	Range 1	Range 2	Range 3	Range 4	Range 5	Range 6	Range 7
Range 1		-.62	-.30	-.55*	-.47	-.06	-.55*
Range 2			-.61***	-.61***	-.62***	-.61*	-.61***
Range 3				-.26	-.08	.00	-.26
Range 4					-.26	-.56	.00
Range 5						-.31	-.26
Range 6							-.56

Note. P-values were calculated by a post hoc test after Conover (1999). Bonferroni adjustment method was used; *p<.05, **p<.01, ***p<.001.

contrast to the training session, participants had to identify 30 SpO_2 ranges in each of the four blocks and no feedback was given after each sequence. The four blocks are illustrated in Figure 2. In addition, the SpO_2 values were generated by a sine function. A smooth function was used, because it was considered to be in line with the fluctuations of oxygen saturation in an actual clinical setting. To account for possible training affects trial 1 and 4 were identical. Moreover, trial 3 was the reversal of trial 1 to examine possible effects of the direction of SpO_2 movement. In trial 2 the sine function was shifted about 2/3 π to the right. For the evaluation of the experimental task each tick, which was not placed in the correct box, that is the row and the column had to be correct, was considered as a wrong answer.

All significance tests were conducted at a significance level of α = .05. Detection rates were calculated as the percentage of correct SpO_2 range identifications for each participant over all 4 trials. To examine possible differences between different SpO_2 ranges, detection rates were also calculated for each SpO_2 range respectively. As an inspection of the corresponding qq-plots revealed deviations from normality a Friedman rank sum test was applied and subsequent multiple comparisons were conducted by a post hoc test after Conover (1999) [26]. The Bonferroni correction was applied, in which the p values were multiplied by the number of comparisons. In addition, it was tested, if different SpO_2 increment sizes did have an effect on the detection rates. Again a Friedman rank sum test was applied, as the corresponding qq-plots did not form a straight line. A post hoc test after Conover (1999) and the Bonferroni correction were used for multiple comparisons as well. To examine possible training effects between trial 1 and 4 the Wilcoxon signed rank test was applied, as the sampling distribution of the differences between scores did not look normal on a qq-plot. Moreover, detection rates between trial 1 and 3 were compared to account for any effect of direction of SpO_2 movement. The Wilcoxon signed rank test was used as well, as the corresponding qq-plot showed deviations from normality. Furthermore, it was of particular interest, if participants could identify an SpO_2 value being either within or outside the target range. Therefore, all given answers were additionally evaluated on a binary basis, whereas only the confusion between SpO_2 values within and outside the target range was treated as an incorrect answer (*inside/outside error*).

4. RESULTS

On average participants could identify in 84% (about 102 of 120 answers) of all 120 SpO_2 values the correct range. The chances to randomly guess the correct box were 1/7 $\approx$ 14%. In 98% (about 118 of 120 answers) of all cases participants could identify either the correct range or its neighbor range. Chances of choosing the correct field or its neighbor with a random guess are 19/49 $\approx$ 38%. To find out which part of the sonification was most ambiguous for the participants, detection rates were calculated for each SpO_2 range respectively (see Table 1). Detection rates of the participants changed significantly over SpO_2 ranges ($\chi^2(6) = 24.96, p < .001$). The results of multiple comparisons are summarized in Table 2. In addition, detection rates were varying significantly as a function of the SpO_2 increment size ($\chi^2(4) = 19.66, p < .001$). An overview of the detection rates for different SpO_2 increment sizes and the post hoc test of multiple comparisons is given in Table 3 and 4 respectively. To further examine, if participants found it particu-

Table 3
Median (Mdn), upper (HQ) and lower (LQ) quartiles for the detection rate of different SpO_2 increment sizes

	Mdn	*LQ*	*HQ*
Two ranges up	20%	5%	35%
One range up	75%	75%	84%
No change	97%	95%	98%
One range down	75%	72%	77%
Two ranges down	20%	20%	35%

Note. Detection rates were calculated as percentage of correct SpO_2 range identifications.

Table 4
Multiple post hoc comparisons of detection rates of different SpO_2 increment sizes

Value 1	Value 2	*p*	*r*
No change	One range up	.003**	-.62
No change	Two ranges up	<.001***	-.62
No change	One range down	.018*	-.62
No change	Two ranges down	<.001***	-.61
One range up	One range down	1	-.06
Two ranges up	Two ranges down	1	-.16
One range up	Two ranges up	.003**	-.61
One range down	Two ranges down	.018*	-.55

Note. Increment sizes: -2 (two ranges down), -1 (one range down), 0 (no change),+1 (one range up), +2 (two ranges up). P-values were calculated by a post hoc test after Conover (1999). Bonferroni adjustment method was used; *p<.05, **p<.01, ***p< .001.

larly difficult to identify SpO_2 ranges above the center, detection rates were compared between SpO_2 ranges above and below the center range. After examination of the corresponding qq-plots, a nonparametric test was chosen, as the data points did not form a straight line. The Wilcoxon signed rank test indicated, that participants detection rates were lower above the center (*Mdn* = 78%) than below the center (*Mdn* = 94%) of SpO_2 saturation ranges (p = .031, r = -.62).

In addition to that, it was of particular interest, if the Shepard tone was a useful choice to convey information about current SpO_2 being below or above the center and the current direction of movement of SpO_2. Of all 720 answers given there was only one case, where a participant mixed up the corresponding SpO_2 ranges below and above the center range. In three cases there was a false evaluation of the direction of SpO_2 movement and in seven cases a change of the SpO_2 range was not recognized. Interestingly all these mistakes were made by one participant. Only participant 3 had a detection rate below 80% (96 of 120 answers). This participant accounted for approximately 37% (40 of 109 incorrect answers) of all falsely identified SpO_2 ranges. Already in the training session participant 3 had together with participant 6 the highest occurring error rate. Overall, participant 3 performed distinctly worse than all other participants. About 6% (about 7 of 120 answers) of the answers of all participants were false, due to an inside/outside error. They accounted for around 39% (43 of 109 incorrect answers) of all incorrect answers. Approximately 84% (36 of 43 inside/outside errors) of all inside/outside errors occurred due to a confusion between range 1 and 2 and around 5% (2 of 43 inside/outside errors) due to a confusion between range 6 and 7. Participant 3 accounted for about 51% of all inside/outside errors. There was no observable training effect, as trial 1 (*Mdn* = 88%), and trial 4 (*Mdn* = 91%) did not differ significantly in their detection rates (p = .371, r = -.26). Moreover, there was no difference between the detection rates of trial 1 (*Mdn* = 88%) and 3 (*Mdn* = 93%), which indicated that there was no effect of the direction of SpO_2 movement (p = .418, r = -.23).

5. DISCUSSION

Overall the results of the listening test are very promising, as the six participants could differentiate seven ranges of SpO_2 saturation well above chance. Although participants received only a short training in advance, they were able to continuously track SpO_2 saturation in each of the four trials. Interestingly the detection rates of all SpO_2 ranges differed significantly from one another. Multiple post hoc comparisons revealed that participants performed better in identifying range 7 than range 1. A reason for this finding might be the design of the sonification. As described above, perceived roughness of the Shepard tone was increased, as soon as SpO_2 values were below the target range (90-95%). On the contrary the acoustic properties of the Shepard tone remained the same, after reaching the upper threshold of the target range. Thus, participants had to recognize the discontinuation of the background noise to detect deviations of SpO_2 above the target range. The fact that values below the target range have been identified more accurately than values above the target range is evidence, that a redundant coding improves detectability. It is possible that participants simply missed the onset or offset of the continuous background noise.

Although this did likely happen on both sides of the target range, transitions below 90% of SpO_2 could still be identified by recognizing the change of roughness of the Shepard tone alone. As the results indicate, participants had greater difficulties to identify SpO_2 ranges in the upper part of the sonification, meaning all SpO_2 ranges above the center range. It is therefore plausible, that participants perceived the sonification of SpO_2 above the center as more ambiguous than below the center. These results underline the importance of redundant coding, to make important thresholds more obvious to the user.

In addition, participants performed distinctly worse in identifying SpO_2 values in range 2 than in all other ranges except range 1. As stated above, the asymmetric design of the sonification probably accounted for participants greater difficulties to detect range 1 in comparison to range 7. This might have also affected the recognition of SpO_2 values in range 2. As participants had to continuously track SpO_2 values the correct identification of a SpO_2 range depended highly on the correct recognition of the previous SpO_2 range. Thus, an increased insecurity concerning range 1 most probably also affected the performance in range 2. Moreover, the detection rate varied as a function of the SpO_2 increment size. More precisely participants had greater difficulties in recognizing the correct change of SpO_2 ranges, if the SpO_2 value jumped two ranges up or down, than if it simply moved one range upwards or downwards respectively. If the preceding SpO_2 value happened to be in the same range, participants performed better than with a preceding SpO_2 value one or two ranges away. This finding might provide an additional explanation for the distinctly worse performance concerning SpO_2 range 2. SpO_2 values in range 2 and 6 were more often preceded by an SpO_2 value two ranges away, than any other SpO_2 range. In fact 50% of all preceding SpO_2 values of range 2 and 6 happened to be two ranges away, thus making it more difficult to identify the correct SpO_2 range. Nonetheless, only detection rates for range 2 were considerably lower than for all other ranges except for range 1. Therefore, it is likely that because of the specific design of the sonification as stated above, participants perceived a greater degree of ambiguity concerning range 1 and 2. As already mentioned, around 6% of all given answers were false, due to an inside/outside error, whereas about 84% of all inside/outside errors occurred due to a confusion between range 1 and 2. This result underlines the already mentioned difficulty to discriminate range 1 and 2. Only in two cases there was a confusion between range 6 and 7, whereas these mistakes likely occurred as an aftereffect. The design of the sonification consequently proved to be useful to inform the listener about SpO_2 being inside or below the target range. On the downside, it appeared to be more difficult for the participants to differentiate between SpO_2 values being inside or above the target range, mainly due to a confusion between range 1 and 2.

The Shepard tone proved to be a useful choice to inform the listener about being below or above the center range, the overall direction of current SpO_2 movement and about deviations outside a critical target range. As already mentioned in the results, only participant 3 made mistakes that disagree with this conclusion. Interestingly participant 3 accounted for around 51% of all inside/outside errors and for about 37% of all falsely identified SpO_2 ranges. Apart from possible differences in individual abilities, the specific design of the listening experiment might contribute to such a distinctly worse performance. As described in the method section, participants had to continuously track SpO_2 values, which were changing every second pulse for 30 times in each block. Therefore, the listening test was highly susceptible to aftereffects. For example, if a single SpO_2 value was missed during the listening test, all subsequent ticks made in the corresponding table were shifted one column to the left. Especially if a SpO_2 value was missed or falsely added at the beginning of a trial this could lead to considerably lower detection rates. This is most probably the reason for such huge performance differences between participant 3 and all the other participants.

Limitations and Prospects

In total six participants took part in the listening test, whereby the sample consisted of students and staff of the Institute of Systematic Musicology at the University of Hamburg. In a subsequent study it would be desirable to have a larger sample, including participants without a musical background. There was no control group and any findings need further corroboration. Moreover, clinicians might interpret the sonification differently, because of a broader medical background knowledge. Also the setting of the listening test differed from a clinical environment, especially as there was little background noise and participants could concentrate solely on listening to the SpO_2 sonification. As for example an anesthesiologist has to divide his attention across different tasks, Paterson, Sanderson, Paterson, Liu and Loeb (2016) tested effects of a secondary task on identification of SpO_2 ranges using an enhanced sonification of the pulse oximeter [27]. Performances for SpO_2 range identification deteriorated more for a LogLinear sonification than for the enhanced sonification of the pulse oximeter, although the difference did not reach significance [27]. This way the applicability of an enhanced sonification of the pulse oximeter can be evaluated under more realistic conditions.

As described above, SpO_2 values for the listening test were generated by using a sine function. It was assumed that a smooth function would provide a more realistic change of SpO_2 over time, but this needs the evaluation of a clinically trained person. By using a sine function, conditions were not identical for each SpO_2 range, as for example the average distance to the previous value differed as a function of the SpO_2 range. This might lead to misleading conclusions, when the sonification is evaluated in terms of each single SpO_2 range. The design of the listening test was very susceptible to aftereffects. As already discussed above, these kind of mistakes probably accounted for a considerable percentage of all mistakes made by participant 3. Therefore, a different design for the listening test might be helpful to prevent bias caused by aftereffects.

The results indicate, that participants found it particularly difficult to identify SpO_2 range 1 and 2. As discussed above, SpO_2 values above 95% could only be identified by the discontinuation of a continuous background noise, in contrast to range 7, where the perceived roughness of the Shepard tone was increased. Therefore, it might be beneficial to increase the perceived roughness for SpO_2 values above 95% as well. This might also contribute to a better detection rate of SpO_2 values in range 2. Alternatively beating could be applied as suggested in [22].

The proposed sonification of the pulse oximeter could be extended to nine different SpO_2 ranges by implementing two levels of roughness above and below the target range. This way clinicians could differentiate between urgent and less urgent deviations of SpO_2 from the target range. This would be similar to the enhanced sonification of the pulse oximeter by Deschamps et al. (2016), where four different SpO_2 ranges outside the target

range were sonified by adding two levels of tremolo to a LogLinear pulse oximeter [19]. However, the need of such a fine grained subdivision (nine different ranges) in the case of oxygen saturation monitoring of neonates needs to be evaluated by a clinically trained person. Furthermore, the sonification principle is designed to be continuous. It would be interesting to see how well the SpO_2 value could be interpreted on a continuous scale.

6. REFERENCES

[1] Hinckfuss, K., Sanderson, P., Loeb, R. G., Liley, H. G., & Liu, D. (2016). Novel Pulse Oximetry Sonifications for Neonatal Oxygen Saturation Monitoring: A Laboratory Study. *Human Factors, 58(2)*, 344-359. https://doi.org/10.1177/0018720815617406

[2] Ziemer T., Black D., & Schultheis, H. (2017). Psychoacoustic sonification design for navigation in surgical interventions. *173rd Meeting of Acoustical Society of America and 8th Forum Acusticum*, Boston, Massachusetts. https://doi.org/10.1121/2.0000557

[3] Vickers, P. (2011). Sonification for Process Monitoring. In T. Hermann, A. Hunt, & J. G. Neuhoff (Ed.), *The Sonification Handbook* (pp. 455-491). Berlin: Logos Publishing House.

[4] Edworthy, J. (2013). Medical audible alarms: a review. *Journal of the American Medical Informatics Association, 20(3)*, 584-589. https://doi.org/10.1136/amiajnl-2012-001061

[5] Sanderson, P. M., Watson, M. O., & Russell, W. J. (2005). Advanced Patient Monitoring Displays: Tools for Continuous Informing. *Anesthesia & Analgesia, 101(1)*, 161-168. https://doi.org/10.1213/01.ANE0000154080.67496.AE

[6] Sanderson, P. M., Liu, D., & Jenkins, S. A. (2009). Auditory displays in anesthesiology. *Current Opinion in Anaesthesiology, 22(6)*, 788-795. https://doi.org/10.1097/ACO.0b013e3283326a2f

[7] Watson, M., & Sanderson, P. (2004). Sonification Supports Eyes-Free Respiratory Monitoring and Task Time-Sharing. *Human Factors, 46(3)*, 497-517. https://doi.org/10.1518/hfes.46.3.497.50401

[8] Ruiz, T. L., Trzaski, J. M., Sink, D. W., & Hagadorn, J. I. (2014). Transcribed oxygen saturation vs oximeter recordings in very low birth weight infants. *Journal of Perinatology, 34(2)*, 130-135. https://doi.org/10.1038/jp.2013.157

[9] Sjberg, F., & Singer, M. (2013). The medical use of oxygen: a time for critical reappraisal. *Journal of Internal Medicine, 274(6)*, 505-528. https://doi.org/10.1111/joim.12139

[10] Habre, W., & Petk, F. (2014). Perioperative use of oxygen: variabilities across age. *British Journal of Anaesthesia, 113*, ii26-ii36. https://doi.org/10.1093/bja/aeu380

[11] Saugstad, O. D. (2005). Oxidative Stress in the Newborn - A 30-Year Perspective. *Neonatology, 88(3)*, 228-236. https://doi.org/10.1159/000087586

[12] Saugstad, O. D., & Aune, D. (2014). Optimal Oxygenation of Extremely Low Birth Weight Infants: A Meta-Analysis and Systematic Review of the Oxygen Saturation Target Studies. *Neonatology, 105(1)*, 55-63. https://doi.org/10.1159/000356561

[13] Lim, K., Wheeler, K. I., Gale, T. J., Jackson, H. D., Kihlstrand, J. F., Sand, C., & Dargaville, P. A. (2014). Oxygen Saturation Targeting in Preterm Infants Receiving Continuous Positive Airway Pressure. *The Journal of Pediatrics, 164(4)*, 730-736. https://doi.org/10.1016/j.jpeds.2013.11.072

[14] Sink, D. W., Hope, S. A. E., & Hagadorn, J. I. (2011). Nurse:patient ratio and achievement of oxygen saturation goals in premature infants. *Archives of Disease in Childhood - Fetal and Neonatal Edition, 96(2)*, F93-F98. https://doi.org/10.1136/adc.2009.178616

[15] Schulte, G. T., & Block, F. E. (1992). Can people hear the pitch change on a variable-pitch pulse oximeter? *Journal of Clinical Monitoring, 8(3)*, 198-200. https://doi.org/10.1007/BF01616776

[16] Loeb, R. G., Brecknell, B., & Sanderson, P. M. (2016). The Sounds of Desaturation: A Survey of Commercial Pulse Oximeter Sonifications. *Anesthesia & Analgesia, 122(5)*, 1395-1403. https://doi.org/10.1213/ANE.0000000000001240

[17] Santamore, D. C., & Cleaver, T. G. (2003). The Sounds of Saturation. *Journal of Clinical Monitoring and Computing, 18(2)*, 89-92. https://doi.org/10.1023/B:JOCM.0000032698.47717.06

[18] Brown, Z., Edworthy, J., Sneyd, J. R., & Schlesinger, J. (2015). A comparison of linear and logarithmic auditory tones in pulse oximeters. *Applied Ergonomics, 51*, 350-357. https://doi.org/10.1016/j.apergo.2015.06.006

[19] Deschamps, M.-L., Sanderson, P. M., Hinckfuss, K., Browning, C., Loeb, R. G., Liley, H. G., & Liu, D. M. K. I. (2016). Improving the detectability of oxygen saturation level targets for preterm neonates: A laboratory test of tremolo and beacon sonifications. *Applied Ergonomics, 56*, 160-169. https://doi.org/10.1016/j.apergo.2016.03.013

[20] Paterson, E., Sanderson, P. M., Paterson, N. A. B., Liu, D., & Loeb, R. G. (2016). The effectiveness of pulse oximetry sonification enhanced with tremolo and brightness for distinguishing clinically important oxygen saturation ranges: a laboratory study. *Anaesthesia, 71(5)*, 565-572. https://doi.org/10.1111/anae.13424

[21] Shepard, R. N. (1964). Circularity in Judgments of Relative Pitch. *The Journal of the Acoustical Society of America, 36(12)*, 2346-2353. https://doi.org/10.1121/1.1919362

[22] Ziemer, T., & Schultheis, H. (2018). Psychoacoustic auditory display for navigation: an auditory assistance system for spatial orientation tasks. *Journal on Multimodal User Interfaces*. http://doi.org/10.1007/s12193-018-0282-2

[23] Ziemer, T. & Schultheis, H. (2018). A Psychoacoustic Auditory Display for Navigation. *24th International Conference on Auditory Display (ICAD 2018)*, 136-144. http://doi.org/10.21785/icad2018.007

[24] Ziemer, T., Schultheis, H., Black, D., & Kikinis, R. (2018). Psychoacoustical Interactive Sonification for Short-Range Navigation. *Acta Acustica United With Acustica, 104(6)*, 1075-1093. http://doi.org/10.3813/AAA.919273

[25] Ferguson, S., Cabrera, D., Beilharz, K., & Song, H. J. (2006). Using psychoacoustical models for information sonification. *Proceedings of the 12th International Conference on Auditory Display*, London, UK. http://hdl.handle.net/1853/50694

[26] Conover, W.J. (1999) *Practical Nonparametric Statistical*. 3rd Edition, John Wiley & Sons Inc., New York, 428-433.

[27] Paterson, E., Sanderson, P., Paterson, N., Liu, D., & Loeb, R. (2016). The effect of a secondary task on identification accuracy of oxygen saturation ranges using an enhanced pulse oximetry sonification: A laboratory study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60(1)*, 628-632. https://doi.org/10.1177/1541931213601143

A SONIFICATION EXPERIENCE TO PORTRAY THE SOUNDS OF PORTUGUESE CONSUMPTION HABITS

Mariana Seiça, Pedro Martins, Licínio Roque and F. Amílcar Cardoso

CISUC, Department of Informatics Engineering
University of Coimbra
Pólo II - Pinhal de Marrocos
3030-290 Coimbra, Portugal
{marianac, pjmm, amilcar, lir}@dei.uc.pt

ABSTRACT

The stimuli for consumption is present in everyday life, where major retail companies play a role in providing a large range of products every single day. Using sonification techniques, we present a listening experiment of Portuguese consumption habits in the course of ten days, gathered from a Portuguese retail company. We focused on how to represent this time-series data as a musical piece that would engage the listener's attention and promote an active listening attitude, exploring the influence of aesthetics in the perception of auditory displays. Through a phenomenological approach, ten participants were interviewed to gather perceptions evoked by the piece, and how the consumption variations were understood. The tested composition revealed relevant associations about the data, with the consumption context indirectly present throughout the emerging themes: from the idea of everyday life, routine and consumption peaks to aesthetic aspects as the passage of time, frenzy and consumerism. Documentary, movie imagery and soundtrack were also perceived. Several musical aspects were also mentioned, as the constant, steady rhythm and the repetitive nature of the composition, and sensations such as pleasantness, satisfaction, annoyance, boredom and anxiety. These collected topics convey the incessant feeling and consumption needs which portray our present society, offering new paths for comprehending musical sound perception and consequent exploration.

1. INTRODUCTION

Auditory display has been writing its history as a scientific field for exploring sound as a medium of communication. In this journey, the main tendency in the research of auditory displays has been to explore its functional side and use sound to accurately represent data. However, a certain criticism marks this tendency, with many resulting experiences characterized as "unpleasant to listen to and difficult to interpret"[1], which hampers the potential of sonifications. Aesthetics has since risen as a potential concern for auditory communication [1][2], through which information can be understood in a designed context and engaging experience.

A recent reflection about aesthetics focuses particularly on the subject perspective [2] and the role of the user and embodied cognition as a subject-position, with aesthetics becoming an essential dimension in sonification design and how auditory information is rendered and presented. Music is inevitably one form of organization; however, beyond a musical experience, sonification should be seen as an experiential, ecological or perception-action aesthetic [2], that the user actively listens to in search for information. As such, the aesthetic dimension is indispensable for the meaning-making process, and cannot be separated from the informational / functional side. The aesthetics of sonification can thus have an essential role in auditory display communication, through which the information is encoded and musically structured by the system, and decoded by the user, exploring the aesthetics of experience, perception and the user as an embodied agent that can interact and explore the system. In this embodied setting, the concept of spatialization appears in creating a virtual or physical environment to explore the tridimensionality of sound [3], and for the user to actively explore while assimilating the auditory object. This exploration follows the phenomenological concept of apprehending a given phenomenon through a bodily experience, as "perception requires action" [4] and an active search in an environment where the body as a whole is inevitably embedded in that action. Our focus lies on the study and exploration of the role of the aesthetic dimension in the perception of auditory displays, studying how do people actively decode the auditory information, assimilating its context and iteratively apprehending its patterns and changes to decode its underlying information.

The first stage of our research focuses on the encoding stage, exploring how we can accurately represent a chosen dataset while composing a musical piece that provides an engaging, listening process, and how it is perceived by users as an auditory phenomenon that conveys information. For this experiment, we explored time-varying data, namely consumption data from SONAE, a Portuguese food retail company, and how its evolution through each day could be musically portrayed. The results, collected through a series of interviews to ten participants, showed some meaningful associations about the data, with emerging themes such as the idea of everyday life, routine, consumption peaks, and aesthetic aspects as the passage of time, frenzy and consumerism. The aesthetic perception also brought emerging sensations, from positive aspects of pleasantness and satisfaction, associated with the cohesiveness of the data representation day after day, to nega-

https://doi.org/10.21785/icad2019.050

tive sensations of annoyance, boredom or anxiety (among others), associated with the constant, steady rhythm and repetitive nature of the composition throughout the days.

The remainder of this paper is structured as follows. Section 2 comprises an overview of sonification projects using time-series, musical structures and embodied meaning-making processes. Section 3 discusses the data processing stage, and the filtering steps carried to gather the final dataset. Section 4 presents the musical mapping and its structure. Section 5 details the user testing conducted through a phenomenological approach, and the analysis of the results. The sixth and final section concludes the paper, listing the findings and future directions to take.

2. RELATED WORK

The concept of time-series with multiple variables is often represented through sonification, considered effective as an appropriate and successful tool to portray data evolution over time. The sound variations and regular patterns that we naturally perceive make most sound objects pertinent channels for conveying temporal dynamics and changes over time [2], and "tend to be more pleasant to the listener"[5].

Several works using time series served as examples from which to build this experiment. The first, called *Quotidian Record* and developed by Brian House in 2012, is a sound creation, engraved in a vinyl record, which translates his daily geographic routine for an entire year. Each place visited is harmonically mapped, translating not only its geographic coordinates (latitude and longitude), but also how he lives it and the time he spends there. Each rotation matches one day, composing a piece which brings out the underlying musical qualities of routine to create a portrait of a person's daily life [6]. The *Climate Symphony* [7] was created by Martin Quinn in 2011, where he used data from the chemical composition of an ice block in Greenland to translate into music the climatic changes endured by the great continental ice sheets. *Living Symphonies* [8] is a sound installation based on the fauna and flora of four ecosystems in the United Kingdom. The authors built a model that reflected the behavior, movement and daily patterns of every being in the wild, translating a network of interactions that formed the ecosystem.

Music is intrinsically connected to the study of auditory displays, serving many times as the main structure for the resulting auditory object. A sonification proposed by Dunn & Clarke in 1999 is described as a collaboration that merges "scientific knowledge and artistic expression" [9], using information of the human DNA's coding units and possible combinations. The project resulted in the *Life Music CD* [10] that achieves the exploration and refinement of a musical aesthetic to translate scientific data. This relationship meets a domain where *ars informatica* and *ars musica* merge together [11], and the goals of sonifications and musical pieces are blended. *Klima-Anlage* is "a walk-in sound installation" [12] that uses climate data from 1950, and a "global climate modeling experiment" that predicts climate variables evolution until 2100. It is an interactive installation, where the user can listen to the data from twelve regions, which includes precipitation data, wind data, radiation data and air temperature data, and also global greenhouse gas concentrations. Sound and video examples of several regions can be found in [13]. Another example is McGee and Roger's *Musification of Seismic Data*, who used straight audification processes to create musical compositions with a variety of timbres, through the "resampling, filtering, granulation, timestretching, and pitch shifting" [14] of seismic vibrations, which resulted in compositions of the February 21st, 2011 Christchurch earthquake and of the 12th January, 2010 magnitude 7.0 Haiti earthquake [15].

Regarding the idea of embodiment and how cognitive processes are influenced by perceptual and bodily experiences, Roddy's research [16] focused on the aesthetic dimensions of sound in the process of meaning-making. To explore different embodied sonic dimensions and the embodied nature of auditory cognition, the author explored vocal gestures, environmental soundscapes and temporo-spatial motion, using human sounds, natural acoustic scenarios and temporal context. Examples for the first two experiments are available online [17], with data-driven compositions such as *The Human Cost*, *Idle Hands* and *Doom & Gloom* which used statistical data of Irelands economic crash between 2007 and 2012.

3. DATA PROCESSING

The dataset used for this project was provided by the food retail company SONAE, as a result of a collaboration with our research group [5][18]. It gathers consumption data from 729 Portuguese supermarkets and hypermarkets of the food retail company SONAE, located throughout the country. The data was collected from the customers' loyalty cards, which can be used to accumulate several discounts and benefits. There are around 6 million active cards, shared by entire families for each household, which is a significant number for the ten million Portuguese population [5]. The data used for this experiment deals with transactions that occurred in 2014, from January 2nd to June 31st.

This dataset was well-fitted to what we intended to work upon, not only for its richness and size in terms of data variables and entries [5], but also for its pertinence in portraying everyday consumption habits and patterns of the Portuguese people.

Each transaction in the database corresponds to one single product, bought in a specific store, at a certain time, and with a particular cost. The products are also grouped within a product hierarchy of the company, with six levels [5]. The categorization that we used for this work was proposed by Maçãs, Martins and Machado [5], which aggregated the products into nine distinct categories: Grocery; Alcohol & Sweets; Health care; Beauty; Clothes; Furniture; House Care; Culture & Leisure and Pets & Nature Care.

3.1. Data Filtering

The database has an average of 5 million rows of transactions for each day, which demanded several grouping and filtering steps to gather a manageable set of data to work upon. Having the spatialization stage and the user exploration possibilities still open for experimentation, there were two possible grouping forms for the data: by a spatial and temporal one. At the spatial level, the transactions could be grouped by district, county and store, allowing the user to explore different levels of depth by listening from the consumption sounds of an entire district to each store separately. The temporal level would focus on the pacing of the data timeline, grouping the transactions by a ten, thirty or sixty minute window for each ten seconds of the composition. This could allow the user to have an overall view of the six-month transactions in a few minutes, or to listen to just one day in the same amount of time.

To maintain a low level of complexity and number of files, we opted in this first stage to gather the data by spatial level, grouping

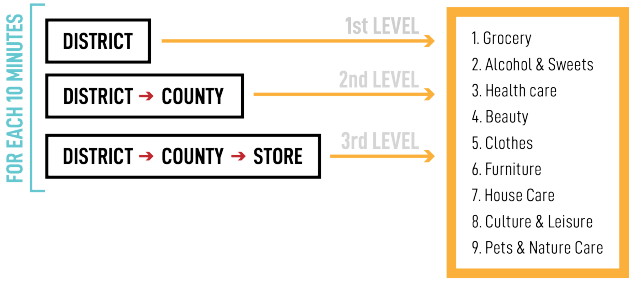


Figure 1: Data organization

the transactions first by category, and then by each 10 minutes for each district, county or store, preferable for the near-future spatialization experiences to explore different spatial depths (see Fig. 1). We also filtered the type of stores, keeping only the department stores of larger dimensions, namely the most known chain called "Continente". For this first experience, we retrieved the files of 12 days, from January 2nd of 2014 to January 12th for 22 districts (18 from Mainland Portugal, 1 from Madeira and 3 from Azores) in the three spatial levels, giving a total of 726 files to use.

4. ORCHESTRAL MAPPING

According to Vickers & Hogg [11][19], sonification and music mutually imply each other in the design of sonifications, which can become, when focusing on the concepts of enjoyment and usefulness, "a mass medium for an audience with expectations of a functional and aesthetically satisfying experience" [1]. The link between science and art thus appear in uncovering the auditory reality of the dataset, surpassing the scientific method using the sound qualities, and specifically the natural "beauty and proportion in music" [20] to shape the listening experience not only for data communication, but even as a means for artistic expression. As such, an early decision made for this research was to devise this auditory experience through a musical form, which would translate the consumption narrative into a musical composition.

One of the main challenges while devising the mapping was how could we create a musical composition where the nine categories of products could be distinguished. Timbre emerged as the parameter with more perceivable variations that could be effective for discerning nine different variables, but even so the challenge of discerning nine instruments in a digital composition still remained.

The works of Grey [21] and McAdams et al. [22] regarding perceptual relationships between musical timbres are references in the study of timbrical similarities and common dimensions, using multidimensional scaling techniques to create spatial representations of timbres. Rentz [23] focused particularly in the perceptual discrimination of orchestral instrument families between musicians and non-musicians, with percussion and brass instruments sweeping the non-musicians attentions for longer, and strings capturing more the attention of musicians.

Thinking about acoustic communication brings the human into the center stage, and how can sound be understood from an human perspective. Truax's aspects [24] of (1). variety and richness of sounds, (2). complexity and the levels of information intended to communicate, (3). and a balance between the first two characteristics and the environment build the concept of an healthy soundscape. In this scenario, the sounds connect and merge with each other, the environment and the user, who is the main interpreter and focus of an acoustic communication.

We brought the notion of an orchestra as the main setup for our experience, with its forming sections naturally dividing the families of instruments, useful not only for organization purposes, but in this case as a method for grouping similar timbres which can be mapped to each category.

PARAMETERS / EXPERIENCES	TIMBRE	PITCH	LOUDNESS	NOTE DENSITY
1	Category	Value of Sales (€)	Value of Sales (€)	Number of Products
2	Category	Value of Sales (€)	Number of Products	(Category)

Table 1: Implemented mapping

Using a parameter-mapping approach to sonification, we associated a set of musical parameters with the most relevant variables in the dataset (see Table 1). Two distinct mappings were made, along with a variation of the first with an added percussive element. The timbre, as the most variable and most naturally distinguishable parameter (we naturally recognize the difference between the sounds of two instruments) was mapped to the category types on both experiences, choosing nine different instruments to represent the nine categories (see Table 3). The instruments were chosen according to symphony orchestra sections (see Table 2), choosing a balanced set of timbres for each section to create an even-tempered sound. The sales value, which represents the amount of money each group of transactions cost, was directly mapped in both experiments to the pitch of the instruments. The pitch was adapted to each category, defining a range of values for each instrument to respect their natural range [25], and also for better distributing the sounds over lower and higher octaves which makes the instruments more distinguishable. To keep a consonant sound, a scale for which to choose the pitches was defined, namely the major (Ionian) scale, reusing part of the work developed in a previous sonification [26].

WOODWINDS	BRASS	PERCUSSION	STRINGS	(KEYBOARDS)
Piccolo	Horns	**Timpani**	Harps	**Piano**
Flutes	Trumpets	Xylophone	**Violin**	Pipe Organ
Oboes	**Trombones**	Marimba	Viola	Harpsichord
English Horn	Tubas	Glockenspiel	Cello	Accordion
Clarinets		**Vibraphone**	Double Bass	
Bassons		Chimes	-	
Saxophones			**Guitar**	
			Bass	

Table 2: Set of chosen instruments divided by orchestra sections

Loudness also changes moderately according to the sales value, within a range from the MIDI values of 77, to maintain an audible minimum value, to the maximum of 127, to emphasize the most sold categories. The choice to map both the frequency and loudness to the sales value, using different ranges of frequencies for each instrument, was made to emphasize the fluctuations between the values across the day, specially the hours with higher sales, and musically reinforce the data flow.

Besides the sales value, the dataset also gathered the number of products traded. As such, we can have a transaction with a high

cost, but only for a single product, or a low-value purchase for a large amount of products. We chose to represent this variable through the note density of the melodies, directly mapping it to the rhythmic figures and their duration.

Categories	Instruments
1. Groceries	Grand Piano
2. Alcohol & Sweets	Violin
3. Health Care	Vibraphone
4. Beauty	Flute
5. Clothes	Oboe
6. Furniture	Trombone
7. House Care	Bass
8. Culture & Leisure	Guitar
9. Pets & Nature Care	Timpani

Table 3: Association between the chosen timbres and the nine categories

In the first mapping, the instruments follow the same beat according to the initial BPMs, marking the tempo with whole, half, quarter or eighth notes (considering each 10 minutes a whole bar), or quarter, eighth, sixteenth and thirty-seconds notes, depending on the number of products. In the second mapping, each 10-minute beat is divided into nine beats representing the category order. For instance, the instrument that represents the first category plays in the first beat, the second in the second, and so forth. This "Arpeggio style", as we named it, does not represent the number of products through the note density, as it is constrained to the 1/9 beats in which each category plays, but through the loudness instead.

For this experiment, a simple interface was designed to browse through the possible days and districts, setting the deeper spatial levels aside to understand the overall patterns. Three software tools are used to produce the sonification: Processing, responsible for the dataset analysis, Max, to generate the musical composition through MIDI, and Ableton Live, to play the composition using VST's plugins. Audio and video examples of both experiments can be found online, of one weekday (January 2nd) and a weekend day (January 5th) for three districts (Lisbon the capital city, Coimbra as a medium-sized district with around 100.000 inhabitants, and Évora with around 50.000 inhabitants) [1].

The process for perceiving the flow of this dataset was also a process of exploration: although we had a notion from a previous rhythmic sonification experience made with this dataset [5], this was also a learning process for us to uncover what we have, to understand how the sales for each place changes in each day, and how that can that change be musically portrayed.

5. LISTENING EXPERIMENT

As an "activity of perceptualization" [27], auditory displays demand a design process that understands how auditory information can be perceived by the user. With the goal of designing an auditory experience, it is fundamental to understand how the user positively engages with the artifact [28]. The users may know in advance that there is an underlying discovery process of the data to undertake, but the way they build that process and develop the auditory narrative depends on how they experience it, what they perceive, and the paths they choose to do it.

Phenomenology emerges as an approach to build this experience, where we can iteratively learn from the user's perspective how he/she perceives it to design a user-centred sonification exploration. Phenomenology, as "the study of experience from the perspective of the individual" [29], focuses on the experiential side of artifacts, and can be a powerful tool to gain insights into the user's actions and motivations, and how they shape their personal interactions with it.

PARTICIPANTS	AGE	GENDER	BACKGROUND
P1	25	M	Biomedical Engineering
P2	57	M	Music
P3	45	F	Music
P4	26	F	Biology & Music
P5	26	M	Design & Multimedia
P6	26	F	Design & Multimedia (with musical training)
P7	43	F	Antropology
P8	46	M	History
P9	27	M	Music
P10	49	F	Physical Education

Table 4: Profiles of the participants for the evaluation

We decided to undertake a phenomenological approach to understand how the sonification is experienced as a phenomenon, uncovering meanings and common themes that the users associate while listening to it. It was not our goal with this type of evaluation to specifically detect if the exact mapping choices are understood, but instead what associations and sensations are perceived during the listening process, and how does each user perceives the sounds and its variations, and builds his/her acoustic narrative. This type of approach would allow us to observe the learning process of each listener, and consequently build our own discovery process of the dataset, creation methods, and its possible paths of exploration while shaping an aesthetic exploration.

5.1. Methodology

Ten participants were chosen to listen to the experience, and were submitted to an interview to document their perceptions. With

[1] Dropbox Link - https://www.dropbox.com/sh/5s2x7vjmjreuyfm/AAACxoZhRb8b8Hh86zrv4f8-a?dl=0

50% female and 50% male, and ages ranging from 23 to 57 years old, we tried to gather participants from different backgrounds (see Table 4), which would enhance the possibilities of a diversified set of perceptions. The interface was simplified for the evaluation as a a continuous composition of the ten days (see Fig. 2), relative to the Coimbra district, which would allow a non-stop flow of the sonification to help building the learning process. A few commands were added, as the space bar to pause and play, providing moments of break in the music to focus on the interview exchanges, the possibility of skipping to the next day or return to the previous day with the left and right arrows, and the choice of changing the tempo (BPMs), from the default 100 BPMs, with the up and down arrows to experience a different scenario. A video example of the 10-day composition can be found online [1].

Figure 2: Screenshot of the interface created for testing

The interview was devised to create a receptive environment for the participants, and a flowing conversation with the interviewer to promote a safe, trustworthy setting to share their thoughts. In this scenario, we tried to ensure, as far as possible, a non-preconceived mindset from us, based on Husserl's *epoché* [30], which describes an attitude where interpretation is suspended to collect variations on the phenomena. Following the three initial hermeneutic rules [30], we presented the phenomenon in a clear, non-invasive way, diminishing the number of visual cues and other stimuli to understand what was immediate to the participants, and leading them to describe what they were experiencing without prejudgment from both parts. With this mindset, the interview was conducted according to the following steps:

1. It started with a minimal **contextualization of the dataset**, disclosing only that the composition represented consumption data from the Portuguese food retail company SONAE. The interface commands were also explained.
2. The descriptive, exploratory concept of the interview, with no right or wrong answers, was also emphasized, encouraging the participant to freely describe what associations and sensations arise during the listening process.
3. The first part of the interview was allocated for the user to freely navigate through the sound, changing or not the controllable parameters, describing to what concepts, scenarios or context the mind was travelling to, with or without relating to the consumption theme.
4. **"Apprehending the Phenomenon"** [31]: at a deeper level, the process of "seeking out structural or invariant features of the phenomena" [30] appeared in the next stage, refocusing the attention of each participant to the consumption context, and guiding him/her in apprehending the phenomenon [31] using his/her own metaphors and limits. In this stage, the intention was for the user to describe the connections he/she was making. We would ask the participant to clarify the emerging terms, sensations and how he/she was composing the narrative meaning, while we would hint, from time to time, new connections of those terms with sound changes and patterns he/she had previously mentioned. Hints would include: refocusing the attention to the numbers in the screen and what could they represent, to the variations of the instruments recognized or the musical characteristics perceived and their possible meaning. In this stage, new listenings were made to access the new paths that were suggested.
5. **"Clarifying the Phenomenon"** [31]: the use of imaginative variation was then included to propose new realities not perceived by the listener, creating new scenarios for reflection, how they were understood, and how could they change the intuitive settings. The clarification of the phenonemon [31] was therefore made with these new intended associations, which ended with the reveal of the mappings parameters to provide a complete scenario and a final discussion of the overall experience.

5.2. Results

After the interviews were conducted, the text was transcribed to identify a set of categories of interest which would provide the overall themes and interpretations perceived by the testers. Three categories were then labeled and distinguished (see Table 5): (1). the interpretations of the domain, to gather the scenarios and realities conveyed by the composition, (2). the overall sensations the sounds would provoke, (3). and the assessment of musical aesthetics, listing the musical parameters and correspondent variations the participants could distinguish. Within each category, we listed the key terms, and how they connected with each other and with parallel finding in other categories.

5.2.1. Domain Perceptions

Although the consumption context was given in the beginning of the interviews, the domain possibilities found surpassed it. The thematic of **everyday life** and **daily routine** appeared in 12 citations of half of the participants, connecting the steady, regular rhythm to a normal day in a person's life. In this routine, words like *habits*, *daily dynamics* and *automatism* were mentioned, with one participant particularly describing a regular day where you "wake up, do your chores, reach noon and a peak of energy, take a nap, have a snack in the middle of the afternoon, and have a drink after dinner". It represents an energy variation, given by loudness variations and the rhythmic density. Participant 2 spoke about the "active day of a person, from 8:30am to 11pm, that is our daily work flow, fast and tiring", joining the idea of movement and obligations to fulfill. The **passage of time**, hinted by the hours in the screen and the steady, repetitive tempo, is described as "the tick-tack of the clock", which conveys the feeling of *construction, evolution*, *productivity*, "that something is working and rising" and *task assessment*, "where we all work to reach the same goal, even doing slightly different things at different rates", with each instrument navigating in the given melodic scale.

The **people's movement**, the way they move across the day or inside the commercial area, was another of the most mentioned themes by four participants, with one comparing the piano from

Domain Interpretations	Occurences	Number of Participants
Everyday Life	12	5
Consumption Peaks	11	7
People's Movement	11	4
Consumerism	10	4
Amount of People	9	4
Average Consumption	8	5
Passage of Time	7	4
Frenzy	5	4
Movie / Documentary	5	3
Daily Search	4	1
Trips	2	2
Culture & Leisure	1	1

Sensations	Occurences	Number of Participants
Annoyance	5	4
Pleasantness	4	2
Satisfaction	4	2
Anxiety	3	3
Lightness	3	2
Fatigue	2	1
Aggressiveness / Austerity	2	1
Sadness	1	1
Boredom	1	1
Claustrophobia	1	1

Assessment of Musical Aesthetics	Occurences	Number of Participants
High and Low Pitches	17	6
Steady Rhythm / Regularity / Time Measure	16	7
Repetition	8	5
Fast Rhythm	8	4
Musical Composition	5	3
Crescendo	5	3
Layer Construction	4	2
Melody	3	2
Martial Style	2	2
Volume / Loudness	2	1
Mantra	2	1
Instrument Distinction	1	1

Table 5: Analysis of the interviews, divided by three categories

2pm to *walking* and two referring to the idea of *paths, mazes*, moving from aisle to aisle by the piano fast pace. The fourth participant even perceived the whole composition as a portrait of the consumers' rhythm, with the low tones representing the constant presence of people in the hypermarkets and the number of people, and the higher, faster tones the peak of people's access, or the ones that stay only for a short time. The focus on the people emerged as well through the link between the **amount of people** with the piano fast notes, the loudness variation and the rhythmic density, with the volume and the number of notes increasing in the peak hours of access and consumption. The context of consumption with **consumption peaks** emerged as such, perceiving "a stronger dynamic as higher consumption", the "hours across the day in which there are more purchases, especially in the evening", and "the rush hours of higher, frantic consumption", with two participants naming the difference between week and weekend days. The word **frenzy** is also mentioned by four participants to describe the shopping environment, particularly in the peak hours, relating to *many people*, *agitation* and even "the non-stop cash register hectically passing products person after person as an assembly line". The low, steady-paced sound was then curiously associated, by half the participants, with the regular amount of products bought, the **standard, average of consumption**, or to the essentials goods that people need to consume everyday, associated with the constant, low-pitched sound that seemed to mark a regular tempo.

Consumerism appears as an underlying subject, mentioned by one participant as an "unceasing daily quest" for what we need or think we need to buy, without calm and reflection. The impulsive attitude is pointed out by three participants, with the high, fast-paced piano representing "not daily products, but those who are highlighted" to gain the consumer's focus, where we think "I don't need it, but the price is nice", revealing a "certain urgency to buy, to consume".

Other surprising themes that emerged include thinking about the composition as a "soundtrack of an alternative movie", or a **documentary or movie imagery**, namely a "documentary about the human life, where they film mass people, mass consumption, mass movements in the search for progress", of "fast-paced life". **Travelling** was also mentioned, associating a "train station, the train dynamics, movement, leaving the monotony behind", and even an "intergalactic journey", associated with the symphonic, technological sound of the instruments.

5.2.2. Sensations

Regarding the sensations that arose from the listening experience, **annoyance** was mentioned by four participants, mainly associated with high-pitched sounds, which would become annoying and strident after a while. Nevertheless, it was characterized as **pleasant** and **satisfying** by four participants, which was associated by listeners as (1). being a "musification instead of sonification" (by a participant whose background is also working with data), (2). being "cohesive", despite its repetitive nature, that conveys the idea of people moving, (3). and providing an enjoyable moment, "that would be wonderful to be playing in a commercial area" and just feels like "going on a ride with no destination, alone, and embracing the moment through the music". At one point in the interview, the last participant that described this last scenario curiously asked if she had to be aware of more details, or if she could just "enjoy it", which can be a hint that the musical piece can provide an engaging listening process. Also in a positive spectrum, the composition was characterized by being **light and soft**, as if it represented "a pleasurable job".

In a negative spectrum, besides annoyance, it gave the idea of **fatigue**, "exhaustion after a day's work, and representing the tiring, repetitive movement of everyday duties". **Sadness** was also

mentioned by a participant when played slowed (20 BPMs), and **boredom** as well, which was felt after a while associated with its repetitive nature. A sensation of **anxiety** was felt by three participants, which was associated by listeners to the fast rhythm and "a more intermittent, higher sound", which would convey the anxious, fast movement during peak consumption hours. One particular participant felt the steady, marked tempo as **aggressive** and "a bit dictatorial, almost as an imposing march", with another participant relating the **march** scenario to the work rhythm. An unnerving feeling of **claustrophobia** was particularly mentioned by one participant, again associated to the faster, high-pitched piano, recalling moments of being in a middle of a crowd, feeling muffled and thinking "I don't want to be here".

5.2.3. Musical Assessment

The assessment of musical aesthetics provided parameters already treated in relation with several domain interpretations and sensations, such as the **acknowledgment of high and low pitches**, a **steady, regular rhythm**, **fast rhythm**, **repetition** and **loudness variations**. Analyzing **musical style**, **composition**, and **melody construction**, was primarily done by the participants with a musical background, with four out of five naturally speaking about the role of each instrument in the composition, with "the winds setting the tempo and rhythm" and each instrument appearing successively as a new layer. It was also characterized as having an oriental, martial style, and even a mantra for having this repetitive, almost *hypnotic nature*. Multiple instruments were naturally distinguished, primarily the piano for every participant and winds for seven participants, with four or five participants with musical background distinguishing more than five instruments, when their attention was redirected to that task.

6. CONCLUSION AND FUTURE WORK

The focus of our investigation lies on the study of the influence of aesthetics in sound perception, primarily in data representation through auditory displays. We focused on how we could compose a musical piece that could be representative of consumption data from the Portuguese food retail company SONAE. Our goal was to consider the sonification experience as a phenomenon that the user perceives and assimilates, iteratively apprehending its patterns and changes to decode its underlying information. We applied a phenomenological approach to evaluate the experience, conducting a series of interviews to assess the users perception.

The evaluation carried out, with the goal of gathering the realities and associations sensed through the listening experience, provided some interesting results, with the consumption context indirectly present throughout the emerging themes. The idea of everyday life, routine, time passing over each day and how we move and carry on our day in a fast, frenetic rhythm are sensations intrinsically connected to the consumption context. The variations of the melody flow throughout a day of consumption are similar to the ones people live in their work day, with peaks of energy, peaks of responsibility, which repeat themselves day after day. Overall, the participants related the musical flow with a more human perspective of how we live each day nowadays in a fast, frenetic and stressful rhythm that dictates our modern routine. The evocation of a documentary or movie imagery about life in the twenty-first century, with mass people moving and working in search for progress, also translates this scenario.

Several suggestions and comments made by the participants defined an initial list of possible refinements and settings to experiment in the future. The amount of negative sensations, as annoyance, boredom or anxiety, could be reduced by exploring the effect of different tempo measures, and balance the differences between the high and low pitches, lowering the maximum range of pitch values to avoid the strident sensation, and lowering the intensity of low-pitched sounds, specially the trombone, which may be masking other timbres. The steady rhythm, with varied speed depending on the time of day, although it may help making the composition feel annoying or boring, it translates the expected repetitive nature of the data, and maybe why people felt the idea of routine and everyday life that is constant. This might be due to the categories with less variations, mapped coincidentally to low-pitched instruments (as the oboe and the trombone), which tend to mark the tempo and stand out. It should be noted that these are just speculative explanations, which demand further testings with a higher sample of interviewees to properly validate the mapping choices.

The emerging themes gave us a set of scenarios to further reflect upon, demanding new iterations on the mapping and how the different variables are balanced within the sound. This experience already demanded an active role of the user in a disembodied form. Further work might include interaction modes, which could provide a better support for active listening, namely a spatialized environment that could allow the user to move through the space, from store to store, and emphasize the idea of moving to search for information, to discover and internalize the auditory phenomenon. In this setting, we intend to explore the expected differences already noticed between stores, as the ones from cities of larger dimension and wider population have more variations across the day and higher levels of sales. We intend to explore how could these differences be perceived while exploring the auditory space, studying the concepts of interaction and perception aesthetics through this embodied process for user engagement.

7. ACKNOWLEDGMENT

Mariana Seiça is funded by Fundação para a Ciência e Tecnologia (FCT), Portugal, under the grant SFRH/BD/138285/2018. We would like to thank the retail company SONAE, and the members of the project *Sonae Viz - Big Data Visualization for retail* Catarina Maçãs and Evgheni Polisciuc for processing the data. We also thank the ten individuals for giving a part of their time in participating and contributing to this experiment.

8. REFERENCES

[1] S. Barrass and P. Vickers, *The Sonification Handbook*. Berlin, Germany: Logos Verlag, 2011, ch. Sonification Design and Aesthetics, pp. 145–172.

[2] P. Vickers, B. Hogg, and D. Worrall, *Body, Sound and Space in Music and Beyond: Multimodal Explorations*. London and New York: Routledge Taylor Francis Group, 2017, ch. Aesthetics of sonification: taking the subject-position, pp. 89–109.

[3] J. G. Neuhoff, *The Sonification Handbook*. Berlin, Germany: Logos Verlag, 2011, ch. Perception, Cognition and Action in Auditory Displays, pp. 63–86.

[4] D. Svanæs, "Interaction design for and with the lived body: Some implications of merleau-ponty's phenomenol-

ogy," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 20, no. 1, p. 8, 2013.

[5] C. Maçãs, P. Martins, and P. Machado, "Consumption as a rhythm: A multimodal experiment on the representation of time-series," in *2018 22nd International Conference Information Visualisation (IV)*. IEEE, 2018, pp. 504–509.

[6] B. House. (2012) Quotidian record. [Online]. Available: http://brianhouse.net/works/quotidianrecord/

[7] M. Quinn, "Research set to music: The climate symphony and other sonifications of ice core, radar, dna, seismic and solar wind data," in *Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland*.

[8] J. Bulley and D. Jones. (2014) Living symphonies. [Online]. Available: http://www.livingsymphonies.com/

[9] J. Dunn and M. A. Clark, "Life music: the sonification of proteins," *Leonardo*, vol. 32, no. 1, pp. 25–32, 1999.

[10] J. Dunn and M. A. Clarke. (1998) Life music. [Online]. Available: http://algoart.com/music.htm

[11] P. Vickers, "Sonification abstraite/sonification concrète: An 'aesthetic perspective space' for classifying auditory displays in the ars musica domain," vol. abs/1311.5426, 2013. [Online]. Available: http://arxiv.org/abs/1311.5426

[12] K. Groß-Vogt, T. Hermann, M. W. Jury, A. K. Steiner, and S. Kartadinata, *Klima |Anlage—Performing Climate Data*. Cham: Springer International Publishing, 2019, pp. 339–355.

[13] K. Gro-Vogt, T. Hermann, M. W. Jury, A. K. Steiner, and S. Kartadinata. (2016) Klima |anlage: Perfoming climate data. [Online]. Available: http://klima-anlage.org

[14] R. McGee and D. Rogers, "Musification of seismic data." The 22nd International Conference on Auditory Display, Canberra, Austria, 2016.

[15] ——. (2012) Haiti earthquake mw7.0 12th january *02010 - ryan mcgee. [Online]. Available: https://soundcloud.com/seismicsounds/haiti-earthquake-12th-january

[16] S. Roddy, "Embodied sonification," Ph.D. dissertation, University of Dublin, Trinity College, 2015.

[17] ——. The human cost: Sonification and irelands economic crash. [Online]. Available: https://stephenroddy.bandcamp.com/album/the-human-cost-sonification-and-irelands-economic-crash

[18] C. Maçãs, P. Cruz, P. Martins, and P. Machado, "Swarm systems in the visualization of consumption patterns," in *24th International Joint Conference on Artificial Intelligence*, 2015.

[19] P. Vickers, *The Routledge Companion to Sounding Art*. Routledge, 2016, ch. Sonification and Music, Music and Sonification.

[20] S. Gresham-Lancaster, "Relationships of sonification to music and sound art," *AI Soc.*, vol. 27, no. 2, pp. 207–212, May 2012.

[21] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *the Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[22] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological research*, vol. 58, no. 3, pp. 177–192, 1995.

[23] E. Rentz, "Musicians' and nonmusicians' aural perception of orchestral instrument families," *Journal of Research in Music Education*, vol. 40, no. 3, pp. 185–192, 1992.

[24] B. Truax, *Acoustic communication*. Greenwood Publishing Group, 2001.

[25] Ranges of orchestral instruments. [Online]. Available: http://www.orchestralibrary.com/reftables/rang.html

[26] M. Seiça, R. B. Lopes, F. A. Cardoso, and P. Martins, "Sonifying twitter's emotions through music," in *Lecture Notes in Computer Science, Music Technology with Swing, Revised Selected Papers of the 13th International Symposium on Computer Music Multidisciplinary Research*, M. Aramaki, M. Davies, R. Kronland-Martinet, and S. Ystad, Eds. Berlin, Heidelberg: Springer, in press.

[27] B. N. Walker and G. Kramer, *Ecological Psychoacoustics and Auditory Displays: Hearing, Grouping, and Meaning Making*. New York: Elsevier Academic Press, 2004.

[28] Z. Bilda, *Interacting: Art, Research and the Creative Practitioner*. UK: Libri Publishing, 2011, ch. Designing for Audience Engagement, pp. 163–181.

[29] S. Lester, "An introduction to phenomenological research," 1999.

[30] D. Idhe, *Experimental Phenomenology*. New York, USA: State University of New York Press, 1977, ch. Phenomena and the Phenomenological Reductions, pp. 29–54.

[31] M. T. Bevan, "A method of phenomenological interviewing," *Qualitative health research*, vol. 24, no. 1, pp. 136–144, 2014.

THE SONIFICATION OF SOLAR HARMONICS (SoSH) PROJECT

Seth Shafer *Timothy Larson* *Elaine diFalco*

University of Nebraska at Omaha
6001 Dodge Street
Omaha, NE, 68182
sethshafer@unomaha.edu

tplarson@sun.stanford.edu

University of North Texas
1155 Union Circle
Denton, TX 76203
elainedifalco@my.unt.edu

ABSTRACT

The Sun is a resonant cavity for very low frequency acoustic waves, and just like a musical instrument, it supports a number of oscillation modes, also commonly known as harmonics. We are able to observe these harmonics by looking at how the Sun's surface oscillates in response to them. Although this data has been studied scientifically for decades, it has only rarely been sonified. The Sonification of Solar Harmonics (SoSH) Project seeks to sonify data related to the field of helioseismology and distribute tools for others to do the same. Creative applications of this research by the authors include musical compositions, installation artwork, a short documentary, and a full-dome planetarium experience.

1. INTRODUCTION

It is a poignant coincidence that acoustical physics is such an intrinsic part of our most prominent celestial object when so much of Western philosophical history connects the cosmos to sound. Our home star appears as a mass of boiling plasma, and it rings like a bell in a sandstorm, generating millions of resonant harmonic modes [1]. By applying the mathematics of spherical harmonics and fluid dynamics we are able to determine various properties of the Sun's internal structure. Although the data used is acoustic in nature, scientists have only very rarely listened to it, despite the fact that sonification of other types of solar data has yielded new scientific insights [2].

After a short introduction to the subject of helioseismology, we describe a collaborative research initiative called the Sonification of Solar Harmonics (SoSH) Project. This project seeks to transform helioseismology into a listening experience for the scientist and nonscientist alike. One of the initial outcomes from the SoSH Project is a software tool for rendering helioseismic data as audible sound. This tool is the most advanced contribution to helioseismic sonification to date and provides the first access to audio generated from the most recent data available. In addition to the SoSH Tool, we describe several creative applications derived from solar harmonics research and future directions for the project. The contents described in this paper including the SoSH Tool, datasets, and additional information can be found at http://solar-center.stanford.edu/SoSH/.

2. A BRIEF PRIMER ON HELIOSEISMOLOGY

The study of oscillations inside the Sun is called helioseismology. In particular, here we shall consider acoustic waves. Turbulent convection near the solar surface excites sounds waves, and the waves with frequencies that resonate form the harmonics. Just as the frequency of a plucked guitar string becomes higher with greater tension and lower with greater thickness, the frequencies of the Sun's harmonics enable us to infer properties of the solar interior such as its pressure and density. And just as any acoustic instrument produces a set of harmonics above a fundamental frequency that combine to create a characteristic timbre, so too does the Sun.

2.1. Spherical Harmonics

The input data for helioseismology are typically velocity images of the Sun, where each pixel gives the speed of that plasma element toward or away from the observer. It is a mathematical theorem that any such image of the Sun's surface can be expressed as a sum over spherical harmonics, which are simply a special set of functions of latitude and longitude. Each of these functions are labeled by two integers: the spherical harmonic degree l and the azimuthal order m. The degree l is ≥ 0, and for each l, there are $2l + 1$ values of m, ranging from $-l$ to l.

One way to understand spherical harmonics is in terms of their node lines, which are the places on the sphere where the spherical harmonics have an amplitude of zero. The degree l tells how many of these node lines there are in total, and the absolute value of the order $|m|$ gives the number in longitude, so the number of node lines in latitude is $l - |m|$. Therefore a spherical harmonic with $m = 0$ has only latitudinal bands, while one with $m = l$ has only sections like an orange. A third integer, the radial order n, tells how many nodes the oscillation has along the Sun's radius. Since only the surface of the Sun is visible to us, all the values of n are present in each spherical harmonic labeled by l and m, although only some of them will be excited to any appreciable amplitude. The total mode, then, is represented as a product of a spherical harmonic and another function of radius, known as the radial eigenfunction. The radial eigenfunction depends on both n and l.

Figure 1 below illustrates modes with degree $l = 5$ and all nonnegative values of m. Modes with $m < 0$ are not included because in a still image they are indistinguishable from modes with $m > 0$. As the spherical harmonics evolve in time, one would see the two signs of m rotate in opposite directions; this is discussed further below. One can see that different spherical harmonics sam-

https://doi.org/10.21785/icad2019.011

ple different latitudes, according to the value of $|m|$ relative to the degree l. Modes with high absolute values of m have their maximum amplitude at low latitudes, whereas lower values extend to higher latitudes.

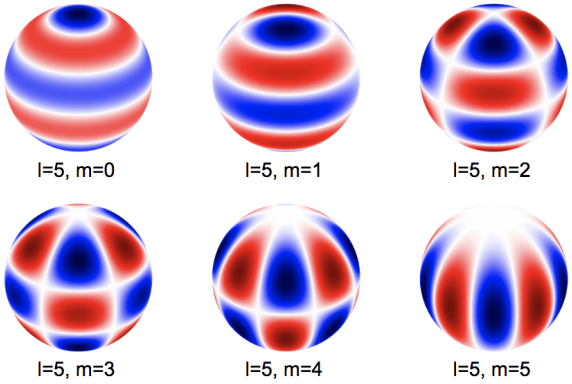


Figure 1: Surface views of the Sun showing all harmonic modes with degree $l = 5$.

2.2. Modeling the Sun's Interior

The harmonics we are able to measure have frequencies ranging from 1000 to 5000 microhertz. For any given value of the degree l, we will find a certain range of values of the radial order n, with frequency increasing as n increases. The next two figures show interior views of the Sun for two different radial orders and degree $l = 5$. For clarity, in Figure 2 we show radial order $n = 3$, although this mode is expected to have a frequency too low to measure. Figure 3 shows radial order $n = 20$, which is easily measured, but at this frequency the nodes along the radius become so closely spaced near the surface that they are difficult to discern at this scale. It is important to realize that we are seeing modes of many different n in each spherical harmonic. For instance, for degree 5 we might measure modes with n ranging from 7 to 28, each oscillating at its own frequency, roughly 140 microhertz apart from each other.

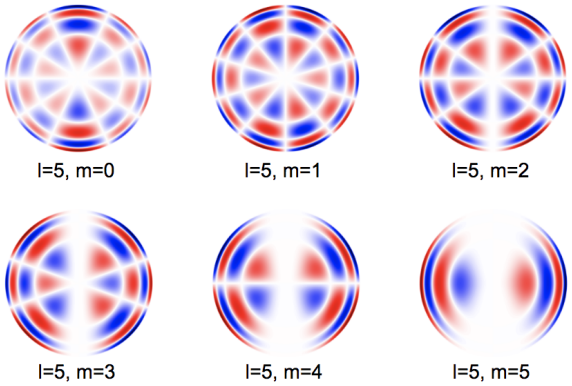


Figure 2: Interior views of the Sun showing harmonic modes with degree $l = 5$ and radial order $n = 3$. A model predicts these modes to have a frequency of 800 microhertz.

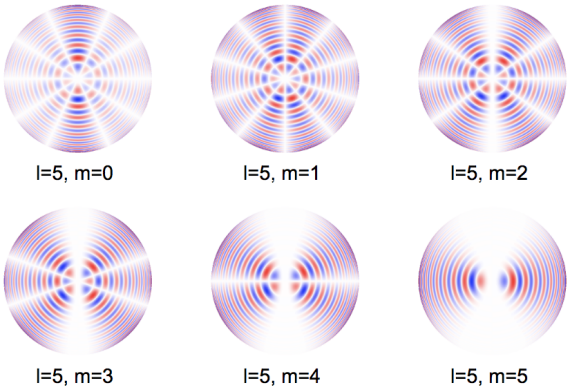


Figure 3: Interior views of the Sun showing harmonic modes with degree $l = 5$ and radial order $n = 20$. A model predicts these modes to have a frequency of 3190 microhertz.

Each mode oscillates with its characteristic frequency, and each samples different depths inside the Sun. At a given degree, high frequencies will penetrate more deeply, while low frequencies are trapped closer to the surface. Likewise, at a given frequency, high values of the degree l will be trapped near the surface, while low values will penetrate almost all the way to the core, with $l = 0$ even reaching the center. This is further illustrated in the next series of figures below, which show a selection of modes with degree $l = 25$ for the same two values of n (see Figure 4). We see that the extent in latitude of the spherical harmonic combines with the extent in radius of the radial eigenfunction to yield a mode that inhabits a particular region of the Sun.

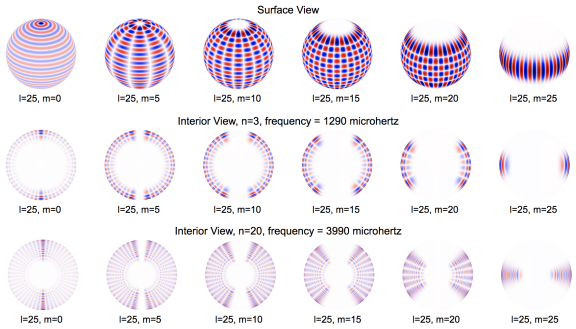


Figure 4: Surface views for degree $l = 25$ and corresponding interior views for $n = 3$ and $n = 20$. The expected frequencies are 1290 and 3990 microhertz.

2.3. Typical Data Pipeline

To determine the frequencies of the Sun's harmonics, a typical instrument might take an image once a minute for 72 days. For each image, we decompose it into its various spherical harmonic components. For each of these components, we form a timeseries of its amplitude. From the timeseries we are able to construct the power spectrum (acoustic power as a function of frequency). It is here that we are able to separate the two signs of m. Without delving into complex analysis and the theory of the Fourier transform, we simply state the two signs correspond to the positive and negative

frequency parts of the power spectrum. Furthermore, the sign of m that rotates in the same direction as the Sun will be shifted up in frequency, while the sign that goes against solar rotation will be shifted down in frequency. Because different modes sample different regions of the Sun, we are able to use their frequencies to determine solar properties as a function of both depth and latitude. For example, we are able to use the frequency splitting in m to measure internal solar rotation.

Once the power spectrum is calculated, the location of peaks will correspond to the frequencies of the modes. The height of each peak tells us the mode amplitude, and the width of the peak tells us how much the oscillation is damped. Each n will have its own peak in the power spectrum.

The relationship between l, n, and frequency is illustrated in Figure 5 below. Plotted on the left is raw power as a function of l and frequency for $m = 0$. As one can see, the modes for a given n form a ridge of power. The right panel is a scatter plot indicating which modes we have been able to successfully fit. In both plots the bottom ridge corresponds to $n = 0$.

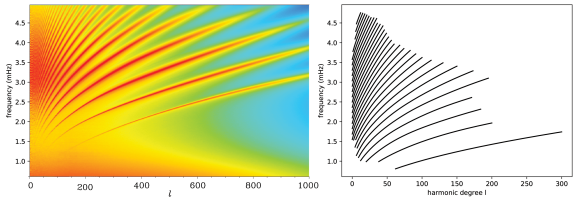


Figure 5: Degree-frequency diagrams. Left panel shows raw power, right panel shows the modes that we are able to fit.

3. OVERVIEW OF THE SoSH PROJECT

3.1. Previous Sonification Efforts

The Sonification of Solar Harmonics (SoSH) Project is the most extensive effort toward the translation of helioseismic data into audible sound, but it is not the first. That distinction goes to Douglas Gough, who created an audio tape for public demonstration using solar data in the early 1980's [3]. More recently, Alexander Kosovichev sonified 40 days of data taken with an instrument in space [4]. Although he had the benefit of computer processing, still only a handful of audio samples were created. Nonetheless, until preliminary investigations by Larson and the current project, these few audio files were the only sonifications of helioseismic data available to the public, and they were used in a wide range of other works. [5]. Although Larson sonified data covering an entire solar cycle, it was not until the development of the SoSH Tool that one became able to sonify helioseismic data interactively.

For completeness, we note that other types of solar data have been sonified over the years, for both artistic and scientific purposes. Examples include the use of solar cycle data by Thorbjoern Lausten [6] and solar radio emissions by Thomas Ashcraft [7]. Several projects have sonified various solar wind data, notably Robert Alexander working with the University of Michigan to use ACE data [8], a UC Berkeley group using STEREO data [9], and Don Gurnett at the University of Iowa using data from the Plasma Wave Instrument onboard Voyager [10]. Finally, Chris Hayward [11] and Florian Dombois [12] conducted work similar to this project to audify geoseismic data.

3.2. Datasets for Sonification

The SoSH Project aims to sonify any solar harmonics dataset. Several such datasets are available for sonification. The first of these comes from the Michelson Doppler Imager (MDI) [13] onboard the Solar and Heliospheric Observatory (SoHO). It was in operation from May 1996 to April 2011 and was the source of data for both the audio files made by Kosovichev and the earlier work of Larson. Similar to MDI and parallel to instruments operating in space, the Global Oscillation Network Group (GONG) has operated a network of six ground-based observatories since 1995 [14]. Both GONG and MDI observe the same spectral line at a cadence of one minute.

In 2010 MDI was superseded by the Helioseismic and Magnetic Imager (HMI) [15] onboard the Solar and Heliospheric Observatory (SDO). HMI observes a different spectral line at a cadence of 45 seconds and remains in operation today. The SoSH Project represents the only sonification of HMI data to date. Although capable of doing so, we have not yet sonified any GONG data.

3.3. The SoSH Tool

The SoSH Project developed a tool to input complex spherical harmonic timeseries and output sonified audio of that data. The only preprocessing needed is to convert the timeseries from FITS (Flexible Image Transport System) to WAV format. This first step represents a pure audification of the data [16]. As discussed below, the subsequent processing done by the SoSH Tool exactly parallels the steps in the scientific analysis. By using fitted mode parameters to filter and transform the data, we produce a sonification.

The strongest of the Sun's harmonics have periods of about 5 minutes, corresponding to frequencies of only about 0.003 hertz. Unfortunately, this is far below the range of human hearing, which is typically taken to be $20 - 20,000$ hertz, although most people are only sensitive to a smaller range. In order to experience the sound of the Sun with our ears, these very low sounds must be scaled to the range we can hear.

3.3.1. Sample Rate Time Scaling

The most straightforward way to do so would be to use the spherical harmonic timeseries we already have in hand and speed them up. But by how much? The answer of course is arbitrary and will depend on your preference, but as long as this choice is applied consistently you will still be able to hear the real relationship between different solar tones.

Let us suppose that we want to transpose a mode in the peak power range at about 0.003 hertz up to 300 hertz; this amounts to speeding up the timeseries by a factor of $100,000$. If we have 72 days of data taken once a minute, it amounts to $103,680$ data points. If each data point becomes an audio sample, the sped-up timeseries would now play in just over a minute. One must also consider the sample rate. Speeding up the original sample rate of $1/60$ hertz by a factor of $100,000$ yields a new sample rate of 1666.67 hertz. Unfortunately, most audio players will not play any sample rate less than 8000 hertz. Assuming this sample rate, our 0.003 hertz mode on the Sun will now be transposed up to 1440 hertz and the timeseries will play in about 13 seconds.

But suppose you want to play it in a shorter time; 13 seconds is a long time to sound a single note, although you might want to do so in some circumstances. You could increase the sample

rate further still, but at some point the mode will be transposed to an uncomfortably high frequency. For example, if we sped up the sample rate so that the duration is only 1 second, the resulting frequency would be over 18000 hertz. To understand the solution to this problem, we must explore the process by which we shall isolate the modes.

3.3.2. *Modal Isolation*

At this point in our processing, playing an unfiltered timeseries would sound just like static, or noise. This is because very many modes are sounding simultaneously in any given timeseries, not to mention the background noise involved in our observation of the modes. Therefore, if we want to isolate a single mode, we have to do some filtering. Luckily, as mentioned above, we have already measured the frequency, amplitude, and width of many modes. We can use these fitted mode parameters to pick out the particular part of the power spectrum corresponding to a single mode, and set the rest of the power spectrum artificially to zero. Once transformed back into the time domain, we can play the filtered data back and hear only the isolated mode.

Properly speaking, we do not compute any power spectra, but rather retain both the real and imaginary parts of the Fourier transform. The input timeseries are also complex because each of them contains both signs of the azimuthal order m (the timeseries for $m = 0$ has only a real part). Because we also perform the transform over the full length of the timeseries rather than windowing, we preserve the maximum amount of phase information present in the original signal. Although of course we cannot hear the phase, it may nonetheless become important when adding different modes together. Finally, we note that the output is strictly real because it now contains only a single m.

3.3.3. *Pitch Shifting and Time Stretching*

Since we are selecting only a narrow range of frequencies, we have the freedom to shift the entire power spectrum down in frequency before we transform back to timeseries. This timeseries will play in the same amount of time as before, but the frequencies in it will be transposed down by the same factor that we shifted the power spectrum. For every power of 2 shifted down, the tone will drop by one octave. This allows for the resulting audio to be set to any arbitrary pitch, solving the problem with uncomfortably high frequency playback.

As long as you use the same sample rate and downshift factor when you sonify every mode, the frequency relationships between them will be preserved. In other words, you will hear the same musical intervals that exist on the Sun.

Conversely, the duration of the resulting sonification can be altered by modifying the sample rate, although one must also multiply the downshift factor by the same amount if they desire to keep the same total transposition factor. Any arbitrary sample rate is permissible. The sample rate modification happens after the inverse transform and uses bilinear interpolation to maintain an internal sample rate of $44,100$ hertz. When writing the output to a WAV file, however, the chosen sample rate is written into the WAV header.

3.3.4. *Using the SoSH Tool*

The SoSH Tool is an open-source system implemented in the visual programming language Pure Data (PD). Upon launching the demo patch "modefilter_standalone.pd" (see Figure 6), you are required to specify a path to your local data directory. The patch will search this directory for ASCII files containing the fitted mode parameters, as well as the WAV files for whatever spherical harmonics you wish to sonify.

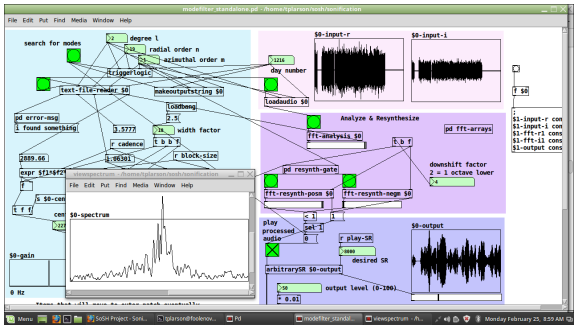


Figure 6: A screen shot of the core processing engine inside the SoSH Tool.

With these files in place, the dataset is now accessible by the day number corresponding to the first day of the timeseries, the spherical harmonic degree l, the radial order n, and the azimuthal order m. By default, the tool is set to use MDI data. If it is successfully able to load the necessary audio files, it will automatically trigger FFT analysis of them. Next, the fitted mode parameters corresponding to the inputs are used to generate a gain array in which full output is allowed for a frequency interval centered on the mode frequency and of width five times the mode width. Full attenuation is set for all other bins. This default width of the pass band is the same as that used for the fitting, but it can be increased multiplicatively in the software by specifying the "width factor."

Once the gain array is generated, the FFT data is multiplied by it and resynthesized. If a downshift factor is specified, the FFT will be shifted down by this amount before the inverse transform.

Beyond the basic sonification of single modes, several ex- tensions to the SoSH Tool enable more advanced functions. One extension processes several modes and additively outputs their results into a single array (see Figure 7). Another extension turns the tools into a kind of multitimbral sampler that can trigger different modes assigned to different MIDI pitches. Yet another extension reads a score from a text file to produce a sequence of modes.

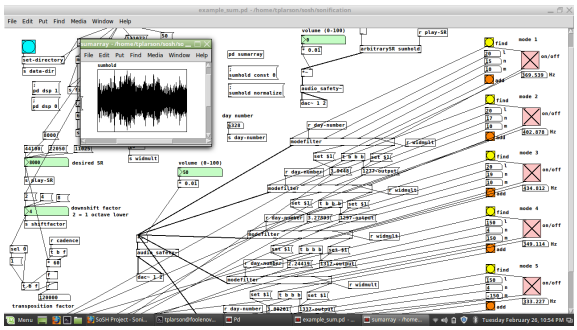


Figure 7: A screen shot of an extension to the SoSH Tool that additively combines modes into a single output array.

The SoSH Tool, datasets, audio samples of sonified data, and

much more can be downloaded from the SoSH website [17].

3.4. Audible Science

The physical phenomenon most accessible to hearing is solar rotation. In just a single timeseries of moderate m, the frequency splitting between the two signs results in a clearly audible beat frequency when they are played together. In a similar fashion, one may also easily hear the so-called small frequency separation, which is sensitive to conditions in the solar core [18]. Both types of frequency difference are known to vary with the solar cycle. However, if we apply an absolute frequency shift rather than a relative one, one may hear the effect of the solar cycle on a single mode by juxtaposing timeseries from two different epochs.

4. CREATIVE APPLICATIONS

While the SoSH Tool is clearly a potential benefit to the scientific community, it also generates possibilities for creative output. Music compositions, installation artwork, a documentary, and a full-dome planetarium experience are detailed below.

4.1. Music Composition

Music composition is one creative outcome from the SoSH Project. diFalco's composition *Helios* (2018), from a set of works collectively titled *Cosmophonia*, is written for orchestra and imagines a journey directly through the center of the Sun beginning at the furthest reaches of the heliosphere [19]. The composition applies the acoustic phenomenon of helioseismology by scoring the natural harmonic series as it relates to the position inside or outside of the Sun. In addition, the primary pitch material is derived from a solar flare recorded on 23 November 1998 by Peter Messmer et al. at the Institute of Astronomy in Switzerland (see Figure 8) [20].

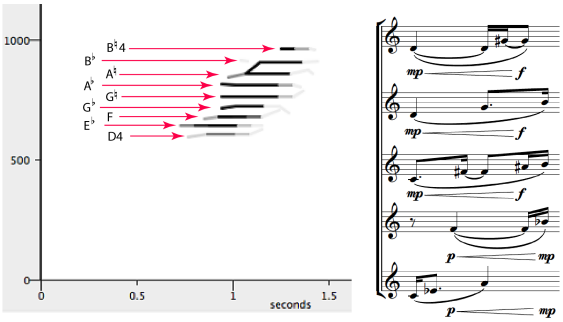


Figure 8: Analysis and orchestration of a solar event used in *Helios* (2018) by Elaine diFalco.

Many possibilities exist to directly implement the audio output from the SoSH Tool into compositions employing electronic means. As noted above, this audio can be manipulated to map to any pitch or stretched/compressed to any duration. One particularly exciting application is using the sonified data at low rate (perhaps even the real-time rate of one data point per minute) to drive other processes like synthesis parameters, sample playback, tempo rate, dynamic contour, etc. In this way, the sonified data can act as an LFO to shape larger-scale parameters.

The authors are collaboratively composing a new work for orchestra and electronics using the SoSH Tool to generate both the acoustic score and the accompanying electronics. Given the large datasets available, the performance will sonify an entire solar cycle to create a kind of "solar concerto."

4.2. Installation Artwork

The application of the SoSH Project to installation based work is particularly attractive. Shafer's multichannel projector installation *Sol Invictus* (2015) uses the raw visual data from the Atmospheric Imaging Assembly (AIA) onboard SDO to display the complex convection happening in the chromosphere (see Figure 9) [21]. The gigantic scale of the solar forces is subverted by the minuscule dimension of the images generated by the five micro projectors.

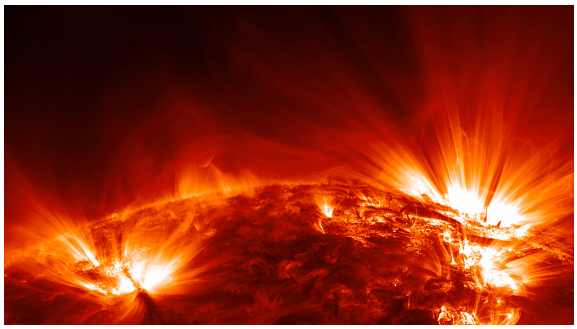


Figure 9: A still from the multichannel projector installation *Sol Invictus* (2015) by Seth Shafer.

4.3. Documentary and Intermedia

The process of creating immersive experiences to understand the abstract datasets of helioseismology is the subject of diFalco's documentary *Immersing* (2018). The film discusses the topic of helioseismology, the complexity of the dataset, and the intermedia practice of creating immersive experiences.

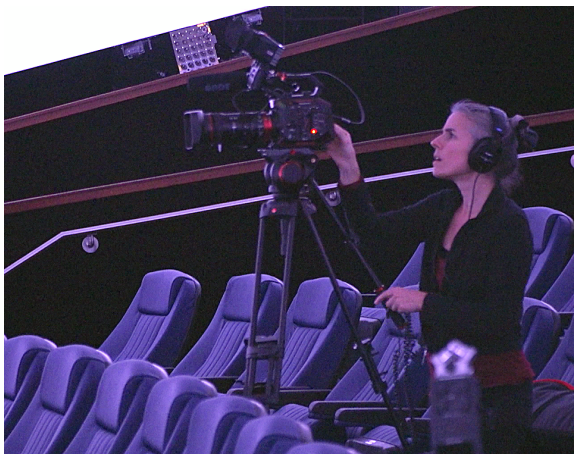


Figure 10: Still from the documentary *Immersing* (2018) by Elaine diFalco.

Another important outcome from the SoSH Project is the SoshPy Visualization Package. Written in Python, this software package was used to generate Figures 1–4 of this paper. It can also plot all three components of the vector velocity associated with any given mode, as well as its energy density. More interestingly, it is also able to plot sums of modes and create animations.

Plans for a large-scale intermedia experience using the SoSH Tool are underway. The full-dome projection theater, commonly used in planetariums, is an ideal platform for the SoSH Project due the immersive possibilities and its attraction to the science-seeking public. Due to the shared use of spherical harmonics for representation, Ambisonic panning derived from the data will be used to direct spatial placement in the full-dome [22]. Further details of this creative activity will emerge as the project progresses.

5. CONCLUSIONS AND FUTURE RESEARCH

The SoSH Project designed a software tool that sonifies helioseismic data from the largest and most recent datasets available. The project is a significant update to and expansion upon earlier efforts to sonify data related to the surface and interior of the Sun.

Current efforts are underway to create an interface between the SoSH Tool and the SoshPy Visualization Package. Both programs will ultimately read from the same "score"; a SoSH extension will create the soundtrack for the SoshPy animations. Going the other direction, we will use SoshPy to trace a trajectory through the Sun and then output the "score" corresponding to the relative amplitudes of modes we specify.

Finally, further creative applications of the SoSH Tool are in development. The full-dome experience described above is in early planning stages and will benefit from community feedback on the project at large and the SoSH Tool specifically.

We hope that the accessibility of the data and now tools to easily sonify that data will lead to new scientific insights related to our home star. In addition, as we attempt to make artwork from these fascinating natural phenomenon, we encourage others to also generate new creative work inspired directly or indirectly from this project.

6. ACKNOWLEDGMENT

The authors would like to thank Jesper Schou for helpful discussions.

7. REFERENCES

[1] A. Graps. (2009) Helioseismology. [Online]. Available: http://soi.stanford.edu/results/heliowhat.html

[2] R. L. Alexander et al., "Audification as a diagnostic tool for exploratory heliospheric data analysis," *The 17th International Conference on Auditory Display*, pp. 1–4, 2011.

[3] D. O. Gough, personal communication, 2018-08-31.

[4] A. Kosovichev. (2008) The singing sun. [Online]. Available: http://solar-center.stanford.edu/singing/singing.html

[5] D. Scherrer. (2010) The sun in music. [Online]. Available: http://solar-center.stanford.edu/art/music.html

[6] T. Lausten. (2000) Sol: Data of two suncycles. [Online]. Available: http://www.sol-sol.de

[7] T. Ashcraft. (2010) Heliotown–radio sun. [Online]. Available: http://www.heliotown.com/Radio_Sun_Introduction.html

[8] R. L. Alexander et al., "Sonification of ace level 2 solar wind data," *The 16th International Conference on Auditory Display*, pp. 39–40, 2010.

[9] J. Luhmann et al. (2010) Uc berkeley–sounds of space. [Online]. Available: http://cse.ssl.berkeley.edu/stereo_solarwind/sounds.html

[10] D. Gurnett. (2018) The university of iowa–space sounds. [Online]. Available: http://www-pw.physics.uiowa.edu/space-audio/index.html

[11] C. Hayward, "Listening to the earth sing," in *Auditory Display: Sonification, Audification, and Auditory Interfaces*, G. Kramer, Ed. CRC Press, 1994, pp. 369–405.

[12] F. Dombois et. al., "Sonifyer: A concept, a software, a platform," *Proceedings of the 2008 International Conference on Auditory Display*, pp. 1–4, 2008.

[13] P. H. Scherrer et al. (2011) The michelson doppler imager. [Online]. Available: http://soi.stanford.edu/

[14] J. W. Harvey et al. (1996) The global oscillation network group (gong) project. [Online]. Available: https://gong.nso.edu/

[15] J. Schou et al. (2018) Helioseismic and magnetic imager. [Online]. Available: http://hmi.stanford.edu/

[16] F. Dombois and G. Eckel, "Audification," in *The Sonification Handbook*, Thomas Hermann et al., Ed. Logos Publishing House, 2011, pp. 301–324.

[17] T. Larson, S. Shafer, and E. diFalco. (2018) Sonification of solar harmonics project. [Online]. Available: http://solar-center.stanford.edu/SoSH/

[18] W. J. Chaplin, *Music of the Sun: The Story of Helioseismology*. Oneworld Publications, 2006.

[19] E. diFalco, "*Cosmophonia*: Musical expressions of astronomy and cosmology," Masters Thesis, University of North Texas, 2018.

[20] P. Messmer. (1998) Acoustic recording of a type iii solar burst recorded on november 23, 1998 between 1140-2280 mhz. [Online]. Available: http://www.astrosurf.com/luxorion/Radio/solar-burst-type3-23nov98-messmer.wav

[21] S. Shafer. (2015) *Sol Invictus*. [Online]. Available: http://sethshafer.com/solinvictus.html

[22] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc*, vol. 33, no. 11, pp. 859–871, 1985.

A RADAR-BASED NAVIGATION ASSISTANCE DEVICE WITH BINAURAL SOUND INTERFACE FOR VISION-IMPAIRED PEOPLE

Christoph Urbanietz[1], *Gerald Enzner*[1], *Alexander Orth*[2], *Patrick Kwiatkowski*[2], *Nils Pohl*[2]

Ruhr University Bochum, Department of Electrical Engineering and Information Technology
Institute of Communication Acoustics[1], Institute of Integrated Systems[2]
Bochum, 44801, Germany
christoph.urbanietz@rub.de[1], alexander.orth@rub.de[2]

ABSTRACT

Sound is extremely important to our daily navigation, while sometimes slightly underestimated relative to the simultaneous presence of the visual sense. Indeed, the spatial sense of sound can immediately identify the direction of danger far beyond the restricted sense of vision. The sound is then rapidly and unconsciously interpreted by assigning a meaning to it. In this paper, we therefore propose an assisted-living device that deliberately stimulates the sense of hearing in order to assist vision-impaired people in navigation and orientation tasks. The sense of vision in this framework is replaced with a sensing capability based on radar, and a comprehensive radar profile of the environment is translated into a dedicated sound representation, for instance, to indicate the distances and directions of obstacles. The concept thus resembles a bionic adaptation of the echolocation system of bats, which can provide successful navigation entirely in the dark. The process of translating radar data into sound in this context is termed "sonification". An advantage of radar sensing over optical cameras is the independence from environmental lighting conditions. Thus, the envisioned system can operate as a range extender of the conventional white cane. The paper technically reports the radar and binaural sound engine of our system and, specifically, describes the link between otherwise asynchronous radar circuitry and the binaural audio output to headphones.

1. SYSTEM OVERVIEW

The goal of this work is to design a tool to support blind or visually impaired people in navigation and orientation tasks. As a general concept, we collect information about the environment that is usually recognized by the visual sense and therefore missing by a technical sensor and convert this collected and processed information to a sensation the user is able to recognize.

There are many products on the market that also use this principle. The app "The vOICe" translates a camera image into an audio signal, whereas the "Orcam" line of products analyze the camera image and translate it to meaningful spoken words. Other tools such as the "UltraCane" or "Live Braille" use ultrasonic detectors and translate object distances into vibrations. A variation of haptic output in the form of pressure on the head in combination with an ultrasonic sensor input is realized by the "Proximity Hat". Additionally, a combination of camera input and vibrating output has been investigated, e.g., by the University of Southern California. Common to all these tools is that they do not need exact maps of the environment. Instead, they rely only on the sensor input, and so they can also be used in unknown environments. For the technical sensing part of our system, we use a radar sensor, and for the output, we choose the binaural acoustic modality.

Radar sensing is a common tool in exploring unknown environments. Most modern cars are equipped with one or more radar sensors to scan across the driving direction, either to identify preceding cars and follow them in an autonomous or half-autonomous driving configuration or to identify dangerous situations such as a rapidly braking car ahead to initiate automatic emergency braking. Radar sensing brings some advantages over competing technologies such as light detection and ranging (lidar) or ultrasonic detection. Ultrasonic detection has a limited distance range and a wide detection beam. Lidar has the opposite characteristics. It can accommodate long distances and has a very pointlike steering region. Radar systems present a compromise between these two approaches. The beam can be focused, and the distance range can extend for several tens of meters. The distance range of a radar system is optimal for our purpose to extend the explorable area compared to that of a white cane. Although the lidar operational wavelength is most similar to the wavelengths used by the human visual sense, radar can extend the exploration of the environment. It can look through fog and even detect glass doors, which can be very challenging with lidar or, in some cases, even with human eyes. Furthermore, we protect the environment from harmful laser emission when using radar instead of lidar. Although the lasers used in lidar devices are claimed to be harmless to human eyes, they can destroy camera sensors as present in many devices such as cameras, smart phones, security cameras or autonomous cars.

For the output part, we use the acoustic modality; therefore, we speak of sonification. For navigation and orientation tasks, an audio channel is often used, but without the presentation of binaural cues. A car navigation system is a good example of this configuration. Although the navigation information is presented visually on a screen, there is the commonly used option to give additional acoustic navigation advice. The main reason for this apparently redundant presentation is to let the eyes focus on the street without the need to look at the screen. Although there are approaches to reduce the time of visual distraction from the street (e.g., head-up displays), the acoustic modality overcomes this issue more rigorously since it can be sensed in parallel. In addition to the factor of convenience and security over the pure visual display, in our case

https://doi.org/10.21785/icad2019.023

of blind or vision-impaired people, the sound modality is more essential to deliver cues for navigation and orientation.

A further step in convenience is the use of natural cues such as binaural localization for indicating directional information, e.g., perceiving sound from the direction to drive or walk instead of explaining the direction in words. Thus, by indicating the directional information by attributes of the sound and not the verbal articulation, we wish to reduce the sound to a more subliminal representation without verbal content for directional information. For simple instructions such as "left" or "right", this approach provides a more subconscious access and avoids unnecessary long speech and therefore might be perceived as a more pleasant hint. For more complex navigation instructions such as "slightly left" or "half right", the use of binaural sound can provide additional cues for more accurate indication of routes or obstacles. Here, the additional binaural cue will be more essential when more degrees of freedom exist for navigation, for instance, for moving in free space. As in most scenarios, the mobility of the subject is restricted to a 2-D plane (e.g., the street level); thus, we restrict the locations of the virtual sound sources to the horizontal plane, i.e., indicating only the azimuth. In this field, there is already some research, e.g., by Geranazzo et al. [1, 2], who investigated human performance in auditory navigation on virtual maps.

As an example of navigation advice, we can use short "ping" tones [3] originating from the direction of interest instead of using full words such as "30 degrees left". Theoretically, the directional information can also be coded in features other than in the natural binaural sound direction. For example, we can generate a beep-tone with the frequency of this tone or its repetition rate coding the direction, but this would not be a natural code on which decoding the brain has trained over its whole life. Therefore, it is assumed that the decoding of the frequency modulation into information specifying a direction might be a relatively difficult task. The repetition frequency is rather used to indicate the object distance, which is difficult to encode in binaural format.

With this binaural technology, we can translate directional information from the radar scan in a natural way to the acoustic sense as we virtually position sound events in a virtual acoustic environment. More precisely, we build an augmented acoustic environment, where the augmented sounds are meant to deliver additional information that is not directly accessible by the user due to the loss of the visual sense. The blind or vision-impaired user relies more than others on his or her hearing sense to manage everyday tasks. Therefore, it is important not to impede this acoustic sense by our tool. Thus, we cannot use simple headphones that would occlude the ears. Instead, we use open-fitting hearing aid technology to supply acoustic information. Moreover, we aim to extract only necessary information from the radar data to create a sparse acoustic output since we want to avoid excessive distraction.

As an interface between the radar input and audio output, we rely on a sparse representation of the data that are of interest. From the point of view of the user, only the first obstacle in each direction is of interest since it is the only object the user would run into if he or she were to head in this direction. The obstacle behind the first obstacle in a given direction will never be reached with straight motion since the user will be stopped by the first obstacle before reaching the one behind it. Therefore, we compress the radar data to a low-dimensional representation, coding only the distance to the first obstacle in every considered direction, which we term the "radar distance profile" and is constrained to only one value per azimuth direction. For this constraint, there are various reasons. On the one hand, this approach immediately reduces the sound potentially delivered to the user to a more restrained level. On the other hand, enough information is provided to guarantee safe navigation when only the nearest obstacle per azimuth is indicated. The user has to avoid an obstacle no matter where it is located in terms of height. It is our task to warn the user about an obstacle in all cases, whether he or she will hit the obstacle with a foot, the torso or the head. In further implementations, there is a possibility to further indicate various heights to allow a more customized reaction depending on the type of obstacle: In the case of a low door, the user can simply cower, but in the case of a wall, the user knows that there is no way to go through the obstacle.

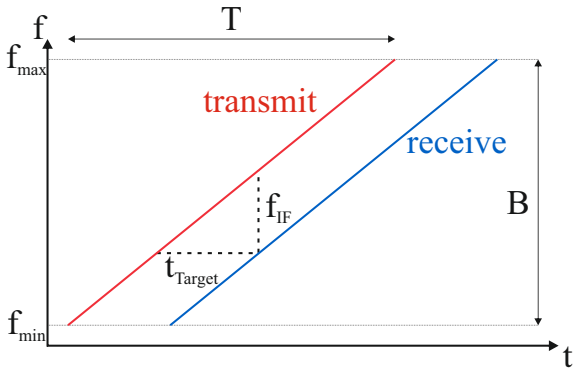


Figure 1: FMCW principle.

The remainder of this paper is organized according to the signal flow in the presented assistance tool. In Sec. 2 the utilized radar technology is presented, and the extraction of the radar distance profile is explained. Sec. 3 then presents the extraction of meaningful sound events from this distance profile, followed by Sec. 4, which gives a deeper insight into the technology of binaural rendering. Sec. 5 reports fundamental aspects of the actual implementation of the system.

2. RADAR SYSTEM

2.1. FCMW Radar Concept

Frequency-modulated continuous wave (FMCW) radar systems use a continuously radiated signal to spread the output power over a period of time, which makes FMCW radar systems easier and cheaper to manufacture than classical pulsed radar systems. The use of a linear frequency ramp enables precise target detection and localization [4, 5].

The FMCW principle is visualized in Fig. 1. A single-frequency radio frequency (RF) signal is generated and radiated by the radar device. This signal's frequency is raised with a linear frequency sweep over a frame of time T. The frequency range covered by this sweep is called the bandwidth B. This radiated RF signal is reflected by one or multiple targets, and the reflection is picked up by the sensor. By mixing the signal (multiplying the momentary signal amplitudes) that is being sent with the reflected one, a so-called difference signal of intermediate frequency f_{IF} is generated. This intermediate frequency relates directly to the distance between the sensor and the reflecting target via the bandwidth and sweep duration.

In a real scenario, the generated intermediate-frequency signal

is sampled with an analog-digital converter. On these data, a fast Fourier transform is performed to compute the amplitude of the signal's frequency components and therefore the target reflections. Therefore, every peak location in this frequency domain representation corresponds to a real target reflection. The distance to a target for a corresponding intermediate frequency f_{IF} is given by:

$$R = \frac{c \cdot t_{\mathrm{Target}}}{2} = \frac{T}{B} \cdot \frac{c \cdot f_{\mathrm{IF}}}{2} \tag{1}$$

where c is the speed of light in the relevant medium.

The maximum distance R_{max} from which a target can be detected is given by (1) with a maximum f_{IF} where the Shannon Nyquist theorem is valid with the sampling rate f_{S} of the analog-digital converter used:

$$R_{\mathrm{max}} = \frac{T}{B} \cdot \frac{c \cdot f_{\mathrm{S}}}{4} \tag{2}$$

The resolution ΔR of an FMCW radar system is solely defined by the bandwidth B used by the system. The resolution is defined as the minimum distance between two equally strong target reflections. If these two targets are closer than the given minimum distance, their peaks in the frequency-domain representation merge into one peak, which makes the two reflections indistinguishable. This distance is based on the 3 dB beamwidth of a target reflection in the frequency domain and is given by:

$$\Delta R = \frac{c \cdot A_{\mathrm{W}}}{2B} \tag{3}$$

where A_{W} is a widening factor based on the windowing function used before FFT computation.

2.2. 2D Scanner

The radar sensor used in this work is an 80 GHz FMCW radar sensor developed at Ruhr-Universität Bochum in collaboration with Fraunhofer FHR. This sensor is capable of extremely precise measurements [6, 7] with a very high bandwidth of up to 25.6 GHz. Its ellipsoid PTFE lens gives the sensor a 3 dB beamwidth of 5°.

To expand this linear measurement system, a rotating metallic mirror is used to steer the radar beam in the azimuthal direction. The mirror is angled 45° to the rotational axis as shown in Fig. 2. This operation deflects the radar beam orthogonally to the rotational axis. The deflecting mirror is rotated by a Trinamic PD42-1141 stepper motor with an integrated controller. For a stable and reproducible measurement with each revolution, a hardware-controlled trigger for the radar system was chosen. The trigger is based on a Hall sensor with an integrated comparator circuit that generates a trigger signal each time a magnet fixed to the rotating mirror enters a specific distance to the Hall sensor. For the compensation of movement of the scanner system between measurements, an inertial measurement unit (IMU) (Bosch BNO055) has been used to correct possible rotational changes. Because of the positioning of system components, the azimuthal range covered is reduced to 270°.

The scanner system is controlled by a Raspberry Pi 3B, which is connected to an external PC by a WiFi connection. The PC is used to control the system configuration parameters and to collect and process the scanner data. In the scanner system, radar and IMU data are collected and sent to the PC as UDP data packages via the Raspberry Pi and WiFi. This process is controlled with a Python script, and data types are conserved over the transmission.

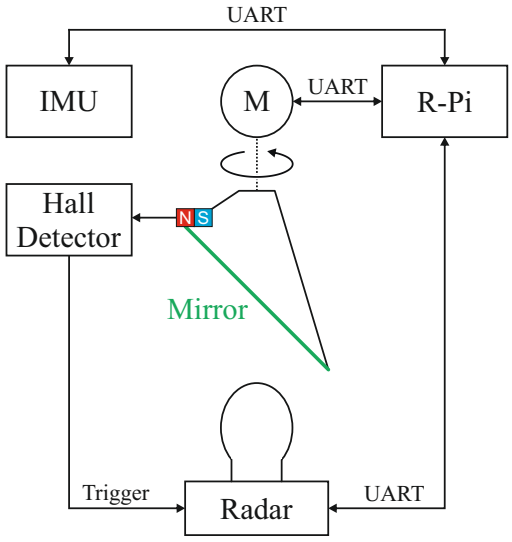


Figure 2: 2D Scanner block diagram.

2.3. Processing

The radar sensor sends its measurement data as a consecutive byte stream once per revolution. This data stream is converted to a two-dimensional data matrix in 16 bit integer format, where one dimension represents the frequency sweeps and therefore the azimuth dimension and the other the captured IF signal. On the IF dimension, a Hann window is applied to suppress sidelobes in the range. On the windowed data, an FFT is performed to represent the data in the frequency domain and therefore the range domain. Because of the high dynamic range of the signal in terms of amplitude, the data are then converted to their logarithmic magnitude.

On these data, target detection can be performed. A static threshold detection algorithm is not suitable for radar measurements because of the high dynamic range of targets as well as clutter. Dynamic threshold algorithms are used called CFAR (constant false alarm rate) [8]. The specific form of algorithm used is OS-CFAR (ordered statistics, Fig. 3), which has been shown to be very robust [9]. This algorithm is used to detect the closest target in each direction. If no target can be detected in a direction, the maximum detectable range of the radar measurement is assumed. This range profile data is then corrected by the IMU data to provide range profile data corresponding to the world coordinate system.

Examples of these radar measurements and the extracted radar distance profile (red) are shown in Fig. 4a. Additionally, we plotted the ground truth positions of the walls (black) into the figure. The measurement was performed during a stepwise walk through a hallway. The discrete positions where a measurement was taken are denoted by R1 to R8. Fig. 4a shows the measurement at positions R3, R6 and R8. In addition to the walls, we put two metal stands as additional obstacles, O1 and O2, into the hallway.

We see that the walls in most cases are well recognized, although they seem to consist of single points. Indeed, the walls in this environment are built in a lightweight construction, and predominantly the girders are seen by the radar system. Caused by the azimuthal spread of the radar beam, the girders are smeared in this direction. Nevertheless, the essential contours and obstacles can be recognized. The middle dataset at R6 gives an example of

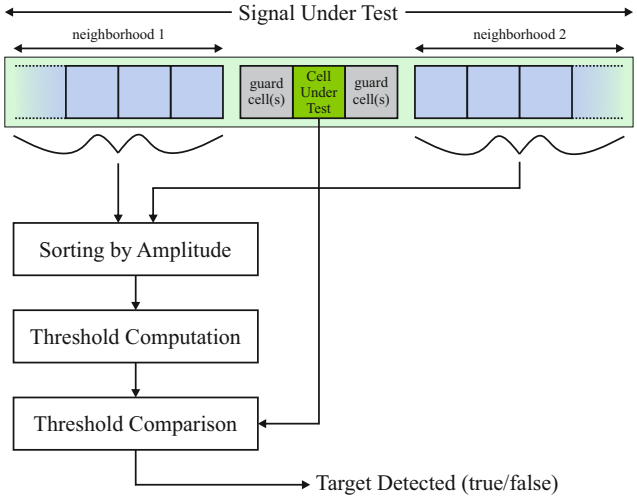


Figure 3: In OS-CFAR, each cell value in a signal under test is iteratively tested if it surpasses a threshold value computed specifically for this cell under test. A neighborhood window above and below the cell under test is selected, with optional guard cells between the cell and the neighborhood. The cell values from the neighborhood windows are then sorted by amplitude, and the value of a chosen rank, for example, the second largest value, is selected as the threshold. Upon this value, scaling factors and a bias can be applied to customize the sensitivity to the given scenario The value of the cell under test is then compared to the computed threshold.

misrecognition. Although there is a solid wall on the left side of the user, the radar sees the nearest obstacle far behind the wall. In this particular case, there was a poster wall made from plain metal at this position. This is a nearly perfect mirror for the radar beam, and therefore, the mirrored wall at the opposite site was recognized as the next obstacle in this direction.

3. FEATURE EXTRACTION

Now that we have the radar distance profile extracted from the measurement, in this section, we describe our approaches to select the information that should be sonified out of the radar distance profile. For the current implementation, we use simple approaches with the aim of a sparse output that can be easily interpreted by the user. In these modes of operation, we do not claim to deliver a comprehensive picture of the environment but only a sparse and helpful augmentation of the natural acoustic environment.

3.1. "Nearest Obstacle" Mode

The main purpose of the first implemented mode is to prevent the user from running into an obstacle. Therefore, the "nearest obstacle" is sonified but only in the case that the "nearest obstacle" is closer than a certain limit. Therefore, the sonification is silent if there is no need for the user to react. More precisely, the radar distance profile is weighted by the viewing direction, and then the weighted minimum is sought, and sound is created from that direction if the target is closer than this certain limit. The distance weighting is applied since both the average and the maximum velocity of human subjects are larger in the viewing direction than in the lateral directions. Therefore, an obstacle at a distance of one meter in front of the user poses more danger than an obstacle at a distance of one meter on the side of the user. The weighting of the distance is performed by

$$\widetilde{d}(\phi) = d(\phi) \cdot (1 + \alpha \sin(|\phi|)) \tag{4}$$

where $d(\phi)$ is the unweighted radar distance and $\widetilde{d}(\phi)$ is the weighted distance in the direction ϕ. Here, $\phi = 0$ is the viewing direction. Therefore, the distance in the viewing direction is not affected by the weighting. Distances on the side are suppressed from sonification since they are projected to longer distances up to a factor $(1 + \alpha)$. The strength of the suppression can be adjusted by the parameter α between 0 and $+\infty$, where 0 means no suppression and ∞ means complete suppression of side obstacles.

In addition to the direction of the nearest obstacle, which is coded by the natural binaural cue, the distance to the obstacle is coded by the repetition rate of the sound, a short "ping" tone as it is commonly used, for example, in parking assistants in cars. A faster repetition means a nearer obstacle and therefore more danger. This association is very natural since the more frequent sound is more imposing and therefore draws more attention as the obstacle grows closer and the danger becomes greater.

Another use case of this very simple mode is a movement along a wall. Sometimes there is the need to walk in parallel along a wall, e.g., if the user is walking down a long hallway. Looking parallel to the wall, the distance to the wall has its minimum at plus or minus 90 degrees. Therefore, we have to hold the orientation in a way that renders the sound laterally to move along the wall without hitting it. The distance to the wall can be easily controlled by the repetition rate of the sound. Although the weighting of the distance will change the effective distances in such a way that the distance at $\pm$ 90 degrees becomes larger, the minimum of the weighted distance will still be at $\pm$ 90 degrees in this scenario. This condition is assured by our choice of the weighting function. To prove this claim, let us assume that the wall is at the left of the user without loss of generality. Then, the minimum of $d(\phi)$ is at +90 degrees looking parallel along the wall. Let us denote this minimum distance by d_1, and hence, the weighted distance in the direction of +90 degrees is

$$\widetilde{d}_1 = d_1 \cdot (1 + \alpha). \tag{5}$$

The raw distance to the wall in any other direction is

$$d_2(\phi) = d_1 / \sin(\phi). \tag{6}$$

The weighted distance in the ϕ direction is then given by

$$\widetilde{d}_2(\phi) = d_2(\phi) \cdot (1 + \alpha \sin(\phi)) \tag{7}$$

$$= d_1 \frac{1 + \alpha \sin(\phi)}{\sin(\phi)} \tag{8}$$

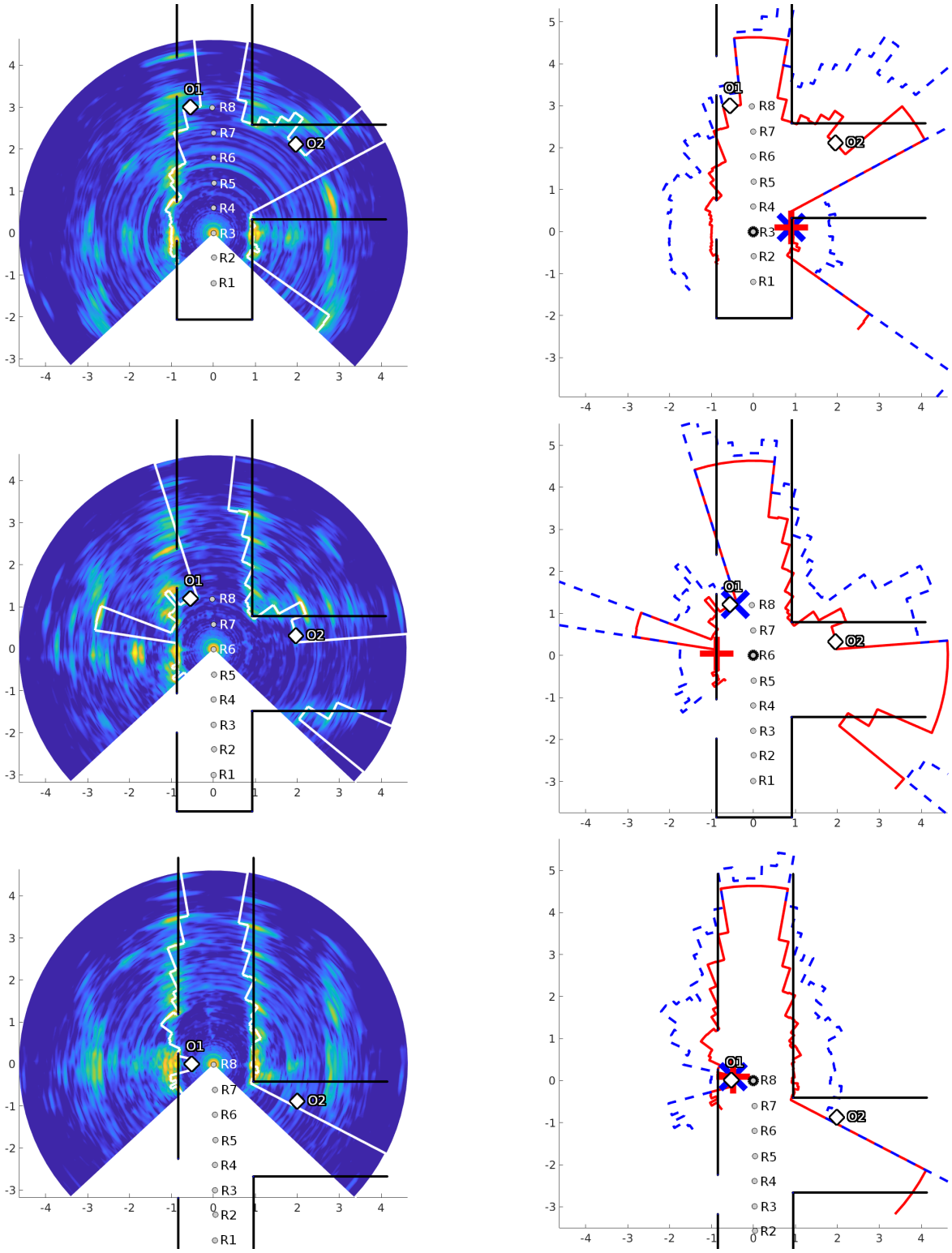
$$= \widetilde{d}_1 \frac{1 + \alpha \sin(\phi)}{(1 + \alpha) \sin(\phi)} \tag{9}$$

and the ratio

$$\frac{\widetilde{d}_2(\phi)}{\widetilde{d}_1} = \frac{1 + \alpha \sin(\phi)}{(1 + \alpha) \sin(\phi)} \tag{10}$$

is always larger or equal to 1 in the area $0^\circ < \phi < 90^\circ$.

Fig. 4b shows the weighted distance with $\alpha = 1$ together with the unweighted radar distance profile. The nearest obstacle based



(a) Radar data and extracted distance profile (white) with ground-truth walls (black) and obstacles (O1, O2)

(b) Weighted (blue dashed) and unweighted (red) distance profile, weighted (blue x) and unweighted (red +) "nearest obstacle"

Figure 4: Examples of measured and processed radar data and user positions from top to bottom: R3, R6, and R8

on the weighted distance profile is denoted by a blue **x**, while the nearest obstacle without weighting is denoted by a red **+**. In many cases, they coincide, but in the case where the user is located at R6, the weighted mode sonifies the obstacle O1, which stands in front of the user and probably represents more danger than the wall at the side where the unweighted distance has its minimum.

3.2. "Ahead Distance" Mode

As a second mode in the current state of development, we implemented a sonification process that gives the user the distance in the gaze direction. Thus, the user is able to scan the room on his or her own volition. The benefit of this "self-operated" mode compared to a comprehensive presentation of the whole environment is the following:

- The user is able to scan the environment at a speed that he or she is able to process the sound stimulation.
- The speed of scanning can vary depending on the complexity of the part of the environment being considered.
- The user can easily select which area is of interest.
- The sound is rather sparse and easy to interpret.
- The distance can be coded in the same way as in the sparse "Nearest Obstacle" mode.

Although the directional coding of the sound is not as essential as in the "Nearest Obstacle" case, we will also use the binaural rendering engine for this mode. In particular, we do so to assure that one single sound is locked to the virtual acoustic environment and does not move in the static environment when the user turns his or her head. This is one aspect of making the virtual acoustic augmentation more realistic.

Since the sound is continuously playing in this mode, the mode is meant only for active exploration purposes. It can be manually switched on by the user if he or she decides to actively look around. The "Nearest Obstacle" mode instead is intended to be an always-on tool that automatically turns silent if not needed but appears automatically if something of interest is happening, i.e., if an obstacle comes too close to the user.

4. BINAURAL RENDERING

In this section, we give a short introduction to binaural rendering using headphones. As an approximation of the sound propagation from a real sound source to the ear of a human, a linear time-invariant (LTI) system can be assumed as long as the sound source and the head have static positions. For slow motions of a walking listener, the approximation is quite precise. The LTI system from the sound source to the left and right ear is called the head-related transfer function (HRTF) or, in the time-domain, the head-related impulse response (HRIR). The influence of the room as reflections of the sound from walls is not part of the HRTF. In particular, the HRTF is often defined as the difference between the sound pressure at the ears and a hypothetical pointlike pressure receiver located at the center of the head [10, 11, 12]. As this is in general a noncausal system, because one ear is almost always closer to the source than the center of the head, an additional delay to the HRIR is appended in technical use to ensure causality. Below, we use the terms HRTF and HRIR for the causal form of the transfer function.

The HRTF can be measured from either a human being or a dummy head [12, 13, 14] or calculated from a model [15, 16, 17, 18]. The accuracy of the HRTF is essential to ensure localization precision. Obviously, the HRTF depends on the relative orientation of the head to the sound source. The impact of the distance can be neglected in the far field, except for an additional delay. Therefore, we need two coordinates to describe the orientation of the head relative to the sound source. Usually, a coordinate system of azimuth ϕ and elevation θ is used, as shown in Fig. 5. In this work, we constrain our discussion to the horizontal plane only with fixed elevation due to the given sonification task.

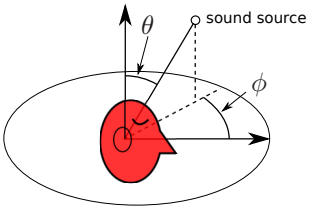


Figure 5: Head-related coordinate system

The HRTF can be utilized to create virtual positioned sound sources using headphones [12, 19, 20, 21, 22] or, as intended in our project, hearing aids, simulating the sound pressure of the virtual sounds at the ears. To position a sound source signal s virtually at a desired position of interest, we pass the signal through the corresponding transfer function to create the output signals y_l and y_r for the left and right ear, respectively. In the time domain, this operation can be performed by a convolution with the corresponding HRIR, i.e., $y_l = s * h_l(\phi, \theta)$ and $y_r = s * h_r(\phi, \theta)$, where ϕ and θ designate the position of the sound source relative to the head. In a real-time environment, the convolution is usually accomplished bufferwise and often performed by fast FFT convolution.

In many practical use cases and in our own use case, the relative sound source position to the head is dynamic, either because a source is moving or, as in our case, the head is rotating. For the latter, head tracking is needed [23] to unlock the virtual sound field from head rotation. Therefore, we no longer have an LTI system. Common practice, however, is to assume a piecewise, i.e., a bufferwise, LTI system. These buffers have to employ varying HRTF filters, and their outputs are crossfaded [24, 25]. This approach can lead to artifacts, especially if the effectively crossfaded HRTF filters differ massively from each other, e.g., when the head orientation changes strongly within one buffer (fast movements) or if the spatial resolution of the HRTF is too coarse in general. Additionally, the total system latency (TSL) for the compensation of head rotation is important for realistic perception [26, 27]. An approach to overcome some of these issues is to render the binaural sound samplewise as described, e.g., in [28, 29].

5. PROCESSING IMPLEMENTATION

As all parts of our assisting device are now known in theory, we present some important aspects of the end-to-end implementation. The main process of creating binaural audio from the radar profile is split into two parts. The first part, the actual sonification algorithm, analyses the radar profile and creates an acoustic scene description from it. In particular, the algorithm creates monaural data containing the sound source signal together with the desired angular position of that sound. The second part is the binaural renderer that produces binaural output from these elements.

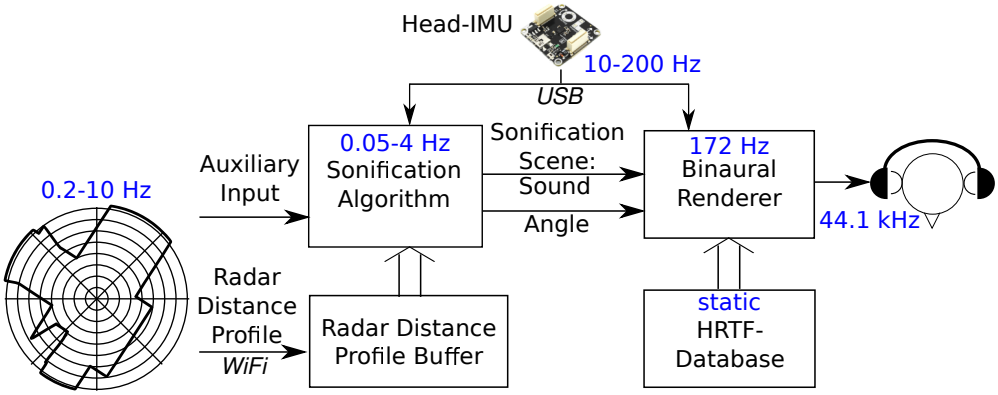


Figure 6: Sonification processing scheme

5.1. Multirate System

Since the sonification process involves many system components, we have to address the various time bases. Fig. 6 shows the blockwise implementation of the audio part and its interface with the other modules of the system. The whole system consists of asynchronous blocks. Whenever radar scanning is available in the form of a radar distance profile, this profile is delivered to the sonification engine. Depending on the type of radar sensor and the mode of operation, the update rate can be between 0.2 Hz and 10 Hz. The IMU delivers updates of the head orientation with an update frequency between 10 Hz and 200 Hz, depending on the IMU sensor. Both the radar profile input and the IMU input are not designed to deliver new data at a constant update rate; rather, the update rate can vary substantially. At the output, we need an audio output stream with a constant sampling frequency, in our case 44100 Hz. At least in our implementation on a Raspberry Pi computer, we apply blockwise audio processing since the platform does not offer enough performance for samplewise real-time processing of binaural audio. On the Raspberry Pi, we use a block length of 256 samples for the audio processing. Thus, an audio block is processed with a repetition rate of 172 Hz, which is fixed due to the fixed audio output sampling frequency. Instead, the sonification block can create acoustic scene descriptions with various durations. A sonification scene can last for a very short time, e.g., if we have a single beep tone that denotes only the nearest obstacle, or can last for very long time, e.g., if we have an algorithm that describes the whole environment. Hence, the sonification block has a variable rate between 0.05 Hz and 4 Hz. Also, in the current implementation, where we have the "Ahead Distance" mode, for instance, there are various sonification scene durations within one sonification mode. One sonification scene, in this case, consists of the "ping" tone with a constant length and a pause that varies in its length depending on the distance to the obstacle.

5.2. Dealing with Asynchronicity

To accommodate these various sampling frequencies in a single system, we use buffers in every connection between the blocks. The radar profile buffer plays a special role because it is the connection between the sonification and radar acquisition. This buffer is used to store the last received radar distance profile to deliver this profile to the subsequent function blocks at any arbitrary time and in the presence of a link failure. The connection between the IMU and binaural rendering engine is similar. The buffer between the sonification and binaural rendering is different since the buffer is a first-in-first-out (FIFO) buffer. Every sound sample is delivered only once through the buffer, and the buffer has two main tasks. On the one hand, it compensates for the difference between the scene length coming out of the sonification block and the audio buffer size used by the binaural renderer. On the other hand, it can deliver data on demand to the binaural renderer while the sonification block is calculating a new sonification scene. This feature is important since the analysis of the radar data and the creation of the sonification scene may take longer than the duration of one audio buffer. Every output scene of the sonification block should be output by the binaural rendering, and the sonification block creates a new sonification scene whenever the delivery of the previous scene to the binaural renderer has started and is paused after creating this scene until it is triggered again. The two blocks operate in independent time bases and are synchronized by this procedure. Indeed, it would be possible to run them completely asynchronously and let the sonification block produce as many scenes as it can. The buffer would then have the task to deliver only whole scenes to the binaural renderer and would always start with the most recent scene. This approach would prevent the buffer from underrunning if the processing of a sonification scene would take longer than the duration of the previous sonification scene. Nevertheless, we did not use this fully asynchronous approach since it would increase the processing costs to a maximum. In our synchronous approach, we simply add silence in the case of a buffer underrun.

6. CONCLUSION

In this paper, we presented a concept of an assisting device for vision-impaired people for navigation and orientation. The device is based on radar input and binaural audio output and was implemented as a research study. The quality of the radar data acquisition was shown in examples. We demonstrated two sonification modes, representing the essence of first subjective preferences from blind and seeing people, who demand a simple interpretability and a sparse sound for an easy-to-access utility. The latter aspect is addressed by restricting the acoustic indications to just the horizontal plane in both presented modes and by paying particular attention to the walking direction of the user in the "nearest obstacle" mode using the weighted radar distance profile. The device is

meant to provide orientation cues additional to, e.g., a white cane and support the user in everyday orientation and navigation tasks. Further end-to-end investigations of the presented system have to be performed with various users to evaluate the helpfulness in realistic scenarios. Depending on these results, we can further tune the modes and algorithms to deliver a more satisfying experience.

7. ACKNOWLEDGMENT

This work is supported by the European Regional Development Fund Nr. EFRE-0800372 NRW grant, LS-1-1-044d as part of the project "Ravis 3D".

8. REFERENCES

[1] M. Geronazzo, A. Bedin, L. Brayda, C. Campus, and F. Avanzini, "Interactive spatial sonification for non-visual exploration of virtual maps," *Int. J. of Human-Computer Studies*, vol. 85, pp. 4–15, 2016.

[2] M. Geronazzo, F. Avanzini, and F. Fontana, "Auditory navigation with a tubular acoustic model for interactive distance cues and personalized head-related transfer functions," *J. on Multimodal User Interfaces*, vol. 10, no. 3, pp. 273–284, Sep 2016.

[3] C. Urbanietz and G. Enzner, "Binaural Rendering for Sound Navigation and Orientation," in *2018 IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*, Mar. 2018, pp. 1–5.

[4] M. A. Richards, *Fundamentals of Radar Signal Processing*. McGraw-Hill Education, 2014.

[5] M. I. Skolnik, *Radar Handbook*. McGraw-Hill Education, 1990.

[6] N. Pohl, T. Jaeschke, and K. Aufinger, "An Ultra-Wideband 80 GHz FMCW Radar System Using a SiGe Bipolar Transceiver Chip Stabilized by a Fractional-N PLL Synthesizer," in *2012 IEEE Transactions on Microwave Theory and Techniques*, vol. 3, Mar. 2012, pp. 757–765.

[7] N. Pohl, T. Jaeschke, S. Scherr, S. Ayhan, M. Pauli, T. Zwick, and T. Musch, " Radar measurements with micrometer accuracy and nanometer stability using an ultra-wideband 80 GHz radar system," in *2013 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet)*, Jan. 2013.

[8] H. M. Finn and R. S. Johnson, "Adaptive detection mode with threshold control as a function of spacially sampled clutter-level estimates," in *RCA Review*, vol. 29, Sept. 1968, pp. 141–464.

[9] H. Rohling, "Ordered statistic CFAR technique - an overview," in *Radar Symposium (IRS), 2011 Proceedings International*, vol. 7, Sept. 2011, pp. 631–638.

[10] V. Pulkki and M. Karjalainen, *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. Wiley, 2015.

[11] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display: Second Edition*. J. Ross Publishing, July 2013.

[12] J. Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*, rev. ed. Cambridge: MIT Press, 1997.

[13] A. Andreopoulou, D. R. Begault, and B. F. G. Katz, "Inter-laboratory round robin HRTF measurement comparison," *IEEE J. of Selected Topics in Signal Process.*, vol. 9, no. 5, pp. 895–906, Aug 2015.

[14] G. Enzner, C. Antweiler, and S. Spors, "Trends in acquisition of individual head-related transfer functions," in *The Technology of Binaural Listening*, J. Blauert, Ed. Springer, 2013, pp. 57–92.

[15] K. Young, T. Tew, and G. Kearney, "Boundary element method modelling of KEMAR for binaural rendering: Mesh production and validation," in *Interactive Audio Systems Symposium*, September 2016.

[16] L. Bonacina, A. Canalini, F. Antonacci, M. Marcon, A. Sarti, and S. Tubaro, "A low-cost solution to 3D pinna modeling for HRTF prediction," in *IEEE Int. Conf. Acoust., Speech and Signal Process.*, March 2016, pp. 301–305.

[17] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe, "Creating the Sydney York morphological and acoustic recordings of ears database," *IEEE Trans. on Multimedia*, vol. 16, no. 1, pp. 37–46, Jan 2014.

[18] F. Brinkmann, A. Lindau, S. Weinzierl, S. v. d. Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," *J. Audio Eng. Soc*, vol. 65, no. 10, pp. 841–848, 2017.

[19] D. R. Begault, *3D Sound for Virtual Reality and Multimedia*. San Diego, CA, USA: Academic Press Professional, Inc., 1994.

[20] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones: Integration of signal processing techniques," *IEEE Sig. Process. Mag.*, vol. 32, no. 2, pp. 100–113, March 2015.

[21] F. Rumsey, *Spatial Audio*. Focal Press, 2001.

[22] S. Carlile, *Virtual Auditory Space: Generation and Applications*. Landes Bioscience, 1996.

[23] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *IEEE Signal Processing Mag.*, vol. 28, no. 1, pp. 33–42, Jan 2011.

[24] A. Kudo, H. Hokari, and S. Shimada, "A study on switching of the transfer functions focusing on sound quality," *Acoustical Sci. and Techn.*, vol. 26, no. 3, pp. 267–278, 2005.

[25] M. Vorländer, *Auralization (RWTHedition)*. Springer, 2007.

[26] D. S. Brungart, B. D. Simpson, and A. J. Kordik, "The detectability of headtracker latency in virtual audio displays," in *Int. Conf. Auditory Display (ICAD)*, 2005, pp. 37–42.

[27] A. Lindau, "The perception of system latency in dynamic binaural synthesis," *Proc. of 35th DAGA*, pp. 1063–1066, 2009.

[28] J. W. Scarpaci, H. S. Solburn, and J. A. White, "A system for real-time virtual auditory space," in *Int. Conf. on Auditory Display*, July 2005.

[29] C. Urbanietz and G. Enzner, "Binaural Rendering of Dynamic Head and Sound Source Orientation Using High-Resolution HRTF and Retarded Time," in *IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Apr. 2018, pp. 566–570.

DIRECT SEGMENTED SONIFICATION OF CHARACTERISTIC FEATURES OF THE DATA DOMAIN

Paul Vickers

Dept. of Computer & Information Sciences
Northumbria University
Newcastle-upon-Tyne, NE1 8ST, UK
paul.vickers@northumbria.ac.uk

Robert Höldrich

Institute of Electronic Music and Acoustics
University of Music and Performing Arts
Inffeldgasse 10/3, 8010 Graz, Austria
robert.hoeldrich@kug.ac.at

ABSTRACT

Like audification, auditory graphs maintain the temporal relationships of data while using parameter mappings to represent the ordinate values. Such direct approaches have the advantage of presenting the data stream 'as is' without the imposed interpretations or accentuation of particular features found in indirect approaches. However, datasets can often be subdivided into short non-overlapping variable length segments that each encapsulate a discrete unit of domain-specific significant information and current direct approaches cannot represent these. We present Direct Segmented Sonification (DSSon) for highlighting the segments' data distributions as individual sonic events. Using domain knowledge DSSon presents segments as discrete auditory gestalts while retaining the overall temporal regime and relationships of the dataset. The method's structural decoupling from the sound stream's formation means playback speed is independent of the individual sonic event durations, thereby offering highly flexible time compression/stretching to allow zooming into or out of the data. DSSon displays high directness, letting the data 'speak' for themselves.

1. INTRODUCTION

In parameter mapping sonification the data values drive the parameters of an audio signal. In contrast, audification involves transposing the frequencies of the data to the human-audible range and occasional filtering to remove unwanted linear distortions (and in rare cases dynamic range compression to remove very large level variations). Therefore, the process maintains a tighter relationship with the data than other auditory display processes which generally rely on mappings to effect the auditory display. These mappings can be low level [1] or more metaphorical [2].

A sonification's directness is a measure of the arbitrariness (in relation to the underlying data) of the mapping [3]. A method exhibiting maximal directness will derive the sound directly from the data (e.g., through the use of direct data-to-sound translations). Low directness arises from more symbolic, metaphoric, or interpretative mappings. Audification is a more direct form of auditory display, the audio being generated entirely by the data.

The method proposed here pursues directness as a design goal so that, as far as possible, the data are allowed to 'speak' for themselves. Any metaphors then arise as contingent properties of the sonification rather than being imposed by the designer. For example, the characteristic sound caused by accentuating data range excursions in §5.3 below assumes its own sonic identity and metaphorical labels may be assigned by (and will vary depending on) the listener. Thus, users may start identifying regions of interest in the data by describing the characteristic sounds they hear. Space does not allow a full treatment of the topic of directness, so interested readers are directed to Vickers's forthcoming chapter that deals with this and related issues in more depth [4].

The proposed method follows a diretct sonification strategy which conserves fundamental properties of (pure) audification, notably the compact temporal support and some aspects of the precise temporal structure of a data set.

1.1. Leveraging the Directness of Audification

Audification a physical process strictly conserves the temporal regime of the source signal and so contains high-frequency components when rapid transients occur in the data. This is advantageous because such transients, which often correspond to points of interest in the data, are also significant features of the audio signal which the human auditory system relies on to identify real-world sounds. Hence, they can be a perceptually salient basis for auditory data exploration [5].

When the data is sampled from a band-limited physical process the audification signal has a one-to-one relationship with the data. In fact, the mapping is, in principle, bijective and fully reversible (at least while the data remains in the digital domain prior to any D/A conversion.) However, even such direct representations can contain misleading features because of the band-limited interpolation of the reconstruction filter of the D/A converter leading to extreme data values being elevated in the audification.

The ideal audification signal has auditory gestalts within time and frequency ranges that are clearly perceptible to a listener [5]. Take a data stream dominated by low frequencies with transients occurring within a range of 1 k data points and with an aperiodic interval of approximately 10 k data points. At a playback rate of 44.1 kHz roughly four of these events will occur each second which is comparable to the number of syllables per second in spoken English and so is suitable for listeners (see Wood [6] for a view of the information aspects of tempo). However, each transient event's duration will be approximately 22 ms appearing as a band-limited impulse with a cut-off frequency at around 50 Hz,

https://doi.org/10.21785/icad2019.043

which is below the most sensitive range of the human auditory system. If the playback rate were raised by, say, a factor of 10–20, the individual impulses would be shifted to a more perceptible frequency range, but at the cost of an indiscernible temporal structure of the impulse series. Thus, pure audification is a trade-off between the macroscopic time scale and the frequency range of the relevant information.

2. DIRECT SEGMENTED SONIFICATION (DSSon)

The DSSon process is regarded as a mapping operation between data domain and sound domain (see Rohrhuber [7]). Because the sonification time domain will often be different from the data time domain (e.g., choosing to listen to a 100 s data set over a period of only 10 s) Rohrhuber proposed superscribing sonification domain variables with a ring to distinguish them from data domain variables, thereby enabling the construction of unambiguous mixed domain expressions. In this scheme the sonification operator $\mathring{\mathcal{S}}$ maps from the given data space $\mathbb{D}$ to the sound signal space $\mathring{\mathbb{Y}}$ as $\mathring{\mathcal{S}} : \mathbb{D} \mapsto \mathring{\mathbb{Y}}$. The relation is more explicit at the level of the variables [8]:

$$\mathring{\mathcal{S}} : x(t) \mapsto \mathring{y}\left(\mathring{t}, x(t); \mathring{p}\right) \tag{1}$$

The sonification signal $\mathring{y}$ depends on $\mathring{t}$ (sonification time), because sound is a temporal phenomenon, on the data $x(t)$ to be sonified which itself is assumed to depend on a data domain time t, and the parameters $\mathring{p}$ of the sonification method which determine how the sonification sounds.

2.1. Sonification Variables

The proposed sonification method uses the variables shown in Table 1 (in Appendix). The sonification parameter set is then given as $\mathring{P} = \{\mathring{\kappa}, \mathring{\Delta}, \mathring{f}_{\text{ref}}, \mathring{\alpha}, \mathring{\beta}, \mathring{\phi}, \mathring{\epsilon}, \mathring{g}(..), \mathring{\gamma}, \mathring{H}\langle..\rangle\}$ with any appropriate subset $\mathring{p} \subseteq \mathring{P}$ being used in the models described below. The meanings of these variables are given in the sections that follow. To distinguish sonification time from data domain time, sonification time variables are given as $\mathring{t}, \mathring{t}_i, \mathring{T}$ and data domain time variables as t, t_i, T.

2.2. General Framework of DSSon

DSSon relies on the assumption that a one-dimensional time-varying data stream, $x(t)$, can be subdivided into short non-overlapping segments of generally different length where each segment contains a consistent portion of application-dependent significant information. Thus, identification of the appropriate cutting points is crucial. For example, if one is interested in the short-term fluctuation of a stock price, the crossing points of the actual stock price with a moving average might be a good choice. We consider a data stream as a time varying signal $x(t)$ expressed as a sequence of sampled values $x(n)$ at a sampling rate f_s. The duration of the data stream is T seconds, hence the sequence $x(n)$ consists of $N = T \times f_\text{s}$ samples. Assuming that the DSSon of the data should last for approximately $\mathring{T}$ seconds (the reason for the duration being approximate is explained below), a time compression factor $\mathring{\kappa}$ is defined by $\mathring{\kappa} = T/\mathring{T}$.

As a first step, the cutting points t_i (the borders between segments $x_i(t)$) have to be determined depending on the application and the specific properties of the data. As a simple example, consider a broadband AC signal. In this case the zero crossing points are a reasonable choice. If the signal contains DC or strong low-frequency components (as is the case with stock prices and the data used in §4) some preprocessing might be necessary. For instance, the trend signal $x_{\text{trend}}(t)$ calculated by a moving average filter can be subtracted from the original data yielding a signal $x_{\text{AC}}(t) = x(t) - x_{\text{trend}}(t)$ which exhibits numerous zero crossings.

Assuming the first cutting point is at $t_0 = 0$ and the last one is at $t_M = T$, a sequence of M segments $x_i(t)$ (or $x_{i,\text{AC}}(t)$ if the low frequency mean or DC component has been removed through preprocessing) is obtained by (2) (see Appendix). Thus, the actual duration of each segment is given by $T_i = t_i - t_{i-1}$. Each data segment $x_i(t)$ is to be sonified as an individual sonic event $\mathring{y}_i(\mathring{t})$ depending on the parameters $\mathring{p}$ of the sonification method at hand and is superimposed to form the final sonification $\mathring{y}(\mathring{t})$. For the sake of simplicity, we skip the explicit dependence of the sonic event $\mathring{y}_i(\mathring{t})$ on the data segment $x_i(t)$ and the sonification parameters $\mathring{p}$ in (3). Note that the individual sonic events $\mathring{y}_i$ might be longer or shorter than the duration of the respective data segment $T_i = t_i - t_{i-1}$ depending on the specific sonification method and parameters. Therefore, the actual length $\mathring{T}$ of $\mathring{y}$ is only approximately equal to the data duration divided by the compression factor: $\mathring{T} \approx T/\mathring{\kappa}$.

The DSSon approach conserves the overall temporal structure of the data as long as the cutting points are chosen appropriately, that is, they are meaningful within the context of the data domain. Since the sonification length of the individual segments is not predetermined by this very general formulation, the resulting auditory display can be adjusted either to focus on the rhythmical structure of the segments' temporal distribution (such as by choosing very short and transient sonic events for each segment and thereby presenting, essentially, a sequence of clicks) or to zoom into the specific data evolution of each segment (e.g., by choosing long sonic events with time-varying properties according to the segment's data values). Note that the latter approach yields a temporal overlap of sonic events of adjacent segments and hence might confound the auditory gestalts originating from the individual segments. In any case, the appropriate choice of the sonification method for the individual segments is crucial for the quality of the DSSon. In the following section, a simple method for segment sonification which is derived from auditory graphing is presented.

3. MODIFIED AUDITORY GRAPHS FOR SONIFYING INDIVIDUAL SEGMENTS

Auditory graphs have been a part of the standard repertoire of auditory display research since its beginning. At its simplest, an auditory graph represents the ordinate value of a data series as the time-varying frequency of a sinusoid with (usually) constant amplitude [9]. An obvious benefit is the straightforward analogy to visual graphs, which makes them readily understandable, at least for sighted users. Flowers [9] recommended using distinct timbres in order to minimize stream confusions and unwanted perceptual grouping. Since auditory graphs usually encode data values as pitch or (fundamental) frequency, harmonic complexes with a small number of partials (around 6–8) and amplitudes in inverse proportion to partial order are recommended instead of pure sinusoids because of the improved pitch salience they are able to produce. Nevertheless, the resulting timbre should be time-invariant to guide the listener's attention to the pitch contour and not obscure the data representation by arbitrary timbral fluctuations. More complex timbres run the risk of evoking categorical associations

with real-world sound that might change at more or less arbitrary data values and therefore confound the intended perceptual continuum of the frequency or pitch range representing the important aspects of the data. If several auditory graphs are to be presented simultaneously spectral overlap between adjacent graphs should be avoided, therefore pure sinusoids might be the better choice in this instance.

To achieve the intended directness, not only must the overall temporal relationship of the segmentation pattern be preserved (as is ensured by the general framework in §2.2), but the sonic events resulting from the individual segments must also display the segments' data evolution as directly as possible. Therefore, a modified auditory graph is proposed as the specific method of segment sonification in DSSon with each segment being treated as an individual graph. We assume segments are derived from zero crossing points (either due to the inherent AC characteristics of the data or after removing the signal average) and exploit the property that each segment starts and ends with data values of negligible magnitude. To accentuate strong deviations from a chosen baseline (such as the average), amplitude modulation derived from the segment's data complements the time-varying pitch progression of the basic auditory graph. Thus, the general form of the sonification signal $\mathring{y}_i(\mathring{t})$ is given in (4) where $a_i(\mathring{t})$ is the amplitude modulator, f_{ref} is the base frequency for the pitch range of the sonification, and $b_i(\mathring{t})$ is a pitch modulator.[1] To include the (previously removed) short-term average value as an overall pitch trend, we explicitly take into account both the mean-free segment $x_{i,\text{AC}}(t)$ and the trend signal at the segment's starting point $x_{\text{trend}}\,(t_{i-1})$ for pitch modulation.

In (5), the magnitude of the segment's data values is used as amplitude modulation and the dilation parameter $\mathring{\Delta}$ determines the length of the sonic event $\mathring{T}_i$ in relation to the duration of the data segment T_i. If $\mathring{\Delta} = \mathring{\kappa}$, adjacent sonic events do not overlap since $\mathring{T}_i = T_i/\mathring{\kappa}$, whereas $\mathring{\Delta} \leqslant \mathring{\kappa}$ results in overlapping events.

Of course, both pitch and amplitude modulation can be parameterized in various ways. For example, if mainly peak or strong deviations from the mean are to be displayed, a power law distortion $\mathring{\phi}$ can be applied to the amplitude modulator a (see 6). If only deviations exceeding a threshold $\mathring{\epsilon}$ around the mean are to be sonified, then a magnitude offset followed by half-wave rectification might be included in the amplitude modulator (7,8). On the other hand, the relative importance of the trend signal x_{trend} and the actual data progression of the segment can be adjusted via non-negative parameters $\mathring{\alpha}$ and $\mathring{\beta}$ (9).

If the stream of segments with positive deviation from the trend should be discriminated from the stream of negative segments, two different reference frequencies f^{+}_{ref} and f^{-}_{ref} could be used. From the above, the general parameterized form of DSSon is given as (10).

[1] In order to allow specific control of timbre, an additional timbre operator $\mathring{H}$ which acts on the sine function has to be considered in the model:

$$\mathring{y}_i(\mathring{t}) = a_i(\mathring{t})\mathring{H}\left\langle \sin\left(2\pi \int_0^{\mathring{t}} \mathring{f}_{\text{ref}} \cdot 2^{b_i(\mathring{t}')} \cdot \mathrm{d}\mathring{t}'\right)\right\rangle$$

$\mathring{H}$ might be implemented as, for instance, waveshaping utilizing Chebyshev polynomials or any kind of additive synthesis. The operator properties itself will depend on the data to be sonified, i.e. $\mathring{H}\langle\sin(..); x_i\rangle$. However, in the case of the modified auditory graph the resulting sonic events consist only of amplitude and pitch modulated sinusoids, hence $\mathring{H}$ can be regarded as the identity function, $\mathring{H}\langle\sin(..); x_i\rangle = \sin(..)$, and will be omitted in the following for the sake of simplicity.

3.1. Modulation of Segment Duration

To relate the duration of sonification segments to some property of the data we can use $\mathring{\Delta}$ not as a constant, but as a function of the segment's data, $\mathring{\Delta}_i$. For instance, if highly peaked segments should be displayed as longer sonic events to display the data distribution in more detail, a monotonically decreasing, concave function of the segment's mean (or other property such as rms, power) or area (or energy) is more suitable for $\mathring{\Delta}_i$.

3.2. Decaying Envelope as Amplitude Modulator

In order to emphasize the rhythmical patterns induced by the temporal distribution of the cutting points, a sharp attack of the individual sonic events is needed. This can be achieved by replacing the amplitude modulator $\left|x_i(\mathring{\Delta}\cdot\mathring{t})\right|$ or the variants in (6) and (7) by an appropriate envelope, for example, $\mathring{g}_i \cdot e^{-\mathring{t}/\mathring{\gamma}}$ or $\mathring{g}_i \cdot \mathring{t} \cdot e^{-\mathring{t}/\mathring{\gamma}}$, where $\mathring{\gamma}$ is the envelope's decay parameter and the gain factor $\mathring{g}_i$ is determined by a specific function of the segment's data values, $\mathring{g}_i = \mathring{g}(x_i)$, e.g., the mean, rms, area, power, or energy of the segment.

4. APPLYING DSSon TO BIOMECHANICAL DATA

We applied DSSon to biomechanical signal data taken from the Functional Readaptive Exercise Device (FRED), a machine designed for physiotherapeutic use to help patients with low back pain [10]. FRED is a modified cross-trainer but which offers minimal resistance (Fig. 1). This creates a situation in which the user has an unstable base of support: when the front foot comes to the forward-most position in its elliptical path, gravity then pulls the foot downward requiring the user to apply compensatory balancing force with the rear foot to control the descent. The goal is to operate the machine with an upright posture in a smooth, controlled manner with minimal variability in movement speed [10].

A rotary encoder in the drive wheel generates a pulse stream representing the instantaneous angular velocity of the wheel. This pulse stream is sampled at 4 kHz into LabChart [11] and is converted to frequency values (i.e., revolutions per second) for ease of display for the user. The stream is then smoothed using a triangular Bartlett filter to remove the steps in the data. The smoothed stream is presented to the user via LabChart (with a zoom level of 50:1) as a means of feedback to help them control their performance (Fig. 2). It has been determined that with the machine in its default configuration (Fig. 1), operating it within a frequency range of 0.2 Hz $\leqslant f \leqslant$ 0.6 Hz results in therapeutic benefit leading to recruitment of the key spinal and abdominal muscles *lumbar multifidus* (LM) and *transversus abdominis* (TrA), and with the biomechanical optimum for maximum benefit being achieved at $f = 0.4$ Hz [10]. At this frequency a complete rotation of the footplates takes 2.5 s, thus requiring a slow and steady pace.

The white area in Fig. 2 shows the user when they are performing inside the required range with the shaded areas denoting frequencies above and below the required range. Fig. 2 shows the user is maintaining a good pace until 27.2 s at which point they slow down dramatically, coming to a brief halt (27.65 s) followed by a sharp corrective acceleration which takes the frequency up to 0.8 Hz followed by a compensatory attempt to slow down, followed by another sharp acceleration, with normal performance being re-attained at around 29.2 s.

Figure 1: The height (difficulty) of FRED's walking path is increased by moving the rear end of the stride rail (A) through the slot (B) towards the edge of the wheel. FRED has five such settings and is shown here in its default (lowest) configuration.

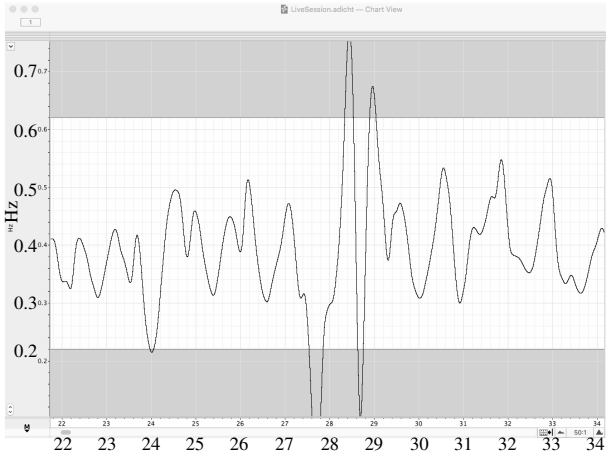


Figure 2: Screen capture of the live scrolling window view in LabChart as seen by a FRED user during an exercise session (axis labels have been superimposed here to aid the reader).

4.1. Features of Interest

During a post-hoc review of performance, the physiotherapist is interested in identifying a number of discrete features in the data sets. The main performance goal is to maintain a walking pace of 0.2 Hz $\leqslant f \leqslant$ 0.6 Hz. While the patient needs to be aware of excursions outside this range during exercise, for the therapist all excursions above 0.6 Hz and long excursions below 0.2 Hz are of interest. If the frequency exceeds 0.6 Hz (Fig. 2) this indicates a loss of control — the machine is running away with the user. However, because it takes a great deal of muscle control to operate the machine slowly, if the frequency momentarily drops below 0.2 Hz and then goes back in range this is of less interest to the therapist as it is still evidence of control — it is a controlled recovery (Fig. 3(b)). But if it drops below 0.2 Hz for an extended period of time (typically half-a-second or more) then this also indicates a lack of control as motion is coming to a stop.

The target range of 0.2 Hz $\leqslant f \leqslant$ 0.6 Hz means that users can demonstrate variability in their average speed while still maintaining acceptable performance. Therefore, for each user, the physiotherapist will additionally determine a maximum deviation from the individual mean as a target range based upon their assessment of the user's current ability and any physical characteristics that might impact upon how well they are able to use FRED. For example, a beginner with reasonable control might be expected to achieve a standard target deviation of 0.15 Hz while someone who is able to keep within the range 0.35 Hz $\leqslant f \leqslant$ 0.45 Hz would have a target deviation of 0.05 Hz. Once the therapist has determined a user's target deviation it is interesting to know at what points they are failing to maintain it.

If someone were able to operate the machine perfectly there would be no variation in their speed and the plot would show a flat line. Therefore, the smoother the plot the less the user's pace is varying. When a user starts to master the required walking technique they begin to exhibit what are known as "flat tops". A flat top is a region of activity lasting approximately 0.5 s or more in which the variation in speed is so small that the curve starts to flatten out. Flat tops typically occur during the portion of a walking cycle after the rear foot has come up from the bottom of the elliptical path and before the front foot descends again. Fig. 3(a) shows a double flat top. At around the 53 s mark the small peak indicates where the user's rear foot has ascended from the bottom of the elliptical path. This is followed by a period of relatively flat speed variation lasting just under 1 s. At around 54.2 s the front foot descends and then another flat top of $\approx$0.7 s occurs.

Because these features require zooming in to see clearly it becomes time consuming to zoom-and-scroll through many data files, so DSSon was applied to FRED data sets to see how well these features could be heard. After discussions with physiotherapists from Northumbria University's Aerospace Medicine and Rehabilitation lab in which FRED is being further developed, the features to be represented were:

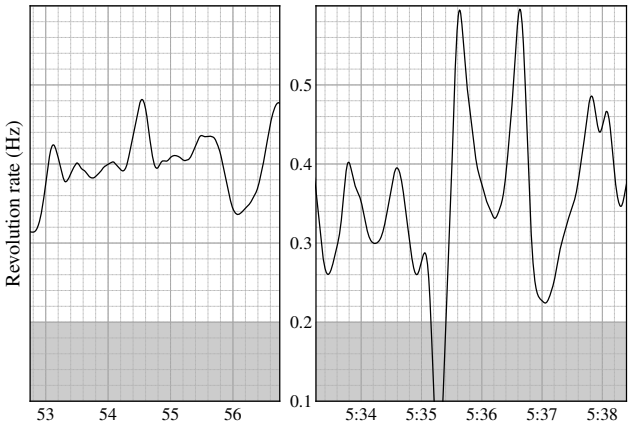
1. Any excursions above 0.6 Hz.
2. Long excursions below 0.2 Hz.
3. Periods outside the user's target deviation range.
4. 'Flat tops' lasting $\approx$0.5 s or longer.

The preprocessing stage involved audifying FRED data streams by simply converting each data point to a signed 16-bit integer and storing the result in a PCM-encoded digital audio file. Because the revolution rate does not exceed 2 Hz (which would be very fast walking) the signal spectrum caused by the speed fluctuations occurring during a full revolution is band limited below 15–20Hz. Therefore, to keep the file sizes small the data extracted from LabChart were first downsampled to $f_s = 100$ Hz prior to audification.

Thus, the time series signal, $x(t)$ in the DSSon method was provided by these audio files. The DSSon method was implemented in a series of MATLAB (for sonification) and Python (preprocessing) scripts (see the project repository [12]).

5. DSSon MODELS FOR FRED SIGNALS

In this section we describe three DSSon models that were applied to FRED data that emphasize the features of interest identified above to varying degrees resulting in differing auditory saliency. DSSon for FRED data is mainly intended to provide an auditory display of users' performance that enables the physiotherapist to conduct a quick analysis during post-hoc review. The DS-



(a) Double "flat top" (b) Controlled recovery

Figure 3: Strong and weak performance. In (a) the user attains two periods of very small velocity deviation. In (b) the velocity drops below target but is quickly recovered back into the target range.

Son parameters might also be individually adjusted by the therapist during the review session in order to concentrate on specific data features. Consequently, it is impractical to evaluate the DSSon display through extensive listening tests based on specific task completion performance and statistical analysis. This kind of evaluation procedure is planned for future work on other application fields. Here, DSSon's properties (benefits and limitations) are demonstrated by comparing data excerpts containing specific features of interest and the resulting DSSon display. Audio files, demonstrating the system output, together with the corresponding data sets used to generate them, can be found in the `examples` and `data` directories in project repository [12] and are listed in Table 2 (see Appendix).

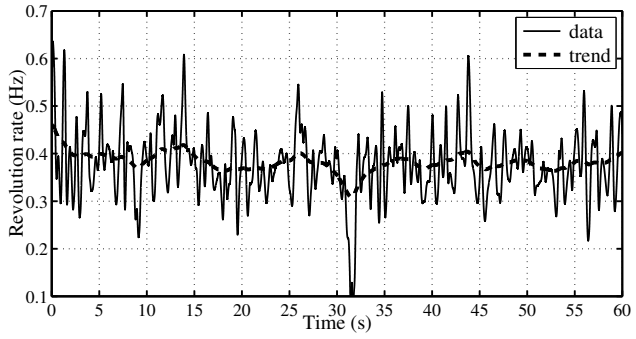


Figure 4: Data stream and trend signal (weighting factor $w = 0.2$) of FRED exercise sessions of user A at the beginning of training (audio file 1).

The first step in DSSon is signal segmentation. For FRED data, the main feature of interest is the deviation of the instantaneous revolution rate from the fixed target value, $x_{\text{target}} = 0.4$ Hz (the biomechanical optimum from above). Hence, an obvious choice for segmentation is to cut the data stream at its crossing points with this target value, that is, extract segments with positive and negative deviation from x_{target}. However, as far as a user is able to maintain a steady revolution rate, even slightly deviating from 0.4 Hz, or shows a slowly varying average revolution rate exhibiting only small excursions, he/she shows sufficient muscle control and therefore gains therapeutic benefit. To account for this fact, we did not use the fixed target value of 0.4 Hz to determine the segments' start and end points, but calculated a weighted mean of the target and the moving average of the data stream, $x_{\text{MA}}(t)$, to obtain the trend signal: $x_{\text{trend}}(t) = w \cdot x_{\text{target}} + (1 - w) \cdot x_{\text{MA}}(t)$.

The data stream and the trend signal (weighting factor $w = 0.2$) of two exercise sessions of the same user are shown in Figs. 4 and 5. The first data stream was recorded in the second week of a six-week training period, and the second was recorded four months after the end of the training period. The data segments are determined utilizing the zero-crossing points of the trend-free signal: $x_i(t) = x(t - t_{i-1}) - x_{\text{trend}}(t - t_{i-1})$ for $t_{i-1} \leqslant t \leqslant t_i$, and $x_i(t) = 0$ otherwise.

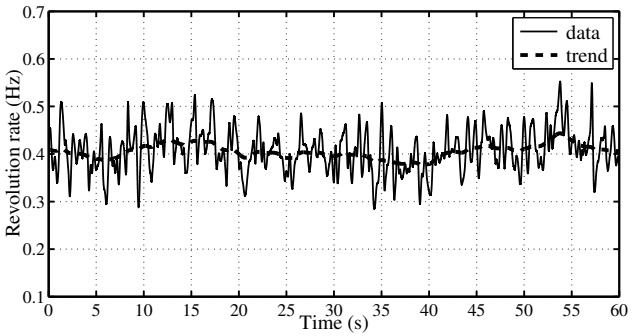


Figure 5: Data stream and trend signal (weighting factor $w = 0.2$) of user A, four months after training (audio file 2).

5.1. DSSon Basic Model

The DSSon basic model uses a time compression factor $\mathring{\kappa} = t/\mathring{t} = 5$ and a dilation parameter $\mathring{\Delta} = T_i/\mathring{T}_i = 5$. This moderate compression factor allows for a rather fast post-hoc review of the data. The sonic events resulting from adjacent positive and negative excursions are displayed at a rate of approximately 8 events per second, that is, a mean revolution rate of 0.4 Hz times (typically) 4 segments per revolution (2 positive and 2 negative excursions) times compression factor $\mathring{\kappa} = 5$. This rhythmical pattern can be easily perceived in detail because it lies quite within the typical range of musical gestures and the individual events do not overlap due to the dilation parameter chosen ($\mathring{\Delta} = \mathring{\kappa}$). In order to better facilitate the discrimination between positive and negative excursions, different reference frequencies for the pitch modulator are employed, specifically $\mathring{f}^{+}_{\text{ref}} = 400$ Hz and $\mathring{f}^{-}_{\text{ref}} = 300$ Hz. To monitor both the individual excursions and the overall trend, both pitch scaling factors are applied $\mathring{\alpha} = \mathring{\beta} = 2$. Amplitude modulation derived from the instantaneous magnitude of the segment's data values is used, that is, the power law distortion factor $\mathring{\phi}$ equals 1. The final model including the parameter values is given in (11).

The model was applied to three FRED data signals, two from user A (audio files `DA1.wav`,`DA2.wav` and one from user B (audio file `DB1.wav` — these audio data files are in the

`Data/Audified` directory in the project repository). Figs. 6 and 7 show the data and trend as well as the spectrogram of the basic DSSon model for a rather poor performance (user B, audio file 3). The user is obviously not able to maintain a stable mean speed at the beginning of the exercise session nor to stay within the range of 0.2 Hz – 0.6 Hz. Large positive excursions are clearly visible at 6 and 15 s in Fig. 6 and result in strong high frequency events at 1 and 3 s (Fig. 7). Sudden slow instants at 11, 45, and 55 s yield prominent low frequency sounds at 2, 9, and 11 s accordingly (Fig. 7). Note that highlighting the trend (of approximately 0.4 Hz) in

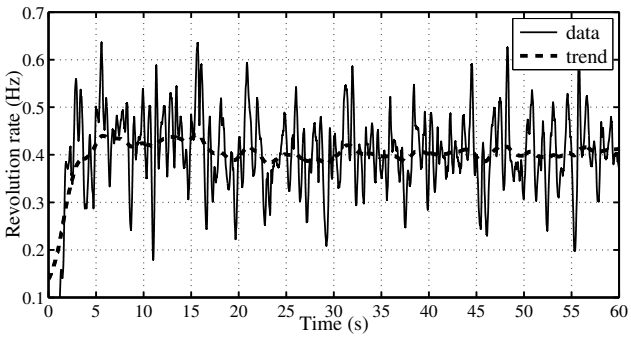


Figure 6: Data stream and trend signal (weighting factor $w = 0.2$) of FRED exercise session for a poor performer (user B).

the sonification (due to $\mathring{\alpha} > 0$) results in an upward shift of the pitch register compared to the range of the reference frequencies.[2] The trend variation results in an overall glissando gliding upward and downward displayed in the spectrogram as the sliding white frequency band framed by the sonic events of positive and negative segments respectively.

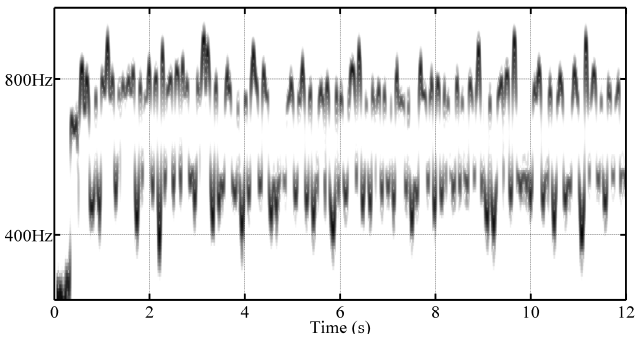


Figure 7: DSSon basic model spectrogram of User B (Fig. 6, audio file 3).

In comparison, the DSSon of the experienced user (Fig. 5, audio file 2) is shown as the spectrogram in Fig. 8. A constant mean rate and regular small deviations resulting in a soft and steady rhythmical pattern dominate this example.

[2]If $\mathring{\alpha} = 0$, the trend data are completely suppressed resulting in a lower pitch register. If $\mathring{\alpha} = 2$ and the trend equals 0.4Hz, then the instantaneous frequencies are multiplied by $2^{\mathring{\alpha}0.4} = 2^{0.8} = 1.75$, resulting in a center frequency of $f = 0.5 \times (300 + 400) \times 1.75 \approx 610$ Hz instead of 350 Hz.

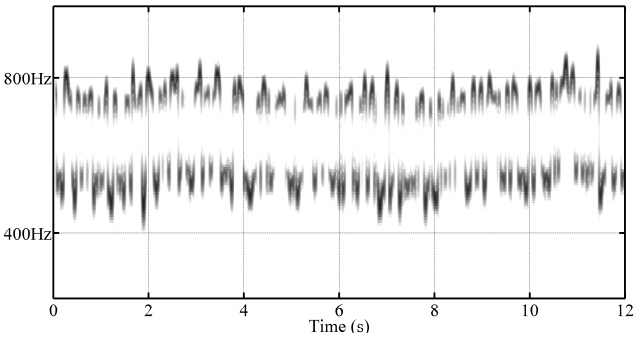


Figure 8: Spectrogram of DSSon basic model for an experienced user (FRED data and trend are shown in Fig. 5) (audio file 2).

5.2. DSSon Individual Target Range Model

The time compression factor $\mathring{\kappa} = 5$ used in the previous examples allows for a quick review of an individual performance. Nevertheless, exploring a collection of FRED sessions consisting of up to five exercise blocks each of 3 minutes duration, would result in a rather time-consuming endeavour and providing sonification with an even larger time compression of $\mathring{\kappa} = 10..20$ is preferable. However, the increased playback speed means that the rhythmical patterns of the sonic events and their pitch contours would become indiscernible if the DSSon basic model with its previous parameter values were employed.

Therefore, the DSSon individual target range model (ITR) suppresses segments whose maximum excursions stay below the target range set for each user individually by the physiotherapist. This is accomplished by a threshold-based amplitude modulator similar to the one proposed in (7) and setting the threshold parameter $\mathring{\epsilon}$ appropriately. Contrary to the amplitude modulator in (7) which displays only the segment's data values exceeding the threshold, one might be interested to listen to the entire segment if its value exceeds the target range at some point. Hence, a threshold-based indicator function combined with the segment's instantaneous magnitude is used as the amplitude modulator $a_i(\mathring{t})$ (13).

To display the remaining segments in sufficient detail, the dilation parameter $\mathring{\Delta}$ is set as $\mathring{\Delta} < \mathring{\kappa}$ yielding potentially overlapping sonic events. Figs. 9 and 10 show the spectrograms of the new model for the two users. $\mathring{\kappa} = 15$ results in a sonification duration of 4 seconds for a 1 minute session, $\mathring{\Delta} = 5$ yielding a threefold overlap of adjacent sonic events. The threshold parameter $\mathring{\epsilon}$ is set to 0.1 Hz for both examples though in practice the therapist would have chosen individual values for the two users according to their level of motor control. All other sonification parameters are set as in the basic model. Note that for the experienced user (Fig. 9), a sparse auditory display is obtained by the new model (audio file 4) whereas a dense sonification with almost constantly overlapping sonic events is caused by the poor performance of user B (Fig. 10, audio file 5).

5.3. DSSon Advanced Model

Both DSSon models presented so far are based on a modified auditory graph of adjacent data segments. They are characterized by a smooth functional relationship between data values and the audi-

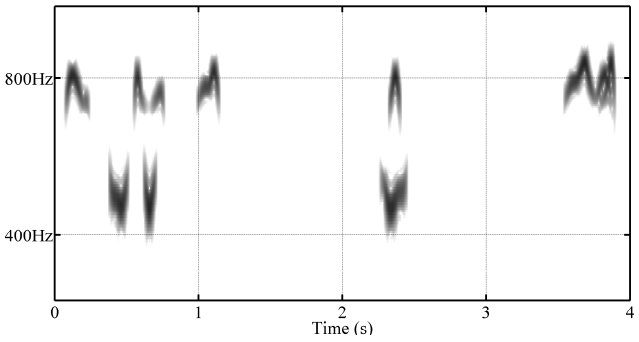


Figure 9: DSSon ITR model experienced user spectrogram (audio file 4).

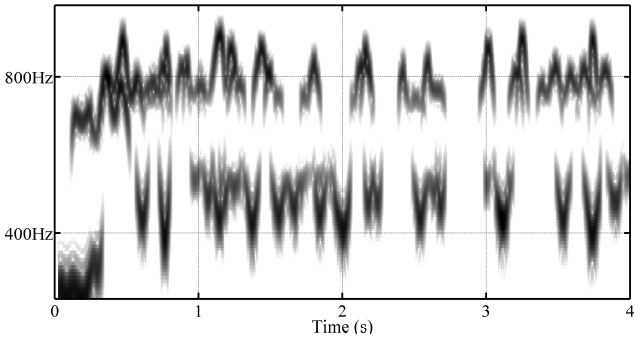


Figure 10: DSSon ITR spectrogram for user B (audio file 5).

tory display which can be easily perceived by the listener. As every segment is sonified by an amplitude and pitch modulated sinusoid, a coherent auditory gestalt of homogeneous timbre emerges. However, the special features of interest mentioned in subsection 4.1 are not displayed saliently except for the ITR model which delivers sonic events only for segments exceeding the individual target range, thereby explicitly displaying feature #3. In order to indicate excursions above 0.6 Hz (feature #1) and below 0.2 Hz (feature #2) prominently, timbre modifications are utilized as an additional sonification parameter. Segments whose maximum excursions cross these limits, are sonified by a fixed harmonic complex (for overshoots above 0.6 Hz) or subharmonic complex (for undershoots below 0.2 Hz) respectively. This is achieved by including the timbre operator $\mathring{H}\langle..\rangle$ in a DSSon advanced model (ADV) as (14).

The auxiliary sonification parameters $\mathring{J}$ and $\mathring{\nu}$ specify the number of partials, hence the bandwidth of the sonic event, and the amplitude attenuation associated with increasing partial order. $\mathring{c}_o$ and $\mathring{c}_u$ are set so as to align the loudness levels of the overshoot and undershoot segment with the basic one (in this case, $\mathring{c}_o = 0.5$ and $\mathring{c}_u = 0.7$). Note that by introducing a non-trivial timbre operator, the additional distinct categories of sonic events will result in a sonification where three auditory streams are likely to be perceived and the coherent gestalts of the previous models become dispersed.

To further accentuate segments of long excursions which predominantly occur for undershoots, a data-dependent transformation of the dilation parameter $\mathring{\Delta}$ is incorporated in the ADV model. For data segments whose maximum excursions stay within specified limits (e.g. $0.2\,\text{Hz} \leqslant f \leqslant 0.6\,\text{Hz}$), the dilation parameter is fixed to $\mathring{\Delta} = \mathring{\Delta}_0$, whereas for overshoot and undershoot segments, the dilation parameter becomes a monotonically decreasing function of the segment's data values, $\mathring{\Delta}_i$, and causes stretched sonic events. As a transformation, we specifically propose the hyperbolic function of the segment's area, that is, the time integral of segment's magnitude $A_i = \int_{t_{i-1}}^{t_i} \big|x(t)\big|\,\mathrm{d}t$ (15). The hyperbolic function translates into a linear dependence of the sonic event's duration $\mathring{T}_i$ on the segment's area A_i, since (15) and $\mathring{\Delta}_i = T_i/\mathring{T}_i$ lead to (16). The additional sonification parameters $\mathring{A}_0$ and $\mathring{\sigma} \geqslant 1$ determine the area threshold and the strength of the dilation transformation respectively. The area threshold should be set to $\mathring{A}_0 = 1/8\pi$ which equals the area of a sine-formed segment of duration $T = 1/(4 \times 0.4\,\text{Hz})$ (the expected duration of an excursion at target revolution rate of 0.4 Hz) and of amplitude 0.2 (magnitude difference between either limit, i.e., 0.2 Hz and 0.6 Hz, and the target rate). Utilizing this dilation transformation yields dominant stretched sonic events for long overshoot and undershoot segments. However, because the amplitude modulator used up to this point ((6) and (13)) delays the loudness peaks of the stretched events, the temporal structure of data segmentation is likely to get obscured. Therefore, an envelope-based amplitude modulation with a rather sharp attack followed by a decay and weighted by the segment's maximum magnitude $x_i^{\max} = \max_t \Big(\big|x_i(t)\big|\Big)$ is considered for overshoots and undershoots in the ADV model (17). The decay parameter $\mathring{\gamma}$ is set to $\mathring{\gamma} = 0.13 \cdot T_i$ which leads the sonic event to end at an amplitude level of -40 dB relative to its maximum. To prevent annoying clicks, a short fade-out portion is further applied at the very end of the envelope. The complete amplitude modulator for the ADV model reads as (18).

We applied the ADV model to FRED data setting the sonification parameters $\mathring{J} = 5$, $\mathring{\nu} = 2$, $\mathring{\sigma} = 1$, $\mathring{c}_o$, $\mathring{c}_u$, $\mathring{A}_0$ and $\mathring{\gamma}$ as mentioned above and the other parameters as in the ITR model. Fig. 11 shows the spectrogram of the ADV model for user B. Note the additional harmonic and subharmonic partials for the overshoot and undershoot segments at 0.4, 1.0, 3.2, 3.7 s, and 0.0, 0.7, 3.6 s respectively (audio file 8). As the experienced user A did not produce any excursions beyond the limits, the ADV model yields the same results as the ITR model (see Fig. 9, audio file 4).

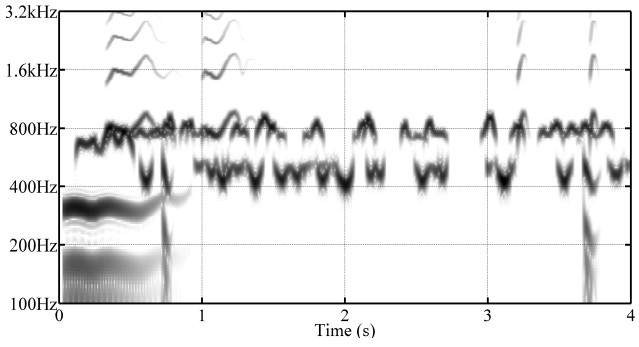


Figure 11: Spectrogram of DSSon ADV model for user B (audio file 8).

6. CONCLUSION

The proposed DSSon method aims to construct a direct sonification strategy for one-dimensional streams of numerical data. To achieve the intended directness, DSSon inherits an important property of other highly direct sonification approaches like audification and auditory graphs, in that it preserves the overall temporal structure of the data stream. DSSon is especially well-suited for data whose size (number of data points), is too small to be suitable for (pure) audification, because the audified sound would be either too short to perceptually decipher data details when using a high playback rate or, otherwise, would be displayed at very low frequencies where the human auditory system lacks good sensitivity.

Höldrich and Vogt's Augmented Audification [5] addressed the same problem domain. To ameliorate the drawback of the output being in too low a frequency range, they applied a data-dependent single side-band modulation to shift audio up by a desired frequency. The problem with this is that the frequencies in the data are scaled linearly resulting in compression of the frequency relationships, thereby destroying the periodicity of harmonic signals. A solution might be to use pitch-shifting which retains the frequency ratios, but this introduces artefacts into the signal and only works well for small shifts.

As the sonification method for the segments is structurally decoupled from the formation of the final sound stream, the playback speed of the entire DSSon signal can be set independent of the length of the individual sonic events offering a wide range of possible time compression/stretching factors and thereby high flexibility for zooming into or out of the data. Even pure audification can be regarded as a special case of DSSon, if every single data point is treated as a segment and sonified by a Dirac impulse weighted by the signed data value.

To ensure maximum directness of the resulting sonification, a modified auditory graph has been proposed as the specific method for sonifying the individual segments. In contrast to common auditory graphs, additional amplitude modulation derived from the segment's data evolution in an application-dependent way is accommodated to accentuate large data values. Furthermore, the reference frequency (and thereby the pitch register) is set individually for each sonic event depending on specific segment properties, for example, positive and negative-valued segments in an AC signal, or an overall trend.

DSSon offers some, albeit limited, potential for real-time applications since a segment's sonic event can generally only be synthesized when its end point is reached and the entire segment is available for deriving parameters of the specific sonification method.

The DSSon framework provides a wide range of application-dependent flexibility (as demonstrated by the different models for post hoc analysis of physiotherapeutic data) while maintaining a high degree of directness of the auditory display in that it succeeds in letting the data 'speak' for themselves. For future work, it is intended to apply DSSon to data from other domains which allow for the precise determination of specific detection or discrimination tasks, so that the DSSon method can be compared with audification and auditory graphs in formal listening tests.

7. ACKNOWLEDGMENT

The authors would like to thank Kirsty Lindsay and Nick Caplan of Northumbria University's Aerospace Medicine and Rehabilitation Laboratory for their advice on the salient information sought by physiologists in the post hoc analysis of FRED exercise data.

8. REFERENCES

[1] G. Parseihian, C. Gondre, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "Comparison and evaluation of sonification strategies for guidance tasks," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 674–686, Apr. 2016.

[2] P. Vickers and J. L. Alty, "Musical program auralization: Empirical studies," *ACM Transactions on Applied Perception*, vol. 2, no. 4, pp. 477–489, 2005.

[3] P. Vickers and B. Hogg, "Sonification abstraite/sonification concrète: An 'aesthetic perspective space' for classifying auditory displays in the ars musica domain," in *ICAD 2006 - The 12th Meeting of the International Conference on Auditory Display*, T. Stockman, L. V. Nickerson, C. Frauenberger, A. D. N. Edwards, and D. Brock, Eds., London, UK, 20–23 June 2006, pp. 210–216.

[4] P. Vickers, "Sonifications sometimes behave so strangely," in *Bloomsbury Handbook of Sonic Methodologies*, M. Bull and M. Cobussen, Eds. New York: Bloomsbury, In preparation, due for publication Dec. 2019.

[5] R. Höldrich and K. Vogt, "Augmented audification," in *ICAD 15: Proceedings of the 21st International Conference on Auditory Display*, K. Vogt, A. Andreopoulou, and V. Goudarzi, Eds. Graz, Austria: Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz (KUG), 2015, pp. 102–108.

[6] S. A. J. Wood, "Speech tempo," in *Working Papers*. Department of General Linguistics, Lund University, 1973.

[7] J. Rohrhuber, "$\mathring{S}$ — introducing sonification variables," in *SuperCollider Symposium 2010*, Berlin, 23–16 Sept. 2010, pp. 1–8.

[8] K. Vogt and R. Höldrich, "Translating sonifications," *JAES Journal of the Audio Engineering Society*, vol. 60, no. 11, pp. 926–935, 2012.

[9] J. H. Flowers, "Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions," in *Proceedings of 11th International Conference on Auditory Display (ICAD2005)*, E. Brazil, Ed., Limerick, Ireland, 6–9 July 2005, pp. 406–409.

[10] A. Winnard, D. Debuse, M. Wilkinson, L. Samson, T. Weber, and N. Caplan, "Movement amplitude on the functional re-adaptive exercise device: deep spinal muscle activity and movement control," *European Journal of Applied Physiology*, pp. 1–10, 2017.

[11] AD Instruments. (2017)

[12] P. Vickers and R. Höldrich, "nuson-DSSon: Direct segmented sonification," July 2017. DOI: 10.5281/zenodo.1007784. [Online]. Available: https://github.com/nuson/DSSon.

9. APPENDIX

9.1. Sonification Variables

Table 1: DSSon Variables, Functions, and Operators

Variable	Description	Value range
Temporal		
$\mathring{\kappa}$	time compression factor	sonification duration $\mathring{T} = T/\mathring{\kappa}$.
$\mathring{\Delta}$	dilation factor	$\mathring{\Delta} \geqslant 0$
Pitch		
$\mathring{f}_{\text{ref}}$	reference frequency	
$\mathring{\alpha}, \mathring{\beta}$	pitch scaling factors	$\langle \mathring{\alpha}, \mathring{\beta} \rangle \geqslant 0$
Loudness		
$\mathring{\phi}$	power law distortion factor	$\mathring{\phi} \geqslant 1$
$\mathring{\epsilon}$	amplitude threshold	$\mathring{\epsilon} \geqslant 0$
$\mathring{g}(..)$	gain function	e.g. mean, rms, …
$\mathring{\gamma}$	decay parameter	
Timbral		
$\mathring{H}\langle..\rangle$	operator for timbral control	e.g., wave shaping, additive synthesis

9.2. Sound Files

Table 2: Example Sound Files

#	Audio file	Description
1	`DSSon_Basic_A_n.wav`	M1, user A — novice
2	`DSSon_Basic_A_e.wav`	M1, user A — experienced
3	`DSSon_Basic_B.wav`	M1, user B — novice
4	`DSSon_ITR_A_e.wav`	M2, user A — exp.
5	`DSSon_ITR_B.wav`	M2, user B — novice
6	`DSSon_ADV_A_n.wav`	M3, user A — novice
7	`DSSon_Adv_A_e.wav`	M3, user A — exp.
8	`DSSon_Adv_B.wav`	M3, user B — novice

Models: M1 = basic model; M2 = individual target range model; M3 = advanced model
Data files used: user A, novice = DA1; user A, experienced = DA2; user B, novice = DB1

9.3. Equations

$$x_i(t) = \begin{cases} x\left(t + t_{i-1}\right) & 0 \leqslant t \leqslant (t_i - t_{i-1}) \\ 0 & \text{else} \end{cases} \tag{2}$$

$$\mathring{y}(\mathring{t}) = \sum_{i=1}^{M} \mathring{y}_i \left(\mathring{t} - \mathring{t}_{i-1}\right) \text{ where } \mathring{t}_{i-1} = \frac{t_{i-1}}{\mathring{\kappa}} \tag{3}$$

$$\mathring{y}_i(\mathring{t}) = a_i(\mathring{t}) \sin\left(2\pi \int_0^{\mathring{t}} f_{\text{ref}} \cdot 2^{b_i(\mathring{t}')} \cdot \mathrm{d}\mathring{t}'\right) \tag{4}$$

$$\mathring{y}_i(\mathring{t}) = \left| x_i\left(\mathring{\Delta} \cdot \mathring{t}\right)\right| \times \sin\left(2\pi \int_0^{\mathring{t}} f_{\text{ref}} \cdot 2^{\left(x_{\text{trend}}\left(t_{i-1}\right) + x_{i,\text{AC}}\left(\mathring{\Delta} \cdot \mathring{t}'\right)\right)} \,\mathrm{d}\mathring{t}'\right) \tag{5}$$

$$a_i(\mathring{t}) = \left| x_i\left(\mathring{\Delta} \cdot \mathring{t}\right)\right|^{\mathring{\phi}} ; \mathring{\phi} \geqslant 1 \tag{6}$$

$$a_i(\mathring{t}) = G\left(\left| x_i\left(\mathring{\Delta} \cdot \mathring{t}\right)\right|, \mathring{\epsilon}\right) \tag{7}$$

$$G(x, \mathring{\epsilon}) = \begin{cases} x - \mathring{\epsilon} & x \geqslant \mathring{\epsilon} \\ 0 & \text{else} \end{cases} \tag{8}$$

$$b_i(\mathring{t}') = \left(\mathring{\alpha} \cdot x_{\text{trend}}\left(t_{i-1}\right) + \mathring{\beta} \cdot x_{i,\text{AC}}\left(\mathring{\Delta} \cdot \mathring{t}'\right)\right) \tag{9}$$

$$\mathring{y}_i(\mathring{t}) = \left| x_i\left(\mathring{\Delta} \cdot \mathring{t}\right)\right|^{\mathring{\phi}} \times \sin\left(2\pi \int_0^{\mathring{t}} f_{\text{ref}}^{+/-} \cdot 2^{\left(\mathring{\alpha} \cdot x_{\text{trend}}\left(t_{i-1}\right) + \mathring{\beta} \cdot x_{i,\text{AC}}\left(\mathring{\Delta} \cdot \mathring{t}'\right)\right)} \,\mathrm{d}\mathring{t}'\right). \tag{10}$$

$$\mathring{y}_i(\mathring{t}) = \left| x_i\left(5\mathring{t}\right)\right| \times \sin\left(2\pi \int_0^{\mathring{t}} \begin{matrix} +{:}400\text{ Hz} \\ -{:}300\text{ Hz} \end{matrix} \cdot 2^{\left(2 \cdot x_{\text{trend}}\left(t_{i-1}\right) + 2 \cdot x_{i,\text{AC}}\left(5\mathring{t}'\right)\right)} \,\mathrm{d}\mathring{t}'\right) \tag{11}$$

$$\mathring{y}(\mathring{t}) = \sum_{i=1}^{M} \mathring{y}_i \left(\mathring{t} - \frac{t_{i-1}}{5}\right) \tag{12}$$

$$a_i(\mathring{t}) = \begin{cases} \left| x_i\left(\mathring{\Delta} \cdot \mathring{t}\right)\right| \max\limits_{\mathring{t}} \left(\left| x_i\left(\mathring{\Delta} \cdot \mathring{t}\right)\right|\right) \geqslant \mathring{\epsilon} \\ 0 \qquad \text{else} \end{cases} \tag{13}$$

$$\mathring{H}\langle \sin(\phi_i(\mathring{t}))\rangle = \begin{cases} \mathring{c}_o \sum_{j=1}^{\mathring{J}} j^{-\mathring{\nu}} \sin\left(j \cdot \phi_i(\mathring{t})\right) & \text{o/shoot} \\ \mathring{c}_u \sum_{j=1}^{\mathring{J}} j^{-\mathring{\nu}} \sin\left(\frac{1}{j} \cdot \phi_i(\mathring{t})\right) & \text{u/shoot} \\ \sin\left(\phi_i(\mathring{t})\right) & \text{else.} \end{cases} \tag{14}$$

$$\mathring{\Delta}_i = \begin{cases} \frac{1}{\mathring{\sigma}} \cdot \frac{\mathring{A}_0}{A_i} \cdot \mathring{\Delta}_0 A_i \geqslant \mathring{A}_0 \\ \mathring{\Delta}_0 \qquad\qquad \text{else.} \end{cases} \tag{15}$$

$$\mathring{T}_i = \mathring{\sigma} \frac{A_i}{\mathring{A}_0 \cdot \mathring{\Delta}_0} \cdot T_i \text{for} A_i \geqslant \mathring{A}_0. \tag{16}$$

$$a_i(\mathring{t}) = x_i^{\max} \cdot \frac{\mathring{\Delta}_i \mathring{t}}{\mathring{\gamma}} \cdot e^{-\left(\frac{\mathring{\Delta}_i \mathring{t}}{\mathring{\gamma}} - 1\right)}. \tag{17}$$

$$a_i(\mathring{t}) = \begin{cases} x_i^{\max} \cdot \frac{\mathring{\Delta}_i \mathring{t}}{\mathring{\gamma}} \cdot e^{-\left(\frac{\mathring{\Delta}_i \mathring{t}}{\mathring{\gamma}} - 1\right)} & \text{o/u/shoots} \\ \left| x_i \left(\mathring{\Delta} \cdot \mathring{t} \right) \right| & x_i^{\max} \geqslant \mathring{\epsilon} \\ 0 & \text{else.} \end{cases} \tag{18}$$

REAL-TIME AUDITORY CONTRAST ENHANCEMENT

Marian Weger[1], *Thomas Hermann*[2], *Robert Höldrich*[1]

[1] IEM, University of Music and Performing Arts, Graz, Austria
[2] Ambient Intelligence Group, CITEC, Bielefeld University, Bielefeld, Germany
weger@iem.at

ABSTRACT

Every day, we rely on the information that is encoded in the auditory feedback of our physical interactions. With the goal to perceptually enhance those sound characteristics that are relevant to us — especially within professional practices such as percussion and auscultation — we introduce the method of real-time Auditory Contrast Enhancement (ACE). It is derived from algorithms for speech enhancement as well as from the remarkable sound processing mechanisms of our ears. ACE is achieved by individual sharpening of spectral and temporal structures contained in a sound while maintaining its natural gestalt. With regard to the targeted real-time applications, the proposed method is designed for low latency. As the discussed examples illustrate, it is able to significantly enhance spectral and temporal contrast.

1. INTRODUCTION

Every sound that we encounter in our daily lives contains information. If the sound is the result of a physical process such as an interaction with our environment, then it contains information on the involved physical objects (e.g., material or geometry), their environment (e.g., room acoustics), and the type of interaction (e.g., hitting or scratching). Pieces of information that are not only restricted to natural sounds but also apply for synthesized sounds are, for example, sound parameters such as frequency or amplitude, as well as their perceptual pendants — here pitch and loudness. If such sound parameters are deliberately modified with respect to some underlying data, as being the case in auditory display and also in music, then even this data is encoded in the sound. Unfortunately, we are not able to perceive the entire information, but only a small fraction of it.

Nevertheless, as an everyday experience, we rely on the auditory feedback of our physical interactions, either consciously, e.g., when shaking a box to guess its contents, or unconsciously, when automatically adapting to the physical structure of the ground while walking. If the auditory feedback (the sonic reaction to physical interaction) is artificially modified, then we speak of *augmented auditory feedback* [1]. It seeks to attain three goals. (1) Add additional information to the sound. This is usually referred to as *Auditory Augmentation* [1, 2, 3, 4]. (2) Modify the information that is already contained in the sound, in order to achieve a change in behavior, e.g., [5, 6, 7]. (3) Enhance the information that is already contained in the sound, e.g., improvement of the Signal-to-Noise-Ratio (SNR).

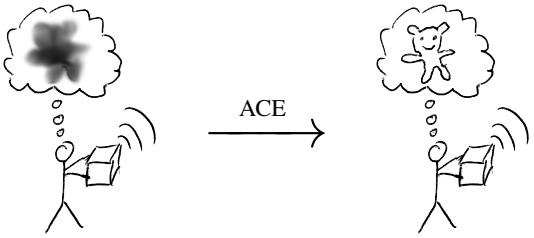


Figure 1: Someone shaking a box to guess its contents from the resulting sound. A task we want to facilitate.

In this sense, we introduce Auditory Contrast Enhancement (ACE) with the objective to enhance relevant sound characteristics in order to facilitate their perception and hence improve the conveyance of the underlying information. This concept is illustrated in Fig. 1. What might be relevant to users, however, depends on their individual activities, as well as on the type and origin of the observed sound. We expect high potential for auditory contrast enhancement where listening is part of a knowledge-making process. Especially when, for example, scientists, engineers, or physicians rely on their ears during professional routines. Even for this limited group of people and their audition-based practices, Supper and Bijsterveld discriminate between at least six different listening modes, depending on the purpose and on the way of listening [8].

One of these practices is *percussion*, a technique where a physical object or body part is actively hit in order to reveal information on its inner structure through the induced auditory feedback. This technique has established in everyday life to locate a good spot for a drill hole in a wall. The passive complement is *auscultation* where a physical object such as a machine or a human body is inspected by passively listening to its sound — usually by using a stethoscope. This tool enhances auditory contrast not only by efficient guidance of the structure-borne sound to the user's ears, but also by amplification of frequency-ranges which are of special interest to the user [9].

We distinguish between two types of auditory contrast. By *inter-stimulus contrast*, we mean the perceived differences between stimuli, which results from juxtaposing them. Inter-stimulus ACE tries to display all aspects in which two or more stimuli differ auditorily. This topic is extensively investigated in our companion paper [10] and will not be covered further here. By *intra-stimulus contrast*, we mean the strengths of peculiarity of a single stimulus. These may be the spectro-temporal dynamics of a sound. By intra-stimulus ACE, we seek to intensify those peculiarities.

https://doi.org/10.21785/icad2019.026

Our goal is to enhance the perception of those sound properties that characterize a sound, while maintaining its original gestalt as good as possible. We assume that this compromise can be achieved by attenuating non-characteristic aspects of the signal, thus leading to reduced spectral, temporal, and informational masking. In the extreme case, a very strong contrast enhancement leads to a cartoonification of the sound, reducing it to only a few very prominent sound attributes. This is conceptually similar to the visual domain where contrast is usually understood as the degree to which areas of an image differ in appearance.

Assuming that a sound is characterized by its unique spectral and temporal structure, an enhancement of this structure may automatically enhance the contrast to other sounds which exhibit a different structure. If, however, two sounds share the same strong characteristics with only minor differences, intra-stimulus contrast enhancement could even suppress those differences, leading to reduced inter-stimulus contrast between both. Such "similarity enhancement" might be useful when searching for similarities between stimuli. Otherwise, inter-stimulus contrast enhancement would be the recommended choice (see companion paper [10]).

In summary, we identify two activities which intra-stimulus ACE should improve: (1) identify the physical sound source, as visualized in Fig. 1. and (2) discriminate between sounds that are different to each other.

The rest of this article is structured as follows. In Sec. 2 we derive an algorithm for real-time intra-stimulus ACE. Spectral and temporal contrast enhancement are individually addressed in Sec. 2.1 and 2.2, respectively. Finally, a general discussion (Sec. 3) as well as conclusions and an outlook on future investigations (Sec. 4) are given. Supplementary material such as the sound examples (Snd.) referenced in the text can be found under the following link: `https://doi.org/10.4119/unibi/2935786`

2. AUDITORY CONTRAST ENHANCEMENT

The main applications that are envisaged for real-time ACE are percussion and auscultation — not so much for medical purposes but more for material testing by ear and auditory observation of mechanical processes such as machines. The targeted sounds therefore include transient interaction sounds and environmental sounds, but not speech or music. The focus on real-time application on auditory feedback makes a low-latency implementation necessary. Furthermore, the sounds resulting from ACE should maintain some degree of naturalness — they should stay within the limits of plausibility with reference to their individual context and the performed action. Even if ACE is only used as a technical tool, we know that "naturalness influences the perceived usability and pleasantness of an interface's sonic feedback" [11]. While development is performed in Matlab, the real-time algorithm will be implemented in SuperCollider and Pure Data to finally be able to run on smartphones or low-latency platforms such as the Bela [12]. Sound recording and playback can be done either with a contact microphone and loudspeaker, or by using a mic-through system (headphones with built-in microphones).

Figure 2 shows the overall block diagram. Output $s'[n]$ is a mix of three signals: (1) the dry input signal $s[n]$ (e.g., coming from a microphone), (2) the output $s_f[n]$ of Spectral Contrast Enhancement (SCE, see Sec. 2.1), and (3) the output $s_t[n]$ of Temporal Contrast Enhancement (TCE, see Sec. 2.2). Their individual gains are parametrized by two linear cross-fades: (1) between $s_f[n]$ and $s_t[n]$ to intuitively tune to the signal dimension of interest, and (2) between this weighted sum and the original signal (wet and dry) for overall strength of the effect.

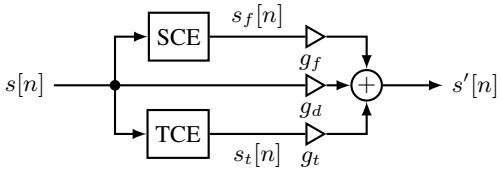


Figure 2: Overall block diagram of real-time ACE.

2.1. Spectral Contrast Enhancement

Yang et al. define spectral contrast as "the decibel difference between peaks and valleys in the [magnitude] spectrum" [13]. They describe several algorithms for spectral contrast enhancement, aiming at two applications: (1) compensation of reduced frequency selectivity in hearing-impaired people, and (2) speech enhancement in noise. One of the easiest methods is to exponentiate the magnitude spectrum by a variable exponent, followed by normalization [14]. This results in a spectral dynamics expansion with respect to the global maximum. Other approaches use linear prediction which works well for speech enhancement where detailed information on the sound source is available [13].

A large group of algorithms is based on an analog circuit proposed by Stone and Moore [15]. In principle, the signal is split into a number of frequency bands which are separately processed by a variable gain amplifier and then summed. The gain of each channel is a weighted sum of its own envelope and the envelopes of four neighboring channels; the latter with negative weights. This weighting is similar to a transversal FIR filter. As result, spectral peaks are amplified while troughs are attenuated. The digital implementation of this algorithm — Yang et al. refer to it as "Cambridge's method" — works as follows [13, 16]:

1. Computation of the spectrum X_k of a (windowed) signal block via Fast Fourier Transform (FFT), with frequency index k.
2. Calculation of excitation pattern P_k — "the representation of a spectral shape in the auditory system" [15]. It resembles a smoothed version of the magnitude spectrum $|X_k|$.
3. The enhancement function E_k is the convolution of P_k with a Difference-of-Gaussians (DoG) function. This is similar to a smoothed 2nd derivative. The DoG function is the sum of a positive Gaussian and a negative Gaussian with larger (here: $2\times$) bandwidth. Convolution runs on a scale which quantifies the number of Equivalent Rectangular Bandwidths (ERB) that fit below a certain frequency — the ERB-rate scale [17].
4. The enhanced magnitude spectrum $|Y_k|$ is then

$$|Y_k| = P_k \cdot \big(|E_k| + 1\big)^{\mathrm{sgn}(E_k)\cdot\rho}\,, \qquad (1)$$

   where $\rho \geq 0$ controls the strength of the effect.
5. Inverse FFT of $|Y_k|$ combined with the original phase values.

While Cambridge's method did not improve speech intelligibility — neither analog nor digital — its high potential in "technical" enhancement of spectral contrast, i.e., increasing differences between peaks and valleys, is evident.

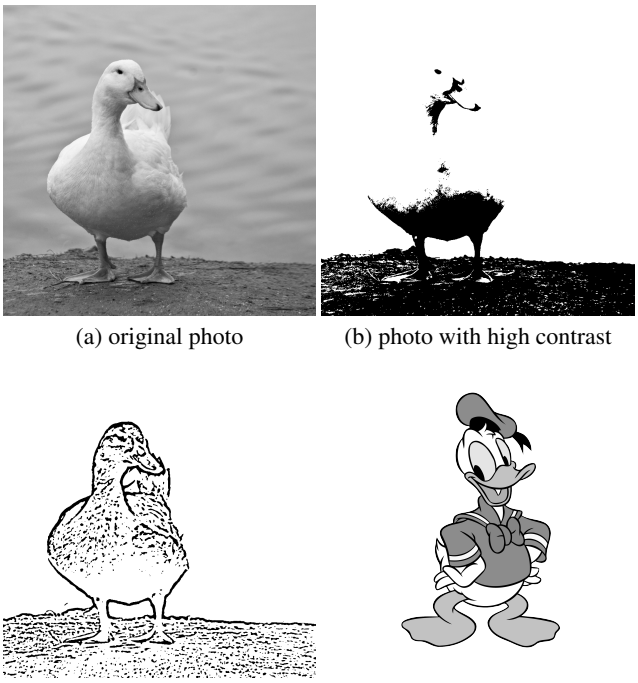
Our auditory system achieves spectral contrast enhancement similar to Cambridge's method. The underlying mechanism is

based on Lateral Inhibition (LI) in the neural networks of the auditory nerves and the auditory cortex [18, 19]. In general, this process can be described as "the suppression of nervous activity at one place in a receptor field as a consequence of the stimulation of adjacent places in this field" [20]. Besides, for instance, the retina and the skin, such receptor fields are also found along the basilar membrane [21, 22]. Kral and Majernik used an artificial neural network to model the effect of spectral contrast enhancement in the auditory system via lateral inhibition [18]. Among their simulated scenarios, three extreme cases are of particular interest. (1) Partly overlapping band-limited noise signals are narrowed in bandwidth and thus separated. (2) Uniform white noise is effectively suppressed. (3) Uniform white noise where a specific frequency-range has been suppressed leads to spikes at the edges of the stopband — the so-called edge effect.

It seems that in general there are two types of spectral contrast: (1) exponentiation relative to the global maximum (we refer to it as spectral dynamics expansion), and (2) lateral inhibition (we refer to it as spectral sharpening). It might be interesting to compare these to the visual domain. Spectral dynamics expansion compares to visual contrast control as shown in Fig. 3b, while spectral sharpening is actually edge detection (see Fig. 3c; the image shows the inverted result) — remember the edge effect demonstrated by Kral and Majernik [18]. In order to achieve something close to cartoonification, as exaggeratedly illustrated in Fig. 3d, we would need a combination of both types of contrast. In vision, this would be an overlay of Fig. 3b and c, e.g., by multiplying or taking the minimum of both images). In the auditory domain, we would take the maximum of both output spectra. The above considerations suggest that both types of spectral contrast enhancement are necessary, depending on the sound characteristics of interest, and therefore need to be implemented for parallel or serial use.

As we target low latency and real-time operation, the use of FFT — the basis for the majority of speech enhancement algorithms — is not possible. For that reason, frequency separation must be achieved by a filterbank, similar to the analog circuit by Stone and Moore [15]. We are therefore restricted to operate on a very limited number of frequency bands. Note, however, that Cambridge's method returns an altered version of the excitation patterns — a signal with significantly reduced frequency resolution. An adequate approximation of the excitation pattern can be obtained by a Gammatone filterbank (GTFB) — a widely used model for the auditory filters [23]. If the filters' center frequencies are equally spaced on the ERB-rate scale (and set to constant bandwidth in parts of the ERB), they simulate an equal spacing on the basilar membrane. The lower bands exhibit a smaller bandwidth in Hz, leading to longer impulse response and group delay. This implies a trade-off between frequency resolution and group delay towards low frequencies, which needs to be taken care of.

The excitation pattern is expressed by the energy distribution across sub-bands, calculated via their channel envelopes. Depending on the implementation of the Gammatone filter, it can also output the imaginary part of the resulting signal, in addition to the real output. An accurate estimation of the signal envelope is then given by the magnitude of the complex filter output. A suitable implementation is the one by Hohmann [24], which is available for Matlab[2], Pure Data[3] and SuperCollider[4]; in the latter case, a small modification of the source code is needed in order to return the imaginary part. We use 60 4th-order filters with center frequencies from 50 Hz to 20 kHz, overlapping at their -4 dB cutoff frequency (as in [25]). During resynthesis, i.e., summation of the processed sub-bands, their different group delays are usually compensated by individual time-delays, in order to reduce ripple in the output spectrum. We circumvent such additional latency by weighting the sub-bands with alternating signs, as proposed by Noisternig [25].



(a) original photo (b) photo with high contrast

(c) photo with edge detection (d) a famous cartoon duck

Figure 3: The photo of a white duck in three versions, and the drawing of a famous cartoon duck.[1]

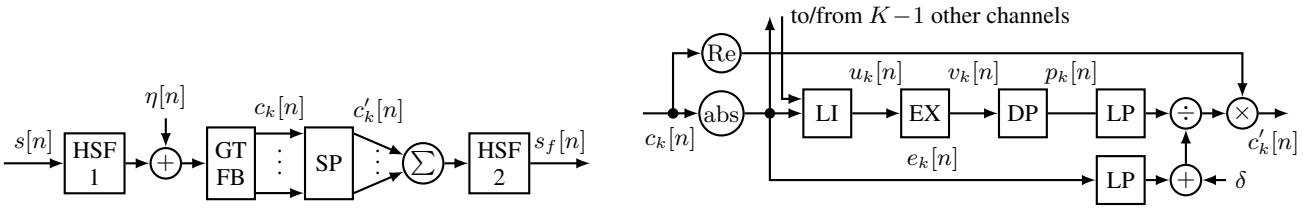
A block diagram of the proposed algorithm for spectral contrast enhancement is shown in Fig. 4. The overall block diagram (Fig. 4a) illustrates the general idea described above. In summary, the input signal $s[n]$ is split into K sub-bands $c_k[n]$ by a Gammatone filterbank with K channels; k is the channel index. The actual spectral contrast enhancement is done within the sub-band processing block (SP). The sum of the processed (real-valued) sub-bands $c'_k[n]$ then forms the enhanced output signal. Within SP, all channels are treated equally. While the Gammatone filterbank accounts for the $1/f$ proportionality of signal energy, this might not be enough for many natural signals which may exhibit even stronger high-frequency loss. This could lead to overly damped high-frequency content in the output. This effect is reduced by a pair of shelving filters (HSF) — one boosting high frequencies of

[1] Fig. 3a-c: Anne Davis, `http://flickr.com/anned/`, Creative Commons Attribution NonCommercial (CC BY-NC) 2.0 Generic License. Fig. 3d: `http://pngimg.com`, CC BY-NC 4.0 International License.

[2] Matlab implementation of the used Gammatone filterbank [24]: `http://medi.uni-oldenburg.de/download/demo/gammatone-filterbank/gammatone_filterbank-1.1.zip`

[3] *Audition library* for Pure Data: `http://lumiere.ens.fr/Audition/tools/realtime/`

[4] *AuditoryModeling* UGens from *SC3 Plugins*: `https://github.com/supercollider/sc3-plugins`



(a) overall block diagram of spectral ACE

(b) one channel of sub-band processing (SP)

Figure 4: Block diagram of spectral ACE.

$s[n]$ before feeding it to the Gammatone filterbank (HSF 1), and another one inverting the effect of the first one by attenuation after resynthesis/summation (HSF 2).

Each channel $c_k[n]$ individually passes sub-band processing as shown in Fig. 4b. First, the sub-band envelope $e_k[n]$ is extracted by taking the absolute value of the complex signal $c_k[n]$. This envelope then successively passes three stages: lateral inhibition (LI, see Sec. 2.1.1), exponentiation (EX, Sec. 2.1.2), and decay prolongation (DP, Sec. 2.1.3). The processed envelope $p_k[n]$ is finally applied to the real part of the sub-band signal $c_k[n]$ by multiplication with the ratio between processed and original envelope (see Eq. 2). Both envelopes are low-pass filtered by a leaky integrator with time-constant $\tau = 2$ ms to suppress disturbing artifacts which occur at high amplitude ratios, especially at low overall volume. For regularization, a small value $\delta = 10^{-5}$ is added to the denominator (assuming audio signals in the range between -1 and 1).

$$c'_k[n] = \mathrm{Re}\{c_k[n]\} \cdot \frac{e'_k[n]}{e_k[n] + \delta} \,. \tag{2}$$

2.1.1. Spectral Sharpening

One problem we see in Cambridge's method (Eq. 1) is that it not only dampens spectral valleys but also amplifies spectral peaks. This uncontrolled amplification of the signal can be avoided by restricting the enhancement function E_k to negative values.

We first define an inhibition term $T_k[n]$ which quantifies the overall energy in the neighboring sub-bands. If it is larger than the energy in the observed band, then this band is attenuated. Calculation of the inhibition term is based on the sub-band envelopes $e_k[n]$ which are low-pass-filtered by a leaky integrator, which leads to $\tilde{e}_k[n]$. The resulting slow attack time suppresses inhibition caused by short spikes in neighboring bands, while the decay adds an aftereffect to the lateral inhibition.

We base the calculation of the neighboring bands' weights on the DoG function as in Cambridge's method. The ratio between the bandwidths of the two Gaussians controls the sharpness of the resulting spikes in the spectrum. As our approach anyway restricts sharpening to the bandwidths of the used filters (which is quite "unsharp"), we reduce the positive Gaussian to a minimum, being a Dirac delta impulse. This way, extreme enhancement (large ρ) would inhibit all frequency bands except those which describe local maxima. The bandwidth of the negative Gaussian is set via its standard deviation σ in ERB-rate.

For the lowest and highest sub-band, neighbors of significant weight are outside the scope of the filterbank. A zero-padding (insertion of zero-valued virtual bands on both sides) would introduce an unwanted edge-effect at the lowest and highest sub-band ($k = 1$ and $k = K$, respectively), similar to the simulation by Kral and Majernik [18]. Therefore, two virtual sub-bands (copies of sub-bands 2 and $K-1$) are introduced as sub-bands 0 and $K+1$, respectively (copying the edge bands themselves would half a potential contrast in those bands). The inhibition term $T_k[n]$ then becomes

$$T_k[n] = \sqrt{\frac{1}{\gamma_k^-} \sum_{i=0}^{k-1} \gamma_{i,k} \cdot \tilde{e}_i^2[n] + \frac{1}{\gamma_k^+} \sum_{i=k+1}^{K+1} \gamma_{i,k} \cdot \tilde{e}_i^2[n]} \,, \tag{3}$$

where $\gamma_{i,k}$ is a Gaussian function, with center frequencies f_c of the filters given in ERB-rate:

$$\gamma_{i,k} = \exp\left(-\frac{(f_{c,i} - f_{c,k})^2}{2\sigma^2}\right) \,. \tag{4}$$

The scaling factor can be omitted, as the weights are anyway normalized for the lower and upper neighbors individually, altogether summing up to 1:

$$\gamma_k^- = 2 \sum_{i=0}^{k-1} \gamma_{i,k} \qquad \text{and} \qquad \gamma_k^+ = 2 \sum_{i=k+1}^{K+1} \gamma_{i,k} \,. \tag{5}$$

This scaling ensures that a signal with equal envelopes, i.e., in which $e_k[n]$ is the same for all k, implies $T_k[n] = \tilde{e}_k[n]$, and therefore leads to unchanged envelopes. Due to the ERB-scaled Gammatone filterbank, this is the case for a pink noise signal which exhibits a magnitude spectrum that is proportional to $1/f$. This relation approximates the decrease in energy towards high frequencies, that is common to many natural sounds. In analogy to Eq. 1, the sharpened envelopes $u_k[n]$ then become

$$u_k[n] = e_k[n] \cdot \min\left\{\left(\frac{\tilde{e}_k[n]}{T_k[n]}\right)^{\rho}, \; 1\right\} \tag{6}$$

The amount of spectral sharpening is set by the parameter $\rho \geq 0$. As the quotient $T_k[n]/\tilde{e}_k[n]$ is restricted to values below 1, any $\rho > 0$ literally suppresses lower quotients.

The effect of spectral sharpening is demonstrated by knocking with knuckles on a wooden plate. Listen to the signal without and with spectral ACE (Snd. 1.1 and 1.2, respectively). Corresponding spectrograms are shown in Fig. 5a-b. Parameters have been set to values which work well for most signals: $\rho = 30$, $\sigma = 3$ ERB, and smoothing time constant $\tau = 7$ ms. It is apparent that the described algorithm effectively suppresses spectral troughs while leaving local maxima as narrowband regions with their original amplitude. In addition, the broadband background noise is reduced to some high-frequency artifacts of the recording which are now

clearly audible. A ρ larger than 30 does not seem to bring any benefit for spectral sharpening; the signal is already reduced to its local maxima. Additional contrast can be achieved by spectral dynamics expansion, as explained in the next section.

2.1.2. Spectral Dynamics Expansion

The goal of spectral dynamics expansion is to attenuate frequency bands with low energy while pulling those with high energy, above a certain threshold value, up to the running global maximum. In contrast to spectral sharpening, this approach should not attenuate broadband regions in the spectrum if they are prominent enough. On the downside, it will suppress even very prominent local maxima if they appear below the threshold.

Spectral dynamics processing is achieved by exponentiation of the magnitude spectrum — inspired by the simple algorithm originally proposed by Boers [14]. In our case, each envelope $u_k[n]$ is scaled with respect to the global maximum of all (smoothed) envelopes (see Eq. 7). As gain-factor, we use the quotient of the smoothed envelope $\tilde{u}_k[n]$ and a fraction of the instantaneous maximum of all smoothed envelopes ($\mu\tilde{u}_{max}$). The exponent $\beta \geq 0$ sets the amount of expansion; $0 < \mu \leq 1$ is the relative threshold. Gain is clipped at $\tilde{u}_{max}/\tilde{u}_k[n]$ so that $u_k[n]$ does not exceed the maximum of all sub-band envelopes.

$$v_k[n] = u_k[n] \cdot \min\left\{ \left(\frac{\tilde{u}_k[n]}{\mu\tilde{u}_{max}[n]} \right)^{\beta} , \ \frac{\tilde{u}_{max}[n]}{\tilde{u}_k[n]} \right\} \quad (7)$$

with the (instantaneous) global maximum

$$\tilde{u}_{max}[n] = \max_k \big\{ \tilde{u}_k[n] \big\}. \quad (8)$$

Listen again to the enhanced signal from the previous section (Snd. 1.2 / Fig. 5b). Additional contrast is achieved by feeding this signal into spectral dynamics expansion (Snd. 1.3 / Fig. 5c). Furthermore, the background noise is gone. The parameters have been set to $\mu = 0.8$ and an extreme value of $\beta = 8$, leading to a spectral gate where values below $\mu\tilde{u}_{max}[n]$ are almost completely suppressed while values above approach the global maximum.

Contrary to spectral sharpening, spectral dynamics expansion can also be used to exaggerate broadband regions in the spectrum. This is demonstrated in Snd. 2.1 and 2.2 with the recording of a vintage printing machine, with noise from a pneumatic system.

2.1.3. Decay Prolongation

Spectral resolution and pitch impression takes time. What if we gave listeners more time to perceive a sound by prolonging it through artificial decay? Such an effect could be achieved in a natural way via reverberation. Dombois and Eckel argue that reverberation might even be used to enhance audifications, as it facilitates discrimination between short transient sounds [26, p. 315]. Koumura and Furukawa examined the effect of reverberation on the identification of material via short impact sounds [27]. They found out that reverberation actually deteriorates material identification; however, after a short while, participants adapted to the reverberation and achieved similar identification rates as with the dry stimuli. It must be noted that the results varied greatly among participants. Furthermore, adaptation to reverberation during speech does not help to identify a following impact sound [28]. Such natural reverberation, of course, is not correlated to the stimulus itself, but just convolves it with an arbitrary impulse response. A completely "transparent" reverberation whose impulse response has a white magnitude spectrum might already lead to better results.

Yet another problem is the broadband spectrum of the transient sounds — any artificial reverberation will therefore mask succeeding parts completely with broadband noise. Even if the resonances are sharpened through spectral contrast enhancement as derived in Sec. 2.1, a short transient signal in a single sub-band still results in a broadband signal at the output. However, if artificial decay is applied to the individual sub-band envelopes, their bandwidths are reduced and more time is given to the listener to gain a pitch impression. The enhanced sub-band envelopes after lateral inhibition and exponentiation may still contain short spikes which are not visible in the spectrogram of Fig. 5b-c, but which would have a huge impact if the sub-band envelopes were decayed as they are. Therefore, the envelopes must be smoothed before decay prolongation. As this further smears the envelopes in time, we instead split them into a transient part and a decay part. Only the decay part receives decay prolongation; both are re-combined afterwards.

We first introduce two simple non-linear low-pass filters based on a leaky integrator. $\mathrm{env_a}$ has a smooth attack but instant decay, while $\mathrm{env_d}$ has a smooth decay but instant attack. $\mathrm{env_a}$ is given in Eq. 9 for an arbitrary input signal $x[n]$ and output signal $y[n]$. $\mathrm{env_d}$ follows the same equation, but with flipped direction of the inequality sign, leading to a naturally-sounding exponential decay.

$$y[n] = \begin{cases} (1-\alpha)|x[n]| + \alpha y[n-1], & |x[n]| < y[n-1] \\ |x_k[n]|, & \text{otherwise} \end{cases} \quad (9)$$

The amount of smoothing is set via the smoothing factor α. A more convenient parametrization can be achieved via time constant τ or -60 dB reverberation time T_{60}:

$$\alpha = \exp\left(-\frac{1}{\tau f_s}\right) = \exp\left(-\frac{\ln(1000)}{T_{60} f_s}\right) , \quad (10)$$

where f_s is the sampling frequency.

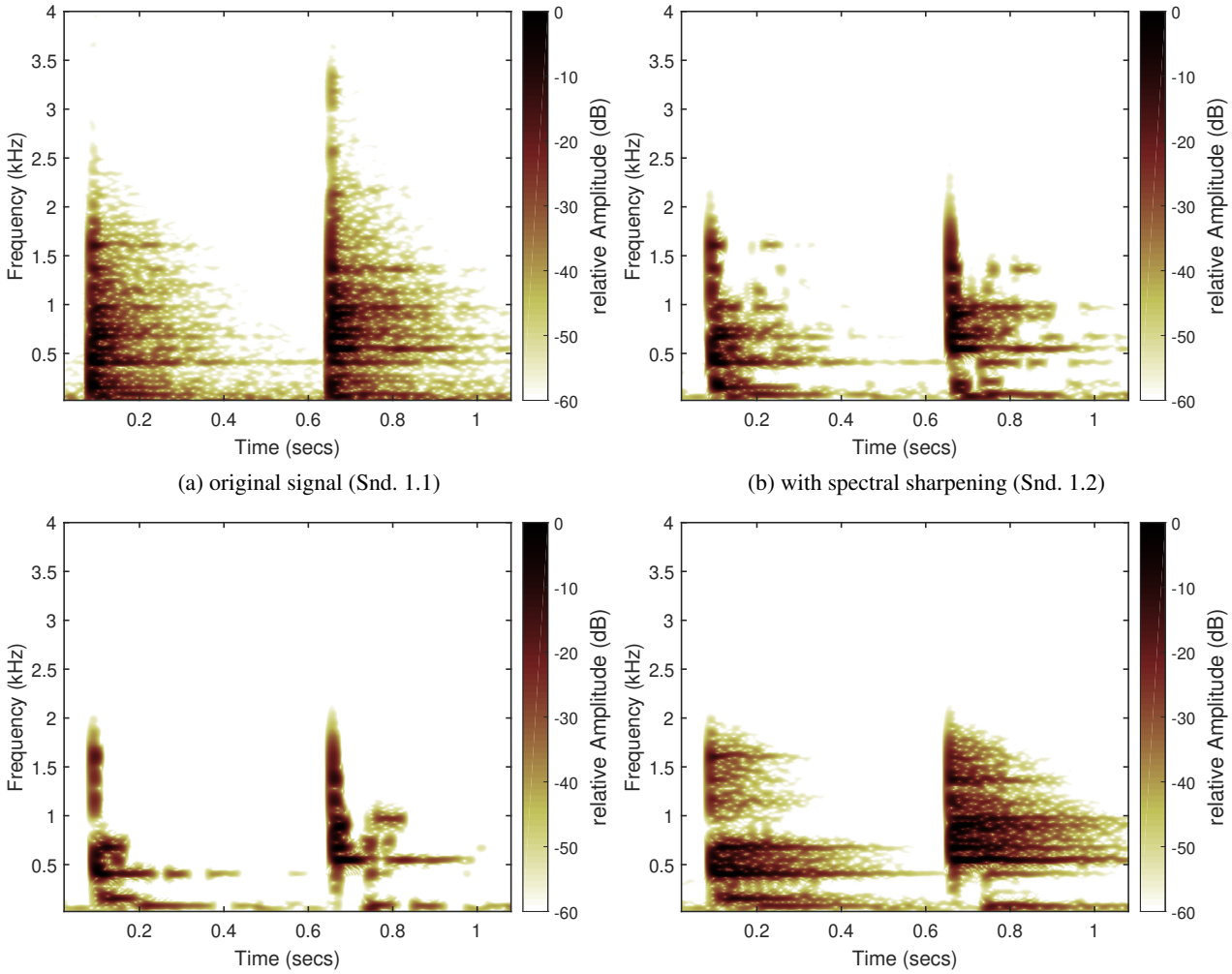
The envelope with smoothed attack $\mathrm{env_a}\{v_k[n]\}$ is fed to decay prolongation, while the residuum ($v_k[n] - \mathrm{env_a}\{v_k[n]\}$) containing only the attack part is added back to the result, leading to the output signal of decay prolongation $p_k[n]$:

$$p_k[n] = \mathrm{env_d}\Big\{\mathrm{env_a}\big\{v_k[n]\big\}\Big\} + v_k[n] - \mathrm{env_a}\big\{v_k[n]\big\} . \quad (11)$$

Due to the normalization with the original envelopes (Eq. 2) the decay is fed by intrinsic signal components of the sub-band signals in the relevant frequency region. In order to supply sufficient signal energy in the case of large SNR combined with long decay prolongation, a pink noise signal $\eta[n]$ is added to the input signal just before feeding it to the Gammatone filterbank (see block diagram in Fig. 4a); at a level below the threshold of hearing, but enough to synthesize literally infinite decay. As internal signal processing on any eligible platform offers at least 32 bit floating-point precision, a noise level of around -96 dBFS is more than enough.

A constant decay time over the whole frequency range leads to an unnatural amplification of high frequencies, as damping usually increases with frequency. We chose a rough approximation by setting T_{60} inversely proportional to the center frequency, but clipped below 1 kHz.

Sound example 1.3 and Fig. 5d show the effect of decay prolongation on the enhanced signal from Sec. 2.1.2 (Snd. 1.3 and Fig. 5 c). For this example, reverberation time T_{60} at 1 kHz was



(a) original signal (Snd. 1.1)

(b) with spectral sharpening (Snd. 1.2)

(c) with spectral dynamics expansion (Snd. 1.3)

(d) with decay prolongation (Snd. 1.4)

Figure 5: Spectrograms of a test sound in 4 conditions: (a) original recording, (b) with spectral sharpening, (c) with spectral sharpening and spectral dynamics expansion, (d) with spectral sharpening, spectral dynamics expansion, and decay prolongation.

set to 0.5 s. The time constant for transient separation was set to 7 ms. It is clearly visible and audible that relevant partials are significantly extended in time.

2.2. Temporal Contrast Enhancement

Temporal contrast enhancement is done for two reasons: (1) to make temporal structures in the sound more prominent, and (2) to compensate latency and time-smearing of the spectral contrast enhancement. Spectral ACE, as described above, always introduces some latency which is small at high frequencies but increases towards lower frequencies. This frequency-dependent group delay is acceptable for steady sounds, but it delays and smoothes any transient, transforming it to something similar to a down-chirp. Due to their broadband spectrum in combination with smoothed lateral inhibition, spectral ACE anyway effectively suppresses all transients. In order to preserve them, they must be detected as fast as possible from the input signal and mixed together with the output of spectral contrast enhancement and delay prolongation.

Transients are detected in real time by the same simple transient detection algorithm that has been used for decay prolongation (see Sec. 2.1.3). A 2nd-order high-pass filter with adjustable cutoff frequency makes the transient detection more sensitive to high-frequency content. $s_h[n]$ is the high-pass-filtered version of $s[n]$. The envelope $e_t[n]$ of the transient part of the signal is estimated via the difference of a slowly decaying envelope $e_{t,d}[n]$ and a slowly rising envelope $e_{t,a}[n]$:

$$e_t[n] = \max\left\{e_{t,d}[n] - e_{t,a}[n] - \nu\ ,\ \ 0\right\} \tag{12}$$

with threshold ν. Envelopes are computed via the two filters $\mathrm{env_d}$ and $\mathrm{env_a}$ that have been explained in Sec. 2.1.3 and Eqs. 9–10:

$$e_{t,d}[n] = \mathrm{env_d}\big\{s_h[n]\big\} \quad , \quad e_{t,a}[n] = \mathrm{env_a}\big\{e_{t,d}[n]\big\}\ . \tag{13}$$

The output signal of temporal ACE, $s_t[n]$, contains only the detected transients with their original amplitude:

$$s_t[n] = s[n] \cdot \frac{e_t[n]}{\mathrm{env_d}\big\{e_t[n]\big\}}\ . \tag{14}$$

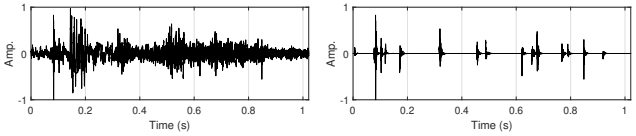


Figure 6: Waveform of the signal without (left, Snd. 2.1) and with temporal ACE (right, Snd. 2.3).

Setting the time constants $\tau_a = 3\,\text{ms}$ for env_a and $\tau_d = 7\,\text{ms}$ for env_a, seems to work well with most of the signals we tested. Threshold ν is adjusted dependent to the overall signal level.

In sound examples Snd. 1.2-1.4, the original transients are smoothed by spectral contrast enhancement. For that reason, the original transients are extracted (Snd. 1.5) and mixed to the enhanced signals. Sound examples 1.6-1.8 are the same as Snd. 1.2-1.4, respectively, but with restored transients.

In Fig. 6 and Snd. 2.3, the effect of temporal contrast enhancement is demonstrated with the machine recording from Snd. 2.1 It is clearly visible that, similar to spectral sharpening, local amplitude minima are attenuated while local amplitude maxima are retained. Note that the algorithm operates on the highpass-filtered version (cutoff frequency set to 4 kHz). The mechanic rattling thus becomes the prominent sound characteristic. A mix with the enhanced signal from spectral dynamics expansion (Snd. 2.2) leads to a spectrally and temporally enhanced signal (Snd. 2.4).

For temporal contrast enhancement, it makes no sense to apply dynamics expansion based on an absolute threshold as for spectral contrast enhancement via exponentiation — this would be a waveshaper, introducing unwanted distortion. The linear crossfade with the dry input signal actually serves as a control for the amplitude of the residuum signal between transients.

3. DISCUSSION

One might notice that the proposed ACE method does not explicitly include spectro-temporal contrast enhancement, e.g., temporal contrast enhancement on a sub-band level. Our hearing system does exactly that via contrast gain control in the auditory cortex, at timescales of about 100 ms [29]. Rabinowitz et al. define spectro-temporal contrast as "the variation in sound pressure in each frequency band, relative to the mean"; a model can be based on the standard deviation of recent sound pressure level [29]. One audible effect is that a harmonic partial which is omitted and then reintroduced may stand out perceptually for a short period of time [30]. While this is certainly a helpful feature, it must be noted that the main objective of such adaptive gain control is to compensate the very limited dynamic range of neurons. We found that spectro-temporal contrast is anyway strong with spectral contrast enhancement alone, e.g., through a possible edge effect in case of a missing partial. Even more so, if smoothing for lateral inhibition is bypassed, together with a large ρ, a clicking transient appears whenever there is a shift of spectral energy from one band to another. Due to the group delay of the filters, however, such a transient would exhibit latency that is unacceptable for short interaction sounds.

For continuous sounds where more latency can be tolerated, it might be interesting to exaggerate amplitude modulations on a sub-band level. For that goal we tried an algorithm which expands the sub-band envelopes individually while preserving their overall envelope trend [31]. While originally designed to exaggerate dissonances, it is capable to enhance also low-frequency amplitude modulations. At a closer look, however, similar results could be achieved by spectral ACE alone.

Concerning spectral contrast, both methods — spectral sharpening and spectral dynamics expansion — are essential. As soon as spectral sharpening has reached its limits (i.e., what is left are local maxima only), spectral dynamics expansion can add additional contrast by suppressing all local maxima below a certain threshold.

In a parallel configuration, spectral sharpening and spectral dynamics expansion can complement each other, producing a cartoonification of the sound. This may be illustrated by the example of human speech: By lateral inhibition, speech is basically reduced to fundamental frequency and formants; consonants are attenuated. While stops/plosives could be recovered via temporal contrast enhancement, sibilants are suppressed. Exponentiation maintains or even exaggerates consonants, including sibilants; however, it has a tendency to suppress formants, so that discrimination between vowels is lost. The solution might be a combination by taking the maximum of both outputs.

Temporal contrast enhancement as implemented here works similar to a transient shaper/designer for music production. The main difference is that we try not to exaggerate transients but to attenuate everything else. A dynamics expansion would conflict with the limited dynamic range of our hearing system, and would also produce an implausible amplification of the targeted interaction sounds. The mix of spectral and temporal ACE works well for these impact sounds sounds, but may produces quite disturbing results for more continuous stimuli such as speech.

4. CONCLUSIONS AND OUTLOOK

We introduced a new method for real-time auditory contrast enhancement, targeting at interactive applications where auditory feedback is used as part of a knowledge-making process. The method is split in two parts — spectral and temporal contrast enhancement — which can be used in parallel to focus on different auditory features. Spectral ACE is achieved in two ways which both are needed for different tasks. While the first approach is based on lateral inhibition and enhances spectral sharpness, the second enhances spectral dynamics via exponentiation. In the visual domain, these would refer to edge detection and contrast, respectively. Crucial for perceptibility of the enhanced sound is decay prolongation which provides a listener with additional time for pitch impression. Transient detection was found to be sufficient for temporal contrast enhancement. First results indicate that auditory contrast can be significantly enhanced by the proposed method.

The next step is to evaluate the multitude of parameters in order to find meaningful ranges and scalings, and ultimately reduce them to only a few intuitive controls. A parameter study is planned to find a compromise, achieving high auditory contrast while maintaining a certain degree of naturalness and plausibility of any auditory feedback. Participants will be rating the plausibility of observed interactions (audition vs. vision) through short video sequences, with different settings of ACE applied to the audio track. Recordings are taken from the Greatest Hits dataset [32], a collection of audio/video recordings of different kinds of objects and materials being hit with a drumstick.

It is further planned to evaluate the presented method concerning its primary target application: percussion. Contrary to the parameter study, interaction will be performed by the participants

themselves. The technical setup can be regarded as a special case of auditory augmentation, similar to the augmented table described in [1, 4]; however, with electronics not hidden but clearly visible, e.g., as a mic-through system. Participants will be asked to identify position and type of concealed physical manipulations (e.g., cavity or thickening) below the visible surface, via percussion with fingers or a hammer tool. Performance with ACE will be compared to the control condition without ACE; qualitative interviews should reveal further implications.

5. REFERENCES

[1] M. Weger, T. Hermann, and R. Höldrich, "Plausible auditory augmentation of physical interaction," in *ICAD*, 2018.

[2] T. Bovermann, R. Tünnermann, and T. Hermann, "Auditory Augmentation," *International Journal on Ambient Computing and Intelligence (IJACI)*, vol. 2, no. 2, pp. 27–41, 2010.

[3] K. Groß-Vogt, M. Weger, R. Höldrich, T. Hermann, T. Bovermann, and S. Reichmann, "Augmentation of an institute's kitchen: An ambient auditory display of electric power consumption," in *ICAD*, 2018.

[4] K. Groß-Vogt, M. Weger, and R. Höldrich, "Exploration of auditory augmentation in an interdisciplinary prototyping workshop," in *Forum Media Technology*, 2018.

[5] S. Papetti and F. Fontana, "Effects of audio-tactile floor augmentation on perception and action during walking: Preliminary results," in *SMC*, 2012, pp. 17–22.

[6] E. Furfaro, F. Bevilacqua, N. Berthouze, and A. Tajadura-Jimenez, "Sonification of virtual and real surface tapping: evaluation of behavior changes, surface perception and emotional indices," *IEEE MultiMedia*, 2015.

[7] J. Maculewicz, C. Erkut, and S. Serafin, "An investigation on the influence of soundscapes and footstep sounds in affecting preferred walking pace," in *ICAD*, 2015.

[8] A. Supper and K. Bijsterveld, "Sounds convincing: Modes of listening and sonic skills in knowledge making," *Interdisciplinary Science Reviews*, vol. 40, no. 2, pp. 124–144, 2015.

[9] P. Y. Ertel, M. Lawrence, and W. Song, "Stethoscope acoustics and the engineer: Concepts and problems," *Journal of the AES*, vol. 19, no. 3, pp. 182–186, 1971.

[10] T. Hermann and M. Weger, "Data-driven auditory contrast enhancement for everyday sounds and sonifications," in *ICAD*, Newcastle, U.K., 2019.

[11] P. Susini, N. Misdariis, G. Lemaitre, and O. Houix, "Naturalness influences the perceived usability and pleasantness of an interface's sonic feedback," *Journal on Multimodal User Interfaces*, vol. 5, no. 3-4, pp. 175–186, 2012.

[12] A. P. McPherson, R. H. Jack, G. Moro, *et al.*, "Action-sound latency: Are our tools fast enough?" in *NIME*, 2016.

[13] J. Yang, F.-L. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Communication*, vol. 39, no. 1-2, pp. 33–46, 2003.

[14] P. Boers, "Formant enhancement of speech for listeners with sensorineural hearing loss," *IPO annual progress report*, vol. 15, pp. 21–28, 1980.

[15] M. A. Stone and B. C. Moore, "Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligibility and quality," *Journal of rehabilitation research and development*, vol. 29, no. 2, pp. 39–56, 1992.

[16] T. Baer, B. C. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," *Journal of rehabilitation research and development*, vol. 30, pp. 49–49, 1993.

[17] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[18] A. Kral and V. Majernik, "On lateral inhibition in the auditory system," *General physiology and biophysics*, vol. 15, pp. 109–128, 1996.

[19] C. Pantev, H. Okamoto, B. Ross, W. Stoll, E. Ciurlia-Guy, R. Kakigi, and T. Kubo, "Lateral inhibition and habituation of the human auditory cortex," *European Journal of Neuroscience*, vol. 19, no. 8, pp. 2337–2344, 2004.

[20] T. Houtgast, "Psychophysical evidence for lateral inhibition in hearing," *JASA*, vol. 51, no. 6B, pp. 1885–1894, 1972.

[21] S. Coren, C. Porac, D. J. Aks, and K. Morikawa, "A method to assess the relative contribution of lateral inhibition to the magnitude of visual-geometric illusions," *Perception & psychophysics*, vol. 43, no. 6, pp. 551–558, 1988.

[22] G. Békésy, "Lateral inhibition of heat sensations on the skin," *Applied physiology*, vol. 17, no. 6, pp. 1003–1008, 1962.

[23] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[24] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[25] M. Noisternig, "Wechselwirkung von lautsprecher-mikrofon anordnungen in fahrzeugen," Dissertation, Graz University of Music and Performing Arts, 2017.

[26] T. Hermann, A. Hunt, and J. G. Neuhoff, Eds., *The sonification handbook*. Logos Verlag Berlin, Germany, 2011.

[27] T. Koumura and S. Furukawa, "Context-dependent effect of reverberation on material perception from impact sound," *Scientific reports*, vol. 7, no. 1, p. 16455, 2017.

[28] ——, "Do speech contexts induce constancy of material perception based on impact sound under reverberation?" *Acta Acustica u. w. Acustica*, vol. 104, no. 5, pp. 796–799, 2018.

[29] N. C. Rabinowitz, B. D. Willmore, J. W. Schnupp, and A. J. King, "Contrast gain control in auditory cortex," *Neuron*, vol. 70, no. 6, pp. 1178–1191, 2011.

[30] Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," *JASA*, vol. 81, no. 3, pp. 700–708, 1987.

[31] M. Hoffman and P. Cook, "Real-time dissonancizers: Two dissonance-augmenting audio effects," in *DAFx*, 2008.

[32] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2405–2413.

HEARING ARTIFICIAL INTELLIGENCE: SONIFICATION GUIDELINES & RESULTS FROM A CASE-STUDY IN MELANOMA DIAGNOSIS

R. Michael Winters

GT Center for Music Technology
Georgia Institute of Technology
Atlanta GA
mikewinters@gatech.edu

Ankur Kalra

Hop Labs
Atlanta, GA
ankur@hoplabs.com

Bruce N. Walker

GT Sonification Lab
Georgia Institute of Technology
Atlanta, GA
bruce.walker@gatech.edu

ABSTRACT

The applications of artificial intelligence are becoming more and more prevalent in everyday life. Although many AI systems can operate autonomously, their goal is often assisting humans. Knowledge from the AI system must somehow be *perceptualized*. Towards this goal, we present a case-study in the application of data-driven non-speech audio for melanoma diagnosis. A physician photographs a suspicious skin lesion, triggering a sonification of the system's penultimate classification layer. We iterated on sonification strategies and coalesced around designs representing three general approaches. We tested each in a group of novice listeners (n=7) for mean sensitivity, specificity, and learning effects. The mean accuracy was greatest for a simple model, but a trained dermatologist preferred a perceptually compressed model of the full classification layer. We discovered that training the AI on sonifications from this model improved accuracy further. We argue for perceptual compression as a general technique and for a comprehensible number of simultaneous streams.

1. INTRODUCTION

Artificial Intelligence (AI) algorithms are becoming an increasingly important part of interacting with computers [1]. Today, almost every major content provider uses machine learning, deep learning, or artificial intelligence more generally to produce their final product.

In spite of the complexity and sophistication that is required to produce a well-functioning AI system, often the information needs to be displayed to a human recipient. In these contexts, an important layer of the AI system is the perceptualization of the machine knowledge. This perceptualization can take many sensory, linguistic, or cognitive forms, and the best way to communicate will depend upon human-factors such as the context, expertise, and task goals.

In this paper, we describe a context where an AI system assists a human in the diagnosis of skin cancer from photographs of suspicious skin areas (lesions). A doctor takes a photograph of a suspicious area on their patient's skin, triggering an analysis phase by the AI system. Once the image has been processed, it generates a sonification that represents what has been sensed/classified in the image—good and bad. The doctor then uses this sound, in addition to other factors such as the patient's medical history, to determine if further tests (biopsy) or treatment is indicated.

We describe our design process for creating sounds for this AI system, which included three sonification designs and a user study with novice listeners. After describing the context around the work, we present the three designs in the order that they were created. We describe the study that we administered and our results, then finish with general design guidelines for working with AI systems that may prove useful in similar contexts.

2. BACKGROUND CONTEXT

Listening has formed a vital component of medical practice. Indeed, auscultation has been considered the first "imaging" technology [2], and the stethoscope is still routinely used by general practitioners. Doctors are trained listeners.

We worked with an algorithm that has been developed to identify melanomas from photos of skin lesions [3]. The algorithm was a deep learning convolutional neural network, and was trained on thousands of images. The algorithm was designed to produce a binary classification output: benign or malignant.

A simple auditory display strategy would be to read out a "benign" or "malignant" diagnosis for a given input image. However, we sought to use a more sophisticated sonification to provide additional information and context. We reasoned that if the sonification targeted the more subtle information behind the course benign/malignant classification, a listener might be able to understand more of the nuance behind the given classification. For example, each image might produce a unique aural signature that helps convey why the algorithm decided on its final classification.

For the purposes of design, we targeted the penultimate layer in the AI system. While the final layer of the network had a binary classification, the layer before that had 1024 nodes, each with an associated weight and image-dependent activation. Although the full system contained hundreds of layers and loops, our choice to use the penultimate layer came from the desire to have the most direct and information rich layer available. This layer also made it easy to use the final classification output.

3. DESIGN PROCESS

In the process of designing the sonification algorithm, we went through several design iterations, which manifested in three distinct design strategies. The three designs all used the penultimate

https://doi.org/10.21785/icad2019.021

layer, but differed in their underlying goal, sound design and mapping strategy.

We made a graphical user interface (GUI) to assist our exploration of the dataset, sampling of the sonification strategies, and our evaluation (See Fig. 1). In the "Training Mode," the GUI displays the image of the suspicious skin region in the upper left, the result of the classification in the upper right, and a graph representing the activations of "benign" and "malignant" nodes on the bottom. For any image, the user could listen to any of the three sonification designs by pressing a button, and control the playback speed using a knob. In the evaluation mode, the image was hidden, and the user diagnosed based only on what they heard.

4. DESIGN #1

The first sonification system used a rapid parameter mapping approach (i.e. granular synthesis) to directly sonify the 1024 nodes in the penultimate layer. Because the data were rendered in an unprocessed format, we nicknamed this design "Raw."

In this approach, the nodes were first sorted by descriptive power. Using all of the images in the dataset, we quantified each node based upon their descriptive power for either benign or malignant diagnosis. Once quantified they were sorted such that the nodes that were most positively associated with the benign images were at the beginning, and the nodes that were most positively associated with the malignant images were at the end. With this ordering in place, each node was mapped to a note whose loudness and duration was determined by the strength of that node for a given image. For example, if a given node had a strength of 1 for a given image, the note assigned to that node would play at full volume for 100ms. If the same node were to have a strength of 0 for a different image, it would have no volume and would have no duration.

In order for each node to be played with most clarity, each node was associated with a unique frequency. These frequencies were evenly distributed across the frequency spectrum according to a logarithmic mapping (i.e., mostly linear in musical note space). This choice allowed each octave to have an equal number of pitches within that octave.

In order to play all of the notes for a given image so that they might be heard, the notes were spaced out in time such that they were triggered in short succession. The amount of time between note onsets was increased in order to produce more clarity of notes, but the amount of time was decreased to limit the total amount of time that a note would ring for.

The notes were played in ascending order from low pitches (generally mapped to benign lesions) to high pitches (malignant lesions), creating a total upward glissando sound as each image played through all of its nodes.

4.1. Analysis

All together, the approach was successful in producing sounds that were different for each image. However, the overall sound was quite chromatic and dissonant due to the closely spaced notes in both frequency space and time.

Furthermore, in order to determine whether a given sound corresponded to a benign or malignant image, it was necessary to do a type of aural weight analysis, where the total amount of low frequency volume was weighed against the total amount of high frequency volume. This process was necessary because each image had a combination of low and high notes that reflected the 1024 nodes of the penultimate layer. Rapidly playing through all nodes did not make the perceptual identification easier. If anything, the resulting effect was to render the choice more difficult. For example, by hearing a mixture of high and low notes (a true reflection of the layer), a listener might be less confident when making a decision regarding whether the image was benign or malignant. By comparison to a simple mathematical number that could be calculated from every image, this approach ultimately seemed to be less useful.

5. DESIGN #2

The second design reflected a desire to make the ultimate decision of benign or malignant more clearly audible, decreasing the amount of learning and time required to make accurate aural diagnosis. To accomplish this goal, we reasoned that the sound of a malignant melanoma should be very clearly different from a benign lesion, and should have sonic qualities of loudness, roughness, dissonance, fear, or in general, "badness." This would contrast to a benign lesion, which would sound more easy-going, clear, consonant and quiet. Furthermore, the goodness or badness of the sound should correspond to the actual certainty that a given lesion would be benign or malignant. Because this design was designed to make the benign or malignant classification clear, we nicknamed it "Type."

With these design goals in place, we drew upon auditory cues from the music emotion literature [4], specifically emotion sonification [5]. In our strategy, a sonic space was modeled that would be controlled by a single "goodness"-to-"badness" dimension that was calculated by summing all of the activations and weights of the first design strategy and applying a scaling based upon the probability of correct diagnosis. This number would be either positive (benign), or negative (malignant), and magnitude would increase linearly with confidence. Because this one dimension controlled many sound parameters, this design was a one-to-many mapping strategy [6].

The timbre used as the basis of the sonification was created using modal synthesis with fixed resonant modes and decay times. In this design, the sound of a benign lesion was a simple timbre that would strike the first note, wait a few moments, and then play a note a perfect fifth above it. The decay time was controlled by the "goodness" of the classification such that a long decay time meant that a lesion was good/benign, and a short decay time indicated bad/malignant. Additionally, the amount of time in between notes was also controlled by the same dimension. The amount of time in between the two strikes was a direct indication of the confidence of being benign. For example, a lesion that was classified with confidence as being benign may have 1.5 seconds of gap between the two sounds, whereas a lesion classified as benign, but with less confidence may have 0.3 seconds of gap between the two notes. Perceptually, benign lesions sounded more "relaxed."

The sound of a lesion that was classified as malignant would sound "bad" using additional auditory cues. Continuing from the benign sound model, the decay time of each of the two strikes would be short, and the time in between the two strikes would be short as well. However, the two strikes were allowed to echo through the sound model, while simultaneously being frequency and amplitude modulated. These modulations, combined with the echo, created a sound that was aggressive, having many attacks in short succession, general roughness, and frequency instability. As

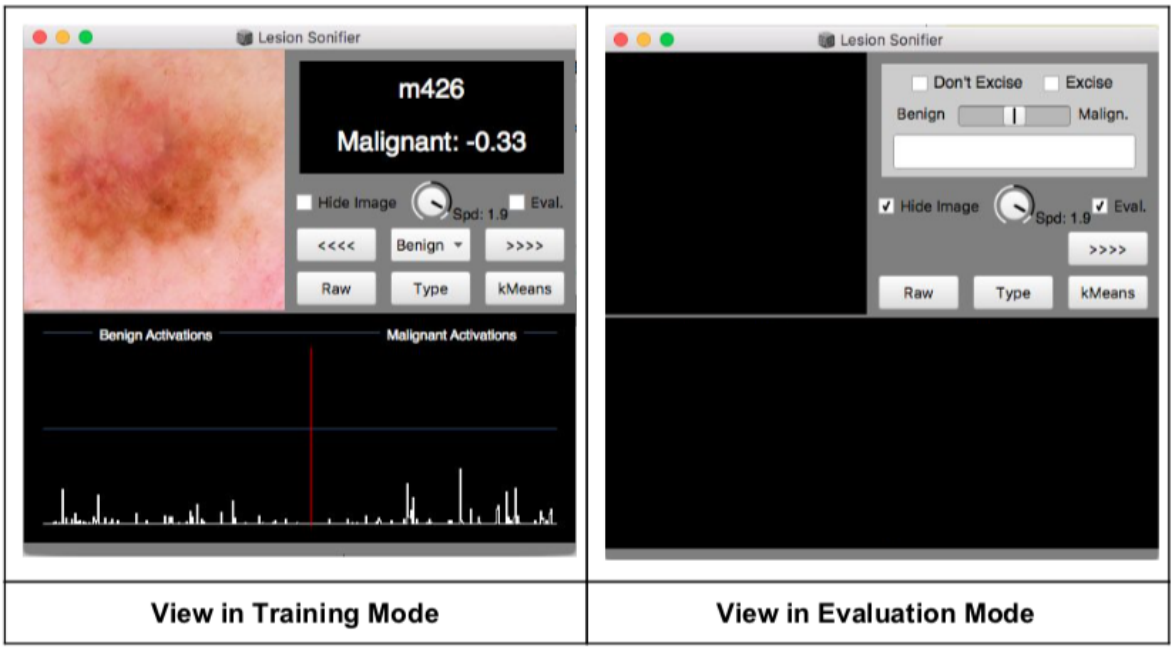


Figure 1: The GUI used for interactively exploring the dataset and sampling the three sonification designs. In the training mode, the sonification strategies are paired with an image of the skin region and AI-output. In the evaluation mode, the user makes decisions based upon the sonification alone.

with the "good" sounds, the cues used for "badness" increased in magnitude when lesions were classified as being more malignant, and decreased in magnitude when lesions were classified as being more benign.

5.1. Analysis

Having created the sonification model that represented a continuous "goodness"-to-"badness" scale, we felt that the sounds themselves were able to communicate these high-level constructs. We reasoned that the difference would be easy to explain to an untrained listener. Furthermore, the emotion-laden auditory cues would contribute to a more "embodied" [7] or tangible sonic character compared to a spoken classification.

However, the design also had weaknesses. By relying upon a single continuous number for sonification, the sound was not able to provide as much detail as the first design. The nuances in the soundscape in this design were not due to subtle differences in the data, but were instead completely determined by the magnitude of a single number. For example, if the user did not like the sound, or didn't want to use it, a simple real number could replace the sound without any information loss. Thus, the sonification in this case might be able to communicate clearly the goodness or badness, but perhaps not provide much (if any) additional unseen information.

6. DESIGN #3

After producing the first two designs, we sought a third design that could capture what we already learned from the first two and produce something in between. The third design would represent some of the subtleties of the underlying 1024 weights, but include clear acoustic cues that would differentiate benign and malignant lesions. Such a design would offer a mid-way point between the first two designs.

The initial idea for the third design came through a brainstorming session with the team members that made the deep learning classifier. We decided to look into ways to intelligently reduce dimensionality down from 1024, without dropping all the way back down to the final binary classification layer. In the end, we used a clustering approach, where a given lesion would be described by its distance to N different cluster centers. By analyzing all of the ground truth data using this method, we determined how descriptive each cluster was for being either malignant or benign in its diagnosis. For example, if a cluster center reliably predicted a malignant diagnosis with 95% accuracy, we reasoned that being close to this cluster center should have a very bad sound. Similarly, if a different cluster center had reliably predicted benign diagnosis with a 95% accuracy, it should contribute a sound that was peaceful and relaxing. Because of the clustering algorithm we used as part of this design, we nicknamed it "kMeans."

Using this approach, we decided to use fewer than 20 cluster centers, and ordered them according to their ability to predict a benign diagnosis. Each of these were then assigned to a pitch, with each pitch being a fourth above the previous pitch in ascending order. By separating the notes in ascending perfect fourths, we were assured that each cluster center would have a unique pitch, and that the overall tone produced would not be associated with any familiar chord (which would include combinations of thirds). Furthermore, by not stacking the notes on 5ths, the overall range from lowest to highest note was smaller and more compact in pitch-space.

For any particular lesion, the underlying data would be the

distance to all cluster centers. However, because a large distance meant that the lesion was not well described by that cluster center, we inverted that value to produce a new parameter: "closeness." Closeness became the variable being sonified, and was mapped to the duration of the note. If a lesion's image was close to any of the cluster centers, the sound of those cluster centers would play for a relatively long time (i.e. up to 2s), compared to clusters that were far.

Because the clusters were ordered according to their ability to predict benign images and stacked in ascending order from a base note, this meant that low notes were (again) associated with benign images and high notes were associated with malignant images. However, after listening to these, we felt that there should be additional cues for cluster centers that were malignant, that would make them not only higher in pitch but also clearly differentiated in timbre. Furthermore, we thought that it would be useful for those notes to also sound more urgent or salient. Therefore, we used a different waveform to frequency modulate the malignant cluster center sounds. The depth of the modulation was fixed relative to the center frequency, but the speed of modulation was greater for cluster centers that were more malignant.

6.1. Analysis

After producing the third design, we felt that we had produced the strongest design yet. By wrapping the design in a GUI, we were able to distribute it to a physician with experience in diagnosis. His feedback was that the sound was able to highlight very minute features in the lesion image. What he was hearing was probably the cluster centers that were not as strongly predictive, and therefore included a slower frequency modulation that might appear in a context of many sounding benign cluster centroids.

7. DEMONSTRATION VIDEOS

We made three demo videos (one for each design) and posted them online.[1][2][3] In each video, a listener uses the GUI (Fig. 1) in Training Mode to hear examples of images from different classification zones. For example, in the video "Design 2 - Type", the user begins by sampling images that have been classified "Malignant 3" (very likely malignant). At 0:42, the user switches to sampling images classified as "Benign 3" (very benign). At 1:04, the user switches to samples classified as "Borderline 0" (equally probable benign or malignant). At 1:28, the user switches to images classified as "Malignant 1" (possibly malignant). Finally at 1:45, the user switches to images classified as "Benign 1" (possibly benign). The corresponding sonification accompanies each new image.

8. STUDY

8.1. Study Purpose & Overview

Given that one physician can become very effective in utilizing the sonification tools, the question arises as to how much practice or training is required for a listener to become proficient in utilizing the sonification output for diagnostic purposes, and whether there is a difference in learnability for the three different sonification approaches. In order to study the learnability of the sonifications, we conducted a training study.

We performed a small controlled ("lab-like") study, to assess the effectiveness and learnability of the sonifications developed in this project. This was a small, initial study focused on the sonification specifically, and not on the entire diagnostic apparatus.

Listeners (not medically trained, but otherwise representative of future medical listeners) were trained to associate sonification sounds to labels (e.g., very bad, neutral, very good), then tested to assess the effectiveness of the training. They also provided subjective feedback about the sonifications.

We were looking at how intuitively the sounds represent the concepts, and how easily the listeners could learn to associate the sounds with the concepts.

Since we developed three novel sonification strategies, all different from each other, the listeners interacted with each of the sonification approaches, in random order. This within-subjects study design allowed us to compare the sonification designs for intuitiveness and training ease. We expected the three sonification approaches to differ not only in how quickly they could be learned, but the way performance evolved with practice.

8.2. Participants

Participants included seven adults (3 male, 4 female), aged 25-35; all had completed a 4-year college degree. These participants were not physicians, but were meant to be somewhat representative of medical students or (young) physicians in many respects (educated, other than medical training).

These participants were recruited in a major US city using a "friends and family" approach. All completed confidentiality agreements, but in any case they did not see any dermatology images, nor were they told anything about the ultimate purpose of the project. Participants were paid $50 for their participation.

8.3. Apparatus

The study was conducted in a commercial office space, generally on weekends and after normal business hours, to maintain confidentiality and ensure a quiet testing environment. Participants interacted with a laptop computer connected to an external monitor and external mouse and keyboard. Participants used high fidelity headphones. A bespoke software program written in SuperCollider provided a GUI through which the sounds could be played and responses recorded. Along with the sound controls, the software presented a word very bad, bad, neutral, good, very good, or a number -3, -2, -1, 0, +1, +2, +3.

8.4. Proceedure

Participants completed a brief demographics form, and executed a confidentiality / non-disclosure agreement. Within one encounter, they completed three sessions, with each session consisting of three blocks of trials. Each block of trials consisted of 21 training trials, followed by 21 testing trials. During the training phase of a block, participants saw the number (or word) and listened to the sound. During the testing phase of a block, participants heard the sound and then selected a number (or word) that they felt represented the sounds "goodness. Data about the responses were recorded for each trial, in every block and session; along with how many times a sound was listened to and the time spent at each stage of each trial.

[1] [Design #1 Example Video:] https://youtu.be/McBoGHIy7qg
[2] [Design #2 Example Video:] https://youtu.be/ay3UpoemiZs
[3] [Design #3 Example Video:] https://youtu.be/4ZZKx9FhYBk

8.5. Results

8.5.1. *Sensitivity & Specificity*

Sensitivity, here, relates to the number of correctly identified malignant lesions. It is also known as the hit rate, and loosely corresponds to the notion of accuracy in the classification task. In the case of diagnosing melanomas, it is important to catch all the malignant lesions. On the other side of the same coin, specificity refers to how often (or how rarely) a benign lesion is correctly classified as benign. Thus, it is also known as the correct rejection rate, and is important since a mis-classification of a benign lesion as malignant can lead to unnecessary tests and stress to the patient. It is clearly desirable, whenever possible, to have both a high sensitivity, and a high specificity. However, in practical applications, it is often necessary to prioritize one or the other of these performance metrics.

In this study, the sensitivity was calculated for each participant, for each sound design, and for each subsection (i.e., each block of trials). Then, the mean sensitivity was calculated across participants (i.e., collapsing on participant), for each subsection and each sonification design. Similarly, specificity was calculated for each participant for each subsection and for each sonification design. Then, mean specificity was calculated for each subsection for each sound design.

The results for mean sensitivity are presented in Figure 2, and the results for the mean specificity is displayed in Figure 3.

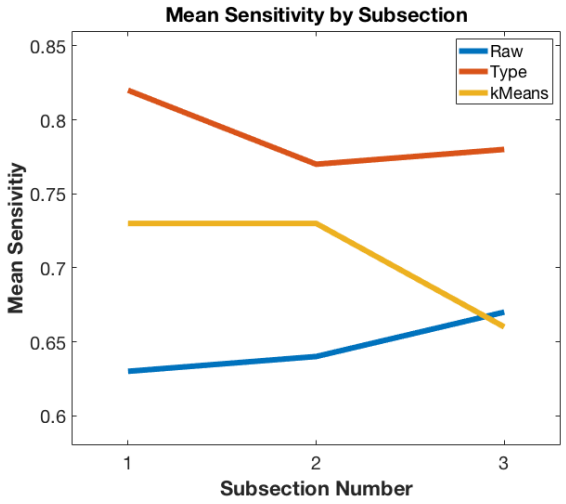


Figure 2: The mean sensitivity of each design accross the three subsections of the study.

8.5.2. *Analysis*

Within-subjects analyses of the comparisons between sessions was one of the primary measures. For our purposes we looked mostly for evidence in support of the sonification intuitiveness, but did not need statistically reliable measures.

Based upon the mean values across participants, we found that the Type sonification had the overall highest sensitivity, and the Raw design had the lowest. The kMeans and Type designs both had comparable specificity, but the Raw design was much lower.

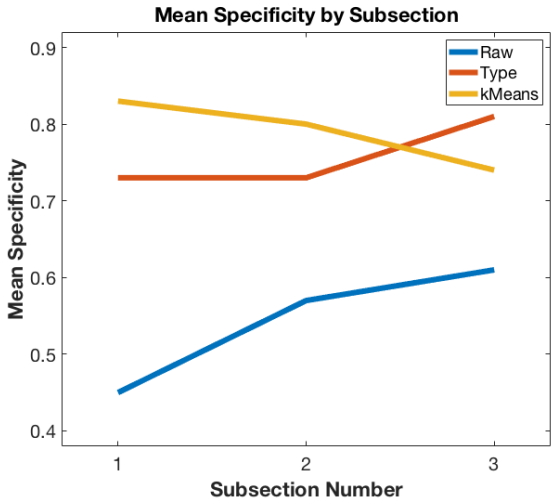


Figure 3: The mean specificity of each design accross the three subsections of the study.

Based upon the mean values across subsections, we found some effects of learning in the Raw design, but not in the Type and kMeans designs. For both Sensitivity and Specificity, the kMeans approach became more difficult with time. For the Type approach, sensitivity decreased somewhat after each subsection, but the specificity increased.

8.5.3. *Additional Data*

In addition to the performance data described above, we collected data about preferences via a post-task questionnaire. Responses for numerical questions were generally provided via a Lickert-style scale, on which 1=Completely DISAGREE; and 7=Completely AGREE.

In general, the small sample size (n=7) means that we are not really able to make statistically reliable conclusions about the preferences data. We can, however, see that in this sample of participants, there was preference for (and dislike for) each of the sound designs. There was no unanimous favorite. There was, however, less support for sound design #1, especially when it came to asking how easy it was to interpret and understand.

9. GENERAL DISCUSSION

9.1. Study

Our study was largely confirmatory of our design intuitions. Namely, the Raw approach was less useful than the Type or kMeans designs, and the Type approach provided the greatest overall accuracy for novice listeners. Although novices performed well with the kMeans approach initially, their performance decreased with time. This change may be due to them hearing more detail and nuance as they learned. A future study with expert dermatologists and images of the lesions, might produce different results.

9.2. Expert Feedback

In addition to the results from the study, our design process included almost daily interactions with a trained dermatologist. This dermatologist was invested in sonification for the domain, and would explore the dataset using the GUI after each design iteration, offering his insights and support as a specialist. From the dermatologist's experience, we learned that the third design was very "precise," often revealing nuances that were also quite subtle in the photograph.

9.3. Sonification as Layer

One of the unexpected outcomes of our work was finding that when we trained the classifier on audio from the third design, accuracy was increased relative to the AI algorithm alone [3]. The process of perceptualizing the information had in a sense formed another compression layer, which removed noise from the data and increased signal. In the future, we think that designing the outputs of an AI algorithm to be interpretable by a perceptual system (such as the auditory system), might be an effective strategy for boosting performance in an AI system.

9.4. Comprehension Guidelines

In our work, we explored different ways of perceptualizing information in the penultimate layer of a AI algorithm. Based upon our experience, we recommend the approach used in our third design. In this design, we applied compression in the form of a clustering algorithm prior to sonification. Although a mathematical algorithm might not be limited by the number of nodes or dimensions it can utilize, the same is not true for the human perceptual system. In our view, a successful compression algorithm will reduce the number of simultaneous streams to a number that will maximize listening comprehension [8]. For example, when sonifying knowledge in a complex AI system, first reduce the information space to a subset of 10-15 dimensions, of which only 2-5 will be prominent for any given input.

10. ACKNOWLEDGMENT

We'd like to thank Dr. Avi Dascalu for his engaged listening and critical feedback during the iterative design phase. We are also grateful for the 7 participants who contributed their data to the study. BosTel Technologies LLC provided funding for this project.

11. REFERENCES

[1] S. Russell and P. Norvig, *Aritificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 2009.

[2] J. Sterne, *The Audible Past: Cultural Origins of Sound Reproduction*. Durham, NC: Duke University Press, 2003.

[3] B. N. Walker, J. M. Rehg, A. Kalra, R. M. Winters, P. Drews, J. Dascalu, E. O. David, and A. Dascalu, "Dermoscopy diagnosis of cancerous lesions utilizing dual deep learning algorithms via visual and audio (sonification) outputs : Laboratory and prospective observational studies," *EBioMedicine*, vol. 40, pp. 176–183, 2019.

[4] P. N. Juslin and J. A. Sloboda, Eds., *Handbook of Music And Emotion: Theory, Research, Applications*. New York, NY: Oxford University Press, 2010.

[5] R. M. Winters and M. M. Wanderley, "Sonification of Emotion: Strategies for Continuous Auditory Display of Arousal and Valence," in *Proceedings of the 3rd International Conference on Music and Emotion*, Jyväskylä, 6 2013.

[6] A. Hunt, M. M. Wanderley, and M. Paradis, "The Importance of Parameter Mapping in Electronic Instrument Design," *Journal of New Music Research*, vol. 32, no. 4, pp. 429–440, 2003.

[7] S. Roddy and D. Furlong, "Embodied Aesthetics in Auditory Display," *Organised Sound*, vol. 19, no. 1, pp. 70–77, 2014.

[8] J. H. Schuett and B. N. Walker, "Measuring comprehension in sonification tasks that have multiple data streams," in *Proceedings of the 8th Audio Mostly Conference on - AM '13*, Piteå, 9 2013.

TOWARD SUPPORTING END-USER DESIGN OF SOUNDSCAPE SONIFICATIONS

KatieAnna E. Wolf

MeasuringU
Denver, CO, USA
katieanna.wolf@gmail.com

Rebecca Fiebrink

Department of Computing
Goldsmiths University of London
London, UK
r.fiebrink@gold.ac.uk

ABSTRACT

In this paper, we explore the potential for everyday Twitter users to design and use soundscape sonifications as an alternative, "calm" modality for staying informed of Twitter activity. We first present the results of a survey assessing how 100 Twitter users currently use and change audio notifications. We then present a study in which 9 frequent Twitter users employed two user interfaces—with varying degrees of automation—to design, customize, and use soundscape sonifications of Twitter data. This work suggests that soundscapes have great potential for creating a *calm technology* for maintaining awareness of Twitter data, and that soundscapes can be useful in helping people without prior experience in sound design think about sound in sophisticated ways and engage meaningfully in sonification design.

1. INTRODUCTION

1.1. Involving End Users in Sonification Design

The design of effective sonifications for a particular type of data and task can be challenging. Many approaches to sonification (e.g., parameter mapping sonification) present seemingly endless possibilities for ways sound may be manipulated in response to characteristics of the data [1]. A large body of work revolves around developing patterns and theories around representing data with sound [2, 3, 4, 5], and sound designers familiar with this work can benefit from such guidance in designing sonifications for end users of sonification systems. However, third-party designers often lack end users' domain knowledge and understanding of the intended use of a sonification. Further, individuals may differ in their interpretation of audio representations of data [6].

For such reasons, user-centred design strategies have become more common in auditory display research [7]. As suggested by Walker and Nees, an effective sonification requires an understanding of the listener's function and goals [5]. In practice, work by Verona and Peres shows that using a "task-based" approach—in which sonifications are designed based on the listener's task rather than based on characteristics of the data alone—was found to increase listeners' accuracy of working with sonifications [8].

However, including end users in the sonification design process is not an easy task. Collaboration between end users and expert sound designers can be labor intensive, and can involve conflicting priorities (as was found between dancers and designers in the design of a dancer movement sonification system [9]). Supporting independent end-user design is likewise challenging: for instance, non-experts may struggle to interpret specialist terminology used in sonification guidelines or design tools (e.g., terms like *frequency* and *timbre*).

We hypothesize that using real-world sounds that everyday people are already familiar with, like sounds in environmental soundscapes (sounds of the weather, animal vocalizations, and other natural sounds) might allow end users of sonification systems to design or refine sonifications while relying on terms they already know (like *bird tweet* or *running water*). Natural soundscapes have additional benefits: they are easily distinguished from background sounds, while still being able to fade out of attention without being tiring or obtrusive. Mauney and Walker found that users listening to such sonifications found the natural sounds to be "relaxing" [10]. Vickers et al. suggest that soundscapes can be "effective communication channels at the same time as being environmentally compatible and less fatiguing" [11].

1.2. Twitter Sonification

Twitter is a micro-blogging social media platform that allows users to broadcast short messages ("tweets") to the world for anyone to view. Twitter can be thought of as a data monitoring platform—each user chooses specific other users who will appear on their "timeline" of recent tweets. While the choice of data itself can be highly customized, the presentation modality of that data cannot. In addition to viewing the timeline on a Desktop or mobile device, Twitter offers sound and visual alerts to notify users of certain events of interest (e.g., a new tweet addressed to the user). Sound alerts can be muted, and the choice of alert sound can be changed, but this is the extent of sound customizability.

When enabled, auditory social media notifications can be obtrusive and—unless they drive the user to an app to view the triggering event—minimally informative. We hypothesize that using ambient soundscapes for Twitter data representation could be used as a form of *calm technology* that engages both the center and periphery of our attention, and is able to move back and forth between them [12]. That is, users could draw their attention to the soundscape and the data it represents when they wish to do so, and otherwise let the soundscape float at the edge of their periphery and maintain passive awareness. Yet it is impractical to pair every day Twitter users with professional sound designers to create such soundscapes, and unworkable—given the highly individual information characteristics of each user's feed—to create a one-size fits-all sonification suitable for all (or most) Twitter users. There

https://doi.org/10.21785/icad2019.046

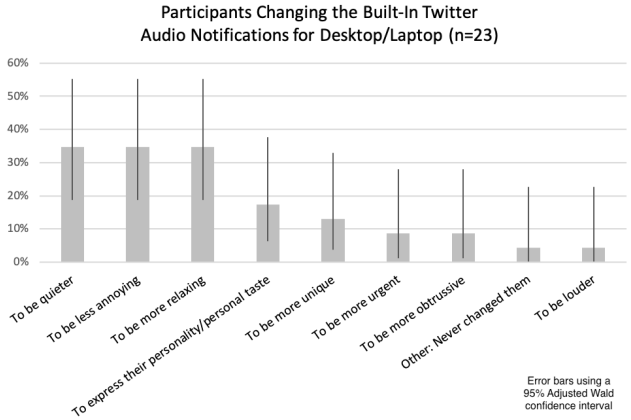


Figure 1: The percent of participants who selected various reasons for changing their Twitter audio notifications on their computer.

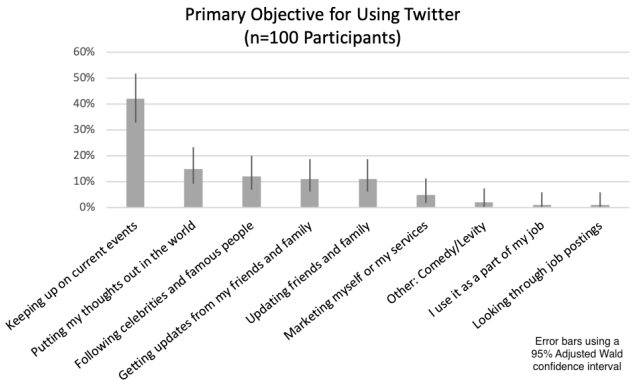


Figure 2: This figure shows the percentage of participants who selected the different primary objective for using Twitter.

fore, the design of appropriate ambient sonifications for Twitter users must leverage automation and/or interaction by the individ ual users themselves.

Some past work has used sound to represent Twitter data, mostly focusing on aesthetic presentations and performances. For instance in both *Tweetscapes* [13] and *I Hear NY4D* [14], real-time sonifications are created that utilize the content and geo-location of Twitter messages. Similarly the *Listening Machine*[1] presented a live sonification of 500 Twitter users around the UK. However, none of these projects provide users with control over the choice of Twitter that is sonified, or over the selection or design of sounds.

In the next section of the paper, we describe results of a survey that reveals how 100 Twitter users currently use audio notifications and what their primary objective is for using Twitter. Then, we de scribe a study of nine people engaging with a new tool for end-user design and customization of soundscape sonifications for Twitter data of interest. In this study, we explore how participants felt about and used soundscape sounds for representing their Twitter data. We discuss how properties of sound, user intention, and per sonal associations impact users' experience of soundscapes. We also discuss users' rationale for sonification design decisions. This work contributes to a better understanding of the utility of sound scapes for creating ambient, personal data displays, as well as a better understanding of how to support end users in designing be spoke soundscape sonifications.

2. SURVEY OF TWITTER USERS

We conducted a survey of active Twitter users to better understand the type of information people seek when they check Twitter, and how and why people use and customize Twitter audio notifications. We posted the approximately five-minute survey on Amazon Me chanical Turk and asked for active Twitter users to "Answer ques tions about your use of Twitter". We payed each participant $0.50.

2.1. Survey Results

We collected responses from 100 self-described active Twitter users. 53 were female, 47 were male, and their ages ranged from 19 to 78 years old (mean = 35.51, σ = 12.05). We collected partic ipants' self-reported information on: the number of years they had been active on Twitter (mean = 4.67, σ = 2.22), the average amount of time they spent on Twitter per day (mean = 68.87 minutes, σ = 61.22), the average number of times they accessed Twitter per day (mean = 8.29, σ = 9.01), the number of accounts they followed (mean = 426.81, σ = 714.50), and the number of Twitter followers they had (mean = 456.22, σ = 829.40).

Of the 100 participants, 64 used their desktop/laptop at least some of the time to access Twitter. Of these 64, 41 never en abled audio notifications and 23 used them at least some of the time. When asked why they didn't enable audio notifications (a multiple-choice questions in which participants were asked to se lect all options that applied), over half (24 participants) indicated that they didn't want to be interrupted, while just under half (18 participants) indicated that they used visual notifications. Other responses included "I don't want to disturb others" (8 responses) and "The information is not important" (6 responses).

We asked participants who enabled Twitter audio notifications to tell us reasons why they had ever changed the built-in sound no tifications on their device (another multiple-choice question asking participants to check all options that applied). Results for the 23 laptop/desktop participants appear in Figure 1. The three main rea sons for changing the audio notifications were: (1) to be quieter, (2) to be more relaxing, and (3) to be less annoying. Each were selected by eight respondents (34.8%).

We also asked participants to specify their primary objective for using Twitter, by selecting an option from a pre-compiled drop-down list or by writing in their own. The results appear in Figure 2 and show "Keeping up on current events" was the most common objective selected by 42 out of 100 participants (the next most se lected was "Putting my thoughts out in the world" with 15). Fi nally, we asked participants to provide an estimate of the amount of time they spent on various activities (writing tweets, responding to other tweets, reading tweets on their timeline, reading tweets on trending topics, and reading tweets on the timelines of other ac counts). While answers varied, each activity drew a response from at least 74 of the 100 participants. "Reading tweets on my home timeline" was the most popular activity overall; 87 participants re ported that they did this at least some percentage of the time, and 33 participants spent at least 50% of their time on this activity.

The survey findings further support our intuition that using

[1] http://www.thelisteningmachine.org

soundscapes to represent Twitter data may be useful, as sound scapes have been found to be relaxing [10] with the potential to be less fatiguing than other sound interfaces [11]. A data presentation modality with the capacity to be less disruptive and slightly more informative than existing audio notifications also seems attractive. These findings have also informed the design and participant se lection of our subsequent study on end-user Twitter sonification design, described in the next section.

3. SOUNDSCAPE SONIFICATION DESIGN STUDY

We next conducted a study to determine how Twitter users feel about using soundscapes to represent Twitter data of interest to them, what may be the benefits and challenges of using sound scapes to represent this data, and how people reason about and de sign with soundscapes when given the ability to design their own sonifications.

3.1. Summary of ESCaper Application

To support the study, we designed a GUI-based web application called ESCaper (Environmental Soundscape Creator). ESCaper allows Twitter users to identify specific Twitter data (e.g., accounts or hashtags of interest) that they are interested in monitoring with sound, and to create a sonification of that data using soundscapes. ESCaper uses the WebAudio API[2] to play and manipulate environ mental sounds selectively chosen from freesound.org. ESCaper provides sound samples for two soundscapes: a Forest (crickets, stream, thunder, frogs, a wolf howling, a nightingale, a fly, and an owl) and a Beach (ambient people, sea wind, waves, flock of seag ulls, single seagull, foghorn, single wave crash, ambient birds).

Users first sign in to ESCaper using their Twitter account, which authenticates our application and gives it read-only access to their Twitter data. The interface then asks users to select the data they wish to passively monitor using soundscape sounds. Specif ically, users select four of the Twitter accounts they are following and additionally specify two keywords or hashtags of interest.

ESCaper uses an automatic mapping technique from previ ous research [15] that restricts how data can be mapped to sound- scape sounds. Specifically, ESCaper allows each of the four se lected Twitter accounts to be mapped to one short-duration ("in stant") sound, such as a wolf howl or seagull call; the playback of a given sound will indicate that the corresponding account has just tweeted. Additionally, ESCaper allows each keyword/hashtag to be mapped to a longer duration ("continuous") sound (such as continuous crickets or ocean waves); one specified aspect of the sound playback (speed, gain, or left-right panning) will then rep resent changes in the number of tweets each second that contain that keyword/hashtag.

ESCaper provides two user interfaces for specifying these mappings from Twitter accounts and keywords to the selection of corresponding sound samples. Interface 1 requires uses to manu ally select a sound to correspond to each selected account or key word. Interface 2 is nearly identical, with two differences: (1) after the user specifies the accounts and keywords of interest, the inter face automatically populates itself with initial choices of sounds for each one; and (2) it includes a 'Randomize Sonification' but ton that, when pressed, randomly assigns a unique sound to each of the Twitter accounts or keywords not currently mapped to a sound. See Figure 3 in the Appendix for an image of Interface 2.

[2] https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API

Once a mapping has been fully or partially specified, users can listen to it by running it on "new" Twitter data and listening to the sound generated from the data. For the purposes of this study, however, we generated the sonification using a pre-recorded Twitter data stream that was identical for each participant (with ac counts and keywords in this historical data re-mapped to the each participant's chosen accounts). This allowed us to keep properties of the data generating the soundscapes (e.g., frequencies of tweets) consistent across all participants, no matter which accounts and keywords participants chose.

3.2. Participants

As ESCaper was designed as a computer application, we used the survey in Section 2 to identify active Twitter users who used Twit ter on their computer at least 30% of the time and also met at least two of the following three criteria: (1) They spent at least 50% of their time using Twitter on a computer; (2) Their primary objective on Twitter was to receive information rather than post information (i.e., they selected either: "Keeping up on current events", "Get ting updates from friends and family" or "Following celebrities and famous people"); (3) They spent at least 50% of their time on Twitter reading tweets rather than writing them.

We contacted 13 of our survey participants who met this crite ria as well as an additional 14 people who took a separate survey we posted with the same screening questions. Of these 27 people, nine consented to participate in our study. Seven were female, two were male, and their ages ranged from 26 to 56 (mean 39). Two participants had some past experience with sound design: one had five years of experience working with oscillators and synthesizers for tone layering and sound mixing, while another had 1.5 years of experience editing audio clips with the Audacity sound tool. Par ticipants' musical abilities ranged from 0 to 35 years of experience.

3.3. Study Procedure

A pre-study survey asked participants which Twitter account they would be using for our study and to list seven Twitter accounts that they were most likely to check on an average day, as well as five hashtags or keywords that they were most likely to search for on an average day or that they had searched for in the past month. We used this information to enable ESCaper to pre-populate its drop-down lists for selecting accounts.

For the study itself, we video chatted with participants for one hour and had them share their screens using the *appear.in* com munication tool. We recorded the screen and audio of each session (with consent from the participant), so we could reference the tran script afterwards. After a brief introduction of the facilitator and the research, the facilitator read the task scenario below:

"Imagine that you are on your computer doing your normal tasks, such as checking your email, reading online articles, online shopping, paying bills, etc. As you are focused on these tasks, you also want to be able to passively monitor specific Twitter infor mation. Your goal with this user study is to design an audio rep resentation of your Twitter data using environmental soundscapes (animal vocalizations, sounds of the weather, etc.) that will allow you to stay informed about the data on Twitter while your main focus is on another task. Please talk out loud and describe your thought process as you interact with the interfaces, with a specific focus on your design process."

The study was designed with both soundscape (Forest, Beach) and interface (Interface 1, 2) as within-subjects variables. Each

participant used Interface 1 once and Interface 2 once, with a dif ferent soundscape for each; we randomly assigned each partici pant an order of the two interfaces and an order for the sound scapes. This enabled each participant to explicitly compare the soundscapes and interfaces.

After the scenario text was read, the participant was told to use the first ESCaper interface until they felt they had created a design that accomplished the task, after which they were encouraged to open a new tab on their web browser and to spend a minute do ing a secondary task with the soundscape in the background. As our primary focus of the study was to observe how participants designed their sonifications with soundscapes (rather than on how accurately they were able to monitor the data) the single minute was for the user to assess and iterate on their initial design. Af ter a minute had passed, the facilitator asked the participant if they thought their soundscape accomplished their goal or if they wanted to make changes. Once they did not want to make any changes, the facilitator then asked the participant the following about their ex perience with that interface in an unstructured interview format:

- Overall, how easy or difficult was it for you to complete the task? (1-7 Very difficult to Very Easy)
- Overall, how enjoyable was it for you to complete the task?(1 7 Not at all enjoyable, to Very Enjoyable)
- How satisfied are you with the sound of your final design? (1-7 Not at all satisfied to extremely satisfied)
- How confident are you that you would be able to use this soni fication for passively monitoring your Twitter data? (1-7 Not at all confident to extremely confident)
- What were you focused on most while you were designing your sonification?
- What was the most challenging part of creating a design using this interface?

Then, the participant was asked to repeat the same task with the second interface assigned to them. The same questions above were asked of the participant upon completing the task with the second interface.

Finally, at the end of the study the facilitator asked a last set of questions comparing the two interfaces and soundscapes:

- How was your experience different between the two inter faces? Which of the two interfaces did you prefer? Why did you prefer that interface?
- What did you like best about the soundscape sounds? What did you like least about the soundscape sounds? Which soundscape did you like best? Why?
- (In Interface 2) How helpful was the starting sonification to your design process? How helpful was the randomization to your design process?
- How would your use of the sounds change if you were going to be listening to them over a long period of time?
- In what contexts could you imagine using this interface? How often could you imagine using this interface? Are there other phenomena on Twitter that you would want to use sound to represent?
- Did you feel like you were able to interpret the Twitter in formation from the soundscapes? What made that easy or difficult to do?

4. SONIFICATION STUDY RESULTS

4.1. Interface and Soundscape Ratings and Preferences

Participants' ratings of difficulty, enjoyment, satisfaction, and con fidence for each task are presented in Table 1. For each ESCaper interface, soundscape, and order of presentation we ran a paired-samples t-test for each rating question and found that there was no statistically significant difference between the ratings for the two soundscapes, between the ratings for the two interfaces, or be tween the ratings for the order in which the tasks were presented.

Table 1 shows each participant's preferred ESCaper interface and soundscape. Six participants preferred Interface 1 and three preferred Interface 2, while six people preferred the Beach soundscape and three preferred the Forest. Five of the six participants who preferred Interface 1 also preferred the Beach soundscape. Running the Fisher exact test for the interface and soundscape preferences we did not find a statistically significant difference.

4.2. Factors Influencing People's Experiences of Using Soundscape Sounds to Represent Twitter Data

Through the interview questions as well as through observations of participants thinking aloud while interacting with ESCaper, we learned about many of the factors that they identified as influencing their experience of hearing Twitter data represented as soundscape sounds. In this section, we present common themes that arose.

4.2.1. Relationships Between Interface Sounds and Real-World Sounds

Several participants mentioned how the sounds in the real world, might interfere with their ability to detect the sounds in the ESCa per interface, or how the sounds in the interface may lead them to react as if the sounds were occurring in the real world.

For instance, when discussing what would cause them to change the sounds in the sonification, one participant stated: *"If there was rain sounds [in the interface] and it was raining out side, then I might switch it because then I might be distracted 'Is that the rain on my computer? Is that the rain outside? Or maybe I am just sick of listening to rain', So then I might switch it to something else"*. While determining which sounds to use in their design, another participant stated: *"This thunder will have to com pete with my outside thunder, my real-life thunder"*. Additionally, two participants specifically reflected that they might react to the fly sound in the interface as if it were actually there in the real world: *"Not too long ago we had a problem with flies in here. So every time I hear a fly I instinctively duck my head..."* and *"I live in Florida and yesterday there was fly in the house. To me it may not be as ambient, it might sound like an actual fly"*.

4.2.2. Personal Associations with Sounds

In addition to participants confusing real-world sounds with the interface sounds (as with the fly samples above), participants also discussed how their personal associations with the sounds played a role in their selection process. Again the fly sample was one that drew a lot of personal associations: *"I hate bugs"*, and *"The fly reminds me too much of my past work... I used to do research with flies. It would just probably make me feel like I am at work"*. Similarly, when selecting forest sounds, one participant ruled out

Participant No.	Task No.	Task Interface	Task Soundscape	Difficulty Rating	Enjoyment Rating	Satisfaction Rating	Confidence Rating
1	1	Interface 1	Beach*	7	7	6	6
1	2	Interface 2*	Forest	7	7	5	6
2	1	Interface 2*	Forest*	7	4	5	6
2	2	Interface 1	Beach	5	4	3	5
3	1	Interface 1*	Forest	4	5	6	7
3	2	Interface 2	Beach*	6	5	4	7
4	1	Interface 2	Beach*	7	7	7	6
4	2	Interface 1*	Forest	7	7	7	7
5	1	Interface 2	Forest	6	6	7	7
5	2	Interface 1*	Beach*	6	7	7	7
6	1	Interface 1*	Beach*	7	6	6	7
6	2	Interface 2	Forest	7	6	5	5
7	1	Interface 1*	Forest	7	7	7	7
7	2	Interface 2	Beach*	7	7	6	7
8	1	Interface 2*	Beach	6	7	6	7
8	2	Interface 1	Forest*	7	7	6	7
9	1	Interface 2	Beach	5	7	6	6
9	2	Interface 1*	Forest*	7	7	7	7

Table 1: Each participant completed two tasks (one for each interface and soundscape). In this table, we present the interface and soundscape they preferred (marked with an '*') and their ratings for each task on: how difficult it was to complete the task (Difficulty), how enjoyable it was to complete the task (Enjoyment), how satisfied they were with the sound of their final design (Satisfaction), and how confident they were that they would be able to use the sonification for passively monitoring their Twitter data (Confidence).

the crickets sample, stating: *"I definitely don't like the crickets. I have tinnitus and it reminds me of the cicadas in my ears"*.

Some participant's personal associations went beyond the sample level as participants mentioned their familiarity with the soundscape as a whole. When asked which soundscape they preferred, one participant responded: *"I mean, I like the beach sounds, people love beach sounds, but I grew up in the country, so I am more used to sounds like [the forest]. It was comforting, it made me think of home, and I like that it was running in the back ground, that is was something that I could do while I am work ing, and it would keep me calm and keep me present."* When ask ing another participant what they liked best about the soundscape sounds, she mentioned the familiarity of the Beach soundscape as it sounded like her home on a Saturday morning. She even specif ically ruled out the birds sound because it didn't sound *"beachy"* enough for her and *"didn't feel as familiar"*. She also thought that the Forest soundscape would work well on a day where she wanted to be alerted, since the sounds are not as familiar to her and would not fade out of attention as easily as the Beach sounds.

4.2.3. Desire to Alert or to Passively Monitor

When asked which soundscape participants preferred, three pre ferred the Forest soundscape, while six preferred the Beach. Those who preferred the Beach soundscape often stated that it would be better for passive monitoring as the Forest was more alerting:

- *"[The beach] is more ambient sound, it's more soothing. It is nice to have in the background because it is more mellow. I feel like with the forest sounds, the animals chirping and wolves howling, it was a bit loud and more distracting, so when I am doing a secondary task I prefer the more mellow in the background."*
- *"For passive monitoring I would much prefer the beach. It is much more passive to me. The forest is, as the forest is, it is alive and active and wants your attention."*
- *"Beach soundscape was soothing. The forest definitely alerts you."*
- *"I am a beach person. I wish there was more variety in the soundscape, but the sounds were pleasant. The forest was more annoying with the animals and the birds, [whereas] these at the beach, they were more peaceful."*
- *"The [beach] was less audibly distracting when I was read ing, because the sounds were softer... The sounds were overall more pleasant. The sounds were more distinct in the forest, less ambient."*

However, similar to the woman in the previous section who would use the forest soundscape on a day when she wanted to be alerted, other participants also mentioned their interest in being able to switch to the forest interface: *"For passive monitoring I would definitely use the beach, but if was in a mode where I wanted to be alerted I would use [the forest] sounds"*. In fact, all par ticipants who preferred the Forest soundscape did so because the sounds were more easily distinguished: *"There was more diversity in the [forest] sounds. I feel like I could actually use them more. They were distinct enough I could actually tell what they were try ing to do"*, *"The sounds were more distinct, more discrete, the mix was good"* and *"I was able to hear all the sounds"*.

4.2.4. Properties of Soundscape Sounds

Eight out of the nine participants stated (while they were thinking out loud or when they were asked about the soundscape sounds) that at least some aspect of the soundscapes were *soothing*, *peace ful*, or *relaxing*. In fact from the quotations above, we can see that one of the main reasons participants felt that the Beach soundscape was better for passively monitoring data was because of its relax ing and soothing properties. Two participants specifically men-

tioned using the Beach soundscape to for meditation/relaxation: *"It's almost like I can meditate to this, relax"* and *"These sounds are almost soothing. They are almost like a relaxation tape"*.

Participants also discussed how the soundscape sounds com pared to the sounds of other interfaces. For instance, one partici pant reflected on the notification sounds used in other applications at her work. She noted that the soundscapes didn't sound like the other devices, which were more *"mechanical"*. She even stated that she had asked someone that day to turn down their notifica tions because of their sounds. Three participants specifically con trasted the soundscape sounds with other application sounds that were "jarring" and "jolting":

- *"The kind of standard g-chat or whatever "ding" is a very attention grabbing sound, but you can't really diversify that too much 'cause if you are just doing higher "ding" versus a lower "ding", that is just not going to work. The natural sounds you have more diversity in it and they are not super alarming, it is not like a big alarm that is going off in your head. For me, the nature sounds, they don't induce any anxi ety whereas some sort of buzzer would jolt my mind."*
- *"[The soundscape sounds] weren't jarring, like alerts, like an alert sound like a bell or chime. It is just something that is going on in the background and if you want you can just tune it out. Pay attention to it when you want to."*
- The third participant stated that for the software she used at work (messaging and video-chatting software), the sounds were more distracting (*"boings" and "ding dongs"*), so she doesn't use them because they are *"jolting"*. She thought that since the soundscapes were more natural that she would like them better and could see herself using them.

Only one participant expressed concern with the soundscape sounds: *"Nature sounds seem generic. Not modern enough"*.

4.3. Participants' Motivations for Sound Selection

The main action in designing the sonification of the Twitter data in ESCaper is to assign particular sounds in the soundscape to either particular Twitter accounts or to Twitter hashtags/keywords. Be low we present the common themes we observed as people thought out loud about why they were choosing particular sounds and as they responded after each task to the question "What were you focused on most while designing your sonifications?"

4.3.1. Associations Between Data and Sound

Five of the nine participants mentioned that certain aspects of the sound were reflective of certain personalities of the Twitter ac counts they were selecting them to represent. For instance, one participant described that she selected the wolf sound for a par ticular Twitter account, as she thought of them as being the pack leader, and she wanted to make sure she could hear that sound as they were Tweeting updates in the upcoming week. Another par ticipant kept the gull sound that was preselected because it *"Might be appropriate for him. He is very different in appearance with [tattoos] and stuff"*. The three other participants very heavily relied on personalities to assist them in making decisions about which sounds to use for the particular accounts:

- *What I was focused on most was really the personalities of the people and also the personalities of the hashtags, kind of like what I think of them. So like for Julian Assange, someone for WikiLeaks, the singing bird was kind of the perfect thing. For a sports reporter to be a seagull, and just be squawking all the time also fit perfectly for Jeff Howe. And Nina just alerting people, just bringing attention to stuff that people may ignore and may not realize is going on, I think is a great use of the foghorn. So I was really just matching personalities and kind of visualizing what these people sound like to me.*
- *I was thinking of the nightingale for Joy Reid because al though she is calm and discusses things calmly, when she gets excited about something she gets happy, and I think this is a very happy sound. So I am going to choose [the nightingale] for her. Wolf I am going to choose immediately for Stonekettle, because he is a veteran, ex-military guy, he is very masculine. So I think Wolf will be very good for him. Joe I think of "Fly in the Ointment". Like the idea of getting under someone's skin, or spoiling something, and since he's a comedian and since he likes to heckle other users, or annoy the crap out of them, I feel like that would be a good representation for him. He doesn't really annoy me, but I feel like that's his person, that he likes to do that to other users. So I feel that will work for him.*
- *The wolf howls to get attention and that is what Donald Trump usually does when he is on Twitter or making his speeches. For Netflix...the owl. The reason why I am going to choose the owl for Netflix is that basically when you want to watch a movie or a show you can watch anytime, but with Netflix people are usually watching it over night hours. So with the owl usually it stays up all night, as the same as the Netflix we'll be able to watch movies over night.*

However, two participants did mention that they had difficulty forming associations to soundscape sounds: *"I don't know if I would associate any animals with this account"* and *"The most challenging thing is with a lot of abstract sounds, making sure I remember what each one is."* These two participants both sug gested that being able to add in their own sounds would help to create a more memorable association to the data. For instance one participant stated *"I would be curious to try [uploading my own sounds]. Like for example the Korean [hashtag], I would proba bly upload a sound clip of either a song, or a snippet of a song, or someone speaking Korean, like that sort of thing. Something that is super duper customized."* Another participant stated: *"Be ing able to add my own music or clips, not necessarily just nature sounds. Nature sounds seem generic. Not modern enough."*

A third participant was very interested in using song clips from television shows and movies to create notifications to Twitter ac counts related to those shows and movies. Specifically, she men tioned having a song from the Star Wars films (the *Imperial March*) be associated with Twitter accounts related to the those films, and having the *Stranger Things* TV show theme song be associated with Netflix Twitter account, which is the streaming service that distributes that show. She believed that because these songs were so familiar to her, she would be able to tune them out and use them for passively monitoring the data.

4.3.2. Importance of Data

Three participants discussed selecting sounds that would be louder or more prominent in the soundscape to represent the Twitter ac counts that they were more interested in detecting:

- *Basically, I was looking to make certain ones stand out. I wanted Raw Story and The Hill to stand out because they are actual news sources, so they would be most likely to be break ing news. The other two were comedy accounts that I just like to follow.*
- *I like Jaclynhill so I gave her a little louder [sound]. The ones I really care about I would give louder [sounds].*
- *Funko does a lot of giveaways if you retweet their stuff, so I want to make sure that I heard that one.*

4.3.3. Ability to Interpret Sounds

Five participants mentioned how the properties of the sounds themselves would affect their ability to detect/interpret the data through the sounds:

- *The first thing I was thinking of was just how distinct the sounds were. So I could tell which ones were which with out any effort... If I can't detect differences in sound then it isn't going to do its job.*
- *For the first couple ones [I was focused] on the louder an imals, because if I am going off and looking at something I would want something that would get my attention.*
- *Those are both birds, so I am going to change one of them. Some of these [sounds] are too similar to the others. With all of the different birds, all of the birds would get jumbled over each other. It would be hard to tell them apart.*
- *Waves is louder so I am going to pick that. I am mainly pick ing it because it is louder than the other options.*
- *I will choose thunder. It is more attention grabbing than crick ets and stream.*

4.3.4. Sound Preference

Finally, most participants ruled out or selected particular sounds due to their preference of the sounds:

- *I would not want to listen to the fly. I hate bugs... I love wolves, so I wanted Funko and wolves to be associated to gether. Things that I like I wanted to hear more of, I know that it would call my attention to more.*
- *Secondary was how pleasant or unpleasant the sounds were, so that is why I eliminated the fly sound.*
- *I was avoiding water sounds. They were not as desirable.*
- *[I focused on] the sounds that I liked and how soothing they were.*
- *I like crickets on a summer night in a field, but this just doesn't feel comfortable.*
- *Oh no - not flies. To me [the fly] sound is annoying*

In some instances, participants were actually drawing on per ceptual properties of the sounds themselves to describe why they preferred or did not prefer certain sounds:

- *I definitely like the thunder more [than the crickets], it is more of a lower rumble, instead of the crickets which are high pitched. I feel like something that is lower is better for me personally. If it is too high pitched, it is drawing all my atten tion. I can't even read a twitter post, because it is captivating too much attention.*

4.4. Creativity and Control in the Sonification Design Process

The main differences between the two ESCaper interfaces were that Interface 2: (1) started participants with each Twitter ac count/keyword/hashtag being pre-assigned a soundscape sound, and (2) contained a 'Randomize Sonification' button to randomly select sounds for data groups that were not yet selected by the par ticipant. By asking each participant which interface they preferred and why, we were able to gain information about how useful these functionalities were to the participants and how these functionali ties affected the participants' design process, with a specific focus on the amount of control participants had on the designs.

Two of the three participants who preferred Interface 2 ex plained that the automated sound selection helped relieve the men tal burden and stress of having to make a decision: *"Sometimes if I am unsure which one to select, the randomization would choose for me and avoid the confusion and be less time consuming by picking it for me"* and *"The suggested layout was easier to use, because I didn't go through and click and decide if I want to lis ten to all of [the sounds]. Overall, it was a lot faster and just less mental energy going into making a decision. It felt very effortless. I preferred how easy and not much energy to pick what I want"*.

However, all of the participants who chose Interface 1 ex plained that being able to make up their own mind and have the creative control was the reason they preferred Interface 1:

- *I would rather make up my own mind based on what I want to hear. I liked having more choices.*
- *[Interface 2] seemed to be a simpler interface, and I didn't actually like it as much because of that. I like to have more options. The ability to be creative. You get some satisfaction when you are listening back to it to know that you put some work into it.*
- *I enjoyed [interface 1] more, even though it was more work for me, I felt more in control and I felt like it was more per sonal.*
- *For someone like me who is picky, Interface 1 would be better.*
- *Seeing them already randomized, it put in my mind that Oh, I can't play around with this.'*
- *That is exactly what I personally as a user want to see: 'Here is everything, make your choices'.*

5. DISCUSSION AND CONCLUSIONS

In this paper we explored how people use and feel about using soundscape sounds for representing social media data, specifically for communicating information about the occurrence and density of tweets of interest.

One of the challenges that participants identified with using soundscape sounds was that they could be difficult to separate from sounds in the real world. Real-world sounds could be mistaken for data, while sounds within the interface could be mistaken for events in the real-world and could trigger reactions from people as if those sounds were really occurring. Additionally, some found that certain sounds were annoying, too alerting, or just difficult to listen to. However, our study showed that when users are given control over the choice of sounds, they can avoid using sounds that they do not prefer in their soundscapes. Also, our participants described that they envisioned taking advantage of the ability to dynamically adjust their sonifications to mitigate these challenges.

For instance, some participants described using the Forest soundscape when they wanted their sonification to be more alerting and using the Beach soundscape when they wanted to more passively monitor their data.

Participants noted that a benefit of soundscapes was the relaxing and soothing quality of the sounds, especially those in the Beach soundscape. Even with these relaxing qualities, some participants still appreciated the distinctness of the sounds, and in comparison to the sounds of other applications, participants found soundscape sounds as less "jolting" and "jarring". Additionally, some participants described the soundscapes as being familiar to them (*"made me think of home"*), which made it easier for the sounds to fade into the background. As described by Mark Weiser and John Seely Brown *"The result of calm technology is to put us at home, in a familiar place. When our periphery is functioning well we are tuned into what is happening around us, and so also to what is going to happen, and what has just happened"* [12]. All of our participants felt that they would use soundscapes as a way to passively monitor their Twitter data in some form or another. In particular, participants often described the soundscapes as being something they could listen to in the background while doing other tasks (working, writing, reading articles, etc.), yet still be able to draw their attention when they wanted. Soundscapes clearly have potential to turn Twitter into a calm technology that is less intrusive and more passive. One participant described this perfectly: *"[The soundscape] is just something that is going on in the background and if you want you can just tune it out. Pay attention to it when you want to."* While a longitudinal study would be necessary to explore the use of soundscapes for passive monitoring of Twitter data over long periods of time, clearly it is worth exploring in the future.

Our study also demonstrated the feasibility of enabling end users to personalize soundscape sonifications representing their data, using a simple GUI interface. This interface gave each user the ability to individually decide what mattered most to them in the design process. We saw some participants prefer to make fewer decisions with the use of Interface 2, while others were interested in the creative process and specifically choosing sounds for particular aspects of the data. Even with only nine participants, we saw that there were several methods participants used to make design decisions including: (1) associating specific personalities or traits of their Twitter data with the sounds, (2) presenting the data they were most interested in more clearly than other data, (3) using sounds that were easiest for each of the participants to individually interpret, and (4) using the sounds that they preferred while avoiding sounds they disliked. Clearly, our participants were able to think about sound design in a sophisticated way, even with little or no experience in sound or sonification design.

6. ACKNOWLEDGMENT

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship under grant number 1148900. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Dr. Wolf is also grateful for the support of the Microsoft Research Womens Fellowship and the Gordon Y.S. Wu Fellowship from Princeton University's School of Engineering and Applied Science in enabling this work.

7. REFERENCES

[1] F. Grond and J. Berger, *The Sonification Handbook*. Logos Publishing House, 2011, ch. Parameter Mapping Sonification, pp. 363–397.

[2] A. de Campo, "A data sonification design space map," in *Proceedings of the International Workshop on Interactive Sonification (ISon)*, 2007.

[3] S. Barrass, "Sonification design patterns," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2003.

[4] J. Anderson, "Creating an empirical framework for sonification design," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2005.

[5] B. Walker and M. Nees, *The Sonification Handbook*. Logos Publishing House, 2011, ch. Theory of Sonification, pp. 9–39.

[6] B. N. Walker and B. S. Mauney, "Individual differences, cognitive abilities, and the interpretation of audiotry graphs," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2004.

[7] S. D. H. Cornejo, "Towards ecological and embodied design of auditory display," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2018.

[8] D. Verona and S. C. Peres, "A comparison between the efficacy of task-based vs. data-based semg sonification designs," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2017.

[9] S. Landry and M. Jeon, "Participatory design research methodologies: A case study in dancer sonification," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2017.

[10] B. S. Mauney and B. N. Walker, "Creating functional and livable soundscapes for peripheral monitoring of dynamic data," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2004.

[11] P. Vickers, C. Laing, M. Debashi, and T. Fairfax, "Sonification aesthetics and listening for network situational awareness," in *Proceedings of the Conference on Sonification of Health and Environmental Data*, 2014.

[12] M. Weiser and J. S. Brown, "Designing calm technology," *PowerGrid Journal*, vol. 1.1, pp. 75–85, 1996.

[13] T. Hermann, A. V. Nehls, F. Eitel, T. Barri, and M. Gammel, "Tweetscapes: Real-time sonification of twitter data streams for radio broadcasting," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2012.

[14] B. Boren, M. Musick, J. Grossman, and A. Roginska, "I hear NY4D: Hybrid acoustic and augmented auditory display for urban soundscapes," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2014.

[15] K. E. Wolf, G. Gliner, and R. Fiebrink, "End-user development of sonification using soundscapes," in *Proceedings of the International Conference on Auditory Display (ICAD)*, 2015.

8. APPENDIX

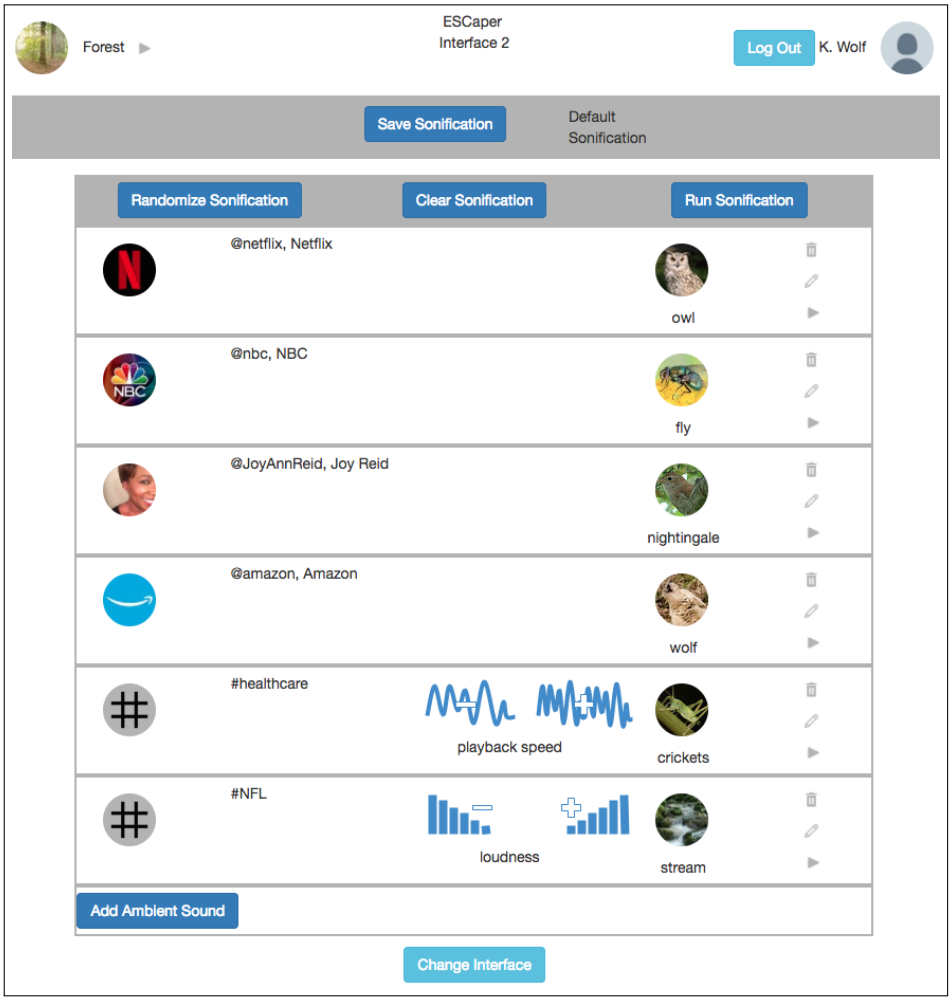


Figure 3: This image shows Interface 2 of the ESCaper application using the Forest soundscape sounds. In this sonification design the Twitter accounts for Netflix, NBS, Joy Reid, and Amazon are represented by the owl, fly, nightingale, and wolf, respectively. The change in the number of tweets including the hashtag 'healthcare' are represented by the change in playback speed of the crickets, while the change of the number of tweets including the hashtag 'NFL' are represented by the change in loudness of the stream.

PSYCHOACOUSTICAL SIGNAL PROCESSING FOR THREE-DIMENSIONAL SONIFICATION

Tim Ziemer

University of Bremen
Bremen Spatial Cognition Center
Medical Image Computing
Enrique-Schmidt-Str. 5, D-28359 Bremen
ziemer@uni-bremen.de

Holger Schultheis

University of Bremen
Bremen Spatial Cognition Center
Institute for Artificial Intelligence
Enrique-Schmidt-Str. 5, D-28359 Bremen
schulth@uni-bremen.de

ABSTRACT

Physical attributes of sound interact perceptually, which makes it challenging to present a large amount of information simultaneously via sonification, without confusing the user. This paper presents the theory and implementation of a psychoacoustic signal processing approach for three-dimensional sonification. The direction and distance along the dimensions are presented via multiple perceptually orthogonal sound attributes in one auditory stream. Further auditory streams represent additional elements, like axes and ticks. This paper describes the mathematical and psychoacoustical foundations and discusses the three-dimensional sonification for a guidance task. Formulas, graphics and demo videos are provided. To facilitate use at virtually all places the approach is mono-compatible and even works on budget loudspeakers.

1. INTRODUCTION

Just like visual displays, auditory displays can serve for various applications in numerous scenarios. Many ubiquitous auditory displays are information-poor alerts and alarms, like the doorbell, horn, siren and alarm clock [1, 2, 3]. Here, the information is binary (on/off). Other auditory displays carry more information, like the Geiger counter, which sonifies radiation on a ratio scale [4] (i.e., a continuous scale with a natural zero). Even more information is sonified in pulse-oximetry [5, 3], where heart rate is presented on a ratio scale and, simultaneously, oxygen concentration on an interval scale (i.e., a continuous scale without a natural zwro). [4] point out that developers of auditory displays have to make sure that relations in the data are heard correctly and confidently by the user. Likewise, [1] state that the perceived information should match the intended message. At the same time researchers face issues when trying to add further information to an auditory display. They experience that orthogonal, i.e., independent, acoustical parameters perceptually interact [6, 7, 8].

Some researchers avoid this issue by leveraging spatial audio. The human listener is able to localize sound sources in three-dimensional space. Hence, sonification for navigation in one-[9, 10], two-[11], and three-dimensional space[12] often employs spatial audio by means of binaural headphone presentation or loudspeaker arrays. Authors report intuitive and successful use, especially in combination with visual guidance. However, the highest localization precision of audible sources is $1 \pm 3^\circ$ along the azimuth in the front [13, 14]. It becomes worse by one order of magnitude towards the sides and in the median plane. Distance estimation has a resolution of decimeters in the near surrounding and degrades drastically with increasing distance. Listeners can distinguish some dozens locations in the horizontal plane, a little less in the median plane and along the distance dimension. For many applications, this spatial resolution is not sufficient. Binaural presentation and sound field synthesis methods further degrade localization precision and may cause additional localization phenomena, like in-head localization, front-back confusion, a vague distance perception, elevation and, sometimes, azimuth errors[15, 16]. To improve sonification in three-dimensional space, [12] added monaural cues as redundant elevation and distance cues.

Other authors suggest mapping multidimensional data completely to monaural sound attributes. The approach is promising as we can distinguish for example between 640 and $4,000$ pitches [17, ch. 7][18, p. 136], 120 loudness steps [17, ch. 7] and 250 sharpness steps (cf. [17, ch. 9] and [19]). [4] realized the need of a set of orthogonal parameters that adequately span hearing perception. [1] recognize that implementing psychoacoustical modeling and synthesis is a challenging task. It is an inverse problem because perceptual sound attributes cannot be controlled directly via signal processing. Only physical sound attributes can be manipulated. If anything, the perceptual outcome of physical parameter magnitudes can be predicted. [7] argue that psychoacoustic models provide no implementable guidelines to achieve this, because they are only valid for certain, mostly static, test signals. Still, [6] formulate two suggestions to solve the problem. The first is to create massive lookup tables to solve the inverse problem by looking up physical audio parameter magnitudes that create the desired magnitudes of all perceptual attributes. The downside of this approach is that continuous, subtle changes in desired sound attributes are created by discontinuous jumps of physical audio parameters, which creates audible artifacts. The second suggestion is the deliberate control of physical audio parameters that could be referred to as *psychoacoustic signal processing*. Physical audio parameters should either be used within ranges at which they affect exclusively one perceptual sound attribute. Or the undesired effect that one physical parameter has on a second perceptual attribute shall be counterbalanced by carefully adjusting other phys-

https://doi.org/10.21785/icad2019.018

ical audio parameters. Psychoacoustic signal processing treats the problem as a forward problem. However, it restricts the sonification designer to appropriate signal attributes and ranges. In that sense, [8] suggests a three-dimensional sonification approach that is based on timbre space in cylindrical coordinates. Elevation is represented by pitch, the radius by brightness, and discrete angles by certain musical instruments, i.e., timbre. Here, pitch is controlled by discrete notes played on the instrument and brightness by low pass filters. He points out the distinction between physical audio parameters and perceptual aspects of sound. However, he experienced that timbres in terms of instrument groups are nominally scaled (categorical), rather than rather than ordinal or even interval or ratio scaled. Hence, the angle in his approach is very vague and there exists no intuitive orthogonal or opposite direction.

In earlier studies we already presented an implementation of psychoacoustic signal processing for two-dimensional sonification [20] and some experimental validation with passive [21, 22] and interactive users [23, 24, 25] in a navigation task. In the paper at hand we modify the approach to enable three-dimensional sonification, e.g., for the sake of three-dimensional navigation. We will first explain the fundamentals of psychoacoustics and then its technical implementation in a sonification. Then, we discuss strengths and weaknesses of the approach. Finally, we give an outlook towards experimental validation, further development steps and potential application areas.

2. PSYCHOACOUSTIC SONIFICATION

For psychoacoustic sonification, principles of auditory perception are implemented in digital signal processing. The fundamentals are briefly summarized in this section. A deeper insight in psychoacoustics is given by [17]. Technical details of psychoacoustic signal processing can be found in [24, 25].

2.1. Psychoacoustics

The auditory system groups parts of incoming sounds so that the attributes of a group and their variations over time can be analyzed. This grouping is referred to as *auditory scene analysis* [26]. Complex tones are considered to exhibit at least five perceptual attributes, which are pitch, loudness, brightness, roughness, and fullness[27, ch. 32.2][28, 29]. Further attributes mentioned in the literature include subjective duration, tonalness and harmonicity [17, ch. 12 and pp. 363f][29]. All these perceived *auditory qualities* are nonlinear functions of more or less all *physical quantities*, which are the amplitude and the temporal and spectral distribution of frequencies. The physical attributes are physically independent, i.e., orthogonal, from one another. But they interfere perceptually. Likewise, the perceptual attributes are psychologically largely independent from one another, i.e., they can be regarded as orthogonal. But several physical attributes may affect them.

Auditory scene analysis: Auditory scene analysis is the psychological organization of sound and is closely related to Gestalt psychology [26, 20]. Portions of sound are integrated into *auditory streams* when they are in fair synchrony, have rather harmonic frequency relations and/or seem to come from the same spatial location. Streams are the auditory counterpart of visual objects. Streams have perceptual attributes like pitch, loudness and timbre. They are sustained over time as long as their components follow the law of continuity, proximity, common fate, timbre and closure, i.e., changes must be gradual and relations of the components must persist to some degree. Listeners can recognize some attributes of a second stream whilst consciously keeping track of another stream. To analyze details, the listener has to switch attention between streams. The presence or absence of a third stream can be noticed, but its details are not heard. Hence, sonification should be perceived as one auditory stream to ensure that all details are audible at once, without the need to switch attention [30]. Additional streams can be used to binary add simple pieces of information, like the plain presence or absence of a state or an item. When the specific task allows to concentrate on one stream at a time and switch attention if needed, two streams can be leveraged as well. Many researchers already highlighted the importance of auditory scene analysis principles and psychoacoustics in auditory display design [1, 5, 31, 8, 32].

Pitch: Perceived pitch is a multi-dimensional quality that consists of rectilinear *height* and circular *chroma*, which repeats every octave[33]. Pitch tends to be a rather linear function of fundamental frequency of harmonic complex tones. However, at fundamental frequencies above about 1 kHz the function becomes nonlinear. Sometimes, pitch is determined by signal period, e.g., in the case of a *missing fundamental*, or by the cutoff frequency, e.g., in the case of filtered peaked ripple noise [17, ch. 5]. Pitch can also be affected by signal amplitude, especially at very high or very low sound pressure levels. Pitch is neither binary nor instantaneous. A pitch strength exists, being generally higher for pure tones and complex tones compared to percussive, inharmonic, or noisy sounds. Pitch perception needs several milliseconds to build up.

Loudness: Loudness is closely related to signal amplitude [17, ch. 8][29, 34]. Increasing the amplitude or amplification gain makes a sound louder. However, different frequencies with equal amplitude tend to create different loudness sensations. Amplitude modulations slower than about 15 Hz are heard as loudness fluctuations, i.e., as *beats* [17, ch. 8 and 10][29].

Brightness: Auditory brightness mainly depends on the spectral distribution. It is considered the main contributor to timbre perception. The sensation of auditory brightness is closely related to auditory sharpness. The first is correlated with the spectral centroid [35]. The latter is explained by partial loudnesses along the Basilar membrane in the cochlea and considers masking effects[17, ch. 6][34, 29]. Shifting a spectral envelope towards higher frequencies makes a sound brighter. So does harmonic distortion, transposition towards higher frequencies, or applying a high-pass or shelving filter.

Roughness: Auditory roughness is considered another aspect of timbre [17, ch. 11][34, 29]. A rough sound is also referred to as being *jarring*, *harsh*, *raspy* or *blurred* [36, pp. 171, 349][27, ch. 32.2]. A pure tone sounds very smooth. Adding a second tone can have three effects that result from the critical bandwidth of the Basilar membrane, which is about 20% of a frequency. When its frequency deviates by more than 20% an interval may be heard, like a third or a fifth. An exception is tonal fusion, which may occur at frequency ratios of $1:2$, $1:3$, $1:4$ etc. Here, the tones may fuse and the additional tone creates the impression of changed timbre in terms of brightness, rather than the impression of an interval. When the frequency deviates by much less than the critical bandwidth, beating and a subtle pitch shift are perceived. Other frequency deviations up to 20% sound rough. The degree of roughness increases with an increasing number of non-beating frequencies within one critical bandwidth, as well as with the number

of critical bands that exhibit roughness. Roughness is easily created by means of amplitude modulations with frequencies between about 15 and 200 Hz, or by frequency modulations with frequencies around 50 Hz.

Many authors like [6, 32, 8] already identified pitch, loudness, sharpness, roughness and beating as potential parameters for psychoacoustic sonification.

Fullness: Fullness is sometimes referred to as *volume* or *sonority*[27, ch. 32.2][37, ch. 2]. Like brightness and roughness, it is an aspect of timbre. A full sound exhibits a broad frequency spectrum. The opposite is a *thin* or *narrow* sound, like that of a sinusoidal frequency or narrow band noise. Increasing the bandwidth, e.g., by means of distortion or frequency modulation, increases the degree of fullness. Decreasing the bandwidth by band pass filters lowers the degree of fullness.

Subjective Duration: Subjective duration only tends to be a linear function of physical duration in the range between 100 ms and several seconds [17, ch. 12]. For shorter sound events one has to divide signal duration by much more than 2 to half the subjective duration. When duration exceeds the capacity of the echoic memory, subjective timing becomes vague.

Further Attributes: Some studies consider *tonalness* and *harmonicity* as additional attributes of sounds [29, 34]. Tonalness ranges from tonal to noisy and depends on the number of frequency components and their amplitude and frequency relations. While a spectrum with discrete peaks sound tonal, a continuous frequency spectrum sounds noisy and loses pitch strength [17, ch. 5]. Pure and complex tones sound tonal and harmonic. The less peaks in a frequency spectrum resemble a harmonic series, i.e., $1:2:3:\ldots$, the more inharmonic it sounds. Inharmonic sounds can have one or multiple, more or less distinct, pitches. Spatial aspects of sound include source location and perceived source extent [38, 14]. The perceived source location is mainly a matter of the head-related transfer function, i.e., interaural level and time differences as well as spectral peaks and notches. These result from the sound propagation from the source to the ears, including deflections around and reflections from ears and torso. The ratio of direct sound to reverberation gives additional distance cues. Less is known about the perception of source extent. It seems to be affected by the coherence of ear signals and the presence of bass frequencies.

2.2. Sonification

In our previous two-dimensional sonification, chroma, loudness, and roughness were leveraged to communicate the direction and distance along two dimensions in Cartesian coordinates. Technical details of the implementation are provided in [25]. This sonification is the basis of our three-dimensional sonification that is described below. Table 1, summarizes the sonification parameters and their effect on the sound attributes, Fig. 1 helps for the understanding of the sound signal, Fig. 2 is a qualitative depiction of the mapping principle. Parameters are explained in the running text before their formulas are presented in Eqs. 1 to 9. Exemplary videos can be found on the first author's YouTube channel[1].

The sonification core is a harmonic *Shepard-tone* [33] with $N = 12$ partials. These partials act as carrier frequencies in terms

[1] See `https://tinyurl.com/ycwmdh8r` for previous 2D sonification and `http://tinyurl.com/y4um5odf` for the new 3D sonification.

direction	attribute	characteristic	function
left	pitch	falling speed	$\phi(\Delta x, t)$
right	pitch	rising speed	$\phi(\Delta x, t)$
up	beats	frequency	$g(\Delta y, t)$
down	fullness	degree	$\sigma(\Delta y)$
front	roughness	degree	$\beta(\Delta z)$
back	brightness	degree	$\mu(\Delta z)$

Table 1: Sonification principle and parameters when the target lies to the left/right/up/down/front/back.

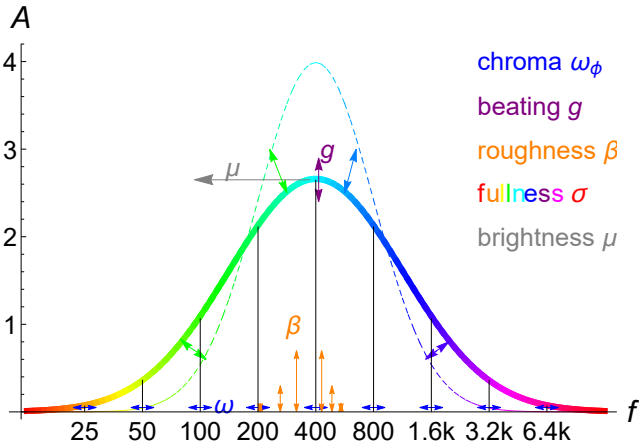


Figure 1: Sonification spectrum and parameters that modify the spectrum according to the direction and distance along the dimensions.

of additive frequency modulation synthesis. Their frequency ratios are $1:2^n$ with $n = 0,\ldots,N-1$, i.e., all frequencies are octaves of their neighbors. The amplitude $A(\phi_n(\Delta xt))$ depends on frequency f. On a logarithmic frequency scale the amplitude is a symmetric envelope that peaks in the center and is tapered off towards the sides. In Fig. 1 the envelope is represented by the colorful curve, the black vertical lines denote the carrier frequencies. The perceived pitch of Shepard tones exhibits only chroma but no height.

At the target x-coordinate the frequencies are steady. At all other x-coordinates the fundamental frequency and, accordingly, all partials move. When the target lies to the right, all frequencies rise. Many people perceive this as a rising pitch, the scientific expression is that chroma moves clockwise. In our implementation the lowest frequency is $f_0 = 3.125$ Hz, so the highest frequency will approach $f_{\max} < f_0 \times 2^{12} = 12,800$ Hz. When reaching this frequency the frequency will be shifted instantaneously back to f_0 Hz and start rising again. This way the Shepard tone creates the auditory illusion of an infinitely rising pitch while in fact it is a cyclic repetition. The envelope ensures that the lowest and highest frequencies become gradually (in)audible while the overall loudness is kept constant. While one frequency rises from f_0 to $f_{\max}$ the exact same spectrum repeats 11 times, i.e., each time a partial increased by one octave. The speed of rising determines the period of the repetition. The further the target x-coordinate lies to the right, the faster the frequencies rise, i.e., the shorter the repetition period and the higher the repetition frequency. To ensure a linear scaling, the repetition period should lie above the

lower limit of linear subjective duration, i.e., 100 ms. However, it is wise to restrict it to even longer periods, like 250 ms, i.e., 4 Hz, where the perceived fluctuation strength peaks [17, ch. 10]. This duration equals the mean syllabic length in speech, which lies between 150 and 300 ms; an order of magnitude that fairly corresponds to the integration time of 150 to 250 ms that has been found in the right non-primary auditory cortex[39].

A target to the left is denoted by decreasing frequencies accordingly, i.e., a descending pitch or counterclockwise chroma movement. Blue arrows in Fig. 1 indicate that frequencies move when the target is to the left or right.

The y-dimension is divided in two. When the target lies above, the envelope is periodically raised and reduced by a gain function $g(\Delta y, t)$. This is indicated by the purple arrow near the envelope peak in Fig. 1. This amplitude modulation is perceived as beating. The further above the higher the amplitude modulation frequency, i.e., the faster the beating. To ensure that the amplitude modulation does not create a roughness impression, the modulation frequency has to lie below 15 Hz. However, as for the pitch dimension, it is wise to keep modulation duration above the 100 ms threshold of linear duration perception and ideally even above the 150 ms integration time of the auditory system. When the target lies below, the envelope is deformed by a parameter $\sigma(\Delta y)$. The further below, the narrower the spectral bandwidth and the thinner the resulting sound. Reducing the spectral bandwidth has a drastic effect on perceived loudness. To balance out this effect, and keep loudness constant, the peak of the envelope is increased as the bandwidth decreases. Fig. 1 shows two exemplary values of σ, i.e., the thick envelope and the thin, dashed envelope. The deformation of the curve is indicated by colored arrows that connect the two curves.

The z-dimension is also divided in two. When the target lies in front, the sound becomes rough. The further to the front the rougher the sound. This is achieved by a frequency modulation of all carrier frequencies. A modulation frequency of $\omega_{\text{mod}} \approx 50$ Hz sounds rough for most carrier frequencies. The higher the modulation index $\beta(\Delta z)$, the rougher the sound. The frequency modulations not only create the impression of roughness, but also slightly increase loudness, create subtle inharmonicity and, at a extreme modulation depths, noisiness. Exemplary sidebands of one carrier frequency are illustrated as orange arrows in Fig. 1. When the target lies in the rear, the brightness is decreased, the further behind the target lies. This is achieved by a shifting the envelope towards lower frequencies by a function $\mu(\Delta z)$ illustrated in Fig. 1 by a gray arrow.

Fig. 2 illustrates how to navigate based on the sound attributes. The target lies in the center of the coordinate system. The sound tells the user where the target is. Accordingly, the symbols along the axes describe how the sound attributes change when moving along the dimensions. We refer to the current location of the user as *cursor*. When the cursor lies to the left of the target, the perceived pitch rises. This is indicated by blue angle brackets. The further to the left, the faster the pitch rises, indicated by the density of the angle brackets. Accordingly, when the cursor is far to the right of the target, the pitch will fall quickly. While approaching the target, the pitch changes more slowly. Finally, at the target x-coordinate the pitch is steady. When the cursor lies far below the target, a quick beating will be audible. While approaching the target the beating becomes slower. Finally, at the target height, loudness is steady, i.e., no beating can be heard. The beats are indicated by a fluctuating curve with a purple envelope that represents the beating frequency. When moving even further up, the

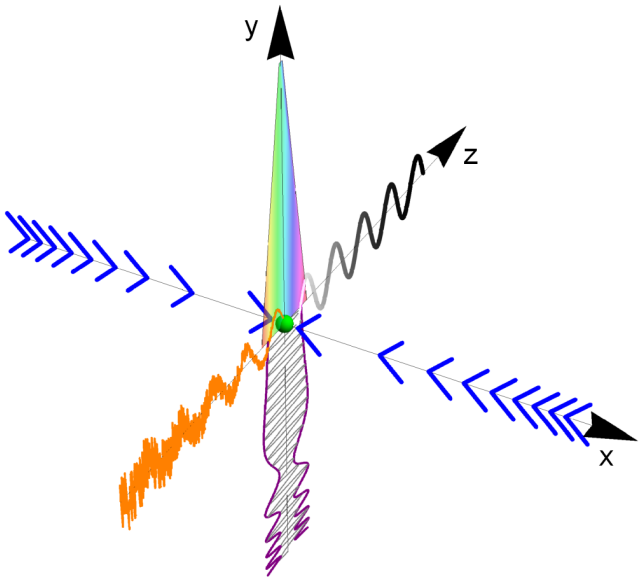


Figure 2: Navigation by sound attributes. The target lies at the origin of the coordinate system. The graphics symbolize how sound attributes change when moving along the corresponding axis.

fullness of the sound reduces more and more. This is indicated by a color spectrum that becomes narrower while the cursor goes up. When the cursor lies far behind the target, the sound is very rough. While approaching the target, roughness decreases, i.e., the sound becomes smoother. In the figure this is indicated by a jagged curve that becomes smoother towards the target. When the cursor lies far ahead of the target, the resulting sound is dull. While approaching the target the sound becomes brighter. This is indicated by the brightness level of the curve. One to three sound attributes can change at once, enabling three-dimensional navigation.

Additional elements that support navigation are added as segregated auditory streams. To enable navigation towards an extended target, a radius around the central target point is defined. The sonification guides towards the central target point. Pink noise is triggered as soon as the target region is reached. Pink noise is chosen because it is more subtle and pleasing than white noise, so it should not be perceived as a sudden alarm but as a calm confirmation sound in the background [25]. Slow beats are practically inaudible if a period takes seconds or even minutes. Likewise, there is not a specific point of maximum fullness that indicates the target y coordinate. Hence, a click is triggered as soon as the target height is reached. This click represents the x-z-plane in target-centered coordinate system. It is perceived as individual auditory stream because it is impulsive and neither belongs to the tonal Shepard tone nor to the noisy pink noise. Earlier studies showed that many novice users tend to trigger this click regularly to confirm that they are still at the target height [24]. Like fullness, roughness is gradual. To some extent, the degree of roughness is relative. A sound can be perceived as mildly rough, when heard after a very smooth sound. However, the same sound can be appear as perfectly smooth when heard directly after a very rough sound. Sometimes, it is difficult to identify the lowest possible degree of roughness, just as it may be difficult to identify the lowest audible pitch or loudness. Brightness also has no distinct minimum or

maximum. Due to the absence of a distinct point of origin, crossing the target z-coordinate is only heard because the sound that used to become fuller suddenly maintains its fullness but starts to become duller. This effect can be heard, but it is not very obvious. Hence, a short major chord is triggered every time the target z-coordinate is reached. This chord represents the x-y-plane. It confirms users that the target z-coordinate has been reached. Again, the chord is an individual auditory stream. Due to its short duration it sounds percussive, just as the click, but tonal. These three additional auditory streams only carry binary information. No attention switch is necessary to interpreted them. Note that the x-dimension is a ratio scale, because the steady pitch at $x = 0$ is an absolute zero. The same is true for the loudness fluctuation at $y = 0$. Roughness may also be considered as ratio scale, because a Shepard tone with octaves only contains no more than one frequency within each critical frequency band. However, brightness and fullness only represent an interval scale, because there is neither an obvious maximum of fullness at $y = 0$ nor of brightness at $z = 0$.

The sonification can be described by the formula

$$a_{\text{out}}(\Delta x, \Delta y, \Delta z, t) = g(\Delta y, t) \sum_{n=0}^{N} \Big[A(\phi_n(\Delta x, t), \Delta y, \Delta z) \times \cos\left[\omega_{\text{car}}(\phi_n(\Delta x, t))t + \beta(\Delta z)\cos(\omega_{\text{mod}} t)\right] \Big] , \quad (1)$$

where Δx, Δy and Δz describe the distance and direction between the current location and the designated target. The formulation is dynamic and real-time capable, so the current location and/or the target can move. It is explained in the same order as in Table 1 and the previous explanations.

The amplitude function

$$A(\phi_n(\Delta x, t), \sigma(\Delta y), \mu(\Delta z)) = \frac{e^{\frac{(\phi_n(\Delta x, t) - \mu(\Delta z))^2}{2\sigma^2(\Delta y)}}}{\sqrt{2\pi\sigma(\Delta y)}} \quad (2)$$

describes the symmetric envelope of the frequency spectrum. It is indirectly modified as a function of Δx, Δy and Δz. The phasor

$$\phi_n(\Delta x, t) = (\Delta x t + \varphi_n) \mod 1 \quad (3)$$

sweeps linearly from 0 to 1 at a frequency that depends on the distance along the x-axis. The larger the distance, the higher the sweep frequency. At negative Δx the sweep periodically decreases from 1 to 0. At a distance of $\Delta x = 0$ the sweep frequency is 0, so the phasor is a constant. In the equation mod is a modulo operation and φ_n is the initial phase of the nth carrier frequency. It is calculated as

$$\varphi_n = n/(N-1) . \quad (4)$$

Each phase has a corresponding frequency which is calculated as

$$f(\phi_n(\Delta x, t)) = f_0 2^{N\phi_n(\Delta x, t)} . \quad (5)$$

The envelope described in Eq. 2 is a Gaussian bell that peaks at the central frequency and tapers off towards the lower and upper frequencies. The phasor, Eq. 3, describes how frequencies move under this envelope. It is the only function of Δx. Perceptually, $\phi(\Delta x, t)$ controls the speed at which the pitch rises or falls.

In Eq. 1,

$$g(\Delta y, t) = \begin{cases} 1 + 0.5\sin(v_1 \Delta y t), & \text{if } \Delta y < 0 \\ 1, & \text{otherwise} \end{cases} \quad (6)$$

is an amplitude modulation. When the target lies above, it periodically modifies the gain of the signal. The modulation frequency is a function of the distance in y-direction. The further away the faster the modulation. The factor v_1 is scales the function to ensure a maximum beating frequency way below 15 Hz. Perceptually, Eq. 6 controls the speed of beating. The term $\sigma(\Delta y)$ in Eq. 2 is defined as

$$\sigma(\Delta y) = \begin{cases} \sigma_0 - v_2 \Delta y, & \text{if } \Delta y \geq 0 \\ \sigma_0, & \text{otherwise} \end{cases} . \quad (7)$$

Again, the term v_2 is just a scaling factor. The constant σ_0 is described later in relation to the brightness. The term $\beta(\Delta z)\cos(\omega_{\text{mod}} t)$ in Eq. 1 describes a nonlinear frequency modulation. The modulation index

$$\beta(\Delta z) = \begin{cases} a\Delta z^b + c, & \text{if } \Delta z < 0 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

increases the further the target lies in the frontal direction. A higher modulation index increases the number and amplitudes of sidebands around each carrier frequency, i.e., new frequencies that are distributed symmetrically around the carrier frequencies. This is perceived as an increasing degree of roughness. A linear function and a power function are chosen because a linear increase starts extreme and then becomes subtle whereas a power function starts subtle but becomes extreme. Adding a constant c creates a sudden roughness jump at $\Delta z = 0$ which makes the target height better audible. The term

$$\mu(\Delta z) = \begin{cases} \mu_0 - \Delta z, & \text{if } \Delta z \geq 0 \\ \mu_0, & \text{otherwise} \end{cases} \quad (9)$$

in Eq. 2 shifts the envelope towards lower frequencies the further the target lies to the rear. This affects the brightness attribute. The further away the lower the brightness. The terms σ_0 and μ_0 have to be balanced carefully. The first has to be large enough to create a full sound at the target height. But it has to be small enough to taper off the signal towards the highest and lowest frequencies. This ensures that the instantaneous frequency shift from f_0 to f_{max} or vice versa creates no audible click due to the discontinuity. This is especially important at low values of μ, where the envelope is shifted towards the lower frequency end. If μ_0 is too low, the sonified range along the z-dimension is too small. If μ_0 is too large the target sound becomes too bright and shrill.

3. DISCUSSION

Our mapping was mainly driven by the need of three dimensions that are

- orthogonal in perception
- integrable (i.e., integrated into one auditory stream)
- linear
- continuous

and that exhibit

- a high resolution
- an audible origin of coordinates (without the need for a reference tone) .

This has been achieved by the described sonification principle. In addition, we tried to make it "sound worse" when the user moves in the wrong direction. On the x-axis the sound near the target feels like balancing. Pitch is slowly going up or down, like tuning of a guitar. However, when moving away from the target x-coordinate the pitch changes more and more rapidly. At the outer end this sounds like a siren that indicates danger. Beating works in a similar fashion. Near the target the beats are slow and gradual. But with increasing distance the beating becomes more hectic and at the outer end it sounds chopped, like an alarm, or the sound of car parking assistants near an obstacle. Roughness is known to be a contributor to auditory annoyance and a negative contributor to the sensory euphony of sound [40, 41]. So the further away, the rougher and less pleasing the sound. We successfully implemented these perceptual sound attributes in our two-dimensional sonification approach [42, 24].

The two new half-dimensions suggested in this paper are based on perceptual sound attributes that have been examined less comprehensively in the psychoacoustic literature. Here, we concentrated on physical, i.e., acoustical considerations. Very near sources tend to sound fuller and brighter compared to remote sources. This is due to near field effects and high-frequency attenuation in air. Low frequencies are not radiated from sound sources that are small compared to the wavelength. Instead, an acoustic short-circuit occurs and the low frequency energy stays in the near field of the source. The low frequencies are only audible in close proximity to the source. Furthermore remote sources sound more dull than nearer sources because heat exchange of short wavelengths in air attenuate high frequency energy. This effect becomes audible at distances in a range of tens to hundreds of meters. So we use one of these two physical effects along the brightness half-dimension and both along the fullness dimensions.

However, this consideration lacks psychoacoustic reasoning. Intuitively, we think that a narrow sound is less pleasing than a full sound. A narrow sound seems artificial, like through a telephone, or cheap loudspeakers. A full sound on the other hand is both warm and brilliant, which is desired at least in room acoustics [14, ch. 6]. But according to the literature a duller sound is more pleasing than a brighter sound [43, 34]. Here, it may be wise to flip the direction and make a sound increasingly bright when moving away from the target.

In our sonification up to three half-dimensions are heard at the same time. With the current mapping roughness and brightness manipulations cannot co-occur, because they occur at different directions on the same axis. The same is true for beats and fullness. In our two-dimensional approach we leveraged beats and roughness for the y-dimension. However, with our new sonification design the bandwidth parameter μ may not only affect fullness but also create beating sensation. This happens as soon as pitch moves and fullness is very low. Reducing the bandwidth to 1 to 4 frequencies the frequency-dependence of loudness sensation becomes obviously audible. As a result loudness fluctuates as a function of pitch. As a solution, we decided using beating and fullness as opposite directions of the same dimension. Now, a user knows that beats of a narrow sound belong to the fullness half-dimension and beats of a full sound belong to the beating dimension.

However, this solution may introduce another issue. Fullness and brightness can co-occur in this constellation. On broadband loudspeakers or headphones, this should not be a problem. The brightness half-dimension attenuates the highest frequencies only, while the low frequencies are kept. Furthermore, the shape of the envelope is kept. As a result loudness decreases together with brightness. The fullness half-dimension attenuates both the highest and lowest frequencies and increases the volume of the frequencies that are left. Here, the fullness does not affect the loudness. Unfortunately, budget loudspeakers may not be able to radiate low frequencies at all. In this case listeners cannot hear whether the lowest frequencies are attenuated or not, i.e., a listener can hardly tell the fullness and the brightness axis apart. In this special case the two half-dimensions perceptually interfere.

4. CONCLUSION

In this paper we discussed independent aspects of complex tones and how to leverage them for multi-dimensional sonification by means of psychoacoustic signal processing. Interpretability, orthogonality and linearity play a crucial role. In earlier works we have been able to derive, implement and examine a two-dimensional sonification that satisfied these criteria. In this paper, we demonstrated how this sonification can be modified and expanded to serve for three-dimensional sonification purposes. A strength of the sonification lies in the interaction. Users instantly hear when they move in the wrong direction and can correct their motion accordingly. Furthermore, the resolution of the sonification is scalable: the highest repetition rate of chroma cycles, beating, etc., can represent distance in the order or micrometers to kilometers. Such a scaling is not straightforward with spatial audio.

5. PROSPECTS

We are in the process of carrying out the same experiments with passive listeners as in [21, 20] to examine whether the half-dimensions are in fact orthogonal. Participants hear several sounds in a row, each representing one out of 16 fields on a map. With the 2D sonification 41% of the targets had been identified correctly, which is much better than chance (i.e., $1/16 \approx 6\%$).

For the new 3D sonification, each participant only evaluates two dimensions at a time, so the experiment can be kept short, the sonification easy to learn, and the results comparable to the earlier study. After some exploration of the system and experiments with 9 users, we decided to implement one of the solutions stated above: Brightness is now increased with increasing distance, to sound worse, i.e., shrill, at a large distance. The lowest degree of fullness is increased, so that even minimum fullness does not create loudness fluctuation. In a passive listening tests with the modified 3D sonification, users recognized over 55% of the target fields correctly. We currently analyze the experiment results and prepare a paper that contains the details of the sonification implementation, experimental conditions, and results.

After that, interactive experiments [25, 24, 23, 21] with the improved three-dimensional sonification will elicit whether the new half-dimensions are perceived as linear. This is a necessity to estimate the exact angle and distance of the target. Furthermore, exposing participants to all three dimensions will elicit whether the amount of information is reasonable or overwhelming for a user, and if a navigation task is manageable or overextending. Furthermore, only interactive experiments can show how comprehensible and effective the pink noise, the click and the major chord are. If interpretable, orthogonal, linear, and not overwhelming, the psychoacoustic three-dimensional sonification is ready for blind guidance in three-dimensional space and other multi-dimensional scenarios. Our team already started to add another sonification that

communicates the direction and distance of an obstacle. This task is demanding because this new sonification has to be integrated into one auditory stream which is segregated from the guidance sonification. Then, the user can focus on the guidance sonification but occasionally switch attention to the obstacle warning sonification that pops up every time an obstacle is closer than a predefined threshold. We are aware that such a six-dimensional sonification is ambitious and we assume a long learning phase. But the outcome is worth trying because such a sonification could communicate a large amount of data even in very complex scenarios and tasks.

We think the sonification has potential to act as an assistive tool in piloting, remote vehicle control, maneuvering and docking of spacecrafts, image-guided surgery, car parking and lane keeping, and as an audio game engine. Note that three orthogonal dimensions do not necessarily have to be spatial dimensions. The dimensions could also be heart rate, oxygen concentration and blood pressure in a patient monitoring task during anesthesia, the number of downloads, citations and mentions of a scientific book in a book metrics app, the average running speed, density and viewing direction of players in team sport modeling interventions or the charging status of gasoline, electricity, and coolant in a hybrid bus. For example our two-dimensional sonification has already been transferred successfully to a pulse-oximetry task [44]. With subtle modification the sonification can be adapted for movement analysis in dance and sports training and stroke rehabilitation, auditory graphs, auditory spirit level and many more. For continuous use over hours, the sonification is certainly too obtrusive and fatiguing. In such cases it it wise to switch it off when it is not needed.

6. ACKNOWLEDGMENT

We thank the students from the CURAT project at the University of Bremen who corrected an error in the sonification calculation and developed useful graphical user interfaces and a game-like environment for demonstrating and testing our sonification.

7. REFERENCES

[1] B. N. Walker and M. A. Nees, "Theory of sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin: COST and Logos, 2011, ch. 2, pp. 9–39. [Online]. Available: http://sonification.de/handbook/

[2] J. Edworthy, S. Lexley, and I. Dennis, "Improving auditory warning design: Relationship between warning sound parameters and perceived urgency," *Human Factors*, vol. 33, no. 2, pp. 205–231, 1991. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/1860703

[3] M. Watson and P. M. Sanderson, "Intelligibility of sonifications for respiratory monitoring in anesthesia," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 45, no. 17, pp. 1293–1297, 2001. [Online]. Available: http://doi.org/10.1177/154193120104501708

[4] S. Barrass and G. Kramer, "Using sonification," *Multimedia Systems*, vol. 7, no. 1, pp. 23–31, 1999. [Online]. Available: http://doi.org/10.1007/s005300050108

[5] J. P. Bliss and R. D. Spain, "Sonification and reliability — implications for signal design," in *ICAD*, Montréal, Jun 2007, pp. 154–159. [Online]. Available: http://hdl.handle.net/1853/50028

[6] S. Ferguson, D. Cabrera, K. Beilharz, and H.-J. Song, "Using psychoacoustical models for information sonification," in *12th International Conference on Auditory Display*, London, Jun 2006. [Online]. Available: http://hdl.handle.net/1853/50694

[7] J. E. Anderson and P. Sanderson, "Sonification design for complex work domains: Dimensions and distractors," *Journal of Experimental Psychology: Applied*, vol. 15, no. 3, pp. 183–198, Mar 2009. [Online]. Available: http://doi.org/10.1037/a0016329

[8] S. Barrass, "A perceptual framework for the auditory display of scientific data," in *International Conference on Auditory Display*, Santa Fe, Nov 1994, pp. 131–145. [Online]. Available: http://hdl.handle.net/1853/50821

[9] D. Black, J. Hettig, M. Luz, C. Hansen, R. Kikinis, and H. Hahn, "Auditory feedback to support image-guided medical needle placement," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 9, pp. 1655–1663, 2017. [Online]. Available: http://doi.org/10.1007/s11548-017-1537-1

[10] F. Nagel, F.-R. Stöter, N. Degara, S. Balke, and D. Worrall, "Fast and accurate guidance – response times to navigational sounds," in *ICAD*, New York, NY, Jun 2014. [Online]. Available: http://hdl.handle.net/1853/52058

[11] A. Vasilijevic, K. Jambrosic, and Z. Vukic, "Teleoperated path following and trajectory tracking of unmanned vehicles using spatial auditory guidance system," *Applied Acoustics*, vol. 129, pp. 72–85, 2017. [Online]. Available: http://doi.org/10.1016/j.apacoust.2017.07.001

[12] T. Lokki and M. Gröhn, "Navigation with auditory cues in a virtual environment," *IEEE MultiMedia*, vol. 12, no. 2, pp. 80–86, April 2005. [Online]. Available: http://doi.org/10.1109/MMUL.2005.33

[13] J. Blauert, *Spatial Hearing. The Pychophysics of Human Sound Source Localization*, revised ed. Cambridge , MA: MIT Press, 1997.

[14] T. Ziemer, *Psychoacoustic Music Sound Field Synthesis: Creating Spaciousness for Composition, Performance, Acoustics, and Perception)*, ser. Current Research in Systematic Musicology. Cham: Springer, 2019, vol. 7.

[15] F. Rumsey, "Spatial audio. binaural challenges," *J. Audio Eng. Soc.*, vol. 42, no. 1, pp. 798–802, 2014.

[16] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013. [Online]. Available: http://doi.org/10.3813/AAA.918643

[17] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*, second updated ed. Berlin, Heidelberg: Springer, 1999. [Online]. Available: http://doi.org/10.1007/978-3-662-09562-1

[18] F. Scheminzky, *Die Welt des Schalls*. Salzburg: Das Bergland-Buch, 1943.

[19] F. Pedrielli, E. Carletti, and C. Casazza, "Just noticeable differences of loudness and sharpness for earth moving machines," in *Proceedings of the European Conference on Noise Control 2008 (EURONOISE 2008)*, Paris, June 2008, pp. 1231–1236.

[20] T. Ziemer, D. Black, and H. Schultheis, "Psychoacoustic sonification for tracked medical instrument guidance," *Proceedings of Meetings on Acoustics*, vol. 30, 2017. [Online]. Available: http://doi.org/10.1121/2.0000557

[21] T. Ziemer, "Two-dimensional psychoacoustic sonification," in *33. Jahrestagung der deutschen Gesellschaft für Musikpsychologie (DGM)*, F. Olbertz, Ed., Hamburg, Sep 2017, pp. 60–61. [Online]. Available: https://www.researchgate.net/publication/319778727_Two-dimensional_psychoacoustic_sonification

[22] T. Ziemer and D. Black, "Psychoacoustically motivated sonification for surgeons," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 1, pp. 265–266, Jun 2017. [Online]. Available: http://doi.org/10.1007/s11548-017-1588-3

[23] T. Ziemer and H. Schultheis, "Perceptual auditory display for two-dimensional short-range navigation," in *Fortschritte der Akustik — DAGA 2018*. Munich: Deutsche Gesellschaft für Akustik, Mar. 2018, pp. 1094–1096.

[24] ——, "Psychoacoustic auditory display for navigation: an auditory assistance system for spatial orientation tasks," *J. Multimodal User Interfaces*, vol. Special Issue: Interactive Sonification, 2018. [Online]. Available: http://doi.org/10.1007/s12193-018-0282-2

[25] T. Ziemer, H. Schultheis, D. Black, and R. Kikinis, "Psychoacoustical interactive sonification for short range navigation," *Acta Acust. united Ac.*, vol. 104, no. 6, pp. 1075–1093, 2018. [Online]. Available: http://doi.org/10.3813/AAA.919273

[26] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[27] A. Schneider, "Perception of timbre and sound color," in *Springer Handbook of Systematic Musiwology*, R. Bader, Ed. Berlin, Heidelberg: Springer, 2018, ch. 32, pp. 687–726. [Online]. Available: http://doi.org/10.1007/978-3-662-55004-5_32

[28] W. Lichte, "Attributes of complex tones," *J. Exp. Psychol.*, vol. 28, pp. 455–480, 1941.

[29] T. Ziemer, Y. Yu, and S. Tang, "Using psychoacoustic models for sound analysis in music," in *Proceedings of the 8th Annual Meeting of the Forum on Information Retrieval Evaluation*, ser. FIRE '16, P. Majumder, M. Mitra, J. Sankhavara, and P. Mehta, Eds. New York, NY, USA: ACM, Dec 2016, pp. 1–7. [Online]. Available: http://doi.org/10.1145/3015157.3015158

[30] J. Anderson and P. Sanderson, "Designing sonification for effective attentional control in complex work domains," in *Proc. Human Factors and Ergonomins Society 48th annual meeting*, New Orleans, LA, Sep 2004. [Online]. Available: http://doi.org/10.1037/e577082012-006

[31] S. Barrass and V. Best, "Stream-based sonification diagrams," in *ICAD*, Paris, Jun 2008. [Online]. Available: http://hdl.handle.net/1853/49945

[32] G. Parseihian, C. Gondre, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "Comparison and evaluation of sonification strategies for guidance tasks," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 674–686, April 2016. [Online]. Available: http://doi.org/10.1109/TMM.2016.2531978

[33] R. N. Shepard, "Circularity in judgments of relative pitch," *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964. [Online]. Available: http://doi.org/10.1121/1.1919362

[34] W. Aures, "Berechnungsverfahren für den sensorischen wohlklang beliebiger schallsignale (a model for calculating the sensory euphony of various sounds)," *Acustica*, vol. 59, no. 2, pp. 130–141, 1985.

[35] E. Schubert and J. Wolfe, "Does timbral brightness scale with frequency and spectral centroid?" *Acta Acustica united with Acustica*, vol. 92, no. 5, pp. 820–825, 2006. [Online]. Available: https://www.ingentaconnect.com/content/dav/aaua/2006/00000092/00000005/art00019

[36] H. von Helmholtz, *On the sensations of tone as a physiological basis for the theory of music*, 2nd ed. London: Longmans, Green, and Co., 1885.

[37] J. Meyer, *Acoustics and the Performance of Music. Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers*, 5th ed. Bergkirchen: Springer, 2009. [Online]. Available: http://doi.org/10.1007/978-0-387-09517-2

[38] T. Ziemer, "Source width in music production. methods in stereo, ambisonics, and wave field synthesis," in *Studies in Musical Acoustics and Psychoacoustics*, ser. Current Research in Systematic Musicoogy, A. Schneider, Ed. Cham: Springer, 2017, vol. 4, ch. 10, pp. 299–340. [Online]. Available: http://doi.org/10.1007/978-3-319-47292-8_10

[39] D. Poeppel, "The analysis of speech in different temporal-integration windows: cerebral lateralization asÔasymmetric sampling in time," *Speech Communication*, vol. 41, pp. 245–255, 2003.

[40] W. Ellermeier, A. Zeitler, and H. Fastl, "Predicting annoyance judgments from psychoacoustic metrics: Identifiable versus neutralized sounds," in *The 33rd International Congress and Exposition on Noise Control Engineering (inter-noise)*, Prague, Aug. 2004.

[41] W. Aures, "Ein Berechnungsverfahren der Rauhigkeit (a procedure for calculating auditory roughness)," *Acta Acust. united Ac.*, vol. 58, no. 5, pp. 268–281, 1985. [Online]. Available: https://www.ingentaconnect.com/content/dav/aaua/1985/00000058/00000005/art00005

[42] T. Ziemer and H. Schultheis, "A psychoacoustic auditory display for navigation," in *24th International Conference on Auditory Displays (ICAD2018)*, Houghton, MI, June 2018. [Online]. Available: http://doi.org/10.21785/icad2018.007

[43] B. Cardozo and R. van Lieshout, "Estimates of annoyance of sounds of different character," *Appl. Acoust.*, vol. 14, no. 5, pp. 323–329, 1981.

[44] S. Schwarz and T. Ziemer, "A psychoacoustic sound design for pulse oximetry," in *The 25th International Conference on Auditory Display (ICAD2019)*, Newcastle, June 2019.

Extended Abstracts

LONDON BUS TUNES: USING SOUND TO IMPROVE THE SAFE NAVIGATION OF LONDON'S BUS SYSTEM

Dr Sara Adhitya

University College London,
Department of Civil, Environmental and Geomatic Engineering
London, WC1E 6BT, UK
s.adhitya@ucl.ac.uk

ABSTRACT

This work-in-progress introduces a proposal to incorporate sound in the passenger navigation system of the London Bus. First, we present the problems of accessibility concerning London's complex bus system. Then, we introduce our proposal of using sound and sonification in particular to aid in the navigation of London's bus system. We explain our sonification strategy and describe a recent preliminary trial of our sonification prototype, implemented as an installation during an accessibility event held by Transport for London at the ExCel centre in London on 19 March 2019. We discuss the feedback obtained from this trial and conclude with proposed future work in terms of both the development of our sonification strategy as well as its implementation in London's public transport system.

1. INTRODUCTION

Consisting of 20,000 bus stops and 800 bus routes, navigating London's bus system can be a complex task for most passengers. Due to the sheer number of connections, getting on the right bus and off at the right stop can be difficult whether you are: unfamiliar with the city in general; travelling a new route; half asleep after a hard day's work; a non-English speaker; suffering from navigational challenges due to memory loss such as dementia; struggling to see in an over-crowded bus; and not least with a visual impairment. While there are many mobile navigation systems on the market developed specifically for the visually-impaired, there are clear benefits for all passengers in improving the overall legibility of the bus infrastructure system.

On 19 March 2019, Transport for London (TfL) [1], the city's integrated transport authority, held its biggest and most accessible transport event entitled Access All Areas at the ExCel centre. This event aimed to showcase innovative ways in which London's public transportation could be made more accessible. Furthermore, TfL has recently prioritised the safety of riding London's buses, as outlined under the Vision Zero action plan of its Transport Strategy [2], with the aim to eliminate serious injuries or death on all London Buses by 2030 [3]. With almost 80% of casualties from accidents involving buses consisting of bus occupants or pedestrians [3], improving the accessibility and legibility of the system clearly has an important role to play in achieving this goal. Thus in this paper, we propose the use of sound to improve the safe navigation of London's bus system.

2. BACKGROUND: Sound and urban transportation

Sound is essential to navigation, particularly when vision is occupied or impaired, yet is an underutilised medium in the design of London's public transport system. While there is an announcement system in place on both tube trains and buses, this is limited to spoken announcements which can be difficult for non-English speaking passengers to understand as well as to hear in the context of noisy environments. Furthermore, psychoacoustic studies indicate that a difference in timbre is needed in order to distinguish between different auditory streams, which makes speech-based auditory cues problematic in crowded environments. Thus we propose the use of non-verbal auditory cues.

Non-verbal auditory cues have proven to be useful not only in helping navigation, but in contributing to the overall soundscape as well as creating a sonic identity for urban transportation systems. One notable example is the Yamanote train line in Tokyo, Japan, which has become famous for its musical tunes used to announce each stop. These tunes have also been recognised to have the added benefit of calming down pedestrian traffic flows and leading to a reduction in accidents [4].

London's bus network spans a much wider area than a tube network can, and thus can reach an even wider population. Yet there is currently no strategy concerning the consistency of its sound design. Thus we propose the development of a sound design strategy using non-speech auditory cues, not only to aid in the navigation of London's bus system, but develop its sonic identity while contributing positively to the overall urban soundscape.

3. THE ISSUES

In navigating any bus system, there are two key points of interchange that can be seen as problematic: a) boarding a bus; and b) alighting at a bus stop. The top three causes of casualties involving buses are due to standing, boarding and alighting [3]. In this section, we discuss the navigation and safety issues during boarding and alighting.

a) Signalling and boarding a bus

When waiting at a bus stop, it can be very difficult to identify which bus is arriving from a distance due to issues of legibility. Often the wrong bus is flagged down unnecessarily

https://doi.org/10.21785/icad2019.077

or at the last minute, leading to abrupt breaking and an increase in accidents. Many road accidents are caused by the bus veering and hitting pedestrians or cyclists. It is not uncommon that passengers board the wrong bus and realise only after it has begun moving.

b) Signalling and alighting from a bus

Once on the bus, it can be very difficult to know when to ring the bell and at which stop to alight. Passenger uncertainty often results in signalling the wrong stop, causing unnecessary breaking. Speaking to the driver whilst still on the move can also be distracting to the driver, and getting up earlier than necessary has been shown to increase the rate of falls. More than double the casualty value of injuries have been shown to occur while standing as opposed to being seated during a collision [3].

4. PROPOSAL

To mediate the navigation and safety problems encountered at these two points of interchange, in this section we propose the use of auditory cues in 2 main ways:

4.1. Indicating which bus is arriving at a bus stop

By announcing the arrival of a bus using sound from speakers within the bus stop, we can help the passenger know when to signal the bus. This can help in the case of inability to see the bus number while ensuring adequate time for the driver to stop. When the bus doors open, the same tune will be played from the bus itself, ensuring that the correct bus is boarded in the case of multiple arriving buses.

4.2. Indicating which bus stop is approaching

While there are currently spoken announcements which announce each stop, these can be difficult to distinguish and understand, particularly if English is not the first language or the background noise is high. By indicating the approaching bus stops using non-verbal auditory cues, we aim to improve the recognition of stops, and thus the timing of signalling and the preparation of alighting.

5. METHODOLOGY

Due to the large dataset of 20,000 bus stops and 800 bus routes, we chose to use the acoustic communication technique of sonification: the representation of data in sound [5]. In particular, we utilised the method of parameter-mapping, involving the mapping one of set of parameters to an audible set of parameters.

With the need to identify each bus and bus stop in the system, we utilised TfL's own identification system of 1 to 3 digit numbers for individual bus routes, and 5 digit numbers for individual bus stops. We then mapped each digit of each number to a specific tone. With 10 different digits to map, we utilised a 9-note blues scale: a chromatic variation of the major scale with a flattened third and seventh [6] which is shown in Figure 1 and can be heard at the following link:
https://www.dropbox.com/s/0d2kzha5cxlppp6/9%20note%20blues%20scale.mp3?dl=0

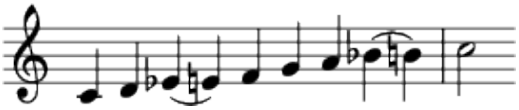


Figure 1: 9-note blues scale [7]

This gave us 10 different pitches which could be used to map each digit, and allowed us to generate a unique 1 to 5 digit auditory cue for each. At this preliminary stage, we utilised acoustic modes of sound production involving improvisation of a tune based on these notes and played on a flute.

For example, Bus 460 would be based on the notes E, G, C and sound like this:

https://www.dropbox.com/s/smnmhrms5x1yxlb/bus%20460%20-%20EGC.mp3?dl=0

whereas Bus stop 58903 would sound as follows: https://www.dropbox.com/s/ouxupq7kxnw49tp/Bus%20stop%2058903.mp3?dl=0

6. INSTALLATION DESIGN

Transport for London invited us to exhibit our proposal at their recent Access All Areas (AAA) exhibition [8], aimed at showcasing more accessible public transport to people of all abilities. Held at the ExCel Centre in March 2019, and with over 1500 attendees on the day, the AAA exhibition was an opportunity to introduce our idea to a large range of people of all capabilities, gauge interest and receive both spoken and written feedback.

Figure 2: Sound installation at the AAA exhibition, ExCel Centre, London, 19 March 2019

We were provided with a 3m by 6m exhibition space in which to communicate our idea. In order to present our sonification proposal with as much context as possible, we created a sound installation in which the participants could imagine themselves in each scenario: a) seated at a bus stop while waiting for a bus; and b) seated on a bus while waiting to alight. The scenarios were displayed in video format on a digital screen provided by the organisers.

6.1. Scenario 1: view from the bus stop

The first scenario showed two buses (Bus 325 and Bus 241) arriving at a bus stop in Stratford, east London. After introducing the sounds of each bus (film available here: https://www.dropbox.com/s/yxwvowqiunr2k93/Scenario%201%20-%20learn%20bus%20sounds.mp4?dl=0) , the participants were asked to identify a particular bus arriving by ear (audio file available here: https://www.dropbox.com/s/as1sy9jx02v93gw/Bus%20241%20is%20arriving.mp3?dl=0).

6.2. Scenario 2: view from the bus seat

The second scenario involved sitting on a bus seat and watching a journey filmed from the front of the bus with the sonified bus stops accompanying the announcement of bus stops. The video can be watched here: https://www.dropbox.com/s/860vr53o7gxag9g/Bus%20241%20Stratford%20-%20Plaistow%20Grove.mp4?dl=0

7. FEEDBACK

The exhibition was clearly not under controlled acoustic conditions, held within a large hall with a large number of other exhibitors in close proximity as well as sound permeating from the main auditorium. However, in spite of this noisy environment, we received a number of comments that the sound of our installation carried well above the various lectures and background chatter. Thus the location of the installation was useful in proving that the sound could be easily heard even in crowded environments, similar to the scenario of catching public transport.

We used the opportunity of having a large range of capabilities at hand to collect feedback from participants at the event. While providing feedback was optional, we obtained 34 survey forms from participants of a variety of capabilities ranging from bus drivers and operators, to the elderly and visually impaired. With respect to each scenario of identifying the bus route and bus stop, we asked participants how easy or hard it was to both recognize and remember the sounds on a scale of 1 to 5.

7.1. Scenario 1

The majority of participants gave an intermediate response for both ease of remembering and recognizing the sounds. with many stating that simply needed the time to learn the tunes. This was to be expected, given that the installation only played each sound twice. Most said that it would help with catching the right bus with the main concern being when more than one bus arrives. Another was audibility if the ambient noise was too high. Several participants claimed specifically that they were struggling to see the bus numbers and that the sounds would help them.

7.2. Scenario 2

Remembering and recognizing the bus stop sounds proved to be more difficult. This was also expected due to the fact that there were more of them to select from and only the selected bus stop was played twice. Again, this was expected to improve by repetition and learning. Furthermore, it would not be expected to learn all the bus stop sounds, but rather the one required.

7.3. Overall Feedback

Finally, we asked participants what they thought of our proposed sonification based on the 9-note blues scale. We received mostly positive but generalized comments such as good, fantastic, nice, pleasant, calming, relaxing, distinct, enjoyable and innovative. Since the range of disabilities at the event were broad, there was some concern for hearing-impaired or 'tone-deaf' people. However, even though some did not think it would help their own navigation due to not being 'musical enough', they still commented on the music making their bus journey more pleasant.

We also collected feedback on what other sounds they would like to hear, and received suggestions such as bells or sirens, animal noises such as dogs barking and birds tweeting, the use of different instruments or lower pitches, and the use of more popular tunes or songs. A number of participants suggested location specific sounds, which would make more sense for more well-known places. There were a few which still preferred spoken announcements, but who specified that they should be clearer and more audible, suggesting the current inadequacies of the announcement system.

8. FUTURE WORK

As a result of the event, we were invited to present our sonification proposal at Transport for London's recent Bus Safety Innovation Challenge, which involved showcasing it to the various bus operators of the London bus network. We are now in the process of discussing the potential implementation of our work with several bus operators on the existing passenger announcement system. This would provide a relatively low-cost approach to testing the impact of the use of sonification in the bus system and we look forward to exploring this in the future.

9. CONCLUSIONS

While there is still much work to be done concerning the choice of sounds, including more controlled acoustic testing with various sectors of the population, the interest shown thus far in our proposed sonification of London's bus network is promising. We are currently in the process of exploring other sonification options, including other types of scales for parameter-mapping, as well as the use of different timbres and rhythms. Given the suggestion for clearer spoken announcements, we also intend to investigate the use of Spearcons [9]. We greatly look forward to the suggestions and feedback of the ICAD community.

10. ACKNOWLEDGMENT

We would like to thank Transport for London for funding our exhibition at the Access All Areas exhibition held at ExCel centre on 19 March 2019. We would also like to thank them for the invitation to present at the Bus Safety Innovation Challenge.

11. REFERENCES

[1] Transport for London, www.tfl.gov.uk, Accessed 3 May 2019

[2] Transport for London, The Mayors Transport Strategy, https://tfl.gov.uk/corporate/about-tfl/the-mayors-transport-strategy, Accessed 3 May 2019

[3] Transport for London, Bus Safety Standard Executive Summary, 2018 http://content.tfl.gov.uk/bus-safety-standard-executive-summary.pdf, Accessed 3 May 2019

[4] A. Richarz, The Amazing Psychology of Japanese Train Stations, CityLab, 22 May 2018, Accessible at: https://www.citylab.com/transportation/2018/05/the-amazing-psychology-of-japanese-train-stations/560822/

[5] W. Gaver, "How do we hear in the world? Explorations in ecological acoustics", *Ecological Psychology* 5, no.4, pp. 285–313, 1998

[6] S. Benward, *Music: In Theory and Practice*, vol. 1, p.39. Seventh Edition, 2003

[7] Wikipedia, Blues Scale, 19 November 2019, https://en.wikipedia.org/wiki/Blues_scale#Nonatonic, Accessed on 3 May 2019

[8] S. Brewster, "Using nonspeech sounds to provide navigation cues," *ACM Trans. Comput.-Hum. Interact*., vol. 5, no. 3, pp. 224–259, 1998

\

SUBJECTIVE ELICITATION OF LISTENER-PERSPECTIVE-DEPENDENT SPATIAL ATTRIBUTES IN A RERVERBERANT ROOM, USING THE REPERTORY GRID TECHNIQUE

Bogdan Ioan Băcilă

Applied Psychoacoustics Laboratory,
University of Huddersfield
Queensgate,
Huddersfield, HD1 3DH,
United Kingdom
bogdan.bacila@hud.ac.uk

Hyunkook Lee

Applied Psychoacoustics Laboratory,
University of Huddersfield
Queensgate,
Huddersfield, HD1 3DH,
United Kingdom
h.lee@hud.ac.uk

ABSTRACT

Spatial impression is a widely researched topic in concert hall acoustics and spatial audio display. In order to provide the listener with plausible spatial impression in virtual and augmented reality applications, especially in the 6 Degrees of Freedom (6DOF) context, it is first important to understand how humans perceive various acoustical cues from different listening perspectives in a real space. This paper presents a fundamental subjective study conducted on the perception of spatial impression for multiple listener positions and orientations. An in-situ elicitation test was carried out using the repertory grid technique in a reverberant concert hall. Cluster analysis revealed a number of conventional spatial attributes such as source width, environmental width and envelopment. However, reverb directionality and echo perception were also found to be salient spatial properties associated with changes in the listener's position and head orientation.

1. INTRODUCTION

Recent developments in the virtual or augmented reality technologies can provide the audience with more realistic and compelling experiences in auditory display applications. For rendering of acoustic scenes in such applications, it would be necessary to plausibly represent the spatial impression in a virtual space, taking into consideration the listener's position and head rotation. This requires a solid understanding of psychoacoustical factors that influence the perception of various spatial attributes.

Spatial impression has been a widely researched topic in in the area of concert hall acoustics, with numerous constructs to subjectively and objectively measure and define it. The early term "spatial impression" introduced by Barron and Marshall [1] was found to be related to early lateral reflections, with a linear dependence on the early lateral energy fraction L_f ($\leq 80\ ms$). This has also been confirmed later on by Blauert and Lindemann [2], where this parameter was found to be strongly correlated with preference. The direction of arrival of a reflection has been also found to affect the spatial impression, with 90º azimuth from the listener generating maximum spatial impression.

Bradley and Soulodre [3] have proposed using the sub-terms ASW (Apparent Source Width) and LEV (Listener Envelopment) for defining the more generic "spatial impression (SI)". ASW is widely known as describing the perceived width of an audio source, being mainly dependent of early lateral reflections. Extensive studies have linked it to the objective measure of IACC (Interaural Cross-Correlation Coefficient). On the other hand, LEV, largely affected by late reflections ($> 80\ ms$) is known to describe how much a listener feels enveloped in the sound field and measurements such as lateral fraction (L_f) and lateral gain (LG_{80}) have been also proposed for predicting and objectively measuring LEV.

The relationship between ASW/LEV and the distance from the source were investigated by Lee [4] putting the objective measures used for predicting these attributes under a new light. Perceived ASW was found to statistically decrease almost linearly as the distance from the source was doubled. Similarly, LEV was found to decrease with doubling the distance, however, with a lower magnitude. Interestingly, it was found that the early sound strength (G_E) predicted the perceived ASW results more accurately, while IACC and L_f would predict them in an opposite direction. In the case of LEV, the late sound strength (G_L) and back/front energy ratio of late sound (B/F ratio) were also found to be more accurate parameters than LG_{80}.

Mason et al. [5] investigated into the perception of head-position-dependent IACC variations. They suggested that when facing forwards, variations in the IACC would cause perceived changes to the width and distance of the sound source and the width of the reverberant environment. However, when facing sideways these variations in IACC will affect the perceived depth of the reverberant environment as well as the envelopment and spaciousness of the reverberation.

Although aforementioned studies on ASW and LEV provide important references for the understanding of spatial perception in a concert hall, they are limited in that they do not take into account the dependency of spatial perception on listener's position and head orientation. While ASW and LEV could be considered to be high-level attributes, the aim of the current study is to define low-level attributes that are perceived depending on the listener's position and head orientation.

To this end, an in-situ elicitation test has been carried out in a reverberant concert hall using the Repertory Grid Technique (RGT), providing fundamental understanding and new insights into spatial perception for future development in 6DOF (6 Degrees of Freedom) auditory display.

https://doi.org/10.21785/icad2019.073

2. EXPERIMENT

2.1. Test Methodology

Evaluating spatial attributes could be a challenging research task due to their highly subjective nature. Especially, as mentioned above, it has been under-researched what kind of auditory attributes are perceived depending on the listener's position and head orientation. Therefore, in order to generate a set of rating scales to be used for future rating experiments, it was considered important that perceived attributes are subjectively elicited first.

Several methods of elicitation have been reviewed and verbal as well as non-verbal methods have been taken into consideration for this experiment. While a graphical elicitation method like the one described by Ford et al. [6] seemed like a good starting point, it was then understood that a non-verbal method would come with limitations in essential areas of the current study, as reported by Mason et al. [7]; The difficulty in representing the reverberation or ambience of a scene or the ambiguity in describing attributes that are not purely location-based, such as envelopment and spaciousness suggest that another elicitation method should be considered

For the purposes of this study, the Repertory Grid Technique (RGT) was used for initial elicitation of perceived spatial attributes as well as an initial quantitative test. The RGT is an elicitation method developed in the 1950s by Kelly [8] as both a qualitative and quantitative testing methodology. The technique was proposed by Berg and Rumsey [9] as a method of elicitation for perceived spatial audio attributes. This method is especially useful for the present research as it helps the generation of personal constructs for describing the audio stimuli, through their comparison, making sure that the subjects were not biased by any provided constructs.

2.2. Experimental Procedure

Six participants, postgraduate students and lecturers from Applied Psychoacoustics Lab, University of Huddersfield, having extensive experience with listening tests and spatial acoustics took part in the experiment.

The test took part in University of Huddersfield's St. Paul's concert hall (average RT = 2.1s; 16m (W) x 30m (L) x 13m (H)). A Genelec 8040A loudspeaker placed in the centre of the stage, playing a male speech was used as an excitation device. Speech was used for its broadband frequency nature and controlled ratio of transient and sustained sound. Ten positions around the hall were tested; Four of them facing the loudspeaker, four facing 90º from the loudspeaker and two facing away from the speaker (Figure 1). The elicitation test using the repertory grid technique consisted of two separate stages:

2.2.1. The elicitation process

In this stage, the participants were asked to compare the test positions with the aim of finding bipolar constructs to describe different aspects of the spatial impression. The stimuli were presented in triads and for each of them they were asked to think about what two of the presented positions had in common and opposite to the third position. These bipolar constructs were noted down and a new triad of positions was then tested, until there were no constructs left to elicit. The participants were asked to move from position to position for the elicitation phase of the test. At least 10 triads were assessed for each participant, covering all the positions. After this stage of the test was finished, the subjects proceeded to doing the second stage, after a short break;

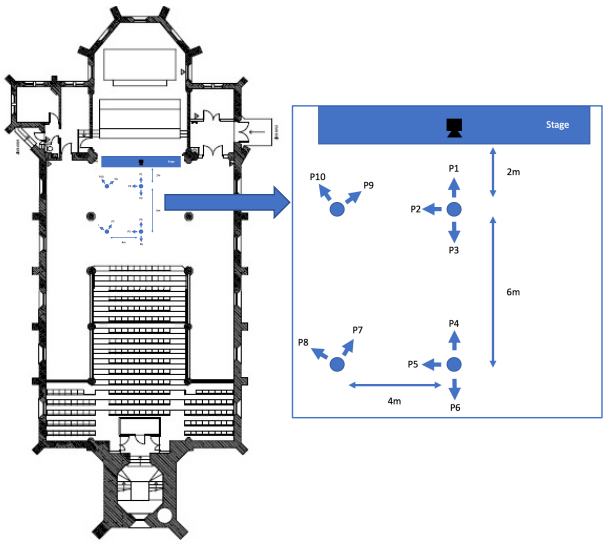


Figure 1. Positions used in the elicitation test

2.2.2. The rating process

After the subjects finished the elicitation process, the resulting bipolar constructs from each participant were arranged in a grid, with a pole on each side of it. Similarly to the experiment done by Berg et al. [9] the participants were presented with the grid and asked to check for consistency with their own vocabulary. They were then asked to walk to each of the positions and rate it for each of their own constructs on a scale from 1 to 5, with 1 corresponding to the left pole and 5 corresponding to the right pole. The order in which the subjects were asked to go to each position was randomized, to avoid any possible bias.

3. RESULTS AND DISCUSSION

An advantage of using the RGT for elicitation is the additional use of a grading system which can help greatly in identifying certain patterns as well as filtering out the relevant information from the less important one. The resultant grids, consisting of a total number of 56 bipolar constructs generated from the elicitation process, along with the associated ratings for each stimulus can be analyzed in multiple ways.

Before any in-depth analysis, it can be observed that there are certain general themes and attributes repeating in each of the participants' responses which put into the same words roughly represent:

- Source width
- Environment width/depth
- Envelopment
- Reverb directionality (Front/Back, Central/Off-axis)
- Echo perception (Clarity/Strength/Direction)

3.1. Verbal Protocol Analysis

Initial analysis of the elicited constructs was carried out by Verbal Protocol Analysis (VPA), a method presented and implemented by Samoylenko et al. [10] for separating verbal descriptors into different categories. Berg and Ramsey [9] as well as McArthur et al. [11] have also used this method for analyzing spatial audio terminology elicited using the RGT. Zacharov and Koivuniemi [12], while working on the development of descriptive language for spatial audio reproduction systems, have also used the VPA method. However, in their study, the Quantitative Descriptive Analysis (QDA) [13] was used for the elicitation process.

In the present research the constructs generated from the elicitation process were analyzed according to the VPA "Level 3" (semantic aspects of verbal units). The terms were divided into descriptive (dfe) or attitudinal (afe) features initially as shown in Figure 2.

Descriptive features were subsequently categorized into unimodal (umd – descriptors referring only to the audio modality) and polymodal (pmd – features that can describe multiple sensory modalities). The attitudinal features were also split into two categories, one expressing emotional or evaluative features (emv – reflecting one's emotions about a sound) and features expressing an element of naturalness (ntl).

After the analysis it was observed that out of the 56 constructs only one (Fig 3. No. 37) was considered an attitudinal feature, with the rest falling into the descriptive category. Out of the remaining descriptive features, all of them were then considered unimodal by the authors, who carried out the VPA based on their own semantic understanding.

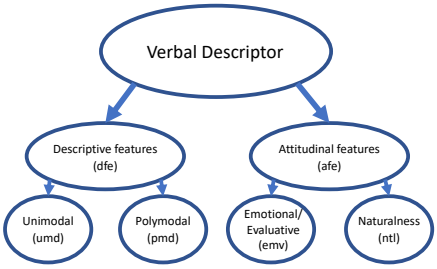


Figure 2. VPA features classification (after Samoylenko et al. [10])

3.2. Cluster Analysis

The cluster analysis of the data obtained was performed using "R" statistical analysis software [14] with the "OpenRepGrid" package [15]. When used on the whole data set the cluster analysis could reveal similarities between attributes and help in identifying repeating constructs that were rated similarly by different subjects. Figure 3 presents the constructs resulted from the elicitation process, along with the dendrogram generated by the rating and clustering process. In their work, Berg and Rumsey [9] carried out the cluster analysis only for the descriptive features. However, in the present paper the one attitudinal feature was left in and analysed with the rest of the descriptors.

This dendrogram was generated by decomposing the entire data set into separate clusters according to the agglomeration distance between the terms. Six clusters were found at an inter-construct distance of 10 (representing 50% similarity) and 10 clusters were found for an inter-construct distance of 7 (equivalent to 65% similarity).

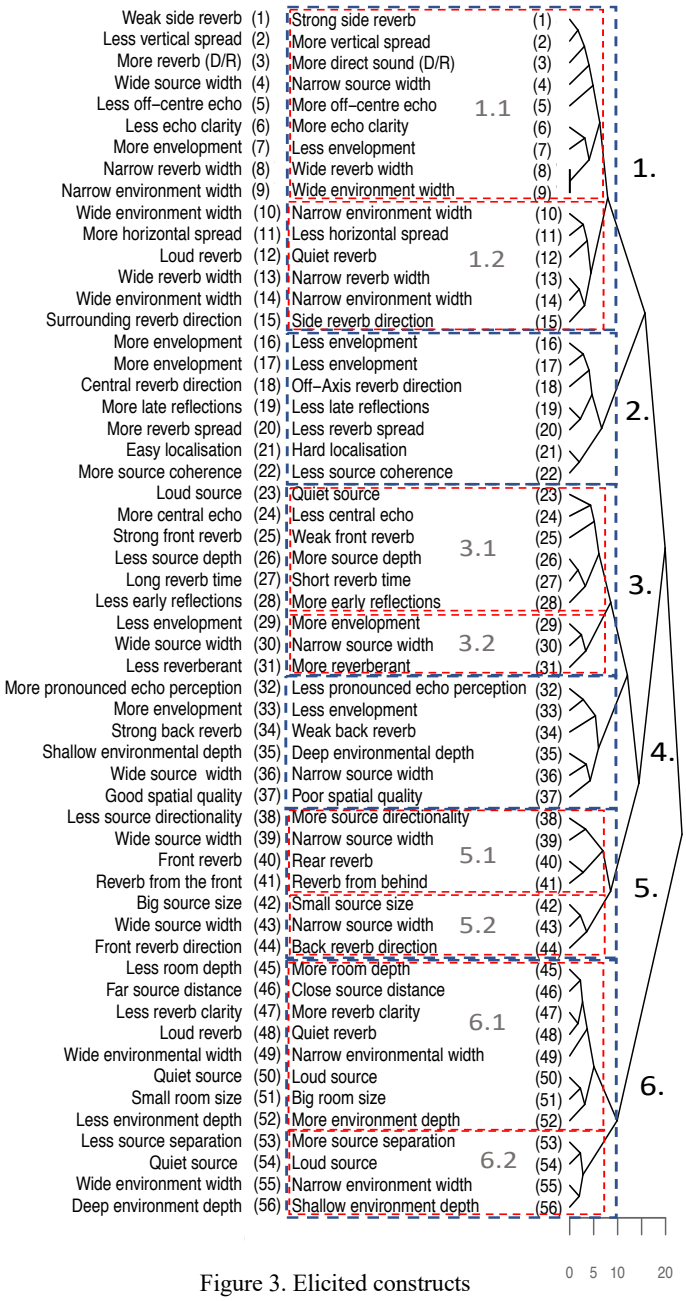


Figure 3. Elicited constructs and cluster analysis

The first resultant cluster from the analysis comprises 15 constructs, which are subsequently divided into a 9-element cluster (1.1) and a 6-element cluster (1.2). It could be observed from cluster 1.2 that the constructs are referring to properties of the reverberation. From cluster 1.1 it could be also observed that the directionality of the reverb and echo can play an important role in the perceived perception of width. *Environmental width* was a term commonly brought up by the participants and can be found in multiple clusters, however, it presents a higher appearance frequency in cluster 1.

In the second cluster some properties of *envelopment* could be seen to correlate to the directionality of the reverb. Terms like late reflections and reverb spread might suggest similar characteristics to envelopment and finally, the source coherence is also affected by changes in the envelopment.

The third cluster brings together elements related to the early reflections' influence over the source perception. While in cluster 3.1 there is no apparent focus on particular attributes, cluster 3.2 can suggest a correlation between the width of the sound source and the perceived envelopment.

Similar to the third cluster, in the fourth one, constructs related to envelopment and environmental depth are linked to the echo perception and reverb directionality. The only attitudinal attribute selected after the VPA ("Spatial Quality") suggests that the preference of a particular position in the room can be influenced by a broad number of attributes.

The fifth cluster brought to attention attributes related to the *source width*. It could be clearly observed from cluster 5.1 as well as cluster 5.2 that the width of a source seems to be influenced by the *reverb directionality*. A narrow source width seems to be corelated with reverb coming from the back and vice versa.

In the sixth cluster it could be observed that participants perceived an *environmental depth*. From both clusters 6.1 and 6.2 it could be observed a correlation between different aspects of room perception like size, width and especially depth, and the loudness of the source.

4. CONCLUSION AND FURTHER WORK

In the present study an elicitation test was carried out for determining the spatial impression attributes perceived in a reverberant room, in the context of multiple listener positions and multiple orientations. Repertory Grid Technique was used for the elicitation process, involving two stages: construct generation and a grading stage.

The responses were first analysed from a semantic point of view by using a Verbal Protocol Analysis, which helped in distributing the constructs into different categories. However, the participants' extensive experience with spatial and concert hall acoustics was reflected in the elicited attributes which were mostly considered descriptive of spatial impression.

Dendrograms created from the repertory grids were analysed by means of cluster analysis and the analysis was carried out on the full data set, for all of the presented positions. While the results do not show a consistent division of the attributes into solid clusters, some patterns could still be observed. Spatial impression attributes such as *source size/width*, *environment size/width and envelopment* were noticed as frequent appearances. However, because of the different listener positions and orientations, attributes such as *reverb directionality* and *echo perception (clarity, strength and direction)* were also found to be associated with the perception of the aforementioned attributes.

While the current analysis of the results presented a general overview of the current spatial impression terminology and brought up new attributes that can influence them, more in-depth analysis has to be carried out on the data set. It is expected that dividing the responses into different sets based on the different positions relative to the source (e.g. facing towards the source vs. facing sideways/away from the source or centre line vs. side line) would bring more insights over the attributes, and more importantly what factors affect them and how.

Ultimately, a quantitative listening test will be carried out for precisely measuring the interaction between the position of the listener and the spatial impression attributes. Furthermore, conventional objective measures for spatial impression will be examined to verify their validity in a multi-position and multi-orientation situation.

5. REFERENCES

[1] M. Barron and A. H. Marshall, 'Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure', *J. Sound Vib.*, vol. 77, no. 2, pp. 211–232, Jul. 1981.

[2] J. Blauert and W. Lindemann, 'Auditory spaciousness: Some further psychoacoustic analyses', *J. Acoust. Soc. Am.*, vol. 80, no. 2, pp. 533–542, Aug. 1986.

[3] J. S. Bradley and G. A. Soulodre, 'The influence of late arriving energy on spatial impression', *J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2263–2271, Apr. 1995.

[4] H. Lee, 'Apparent Source Width and Listener Envelopment in Relation to Source-Listener Distance', presented at the Audio Engineering Society Conference: 52nd International Conference: Sound Field Control - Engineering and Perception, 2013.

[5] R. Mason, C. Kim, and T. Brookes, 'Perception of head-position-dependent variations in interaural cross-correlation coefficient', in *Audio Engineering Society Preprint*, Munich, Germany, 2009, vol. 7729.

[6] N. Ford, F. Rumsey, and T. Nind, 'Evaluating Spatial Attributes of Reproduced Audio Events Using a Graphical Assessment Language - Understanding Differences in Listener Depictions', presented at the Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality, 2003.

[7] R. Mason, N. Ford, F. Rumsey, and B. de Bruyn, 'Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction', in *Audio Engineering Society Preprint*, Los Angeles, USA, 2000, vol. 5225.

[8] G. A. Kelly, *The Psychology of Personal Constructs (Norton, New York)*. 1955.

[9] J. Berg, 'Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique', *J Audio Eng Soc*, vol. 54, no. 5, p. 15, 2006.

[10] E. Samoylenko, S. McAdams, and V. Nosulenko, 'Systematic Analysis of Verbalizations Produced in Comparing Musical Timbres', *Int. J. Psychol.*, vol. 31, no. 6, pp. 255–278, 1996.

[11] A. McArthur, M. Sandler, and R. Stewart, 'Accuracy of Perceived Distance in VR Using Verbal Descriptors', presented at the Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, 2019.

[12] N. Zacharov and K. Koivuniemi, 'Unravelling the Perception of Spatial Sound Reproduction: Language Development, Verbal Protocol Analysis and Listener Training', presented at the Audio Engineering Society Convention 111, 2001.

[13] H. Stone, J. Sidel, S. Oliver, A. Woolsey, and R. C. Singleton, 'Sensory evaluation by quantitative descriptive analysis', *Food Technol.*, 1974.

[14] R Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017.

[15] M. Heckmann, *OpenRepGrid: An R package for the analysis of repertory grids*. 2018.

EXPLORING THE INTERFACE EFFECT IN DISTANT SONIFICATION

Iain Emsley

School of Media, Film and Music,
University of Sussex,
Falmer, BN1 9RH, United Kingdom
I.Emsley@Sussex.ac.uk

ABSTRACT

I introduce ongoing research into the method that I am calling distant sonification as a response to understanding abstractions created through computational reading. My aim is to explore the interface effect and situate it in sonification and media theory. Discussing existing prototypes, I contextualise the visible interfaces within the wider design models, such as patterns, and computational materiality. Reflecting on experiments in media specific analysis, I suggest that there are different models with their own specificities that are brought together to create the interface. They might exist separately or are combined to create a wider effect that I explore through models and grammars. I suggest that there are different models with their own specificities that are brought together by humans and machines through layers.

1. INTRODUCTION

Culture is becoming increasingly digital and requires new forms of practice and reading. Computational practices - such as distant reading [1], viewing or listening – use abstractions, such as maps and graphs, to analyse culture. Moretti's provocation that "distance… is a *condition of knowledge*: it allows you to focus on units that are much smaller or larger than the text" [7] suggests a strategy of not reading. Algorithms and data structures are central to this remediation of culture.

My research uses a method that I am calling distant sonification. It is a method of listening to these abstractions to aid interpretation and exploration. In particular, I want to develop Berry and Fagerjord's [14] view of critical Digital Humanities to explore the materiality of the medium's role in the sonification of cultural data. As the computational enables and remediate culture, I use sonification as a critically reflexive practice as well as representational method.

My research questions are:

1. How might the interface effect exist in sonification?
2. How might materiality affect sonification?
3. What role might sonification take in revealing its grammars and models?

In this paper, I discuss ongoing research into how the interface effect affects the sonification of cultural data. Using experimentation as a pragmatic approach, I advance a theory that the final abstraction is created from models that are remediated by other models, such as design considerations. I raise questions about what we might consider an interface and how it might be reflected as a construction.

Beginning through outlining the theoretical approach that I take; I reflect on the early experiments that explore distant sonification as a critical practice. From this, I begin reconsidering the design decisions being made and how these create an effect. I present experiments in using sonification as a tool for media specific analysis to raise questions about the hidden models. In the final section, I reflect on role of the computational in sonification and move towards the questions that this raises for future work.

2. A MEDIUM BASED APPROACH

In this section, I want to reflect on the role of the computational medium in sonification. My current focus is on Galloway's [3] interface effect, where it is transformation and transformative. Interfaces might be either software interfaces or the User Interface (UI), either graphical or aural. Building on design and listening approaches [19], my focus is on the materiality of the computational. I particularly want to explore the idea of the models and grammars that combine to create the effect and develop an argument that through understanding their role in the abstract and concrete models.

I explore the possibility that understanding the models supports the transformations. Grossman [4] suggests sonification is an extension of the human. Through considering the materiality, I want to use models to move from perception to consider how sonified cognitive models create and are part of an interface effect. Through considering this effect, I suggest that listening with *machines* becomes listening *with* machines.

I want to consider the models and grammar through their semantics and syntax. Vickers [2] raises questions about different design grammars and how they are interpreted by different actors within the development of the sonification. I use this to consider how these affect the transformation from data into sound within the digital medium. The models that both transform and are transformed into the abstraction and how the machine understands culture in its own milieu are constrained by the design considerations and grammars, themselves limited by syntax and semantics. Through this, I raise questions about how sonification might be used to understand these changes as a critical practice and as a reflexive critique. A critical question might be how might the object that is a sonification be de-reified? Instead of hearing it as an object, how might it be considered as an interface effect?

https://doi.org/10.21785/icad2019.072

I show how layers of models combine and are altered by processes to create a façade. Taking a medium based approach, I argue that these layers can be interrogated to reveal hidden machine-based models created through remediation. I see interaction as a way of questioning the given options, leading to questioning the materiality of the medium.

3. USER INTERFACES

In this section, I consider the developed prototypes to explore graphical interfaces - the dashboard and the live editor. Both use different interfaces on a common architecture, shown in Figure 1, and shared data sources, the Early English Books Online [17] and Russian Twitter troll data [16]. I consider the interface as an effect of its underlying mediations but also through external considerations, such as purpose.

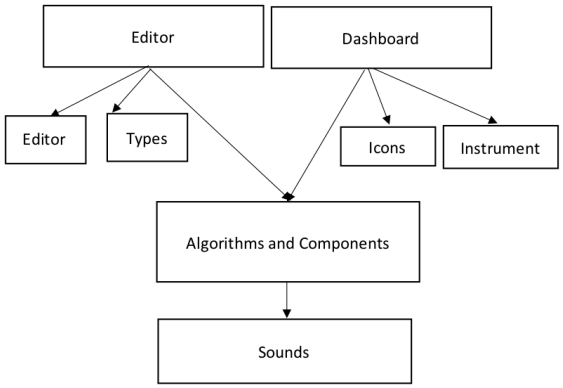


Figure 1. Diagram of the Distant Sonification Components Architecture

3.1. Dashboard Interface

The first developed prototype is a dashboard, reflecting Manovich's cultural analytics [19] approach. The interface displays different facets to the underlying data, such as histograms and line graphs. Icons suggest the shape of the audio to appear to orient the listener. The initial theory is that the icons and components suggest limitations that are designed into the User Interface.

Each component provides a view for the data. These might be altered through widgets, such as filters, to identify particular elements of interest. The interaction that these filters support allow for user to alter the sonification for their own purposes. Two components focus on one particular hashtag which can be chosen through a filter. The first counts the main tag's appearances and those with an edit distance of two symbols between them. Processed into an abstraction of the term and its associated count, it is presented as a histogram. This sonification design suggests an additional interface that reflects design constraints. The second sonification takes the results from the same filter and it maps the initial position of the tag within a text string and maps this to a pan position based on the screen size. The first component is bound by accepted conventions of the histogram and its intentions, where the second uses the text position in the data to allow the listener to make their own interpretation.

Through this, I might begin to think about how the interface effect is created. Hermann and Hunt [5] argue that interactive sonifications might be regarded as a type of virtual instrument that supports the interaction. Initial observations suggest that the dashboard allows the listener to create their own understanding through assigning a frequency to an event. Interaction allows a human model to work with the designed components, creating an apparent reflection of both worlds but with differing semantics, limited by a design context. These initial observations suggest critical angles to approaching and thinking with these models through the presented artefacts and their design.

3.2. Live Editor Interface

The second prototype is an interface that supports the interaction with the processes within a live environment. Replacing the icons with the Ace editor [20], it can either show an empty editor or it can be given an algorithm from a library used behind the dashboard icons. When the mouse is moved away from the editor, it runs the code in the browser and stores the change onto the disk.

The interaction works with the revealed code and the editor's syntax model. By removing the icons, the editor and language become interfaces, reflecting Blackwell and Aaron [6]. The grammar of the library used potentially limits the interactions through its semantics. The underlying software is revealed but it requires the ability to read and reason with them. Unlike the dashboard, the code is able to be altered so embedding the changes at a deeper level, reinforcing the sense of the working with the machine. This reveals the library which might be considered not only as a way of making but itself is a model and is designed for a purpose. Showing the code requires the listener to understand the components created from the algorithms and the language that they are written in, such as JavaScript.

Replacing the icons with an editor alters the human and machine relationship. Where the icons project a set of options, the live editor provides a closer relationship to the machine. It not only requires different forms of reading but also an understanding of the language and library's own models and grammars as part of the generative process that also constrains. The editor seemingly removes an abstraction but demands a deeper understanding of the code and the machine, showing it to be an abstraction for a language through modes. Removing one model allows the coder to create a new one using the available grammars and to place their own context into the new code, provoking questions regarding Tanimoto's concept of liveness [11]. It raises questions about the effect of languages and designs that they impose on the environment and are imposed on them by semantics and syntax as a new abstraction.

The ongoing research will explore these issues through writing a constrained language for the existing sonifications and to consider the role of design.

4. MEDIA SPECIFIC SONIFICATIONS

In the previous section, I discussed the existing prototypes and began uncovering the models and grammars within them. In this section, I want to think about sonification as

critical practice for understanding how the computational remediates itself to present different interfaces. This approach builds on Hayles's [8] consideration that print is flat and code is deep to explore how the medium's specificity and materiality can be sonified. I also want to think about it as a reflexive experimental practice.

4.1. Sonifying mark-up

An early experiment examines sonification as a critical practice to explore changing mark-up elements when a page is interacted with. The browser is an everyday manner of reading HTML and it also mediates the structural and style elements. The aim is to create a note each time that an event alters the HTML markup so that we can begin to explore what has taken place. These events may be human derived or from the web page. From this, we can begin to question how the change took place and what it might mean. Is a new component, such as a form loading, or are elements being removed, such as in infinite scrolling pages?

The probe is implemented as a web extension for the Firefox browser [12]. Using JavaScript's MutationObserver API, the extension listens for changes to the Document Object Model and maps these to a note. The sonification uses the Web Audio API so that a single JavaScript file can run in the browser. The sounds are microtones to suggest the speed of the changes. Changes to attributes, such as the accessibility attributes when a button is activated, are given a tone of 340.25Hz and detuned. The event also listens to the number of elements changed and whether these are being added or removed. This direction is used to calculate new frequencies from a base of 260.25Hz if added or 440.3Hz if removed, using the calculation for new notes to change frequency according to the number of changes.

As the browser code is event triggered, it needs to consider running in near real time. This places a constraint on the choice of timing mechanism. By allowing the event to drive the sonification, the machine is made more prominent and suggests that the sound is being made in near real time. The timing relies on synchronizing the various audio components for one event. This means that time is a model of sonification in itself, but it is a linking mechanism for other models.

The decision to use the AudioContext time allows the data to be used in near real time to demonstrate the amount and type of change that is taking place. The intention is to demonstrate the fast pace of largely invisible changes to suggest the addition of new components, such as forms, or changes to the markup to think about the materiality of the markup language. It does raise the question of what is being the sonified: the changed markup by the website or the browser?

The sonification is triggered by changes to the structural model. It shows that the machine part of the interface may change through intended, or otherwise, interaction in the browser, such as clicking a button or scrolling. The code relies on a language API and there is a map between the type of change and the note emitted. Although the data and transformation are both machine languages, a human mapping is required to create the relations with the endpoint. The sonification suggests that the interface might not just respond to the human action but also a machine driven one. Set into a browser extension, the sonification reveals the browser mediating a changing model.

4.2. Sonifying types

Having explored the interface as components, I turn to the syntax of the distant sonification abstraction. The second experiment explores the way that machine represent cultural data. I want to go beyond the algorithm to understand their effects. As data is being remediated in the initial components, it has to be represented in the algorithms.

The data is converted from textual strings into machine readable objects that can be manipulated. The approach taken echoes languages such as Sonnet [7] and Caitlin [9]. These languages show the operation of a running programme, the type language is aimed at helping the listener understand the way that the cultural form is altered by the algorithms. Yet it goes further than just the show the algorithm but how it represents the data for its operations.

Using reflection, an internal function library tests builds a tree of both time and types with the function that called and uses this to sonify the types, such as array, object, or float.

The library is a simple JavaScript file that use the Reflection API and Web Audio for sonification. Similar to the web extension, this library is designed to run within a coding environment. The live editor in section 3.2 is used to test and run the library, storing the models to allow me to read the underlying data later.

It shows the way that the abstract model presented as the interpretation is itself a mediated model. The second model is the human linking of time which is used here for the analysis. There are to options. Either the sound can be sequential, going through the objects as they appear in strict order of appearance or the nature of the objects can be used to simulate their interactions.

As an example, the filter component was sonified. This component takes a string for the URL to fetch data from and presents the received data as an array. As the array is iterated, it reveals the objects that are converted into numbers before being sonified. As the data is being iterated, it is being tested by the filter and only parts of it sonified. A further extension to this is to test for a named or anonymous function to reveal more about the design.

Through both experiments, we see that the materiality of the digital alters the sonification. In the first, we see the markup changing to reveal or hide components. It also reveals the way that websites alter their representation to support engagement. In the second, we can see the type and time models and the algorithms that alter these through processing. Using a filter to allow for the human model to suggest what should be sonified but revealing the types shows the changes and how the data structures changes through the processes. I want to think about these experiments as a critical practice, to not only test the theory but also how one might think about the practice of sonification.

5. DISCUSSION

The computational object itself is part of the interface effect. The remediation of the data into the abstraction – the map or the graph - suggests that it is constructed of grammars and models that require critical reading.

Through such a reading, the dashboard becomes an assemblage of components, revealing a series of data processes. It is a combination that is reliant on the design

associations, such as the method of analysis or within the components shown. The component itself can be rethought as bringing an analytical model into being. The given interaction provides a way of using these models as experiments themselves in using parts of the model. In the editor, the interface alters from visual to one of code and the editor itself. The language given is an interface to the operating system and with the libraries. It is a more subtle one that relies on a deeper understanding of the computational and hints at the alteration of the human and machine relationship. Interaction suggests a controlled use of grammars to create a new model for sonification using the interfaces.

The reflexive use of sonification to analyse the medium begins a reading the abstraction as a construction. I use this reflection to build on Hogg and Vickers's [10] consideration that even pure data needs mapping. As the mark-up is transcoded into a sound model, the browser extension uses a simple mapping between the type of event and the sound. Although it has no transformation of state, the sound requires a form of link between an attribute in the data defined by a designer and a frequency. The focus on types shows the medium making its own changes and on what is being transformed from the original data to the sonification. The abstraction used is itself a model of types, times and data that is brought into being through both software processes and the design constraints. A media specific approach begins to reveal the digital grammars that are used to construct the abstraction. The computational object requires intervention to sonify it.

The mapping decisions have a role in the consideration of how the interface effect is created, from the type of interaction to the type of computational models that are created. I contend that these considerations supply the conditions for interpretation. By understanding the materiality of the computational reading, we can begin to understand and (re)create it in different ways.

Audio is also part of the interface, though perhaps under theorised in this context. In work providing access to artefacts for visitors with a visual impairment [15], a tablet was used to provide audio information in response to being touched. A sonification alerts users to the activated button and before the voice played. Although screens and paper interface exist, they become invisible through the haptic and audio process. They create an interface through remediating events, models and concepts into sound, reflecting audiation [21]. The aural responses raise questions about the emitted sound as a central concern for sonification.

6. CONCLUSION

I present distant sonification as a method to understand digital culture. Using existing prototypes, I contextualise the visible interface within the wider design models, such as patterns, and the materiality of the computational. The specificity shows the machine creating its own structures that are remediated. I suggest that there are different models with their own specificities that are brought together by humans and machines through layers.

7. REFERENCES

[1] F. Moretti, *Distant Reading,* London, UK: Verso, 2013.

[2] P. Vickers, "Ways of Listening and Modes of Being: Electroacoustic Auditory Display", *Journal of Sonic Studies*, vol. 2, 2012, arXiv:1311.5880

[3] A. Galloway, *The Interface Effect*, Cambridge, UK: Polity, 2012

[4] J. Grossman, "From Metaphor to Medium: Sonification as Extension of our Body" in *Proc. of International Conference of Auditory Displays (ICAD)*, Washington D.C., USA, 2010, pp 145-152.

[5] T. Hermann and A. Hunt, "The discipline of interactive sonification". In *Proc. of the International Workshop on Interactive Sonification*. Bielefeld, Germany. 2004.

[6] A.F. Blackwell and S. Aaron, "Craft practices of live coding language design" in *Proc. of First International Conference on Live Coding*, Leeds UK, 2015, pp 12-22.

[7] D.H. Jameson, "Sonnet: Audio-enhanced monitoring and debugging" in *Auditory Display*, G. Kramer, Ed. Vol. XVIII. Santa Fe Institute, Studies in the Sciences of Complexity Proceedings. Reading, MA: Addison-Wesley, 1994

[8] N. K. Hayles, "Print Is Flat, Code Is Deep: The Importance of Media-Specific Analysis". P*oetics Today 25*, 2004, pp 67–90. https://doi.org/10.1215/03335372-25-1-67

[9] P. Vickers and J.L. Alty, "CAITLIN: A musical problem auralisation tool to assist novice programmers with debugging" in *Proc. of International Conference of Auditory Displays (ICAD)*, Palo Alto, USA, 1996

[10] B. Hogg and P. Vickers, "Sonification Abstraite/ Sonification Concrète: An "Aesthetic Perspective Space" for Classifying Auditory Displays in the Ars Musica Domain" in *Proc. of International Conference of Auditory Displays*, London, UK, 2006

[11] S.L. Tanimoto, 2013, May. A perspective on the evolution of live programming. In *Proceedings of the 1st International Workshop on Live Programming* IEEE Press, pp. 31-34

[12] https://github.com/iaine/mutate

[13] F. Moretti. *Graphs, Trees, Maps: Abstract Models for Literary History,* London, UK: Verso, 2007

[14] D.M. Berry and A. Fagerjord. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge, UK: Polity, 2016

[15] I. Emsley, T. Graven, N. Bird, S. Griffiths, J. Suess and L. Shaw, "Please Touch the Art: Experiences in Developing for the Visually Impaired". *Journal of Open Research Software*, vol. 7 no. 1, p.4, 2019, http://doi.org/10.5334/jors.231

[16] https://github.com/fivethirtyeight/russian-troll-tweets/

[17] https://github.com/textcreationpartnership/Texts

[18] L. Manovich, "Cultural analytics: visualising cultural patterns in the era of "more media"". *Domus* 923 March 2009

[19] F. Grond and T. Hermann, "Interactive Sonification for Data Exploration: How listening modes and display purposes define design guidelines", *Organised Sound*, vol. 19 no. 1, 2014, pp.41-51.

[20] https://ace.c9.io/

[21] G. Kramer "Some Organizing Principles for Representing Data with Soundin *Auditory Display*, G. Kramer, Ed. Vol. XVIII. Santa Fe Institute, Studies in the Sciences of Complexity Proceedings. Reading, MA: Addison-Wesley, 1994

AUDITORY DISPLAYS FOR AUTOMATED DRIVING - CHALLENGES AND OPPORTUNITIES

Pontus Larsson

Interactive Sound Quality
Volvo Car Group
Göteborg, SE-405 31, Sweden
pontus.larsson.3@volvocars.com

Justyna Maculewicz

User Experience Centre
Volvo Car Group
Göteborg, SE-405 31, Sweden
justyna.maculewicz@volvocars.com

Johan Fagerlönn

RISE Interactive

Piteå, SE-941 63, Sweden
johan.fagerlonn@ri.se

Max Lachmann

Pole Position Production
Galtabäcksvägen 11
Bromma, SE-168 55, Sweden
max@pole.se

ABSTRACT

The current position paper discusses vital challenges related to the user experience design in unsupervised, highly automated cars. These challenges are: (1) how to avoid motion sickness, (2) how to ensure users' trust in the automation, (3) how to ensure usability and support the formation of accurate mental models of the automation system, and (4) how to provide a pleasant and enjoyable experience. We argue for that auditory displays have the potential to help solve these issues. While auditory displays in modern vehicles typically make use of discrete and salient cues, we argue that the use of less intrusive continuous sonic interaction could be more beneficial for the user experience.

1. INTRODUCTION

The interest in automated road vehicles has been ever-increasing during the past few years. Reasons for the hype around Automated Driving (AD) may be its potential to bring positive societal effects in terms of reduced environmental impact, improved traffic safety, and more efficient mobility [1]. In addition to this, and regarding the specific appeal to the drivers/users, AD technology may allow people to be more productive, comfortable and relaxed during their daily commutes and other travels [1, 2].

Cars currently on the market offer low levels of automation still requiring human supervision, but we will likely see highly automated cars in the near future [2]. In fact, Waymo is already offering "robotaxi" solutions today, albeit in a limited setting [3]. Users of such vehicles are now considered passengers rather than drivers. Therefore, these AD vehicles bring possibilities to create completely new types of experiences for users. However, they may also introduce new types of problems, such as motion sickness and a lack of trust in the automation.

Sound may be a suitable medium for forming the user experience in AD vehicles. Sound can provide information to the users even if their eyes are closed, inform users continuously and subconsciously, and efficiently affect their emotional state.

In the current position paper, we will discuss the use of sound for the purpose of reinventing the in-car user experience when we go from manually driven- to highly automated vehicles. We will present a set of challenges that we consider vital to this area of research and innovation from the perspective of the automotive industry. The paper is intended to form a foundation for further research activities within our recently initiated research project "Sonic Interaction In Intelligent Cars" (SIIC) [4] and builds on initial investigations within this project. The paper also builds on knowledge from our workshop held at ICAD 2018 with the same title [5]. The intentions of the 2018 workshop were to build a new community for interactive sounds for AD that bridges the auditory display community with the automotive user interface community, and to discuss and exchange ideas within the field of AD as well as to explore promising directions for future work. Hopefully, the current paper will also inspire continued work in the directions set out by the 2018 workshop.

2. BACKGROUND

Driving Automation is currently one of the big trends within the automotive industry today along with electrification and new types of mobility services and solutions. A range of car manufacturers currently offer SAE (Society of Automotive

https://doi.org/10.21785/icad2019.038

Engineers - an automotive standardization body) Level 1-2 automation functions in their cars. With these functions, the driver still has the responsibility to supervise the automation and take over driving when needed [6]. In other words, the driver cannot perform secondary tasks such as reading - or even take their eyes off the road - while this type of low level automation is active. The next generation of automated cars aimed at reaching Level 3 or 4 automation [6] are currently being developed. With Level 4 (L4) automation engaged, the driver no longer has to supervise the automation and automation will not rely on the drivers' ability to take over driving when the automation reaches its operating boundaries - which potentially makes it safer than L2-3 automation [7]. L4 automation is also highly desirable from the user perspective since the user can truly make use of the time freed up by the automation [8].

L4 automation or "unsupervised AD" (as we will refer to it hereafter) provides new ways of interacting with cars and the entire experience of them may be drastically different compared to that of a traditional, manually-driven car. For example, when drivers are placed in a new role where they neither have to control nor continuously supervise the system, their workload is reduced in a great manner [9], which causes a decrease in situational awareness [10]. The driver - or rather passenger/user - no longer receives regular feedback from the car and the driving style is likely not same as her/his [11]. The user does not have to pay attention to vehicle-related visual displays anymore and is enabled to freely carry out non-driving tasks [12]. This situation introduces a lot of freedom and calls for new ways of designing the user experience of the car [13].

However, even if unsupervised AD cars are brought to the market, their success are contingent on that users are willing to use and adopt this new technology [2]. Users need to feel that they can trust the AD technology [14,15], they need to feel that it is comfortable and safe to use it, they need to perceive it as being more useful than their current mode of transportation [14], and they need to enjoy using it, in order for them to accept unsupervised AD and eventually adopt it [15]. This stresses the importance of performing user centred research and development in the area of unsupervised AD. Previous research on user experience of automated vehicles have focused mainly on supervised AD, while there is a lack of work focusing on the possibilities and challenges involved in unsupervised AD.

Within supervised AD, it has been found that trust, feeling of safety and acceptance can be influenced by the users' interaction with and experience of the car via its human-machine interfaces and it is likely that this will be possible also for unsupervised AD [16,17]. An as of yet rather unexplored area of research is to use sound, or sonic interaction, as a means of communication between the user and the AD car. Sound has unique advantages in comparison to e.g. pure visual communication. For instance, since our hearing is omnidirectional, sound can convey information no matter where the user has his/her visual focus. This feature of sonic interaction may prove to be especially useful in an unsupervised AD context where the user might have his/her visual focus anywhere (he/she might be e.g. reading a book or looking at the passing landscape) and not at the vehicle's visual displays [18]. For example, the study by Gang et al. [18] approaches the topic of trust in AD with an auditory solution presenting necessary information through spatially located abstract earcons. Their goal was to present desired information without causing alarm or compelling people to act.

Also, it is well-known that sound can easily catch attention and change the emotional and physiological state of the driver [19]. These properties make sound suitable as warning signals and alarms, which is one of the most common types of sonic interactions in cars today (e.g. collision alerts, belt reminder etc.). Sound design requires a lot of consideration especially when it comes to such warning sounds, which should result in appropriate and sometimes quick reactions [19-23]. For unsupervised AD, these types of attention-grabbing sounds triggered by discrete events may not be as useful as for manual driving and supervised AD since quick driver (user) reactions are not likely to be requested by the AD system. The traditional type of discrete sounds may also be too intrusive and annoying - and especially so when the user is really not involved in driving the car.

An alternative type of sonic interaction design deals with the manipulation of sounds which are already part of the soundscape. For instance, Fagerlönn, Lindberg and Sirkka [24] investigated the possibility of manipulating the sound from the in-vehicle radio to provide early warnings. Similarly, Nykänen, Lopez and Toulson [25] investigated the usefulness of various strategies, such as manipulating music content, to help drivers keep a steady speed.

Yet another design alternative to adding discrete and salient auditory cues is continuous sonic interaction, which builds more proactive interaction between a car and a user, rather than just presenting information in a reactive manner. Continuous sonic interaction in the current project application refers to an auditory display that, based on continuous input signals (obtained from control actions by the users or other input signals), provides concurrent auditory information about the resulting state or response of the system [26]. Cars with functions for unsupervised AD are equipped with an abundance of sensors and related processing units which enable them being able to react properly to the surrounding. The streams of data from the sensors could however also be used for creating such continuous sonic interaction for the cars' users, possibly making the AD system more transparent, intuitive, useful and engaging. An example of this type of auditory display from the AD domain was suggested by Bazilinskyy, Larsson & de Winter [27] who in an on road study explored using a subtle, continuous sound for which the level was mapped to the distance to other, leading vehicles and another sound which was in similar manner continuously informing about the ego vehicle's lateral position in the lane.

As opposed to the traditional discrete auditory signals, the continuous type of auditory display (sometimes also referred to as sonification) is believed to be better matched to humans who have evolved to act and control their environment in continuous fashion, and the auditory responses in everyday interactions tend to involve nuanced feedback depending subtly on human actions [26]. Continuous interaction also promotes closed-loop relationship which creates a higher level of perceived cooperation between human and machine, which can improve the understanding of the machine and have the potential to increase a user's sense of engagement [26]. Still, despite having these potential benefits, this type of sonic interaction likely needs to be carefully designed in order for it to be perceived as pleasant and enjoyable - in turn a prerequisite for its perceived usefulness and user adoption.

Thus, there may be new ways of using sound in unsupervised AD, but exactly which roles, if any, will sound play in this novel type of human-machine interaction? In the next sections we will go into detail of the aspects and challenges of the user experience in unsupervised AD cars that we foresee will be of interest to UX designers from the automotive industry perspective and how auditory displays possibly could play a role in regards to meeting these challenges.

3. SOUND TO REDUCE MOTION SICKNESS

One of the main arguments for autonomous vehicles is that they will allow users to spend their time in a more productive way. For example, the Concept 26 by Volvo Cars [8] suggest that the driver should, when having delegated the driving task to the autonomous driving function, be able to either relax or "create" (meaning: make calls, write emails or watch films and TV shows, etc.). Similar ideas are envisioned by other companies' designers, e.g. in the InMotion concept by NEVS [39], where the seats and interior can be arranged for either privacy, social interaction or work-related presentations, or the Volvo 360c concept which has similar features for socializing, working or even sleeping [40].

However, there is a growing concern that the possibilities for performing non-driving-related in-car activities will be limited due to the risk of motion sickness (kinetosis) [30, 31]. Motion sickness is a condition characterized by symptoms of nausea, dizziness, fatigue and other types of physical discomfort [30]. According to [30] there are three main factors leading to the condition of motion sickness; conflict between visual and vestibular inputs, loss of control over one's movements, and the reduced ability to anticipate the direction of movement - and all these factors may be present in an unsupervised AD scenario. There are basically two different ways of reducing motion sickness: 1) Allow occupants to anticipate the future motion trajectory and 2) Avoid incongruent self-motion cues. This implies that if it would be possible to cancel out the vestibular signals caused by the vehicle motion, when a passenger is looking down, that would reduce motion sickness.

Solutions to the motion sickness problem have been suggested by e.g. Waymo [32], Uber [33], and researchers at UMTRI [34]. UMTRI's and Uber's solution suggests providing the user with artificially generated stimuli that would reduce sensory conflicts and giving the user cues to be able to predict the car's movements – in Uber's case primarily a combination of visual-, haptic-, and airflow stimuli. While these types of stimuli could be efficient in reducing visual/vestibular conflict, the required displays would likely be quite expensive and cumbersome to integrate in a production vehicle. Using sound reproduced by an audio system, readily available in most production cars today, would be a more practical display. Uber suggests giving the user audio prompts (speech or tonal) or visual indications of upcoming maneuvers that could inform the user when to look up to avoid visuo-vestibular conflict. While this solution could potentially be efficient in reducing motion sickness it could also be quite annoying and thus result in a poor user experience.

Instead of providing audio prompts, one could in a more continuous fashion sonify the car's maneuvers slightly in advance to them happening so that the passengers know when to look up at the road. This solution would thus have the same effect as the proposal by Uber referred to above, but could be better from a UX perspective and feel more natural - e.g. as an enhanced engine sound. We label this solution the *Sonic Ghost Mode*, (in analogy to ghost mode used in computer games, see Figure 1) and note that it may be useful also for improving trust since it informs the user of the automation's intentions (see next section for more elaborations on this matter). For examples on how this may sound as well as other examples related to AD sonification, see [4].

Figure 1: Visual ghost mode in Rallisport Challenge 2 [35]. In this case, the intention is to visualize the difference between two race runs.

Given that humans can experience self-motion as a result of being exposed to certain types of sound [36, 37], such sound *synchronized* with the car's movement could be another potential solution to the motion sickness problem. The idea would be that the added sound compensates for the lack of visual stimuli when the user is looking down to read a book or similar and vestibular stimulation is prominent. A similar idea is presented in [38].

Yet another possibility would be to use sound to reduce or even cancel the vestibular signals. This may seem far-fetched but the idea is based on the findings that both air conducted (AC) and bone conducted (BC) sound can affect the vestibular system [39]. For example, in an experiment referred to in [39] a constant 500 Hz tone was shown to cause postural deviation. If sound can be delivered to the passengers when that cause a vestibular response opposite to the response caused by the car, it may result in a reduced total vestibular signal (similar to active noise cancellation). If that sound is presented when passengers look down it could potentially reduce the visual-vestibular conflict and thus also motion sickness.

4. SOUND TO INCREASE TRUST AND ACCEPTANCE

Trust can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [40]. In the case of vehicle automation, the agent would be the vehicle itself or the part of the vehicle that the user identifies as responsible for the automation. In the lower levels of automation (up to level 3 [6]) the most severe trust-related risk is "overtrust", i.e. when a person believes that the automation has better capabilities than it actually has. For higher levels of automation overtrust is less of a problem since automation in these levels by definition never should rely on a driver's intervention.

The most obvious trust-related risk for high level (L4/5) automation vehicles is that of under-trust - and that people will

not use it due to the fact that they do not trust them. Recent studies have found that many people would be afraid of riding in an automated vehicle [41], and a way to increase trust is through the design of the vehicle itself and its user interface.

For example, including human-like features in the interface, anthropomorphism, has been shown to increase trust in automation [15]. Anthropomorphistic features can guide users in their assessment of whether a machine is dangerous. By expressing human intelligence, friendliness and care, the design may increase the feeling of trust. Figure 2 shows an example where anthropomorphism has been used in the external vehicle design to amplify trust during interaction with self-driving cars - another similar concept can be found in [42].

Research has been shown that user's ability to estimate the predictability of the machine's behaviours affects trust [44]. A study by Helldin et al [17] showed e.g. that visual representation of a low level automated car's uncertainty (i.e. how sure it is of its ability to drive automatically) leads to a better calibrated trust. For a high level automation vehicle, it is reasonable to believe that it would be perceived to be more capable of driving by itself when it seems able to think and sense its surroundings than when it just gives an impression of "mindless machinery" [15].

Examples of when this type of representation is shown visually to the backseat passengers in the driverless shuttle service that is being piloted currently can be found in e.g. [45]. Similar type of information can of course be given through other modalities than visual ones. Given that passengers potentially will not, or even would not like to, have their eyes directed to one display, it seems more reasonable to use an auditory display to increase trust.

It is however not obvious what the best type of auditory information would be. One could speculate that event-driven audio prompts (e.g. chimes) are efficient in informing the driver of the vehicle's action and abilities but will quickly become annoying since they may occur quite often.

Figure 2: Anthropomorphism - Visual design concept by Semcon [43].

Speech messages could provide rich information and also give anthropomorphic features to the automation. Considering the fast development of speech assistant technology today, verbal interaction will most likely have a role in future highly automated cars. But again, providing information about ordinary actions through speech may be too intrusive. Furthermore, designing anthropomorphistic features using non-verbal sound is certainly a possibility, which was recently demonstrated by Collins and Dockwray [46]. We therefore hypothesize that a more continuous and subtle sonification of the car movements, intentions and abilities would be more efficient in inducing the appropriate degree of trust in the user.

Trust can be seen as part of the wider scope of user acceptance – naturally also crucial for user adoption of automated vehicles [14]. Among many things, acceptance is contingent on ease of use, usefulness and enjoyment [14] and we believe that sound can play a role in increasing acceptance of autonomous vehicles by making them more useful and comfortable. Consider for example the use case of stop-and-go traffic (low speed queueing) or other situations where the vehicle brakes and/or accelerates frequently which may cause the user to look up to see what is happening. If sonification would provide subtle information to the user on what is going on in traffic as in the Sonic Ghost Mode described in previous section, the user would not be triggered to look up each time he/she experiences sudden motion cues. Or, as we discussed in previous section, the sonification could also aid the user in knowing *when* to look up in order for him/her to avoid motion sickness. Therefore, e.g. performing eyes-off-road visual tasks may be perceived as more comfortable and less annoying with sonification added. There are several other similar situations when sonification could aid in "relaying" information about the driving scenario to the user without being overly intrusive or annoying (change of route, roadworks ahead, time to handover etc.) which in turn could increase comfort and the perceived usefulness of automation.

5. SOUND TO IMPROVE USABILITY AND MENTAL MODELS OF AD

To be able to predict possible actions and their consequences, users create mental models of the situations which they partake in. If a systems' behaviour corresponds to user expectations, encapsulated in a mental model, it heightens trust and provides a more positive user experience [47]. When a user approaches a new system, he or she builds a mental model based on previous experiences which might not be applicable in these new situations. However, with a proper user interaction design a user can be provided with a level of information, which could help to build more appropriate mental models and foresee system's behaviour. Therefore, transparent interfaces adapted to the mental system of the user are a prerequisite for the user to be able to develop necessary situation and system awareness in interactions with the automated system [48].

A use case that has already been identified as critical during lower levels of automation is handover of control from the AD system to the human. Supporting the user in this situation by creating awareness of the system's state and providing the user with a correct mental model of the system is however also important for higher levels of automation as long as the technology allows for multiple levels of automation (i.e. it will obviously not be important in vehicles where only one level of AD is present, such as in "robotaxi" vehicles, e.g. [45]). In unsupervised AD, the transition to manual driving can be challenging to handle since the intended "driver" may be in very different states ranging from being asleep to fully aware of the traffic situation. According to Strabala et al. [49], to perform successful handover one needs to agree that handover will happen, establish timing of the handover and decide how the process will be performed.

Based on this, we hypothesize that the handover situation requires some preparation to transfer a driver from non-driving towards the driving context. The vehicle-driver interaction should gently prepare a driver for handing over control to the car or vice versa. The cooperation between a car and a driver should be seen as partnership and handovers in both directions should happen by mutual agreement and in the right moment. The preparation probably needs to be adaptive to the current state of the intended driver. In some cases, the driver will need minutes or more to prepare, while other situations may require just a few seconds. In a previous research project Methods for Designing Future Autonomous Systems (MODAS) [50], a driver interface for AD was designed with professional drivers. When possible, the system provided information about upcoming handovers hours in advance during the driving route (see Figure 3).

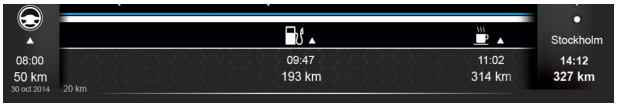


Figure 3. A design concept from the project MODAS [50]. A visual timeline (blue line) presented at the top of the windscreen indicates when a handover may be necessary.

Sonic interaction accompanying the handover process could be a useful part of a supportive multimodal user interface and help the user in developing a correct system/automation mode awareness and mental model of the system. Using sonic interaction, the user can perceive the information given by the user interface even with eyes closed and allows the user to keep his/her eyes on the road/traffic during the time of the actual handover.

6. PLEASANT AND ENJOYABLE SOUND

Even though a particular sound design might be highly useful (e.g. it reduces motion sickness, increases trust, induces correct mental models etc.), it is likely that the sonification and the AD system as a whole needs to be aesthetically pleasing and induce a sense of joy-of-use [51] to make the user engage the AD system for an extended period of time, and for the user to prefer sonification over more traditional means of signalling (visual displays, traditional sound chimes etc.). Apart from the advantages identified earlier in this proposal, continuous sonification could also be used in to induce certain moods [52] in similar ways as is being done within cinema and computer game sound design. Sound can in this way be used to soothe the user and make them simply enjoy the ride. Moreover, a gradually built up sonic atmosphere can be used to gradually increase attention and awareness of the user when, for example, the autonomous drive is about to reach its operational domain limits and the user is supposed to take over driving. Continuous sonification also makes the automation user interface more responsive and adaptive to user behaviour/reactions which could make the whole experience more balanced and pleasant.

While the field of designing efficient traditional sounds for in-car applications is quite well understood, research regarding how to design the above-described adaptive sonification-based displays that people enjoy using is scarce [53].

In line with what is suggested by [53], we believe that employing ideas and methods from the areas of Design Thinking and User Centered design could be one way of improving the overall user experience of auditory displays.

Using guidelines and praxis from the art of sound design for movies or computer games could be another way to understand how appealing, aesthetically pleasing continuous sonic experiences intended for automotive information displays should be designed. For example, [54] proposes a simple method for evaluating computer game soundscapes and devises design guidelines for how to heighten immersion and reduce listening fatigue - these might be applicable to an in-car context as well.

7. CONCLUSIONS

The development in automation has the potential to completely redefine the usage and user experience of road vehicles, especially when the technology allows unsupervised driving. In this paper we have presented a set of design challenges that are central to facilitating the successful introduction of highly-automated cars. These challenges are: (1) counteract motion sickness, (2) increase users' trust and acceptance, (3) improve usability and support the formation of accurate mental models, and (4) provide a pleasant and enjoyable experience. Furthermore, we argue that the utilization of auditory displays is a promising way to meet these challenges. However, while auditory displays in vehicles typically make use of discrete and salient cues, we argue that the use of less intrusive continuous sonic interaction can be a more successful strategy to facilitate a positive user experience. This will be investigated in recently started project [4] and the sonication solutions will be evaluated with users in a virtual environment and in a test car. This car has systems installed that enable experiences of high-level automation in realistic traffic environments. This is made possible by a "Wizard of Oz" setup, where a test leader/driver monitors the vehicle and can make corrections if necessary without the test person's awareness [55].

We hope that the project's challenges, along with the arguments supporting them, can inspire other researchers and practitioners to engage in the research and development of new types of auditory displays for self-driving vehicles.

8. ACKNOWLEDGMENT

The work presented in this paper is part of the currently ongoing project Sonic Interaction in Intelligent Cars (SIIC), funded by the Swedish partnership programme Strategic Vehicle Research and Innovation (FFI), D. Nr. 2018-02730.

9. REFERENCES

[1] D. Watzenig, M. Horn, *Automated Driving: Safer and More Efficient Future Driving*, Switzerland: Springer, 2017.

[2] P. Larsson, *User Experience of On-demand Autonomous Vehicles - Part 1: Background and User Experience framework*, Göteborg, Sweden: Ictech, 2018. Available:

https://ictech.se/om-ictech/artiklar/user-experience-of-on-demand-autonomous-vehicles/ .

[3] J. Fingas, "Waymo launches its first commercial self-driving car service," *Engadget*, May 12, 2018. [Online]. Available: https://engt.co/2CtIfTT.

[4] "Sonic Interaction in Intelligent Cars project website" [Online]. Available: https://siicproject.wordpress.com /

[5] J. Maculewicz, F. Hagman, M. Jeon, "Workshop 1: Sonic Interaction in Intelligent Cars," [Online]. Available: http://icad2018.icad.org/workshops-and-tutorials/.

[6] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_201806," [Online] Available: https://www.sae.org/standards/content/j3016_201806/

[7] Thatcham Research. "Regulating automated driving: The UK insurer view," [Online] Available: https://www.abi.org.uk/globalassets/files/publications/public/motor/2017/07/regulating-automated-driving/

[8] "Volvo Cars Concept 26" [Online] Available: https://www.volvocars.com/se/kop/teknik-och-tjanster/uppkopplad-bil/intellisafe/autonom-korning/concept-26

[9] D. Beattie, L. Baillie, M. Halvey, and R. McCall, "What's around the corner?: enhancing driver awareness in autonomous vehicles via in-vehicle spatial auditory displays," in *Proc. of the 8th Nordic conference on human-computer interaction: fun, fast, foundational,*, Helsinki, Finland, October 26 - 30, 2014, pp. 189-198.

[10] G. H. Walker, N. A. Stanton, and M. S. Young, "The ironies of vehicle feedback in car design," *Ergonomics*, vol. 49, no. 2, pp. 161-179, 2006.

[11] S. Kraus, M. Althoff, B. Heißing, and M. Buss, "Cognition and emotion in autonomous cars," in *IEEE Intelligent Vehicles Symposium*, Xi'an, China, 3-5 June 2009, pp. 635-640.

[12] M. M. Moore, and B. Lu, "Autonomous Vehicles for Personal Transport: A Technology Assessment," *SSRN*, June 2, 2011, Available: https://ssrn.com/abstract=1865047

[13] I. Politis, D. Szostak, A. Meschtscherjakov, S. Krome, M. Tscheligi, R. A. Ratan, R. Mccall, "Experiencing Autonomous Vehicles: Crossing the Boundaries between a Drive and a Ride," in *Proc. of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing System*, Seoul, Republic of Korea — April 18 - 23, 2015, pp. 2413-2416.

[14] S. Nordhoff, B. van Arem, N. Merat, R. Madigan, L. Ruhrort, A. Knie, and R. Happee, "User Acceptance of Driverless Shuttles Running in an Open and Mixed Traffic Environment," in *Proc. of the 12th ITS European Congress*, Strasbourg, France, 2017, ITS European Congress, Strasbourg, France, June 19-22, 2017, Paper ID TS27.

[15] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of Experimental Social Psychology*, 52, 113-117, 2014.

[16] P. Larsson, E. Johansson, M. Söderman, and D. Thompson," Interaction design for communicating system state and capabilities during automated highway driving," *Procedia Manufacturing*, vol. 3, pp. 2784-2791, 2015.

[17] T. Helldin, G. Falkman, M. Riveiro, and S. Davidsson, "Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving" in *Proc. of the 5th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications*, ACM, October 2013, pp. 210-217.

[18] N. Gang, S. Sibi, R. Michon, B. Mok, C. Chafe, and W. Ju, "Don't Be Alarmed: Sonifying Autonomous Vehicle Perception to Increase Situation Awareness," in *Proc. of the 10th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications*, ACM, September 2018, pp. 237-246.

[19] P. Larsson, and D. Västfjäll, "Emotional and behavioural responses to auditory interfaces in commercial vehicles," *Int. J. of Vehicle Noise and Vibration*, vol. 9, no. (1-2), pp. 75-95, 2013.

[20] R. Graham, "Use of auditory icons as emergency warnings: evaluation within a vehicle collision avoidance application," *Ergonomics*, vol. 42, no. 9, pp. 1233-1248, 1999.

[21] R. Gray, "Looming auditory collision warnings for driving," *Human factors*, vol. 53, no 1, pp. 63-74, 2011.

[22] C. Ho, and C. Spence, "Assessing the effectiveness of various auditory cues in capturing a driver's visual attention," *J. of experimental psychology: Applied*, vol. 11, no 3, 2005, pp. 157–174.

[23] J. Fagerlönn, and H. Alm, H. "Auditory signs to support traffic awareness," *IET Intelligent Transport Systems*, vol. 4, no. 4, 2010, pp. 262-269.

[24] J. Fagerlönn, S. Lindberg, and A. Sirkka, "Graded auditory warnings during in-vehicle use: using sound to guide drivers without additional noise," in *Proc. of the 4th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications*, ACM, 2012, pp. 85-91.

[25] A. Nykänen, M. Lopez, and R. Toulson "Safe and Sound Drive: Design of Interactive Sounds Supporting Energy Efficient Behaviour," in *Interactive Audio Systems Symposium*, York, 23 September 2016.

[26] Y. Visell, R. Murray-Smith, S. A. Brewster, and J. Williamson, "Continuous Auditory and Tactile Interaction Design," in Franinović, K., & Serafin, S. (Eds.), *Sonic interaction design*. London, UK: Mit Press, 2013, pp. 77 - 124.

[27] P. Bazilinskyy, P. Larsson, and J. C. F. De Winter, "Continuous auditory feedback for displaying automation status, lane deviation, and headway in a heavy truck," Abstract presented at the IJDS symposium '*Driving the Intelligent Vehicle*'. Haarlem, the Netherlands. 2017 Available: http://www.ijdssymposium.eu/upload/presentation%20Bazilinsky.pdf

[28] "NEVS presents InMotion Concept" [Online] Available: https://www.nevs.com/en/media/press-releases/nevs-presents-inmotion-concept/

[29] "360c: A new way to travel" [Online] Available: https://www.volvocars.com/intl/cars/concepts/360c

[30] C. Diels, and J. E. Bos, "Design guidelines to minimise self-driving carsickness", *Automated Vehicles Symposium*, Ann Arbor, Michigan, 2015. Available: https://www.researchgate.net/publication/280307548_Design_guidelines_to_minimise_self-driving_carsickness

[31] M. Sivak, and B. Schoettle, *Motion sickness in self-driving vehicles*. Ann Arbor, Michigan: The University of Michigan, Transportation Research Institute Report no. UMTRI-2015-12. Available: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/111747/103189.pdf?sequence=1&isAllowed=y

[32] D. L. Larner, and J. S. Russell, Waymo LLC, "Method and system for determining and dynamically updating a

route and driving style for passenger comfort," *U.S. Patent Application* 15/286,153, 2018.
[33] M. Sweeney, and E. Bartel, Uber Technologies Inc,. "Sensory stimulation system for an autonomous vehicle." *U.S. Patent Application* 15/651,878, 2017.
[34] M. Sivak, and B. Schoettle, University of Michigan, (2018). "Universal motion sickness countermeasure system," *U.S. Patent* 9,862,312, 2018.
[35] Microsoft Game Studios, *RalliSport Challenge 2*, [video game], 2004.
[36] A. Väljamäe, P. Larsson, D. Västfjäll, and M. Kleiner, "Auditory landmarks enhance circular vection in multimodal virtual reality," *J. of the Audio Eng. Soc*., vol. 57, no. 3, pp. 111-120, 2009.
[37] A. Väljamäe, P. Larsson, D. Västfjäll, and M. Kleiner, "Sound representing self-motion in virtual environments enhances linear vection," *Presence: Teleoperators and Virtual Environments*, vol. 17, no. 1, pp. 43-56, 2008.
[38] B. Kania, Fountainhead LLC, "Apparatus and method for relieving motion sickness," *U.S. Patent* US6443913B1, 2002.
[39] G.M. Halmagyi, I.S. Curthoys, J.G. Colebatch, and S.T. Aw, "Vestibular Responses to Sound, " Ann. N.Y. Acad. Sci., vol. 1039, pp. 54–67, July, 2005.
[40] J. D. Lee, K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50-80, 2004.
[41] E. Edmonds, "Three in Four Americans Remain Afraid of Fully Self-Driving Vehicles," [Online] Available: https://newsroom.aaa.com/2019/03/americans-fear-self-driving-cars-survey/
[42] "Jaguar Land Rover's virtual eyes look at trust in self-driving cars" [Online] Available: https://media.jaguarlandrover.com/en-gb/news/2018/08/jaguar-land-rovers-virtual-eyes-look-trust-self-driving-cars?q=&start=0&brand=corporate
[43] "The Smiling Car - a concept by Semcon" [Online] Available: https://semcon.com/semcon_smilingcar-_youtube/
[44] C. Gold, M. Körber, C. Hohenberger, D. Lechner, and K. Bengler, "Trust in automation–Before and after the experience of take-over scenarios in a highly automated vehicle," *Procedia Manufacturing*, vol. 3, pp. 3025-3032, 2015.
[45] "Waymo 360° Experience: A Fully Self-Driving Journey" [Online] Available: https://www.youtube.com/watch?v=B8R148hFxPwp
[46] K. Collins, and R. Dockwray, "Tamaglitchi: A Pilot Study of Anthropomorphism and Non-Verbal Sound," in *Proceedings of the Audio Mostly 2018,* Nottingham, UK, 2018.
[47] D. Gefen, E. Karahanna, and D. W. Straub, "Inexperience and experience with online stores: The importance of TAM and trust," *IEEE Transactions on engineering management*, vol. 50, no. 3, pp. 307-321, 2003.
[48] I. Wolf, "The interaction between humans and autonomous agents," in *Autonomous Driving*, pp. 103-124, Berlin Heidelberg: Springer, 2016.
[49] K. W. Strabala, M. K. Lee, A. D. Dragan, J. L. Forlizzi, S. Srinivasa, M. Cakmak, and V. Micelli, "Towards, seamless human-robot handovers," *J. of Human-Robot Interaction*, vol. 2, no.1, pp. 112-132, 2013
[50] S. Krupenia, Methods for Designing Future Autonomous Systems (MODAS), Stockholm, Sweden: Vinnova, Report/project no. 2012-03678. Available: https://www.vinnova.se/contentassets/5cb2a63a302b4e859e642ff20ea86550/2012-03678_en.pdf
[51] J. Nielsen, "User empowerment and the fun factor," In *Funology*, pp. 103-105, Dordrecht: Springer, 2003.
[52] R. M. Winters, and M. M. Wanderley, "Sonification of emotion: Strategies for continuous display of arousal and valence," in *3rd International Conference on Music & Emotion,* Jyväskylä, Finland, June 11-15, 2013.
[53] S. D. H. Cornejo, "Towards Ecological and Embodied Design of Auditory Display," in *Proc. 24th Int. Conf. on Auditory Display*, Michigan Tech, MI, June 10-15, 2018.
[54] B. Jacobsen, "How to maintain immersion (+ reduce repetition & listening fatigue) in game audio", *A Sound Effect.* [Online] available: https://www.asoundeffect.com/game-audio-immersion/
[55] K. Osz, A. Rydström, V. Fors, S. Pink, R. Broström, " Building Collaborative Test Practices: Design Ethnography and WOz in Autonomous Driving Research, " Interaction Design and Architecture(s) Journal - IxD&A, no. 37, pp. 12-20, 2018.

SURFING IN SOUND: SONIFICATION OF HIDDEN WEB TRACKING

Otto Hans-Martin Lutz [abc], *Jacob Leon Kröger* [ac], *Manuel Schneiderbauer* [ad] *and Manfred Hauswirth* [abc]

[a] Weizenbaum Institute for the Networked Society, Berlin
[b] Fraunhofer FOKUS
Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany
{otto.lutz, manfred.hauswirth}@fokus.fraunhofer.de

[c] Technische Universität Berlin [d] Humboldt-Universität Berlin

ABSTRACT

Web tracking is found on 90 % of common websites. It allows online behavioral analysis which can reveal insights to sensitive personal data of an individual. Most users are not aware of the amout of web tracking happening in the background. This paper contributes a sonification-based approach to raise user awareness by conveying information on web tracking through sound while the user is browsing the web.

We present a framework for live web tracking analysis, conversion to Open Sound Control events and sonification. The amount of web tracking is disclosed by sound each time data is exchanged with a web tracking host. When a connection to one of the most prevalent tracking companies is established, this is additionally indicated by a voice whispering the company name. Compared to existing approaches on web tracking sonification, we add the capability to monitor any network connection, including all browsers, applications and devices.

An initial user study with 12 participants showed empirical support for our main hypothesis: exposure to our sonification significantly raises web tracking awareness.

1. INTRODUCTION

Web tracking collects information about a particular user's activity on the World Wide Web. It is widely used, with some form of web tracking found on 90 % of common websites, and on 60 % of websites with highly privacy-critical content [1]. Although complex and extremely diverse, the ecosystem of web trackers is dominated by a small number of companies, notably by Google, Facebook and Amazon, who are inconspicuously present as third-party data collectors on many websites [2]. Recent empirical results suggest that third-party scripts owned by Google alone are present in about 80% of web traffic of the top 600 websites, and are used in a tracking context in about 40 % [3].

Since a person's browsing behavior reveals insights into his or her personality, habits and sensitive aspects such as financial and medical situation or political views, web tracking may constitute a serious privacy threat [4]. Even though web tracking is seen unfavorably by the majority of internet users due to privacy concerns [5], they do not understand the full extent, the methods and possibilities of online behavioral tracking [6].

Web tracking is invisible to the user by design. Studies show that there is no sufficient awareness of web tracking [7]. We use sonification of clandestine web traffic to tracking providers as a means of raising awareness for online privacy issues. If visualization is used instead for the same objective, users must divert their visual attention from their primary task (surfing the web). Using the auditory domain, we can simultaneously communicate information in a different modality, which provides additional attention and workload resources [8]. Furthermore, sonification is suitable to present temporal data in real-time and can be shaped to convey emotional content [9, p.11, p.92].

Our contribution is a sonification-based approach to raise user awareness of web tracking which extends the possibilities of existing approaches like *Soundbeam* by Hutchins et al. [10]. We describe a framework for live web tracking analysis and conversion to OSC[1] events, which can be used to monitor web tracking on any network connection – across all kinds of browsers, apps and devices. We discuss our system, sonification and sound design. Finally, we present results of an initial user study with 12 participants. We found empirical support for our main hypothesis: exposure to the sonification significantly raised web tracking awareness.

2. RELATED WORK

There is a comprehensive body of work on using sonification for network traffic monitoring to achieve higher situational awareness in a network operations center (e.g., [11, 12], systematic overview in [13]). In this context, users are network security specialists which use the auditory modality as supplementary resource to achieve their objectives in pattern, anomaly and intrusion detection. The scope of our approach, however, focuses on the average user who, in contrast to network operations professionals, is often unaware of the extent of web tracking [6]. Here, awareness refers to a general consciousness on the prevalence of web tracking. Sonification of web tracking can increase this awareness as it

This work has been funded (in part) by the Federal Ministry of Education and Research of Germany (BMBF) under grant no. 16DII111 ("Deutsches Internet-Institut").

[1] Open Sound Control, a network-based protocol for sound and media control: *http://opensoundcontrol.org*

https://doi.org/10.21785/icad2019.071

provides immediate auditory feedback to the user while he or she is browsing the internet.

Soundbeam [10] sonifies third-party connections extracted by Mozilla Lightbeam, a plug-in for the Mozilla Firefox browser. It sends data on intentionally visited websites and unintentionally visited third-parties (e.g., analytics or advertisement providers) to the SuperCollider synthesis engine via OSC. Soundbeam is designed for ensemble performance. Several users can run the software on different computers in the same network. When user B encounters a third-party element that has been identified by user A before, it is sonified for both users. This is intended to "highlight both the ubiqitousness and interconnectedness of tracking" [10].

Another related project is an earcon-based sonification of internet security threats for vision-impaired users [14]. Here, warning sounds that convey their intended meanings with little-to-no user training (e.g., casting a fishing reel to warn about a phishing attack) were used to notify users about security threats while browsing on a screen reader.

3. FRAMEWORK DESIGN

Our software runs in the background while the user is browsing the web. The framework comprises four stages: (1) monitoring network traffic, (2) filtering for connections to known web trackers, (3) extracting different kinds of tracking-related events, and (4) sending these events to the sound generator via OSC (see Figure 1).

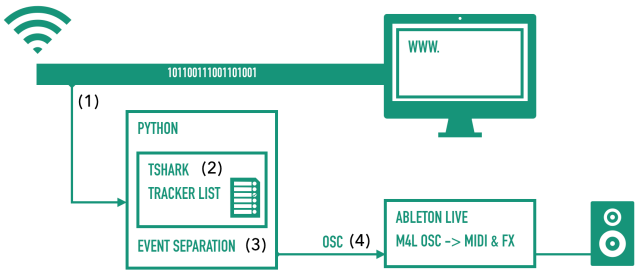


Figure 1: System overview

In the prototyping phase, we used Ableton Live [16], with a Max for Live OSC receiver for sound synthesis. We aim to switch to cross-platform (Linux supported) open source software in the future.

3.1. Implementation

In order to be able to intercept any network connection, we use Python to create several instances of TShark processes, a text-based version of the network protocol analyzer wireshark [15]. These processes listen to the traffic of the selected network connection. They are configured with filter lists of web tracker IP addresses, so only traffic to these addresses is analyzed in the following steps.

Tracker identification: Connections to tracker services are detected by tracker identification lists available from different sources (e.g., whotracks.me [17], easyList [18], or generated from Mozilla Lightbeam). Each list has benefits and disadvantages. For our prototype, we used a semi-automated approach, accessing all Alexa Top 50 Websites International and Germany [19] with Mozilla Lightbeam running in the background and exporting the list of third-parties accessed. When testing the lists by browsing random websites, this semi-automatically generated list caught more third-party connections than the whotracks.me list. On the other hand, the whotracks.me list supplies a differentiation between different categories of third-parties (e.g., advertising, analytics, content delivery networks), which can provide a clearer picture of the intentions behind the third-party connection. We aim to systematically compare different tracker lists in the future.

Event separation: We configured TShark to listen to ports 80 (HTTP) and 443 (HTTPS) of the IP addresses generated from the tracker lists. We spawned separate TShark processes: a) monitoring establishment of a connection (SYN events) and b) monitoring data transferred to trackers (GET / TLS application data events). We further filter the SYN events by connections to the top 10 most prevalent trackers to further accentuate these acoustically (see Section 3.2).

All these events are stored in buffers and then sent out via OSC. As sound events which happen in close temporal proximity are not discernible anymore (precedence effect) [20], we send out the buffered events with a short pause in-between. In a heuristic pre-test, a pause of 70 ms turned out to provide the best balance between discering single events and an overall coherent impression.

3.2. Sound design

The overall purpose of our approach is raising awareness, creating interest and stimulating tought on the topic of web tracking. The auditory representation is designed to show the amount of web tracking in the background, raise interest and convey some degree of danger in order to feature the associated privacy concerns. Not only the amount of tracking is important, but the fact that a group of very few companies are present on most websites. Therefore, we aim to disclose the oligopoly of these companies as well.

When a connection to one of the top 10 tracking companies is established, we present an audio recording of the company's name in a whispered manner. Reverb is added to the whispers to intensify the spatial and suspicious impression, as a reference to the intrusion on privacy. Some of the companies are well known to users (e.g., Google, Facebook), others are less known (e.g., ComScore, criteo). The whispered names are supposed to stimulate questions about these companies as well.

Each data transfer event is presented with a short sound event. The following sound variations were designed for comparison regarding users' perception in terms of interest, curiosity, danger, and fear. We aimed to design our sounds in a way to reflect either power or fragility to convey both the power of tracking and the hidden, brittle quality it has as well. The powerful and fragile sounds were designed both in a musical and an abstract sound variation. Their numbers correspond to the sequence used in evaluation.

1. powerful and musical: low cello and tuba
2. fragile and abstract: granular synthesis
3. powerful and abstract V1: deep bleeps
4. fragile and musical: piccolo flute and violine
5. powerful and abstract V2: like V1, added delay

A video containing both an impression of the sonic experience with our system while surfing and examples of all sound variations can be found at *http://s.fhg.de/SonificationICAD2019*.

3.3. Comparison to existing approaches

Our approach of monitoring the internet traffic itself instead of relying on the Lightbeam browser plugin extends the capabilities of Soundbeam by:

- supporting all browsers and combinations of ad / tracking blocker plug-ins.
- supporting monitoring of any physical or virtual network connection on the host computer. This enables monitoring traffic generated not only by web browsing but by apps as well.
- supporting monitoring the traffic of any device (e.g., laptop, smartphone), if we open and monitor an ad-hoc wireless network that this device connects to.
- usage and comparison of different tracker blocking lists.
- conveying the name of the tracking company by whispers.

As we have no means of identifying which addresses or links the user wants to visit, our approach does not support differentiation between intentional website visits and third-party connections. Therefore, the quality of the tracker identification list is an essential factor for a reliable result.

For now, we do not support ensemble performance as we currently aim to make an individual user aware of the tracking he or she personally is subjected to. To create a multi-user experience, the capability for sending OSC events to different computers in the network can be added to our framework.

4. EVALUATION

4.1. Study design and hypothesis

We conducted an initial user study with 12 participants (6 male, 5 female, 1 no gender stated) with an age range between 23 and 36 years, mean age was 28.9 years. In a within-subjects design, we presented the recordings of five different sound variations in a classroom setting. Each recording represented the sonification of accessing the same website. It showed the actual sonic experience while surfing, consisting of several single bleeps occurring shortly after each other. Whispering of the tracker names was muted in order to set focus on the tonal quality of the sonified events. After each sound variation, participants filled out a questionnaire regarding the perceived emotional qualities of the respective sonic experience. We asked participants to rate their overall auditory impression of the sound playback (as if visiting a website), not the single sound elements. At the end, we presented all sound variations again and asked participants to state their favorite.

For the emotional qualities of the sounds, we asked participants to rank each sound between the following poles on a four-point likert scale. For statstical analysis, we assigned the numbers (-2,-1, 1,2) to the scale items.

- innocent (-2) to dangerous (2)
- relaxing (-2) to frightening (2)
- boring (-2) to interesting (2)
- indifferent (-2) to curious (2)

As we designed the system to raise awareness, our main hypothesis is that the awareness regarding web tracking gets higher after exposure to the sonification. We assessed awareness before and after the sonification experience each with a five-point likert scale (low, rather low, medium, rather high, high).

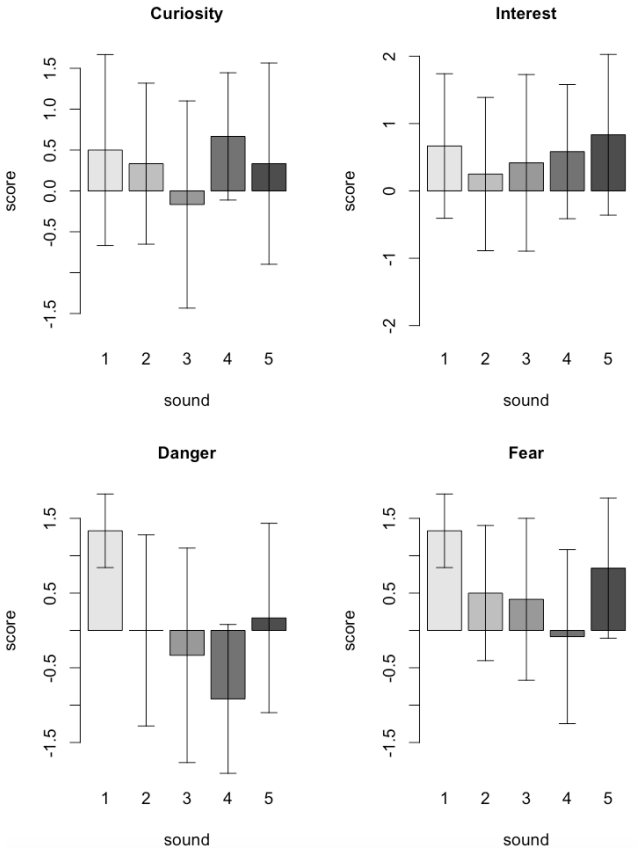


Figure 2: Emotional content of the sound variations. Error bars in plot: +/- one standard deviation

4.2. Results

As Shapiro-Wilk normality tests showed that normal distributions cannot be assumed in our sample, we performed a one-sided Wilcoxon signed rank test with continuity correction (see [21, p. 977]) to assess the differences between awareness scores prior to and after exposition to the sonification. The test results support our main hypothesis: Awareness levels were significantly higher after exposure to the sonification than before ($mean_{before} = 0.75,\ mean_{after} = 1.25,\ p = 0.024,\ r = -0.652$).

Results on the emotional qualities curiosity, interest, danger and fear were less distinct and not significant (see Table 1 and Figure 2). Hence, all statements on the emotional qualities of the sounds are descriptive only. For sound 1 (low cello and tuba), danger and fear ratings were both high in mean and with a smaller standard deviation compared to the other sounds. Interestingly, sound 4 (piccolo flute and violin) was perceived least dangerous, but raised the most curiosity. Sound 1 was stated most often as favorite (five times), followed by sounds 4 and 5 (three times each).

Sound variation:	1	2	3	4	5
mean(curiosity)	0.500	0.333	-0.167	0.667	0.333
sd(curiosity)	1.168	0.985	1.267	0.778	1.231
mean(interest)	0.667	0.250	0.417	0.583	0.833
sd(interest)	1.073	1.138	1.311	0.996	1.193
mean(danger)	1.333	0	-0.333	-0.917	0.167
sd(danger)	0.492	1.279	1.435	0.996	1.267
mean(fear)	1.333	0.500	0.417	-0.083	0.833
sd(fear)	0.492	0.905	1.084	1.165	0.937

Table 1: Sound variations: Means and standard deviations of emotional content scores

5. DISCUSSION

The initial user study has limitations: Most notably, as the sounds were presented in a classroom setting, a sequence effect is expected. Future studies will benefit from individual presentation via headphones and randomisation of the sound variations. Adjectives of the emotional quality poles were not selected from standardized test batteries on emotional content. Additionally, the sample size of 12 participants was quite small. Nevertheless, some effect of the sonification experience on web tracking awareness could be shown.

6. FUTURE RESEARCH

As our initial results are encouraging, we will continue and extend our work in the following ways: First, we aim to set it up in a way that supports connecting a user's own device (laptop, smartphone) to a special wireless network we provide and monitor. By this, we allow users to explore the tracking sounds of their own browser or app configuration. We are also looking into porting the framework to a small computer like the Raspberry Pi [22]. This can ease the usage of our system in installations in public. Then, we plan to conduct a larger user study that assesses the impact of our approach to web tracking awareness in the field.

Future research questions regarding sound design are manifold: We aim to disclose not only the amount of web tracking, but the oligopoly of the tracking companies as well. So far, we approached this with the tracker name whispering when connecting initially. In future, we want to design signature sounds for each company, so the corresponding single events can be linked to these companies. Another significant step is moving on from producing the sounds in Ableton Live to a model-based sonification. Additionally, incorporating the spatial domain can help conveying tracker parameters by placement in the virtual room.

7. ACKNOWLEDGMENT

The authors want to thank Jan Maria Kopankiewicz for his support with implementation.

8. REFERENCES

[1] S. Schelter and J. Kunegis, "On the Ubiquity of Web Tracking: Insights from a Billion-Page Web Crawl," pp. 53–66, 2016. [Online]. Available: http://arxiv.org/abs/1607.07403

[2] S. Macbeth, "Tracking the Trackers: Analysing the global tracking landscape with GhostRank," Cliqz GmbH, Tech. Rep., 2017.

[3] A. Karaj, S. Macbeth, R. Berson, and J. M. Pujol, "WhoTracks.Me: Monitoring the online tracking landscape at scale," pp. 1–15, 2018. [Online]. Available: http://arxiv.org/abs/1804.08959

[4] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, 2015.

[5] K. Purcell, J. Brenner, and L. Rainie, "Search engine use 2012," *Search*, 2012.

[6] T. Bujlow, V. Carela-Espanol, B. R. Lee, and P. Barlet-Ros, "A Survey on Web Tracking: Mechanisms, Implications, and Defenses," *Proceedings of the IEEE*, vol. 105, no. 8, pp. 1476–1510, 2017.

[7] W. Thode, J. Griesbaum, and T. Mandl, ""I would have never allowed it": User Perception of Third-party Tracking and Implications for Display Advertising," in *Proc. International Symposium on Information Science*, 2015.

[8] C. D. Wickens, "Multiple resources and mental workload," *Human factors*, vol. 50, no. 3, pp. 449–55, 2008.

[9] T. Hermann, A. Hunt, and J. G. Neuhoff, *The Sonification Handbook*, 1st ed. Berlin: Logos Publishing House, 2011.

[10] C. Hutchins, H. Ballweg, S. Knotts, J. Hummel, and A. Roberts, "Soundbeam: A Platform for Sonyfing Web Tracking," *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 497–498, 2014.

[11] M. Ballora, N. Giacobe, and D. Hall, "Songs of cyberspace: an update on sonifications of network traffic to support situational awareness," *Proc. SPIE Defense + Commercial Sensing*, vol. 8064, pp. 1–6, 2011.

[12] M. Debashi and P. Vickers, "Sonification of network traffic flow for monitoring and situational awareness," *PLoS ONE*, vol. 13, no. 4, pp. 1–31, 2018.

[13] L. Axon, S. Creese, M. Goldsmith, and J. R. C. Nurse, "Reflecting on the Use of Sonification for Network Monitoring," *Proc. SECURWARE 2016*, pp. 254–261, 2016.

[14] A. Siami Namin, R. Hewett, K. S. Jones, and R. Pogrund, "Sonifying Internet Security Threats," *Proc. 2016 Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 2306–2313, 2016.

[15] https://www.wireshark.org, [Accessed 20.04.2019].

[16] https://www.ableton.com, [Accessed 20.04.2019].

[17] https://github.com/cliqz-oss/whotracks.me, [Accessed 20.04.2019].

[18] https://github.com/easylist, [Accessed 20.04.2019].

[19] https://www.alexa.com/topsites, [Accessed 20.04.2019].

[20] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "A Precedence Effect in Sound Localization," *The Journal of the Acoustical Society of America*, vol. 21, p. 468, 1949.

[21] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R*. SAGE Publications, 2012.

[22] https://raspberrypi.org, [Accessed 20.04.2019].

DESIGNING ADAPTIVE AUDIO FOR AUTONOMOUS DRIVING: AN INDUSTRIAL AND ACADEMIC-LED DESIGN CHALLENGE

Doon MacDonald

Swansea University
Swansea
SA1 8EN, Swansea, UK
d.g.macdonald@swansea.ac.uk

ABSTRACT

The paper discusses a design challenge around the use of adaptive audio to support experience and uptake of autonomous driving. The paper outlines a collaboration that is currently being established between Researchers at Swansea university and a major OEM that is set to examine user-centred approaches to designing audio that enhance and enrich human-experience with driving.

The paper outlines the potential collaboration and describes how we will address the challenge to designing adaptive audio for unsupervised /autonomous driving. The paper outlines the research question we will address and how we will apply a tool/method that supports rapid prototyping for novice designers alongside addressing ideas around aesthetics in the interface and relationships between sound as a means for communication and as experience.

1. INTRODUCTION

The enthusiasm around autonomous cars is on the increase. This exciting technology has the potential to nurture positive societal changes, including reduced environmental impact, improved traffic safety and more efficient mobility. Additionally, using Autonomous Driving (AD) technology might support commuters by allowing them to be more productive to and from the commute to work (time to do work, for example). However, the introduction of highly automated cars will require a re-definition of the car-driver interaction. The ongoing technological development will put completely new demands on the design of interactions inside of the car, in order to support the driver in his or her role and to create an appropriate driving experience.

The relationship between the human and the car becomes a vital factor and is more important than ever when it comes to the trust and uptake of autonomous vehicles. The Car will be in control and so, it is only prudent to ask, what role the driver will take when the car is making the decisions? how will the driver (or, end-user) trust the car and perceive it as intelligent enough? what can we, as designers, do to enable a comfortable and safe experience for the end-user?

Even if unsupervised AD cars are brought to the market, their success will be down to the willingness of users to accept and adopt this new technology. Users need to feel that they can trust the AD technology [1, 2] they need to feel that it is safe to use it, they need to perceive it as being more useful than their current mode of transportation [1] and they need to enjoy using it, in order for them to accept unsupervised AD and eventually adopt it [2]. The user does not have to pay attention to vehicle- related visual displays any more and is freely enabled to carry out non-driving tasks [3]. This situation introduces a lot of freedom and calls for new ways of designing the user experience of the car [4].

> As control is shifted away from the driver and vehicles become autonomous, there is a limit to the enjoyment felt. This is because the travelling experience for the driver is not taken into account [5]

It is fair to argue that, with the adoption of autonomous cars visible on the horizon, the car industry needs to explore important questions that focus on the human: trust, experience, up- take, acceptability and accessibility.

2. ACADEMIA AND INDUSTRY

Researchers from the CHERISH-Digital Economy Centre at Swansea University, UK and the OEM are developing an important relationship in order to address this challenge. Specifically, researchers at both institutions will explore the role of sound in creating a valuable user experience, with a focus on how adaptive sounds can be designed and implemented in order to support the relationship between the user and the car in a given driving scenario.

Researchers from the OEM and Swansea met at when the author presented SoundTrAD (a method and tool created by the researcher) to a driving scenario [6]. SoundTrAD is a tool that enables a designer to create prototype auditory displays and adopt a user-centred approach to the design. The tool is based on ideas and principles from Soundtrack composition. It enables a systematic approach to prototyping audio for a given scenario whereby the story and aesthetics and the use of sound as both communication and experience are important design considerations. SoundTrAD enables designers to blend different audio, test different use cases and rapidly prototype auditory displays. More is discussed on this in section 4.0.1.

A relationship was formed because the OEM and CHERISH-DE both share a human-centred approach to design whereby human values remain at the heart of any technical innovation. The CHERISH-DE centre (CHERISH is an acronym for ‘challenging

https://doi.org/10.21785/icad2019.079

Human Environments and Research Impact for a Healthy and Sustainable Digital Economy') is tasked with designing for the human, taking on human values when it comes to advancements in technology. Autonomous vehicles will be on our roads soon so together industry and academia can partner in order that the human experience is kept in the loop.

The shared design approach will take into account the needs of the user from the start of the design process and this is something that is important to both institutions. To exemplify, [7], acknowledged the need for industry and academia to work together, by observing that 'empirical Studies that follow industry projects, from design intention to the end user-experience, are scarce' [and that] 'this knowledge can enable the HCI community to better support industrial practices and aid in bridging the exisiting Industry-Academia gap'. They later go on to write that 'by cross-disciplinary research cases, we believe the discourse between industry and acdemia can be improved, and best practices for user experience can evolve'. [7]

3. RESEARCH QUESTION AND HYPOTHESES

There are many benefits to using Sound in the car interface. Sound, for example, can be used to provide feedback and represent information that might normally be associated with a visual channel [8]. This is particularly relevant to a potential autonomous driving scenario whereby the passenger might not be paying attention to decisions the car is making. Sound can be used for warning and alerting and sound can also represent information on many channels. Researchers are enthusiastically exploring the use of sound in AD as it is set to play a vital role. For example, [9] looked at spatial and multichannel audio in driverless vehicles to communicate the intended actions of the vehicle to the user. The application of Earcons and Auditory icons were explored by [10, 11] and specific applications were examined by [12], who looked at the use of sound to represent and warn of speed and [13, 14] who looked at the use of sound to represent fuel efficiency.

The specific design challenge of creating adaptive audio in relation to in-car interfaces is an area that opens up questions around the direct role of the user in a given driving scenario. Sound that can adapt in accordance to data (from the car or interaction from the user), adds the human and the car into the loop and supports the 2-way relationship between the car and the human, in as far as the real-time reaction and interaction from the user becomes an important design factor. For example, passenger information such as fatigue and levels of attention (becoming accustomed to the sound), become control variables for sound design. Furthermore, questions can be addressed concerning the emotional state and experience of the user and their subsequent trust and acceptance of the vehicle. We can design for playfulness and interaction to keep attention and engagement, or even consider the customization of the soundscape to individual people. Car information such as infotainment, other in-car noises, external sounds, external factors such as time of day, bikes, pedestrians, can all all be taken into account when it comes to the audio design for the car.

Together the OEM and Swansea University will address the following research question: Can adaptive sounds increase perceived intelligence, safety, usability and experience in unsupervised driving? The following hypotheses (H) will be tested through a series of user-centred studies.

- H1: Adaptive Sounds can increase usability in unsupervised driving.
- H2: Adaptive sounds can enhance experience (comfort) in unsupervised driving.
- H3: Adaptive Sounds can increase a sense of intelligence and safety within the car in unsupervised driving.

4. UPCOMING STUDY

Collaboratively researchers will test these hypotheses. The first stage of the study will involve asking end-users to discuss their journeys to work in order to refine some use cases. This should enable the application of some real-world scenarios and refinement of parameters to work to. Following this, prototypes will be created that use adaptive sounds (effected by car data and user input) that can be tested and iterated.

Various methods and tools will be employed in order to apply and test these scenarios including:

- WOz Cars
- Mule Car
- Virtual Reality Scenario
- The SoundTrAD Tool (see Section 4.0.1 below)

4.0.1. SoundTrAD System

As referenced in Section 2, SoundTrAD is a tool and method that allows novice designers to prototype audio. Different sounds can be designed, blended and iterated given a specific use case. It is based on principles from Soundtrack Composition and so supports aesthetic and gamification as design considerastions[6]. Technically, SoundTrAD is programmed in Max/MSP[1] and Processing[2] and sends information about audio events through Open Sound Control (OSC) messages[3]. Designers can map out a given scenario and map sounds to different events on a time line. The events can be re-ordered in real-time and sounds subsequently tested for masking and suitability. Importantly, there is potential for SoundTrAD to be integrated (or used in alignment) with Unity[4] (via OSC), thus creating the potential for working with a VR driving simulation. This is something researchers hope to explore.

5. OUTCOMES

From this collaboration and study a set of design guidelines will be created alongside a set of prototype examples that help industry and academia further understand the role that adaptive audio can play in autonomous driving. **We hope to further nurture a meaningful industrial and academic relationship in order to address important questions around the role of the user in future autonomous cars. Furthermore, we hope to develop and apply SoundTrAD as a useful design tool.

6. ACKNOWLEDGMENT

The CHERISH Digital Economy Hub, Swansea University

[1] https://cycling74.com
[2] https://processing.org
[3] http://cnmat.berkeley.edu
[4] https://unity.com

7. REFERENCES

[1] S. Nordhoff, B. Van Arem, N. Merat, R. Midigan, L. Rohrort, A. Kniw, and R. Happee, "Acceptance of driverless shuttles running in an open and mixed traffic environment." in *Proc. of the 12th ITS European Congress.*, 2017.

[2] A. Waytz, J. Haefner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of Experimental Psychology*, vol. 52, pp. 113–117, 2014.

[3] I. W. P. F. Cars, "Autonomous vehicles for personal transport: A technology assessment."

[4] NHTSA, "Automated vehicle policy on levels of automation and considerations for research progre," https://www.nhtsa.gov/press-releases, 2013.

[5] D. Beattie, L. Baillie, M. Halvey, and R. McCall, "Maintaining a sense of copntrol in autonomous vehicles via auditory feedback," in *the Fourth Workshop on the Perceptual Quality of Systems*, 2013.

[6] D. MacDonald and T. Stockman, "Soundtrad, a method and tool for prototyping auditory displays: Can we apply it to an autonomous driving scenario?" in *Proc. of ICAD 2018, Meeting of the International Conference on Auditory Display*, 2018.

[7] I. Pettersson, L. Hylving, A. Rydström, and D. Gkouskos, "The drive for new driving interfaces: Researching a driver interface from design intent to end-user expereince," in *in Proc. of NordCHI'16*, 2016.

[8] M. Jeon, S. FakhrHosseini, E. Vasey, and M. Nees, "Blueprint of the auditory interactions in automated vehicles: Report on the workshop and tutorial." in *Proc. of AutoUI'17*, 2017.

[9] L. Baillie, M. Halvey, and R. McCall, "What's around the corner? enhancing driver awareness in autonomous vehicles via in-vehicle spatial auditory displays," in *Proc. NordCHI'14*. ACM, 2014.

[10] P. Larsson, A. Opperud, k. Fredriksson, and D. Västfjäll, "Emotional and behavioural response to auditory icons and earcons in driver-vehicle interfaces," in *Proc. 21st International Technical Conference on Enhanced Safety of Vehicles*, Germany, 2009.

[11] P. Larsson, "Tools for designing emotional auditory driver-vehicle interfaces," *Auditory Display*, vol. 5954, pp. 1–11, 2010.

[12] J. Hammerschmidt and T. Hermann, "Slowification: An in-vehicle auditory display providing speed guidance through spatial panning," in *Proc. of ISon 2016, 5th Interactive Sonification Workshop*, 2016.

[13] M. Nees, T. Gable, M. Jeon, and B. N. Walker, "Prototype auditory displays for a fuel efficiency driver interface," in *Proc. of the 20th Int. Conf. on Auditory Displays*, 2014.

[14] S. Landry, D. Tascarella, M. Jeon, and S. Fakhr Hosseini, "Listen to your drive: Sonification arhitecture and stretegies for driver state peformance," in *AutomotiveUI Adjunt Proceedings '16*, 2016, pp. 225–228.

AUDIO GUIDANCE FOR OPTIMAL PLACEMENT OF AN AUDITORY BRAINSTEM IMPLANT WITH MAGNETIC NAVIGATION AND MAXIMUM CLINICAL APPLICATION ACCURACY

Ognjen Miljic, Zoltan Bardosi, Wolfgang Freysinger

4D Visualization Laboratory
University ENT Hospital
Medical University of Innsbruck, Austria
ognjen.miljic@student.i-med.ac.at

ABSTRACT

For patients with ineffective auditory nerve and complete hearing loss, Auditory Brainstem Implant (ABI) [1] presents diversity of hearing sensations to help with sound consciousness and communication.
At present, during the surgical intervention, surgeons use preoperative patient images to determine optimal position of an ABI on *cochlear nucleus* on brainstem. When found, the optimal position is marked and mentally mapped by the surgeon; Next, the surgeon tries to locate the optimal position in patient's head again and places the ABI. The aim of this project is to provide the surgeon with maximum clinical application accuracy guidance to store the optimal position for the implant, and to provide intuitive audio guidance for positioning the implant at the stored optimal position. By using three audio methods, in combination with visual information on Image-Guided Surgery (IGS), surgeon should spend less time looking at the screen, and more time focused on the patient.

1. INTRODUCTION

This work presents a dynamic audio feedback system for positional guidance in real-time during surgical procedure of ABI placement. ABI is a solution for individuals with hearing loss due to an ineffective auditory nerve, and it is implanted on *cochlear nucleus* which is located on the anterior part of the brainstem. Bypassing both, the inner ear and the auditory nerve ABI stimulates the *cochlear nucleus* and provides the patient with a hearing sensation, which can improve communication and consciousness.
At present, during the surgical intervention, surgeons use preoperative patient images, usually including Magnetic Resonance Imaging (MRI) and/or Computed Tomography (CT), to determine the optimal position for the ABI on the *cochlear nucleus*.
The main purpose of this project is:

- Providing support for better spatial accuracy in preoperative planning of the ABI implementation
- Remembering the spatial position that is localised with an Electronic Auditory Brainstem Response (E-ABR)
- Final positioning of the ABI with maximum possible accuracy in the appropriate position

During the IGS, location of the surgical probe is visualized by tracking and mapping its location to the pre-operative model of patient anatomy. Nevertheless, the main issue during IGS is that the surgeon often has to divert attention from patient to the screen (navigation system). By using auditory cues, surgeon should spend less time looking at the screen and more time focused on the patient. Combining audio guidance signals with existing IGS visual information, may result in greater accuracy when locating a given target in a 3D volume [2].

2. METHODS

The proposed audio guidance system is comprised of the following elements: (i) *NavABI* software for IGS developed in house - basis of this software is Rhinospider Technology [3] developed by the University Hospital for ENT at the Medical University of Innsbruck (ii) custom NavABI audio plug-in software developed using OpenAL(Open Audio Library) software interface to audio hardware (iii) Electromagnetic Tracking System - Aurora NDI [4] (iv) SoundWear Companion speaker BOSE - 2.0 wearable Bluetooth speaker for presenting audio guidance.

In this work three different types of audio guidances are used. The main difference between following methods lies in complexity and cognitive effort used to understand audio signals. From the simplest, Pulsed Tone sonification distance guidance that most of the participants are familiar with (car parking assistant), over Signal To Noise sonification (participants should recognize it as tuning the old radio) which also represent distance information, to the method designed for guidance in all three axes of Euclidean space by using three different perceptions of the sound Pitch, Loudness and Duration.

To avoid perceptual inaccuracies on the part of the listener, amplitude scale (frequency as a function of amplitude) according to the Fletcher-Munson Curve [5] is taken into account for each tone in all three methods.

2.1. Pitch, Loudness and Duration Sonification (PLD)

This solution is based on the idea that three different perceptions of the sound (Pitch, Loudness and Duration) guides the surgeon along the *X, Y* and *Z* axes of the operating table.

https://doi.org/10.21785/icad2019.075

t marks the desired target location in the operating table space, t_x, t_y and t_z mark the projections of the target to the *X*, *Y*, *Z* axes. Similarly, p_x, p_y and p_z marks the projection of the probe to the table axes. The projected signed distance between the target and probe along axis $\bullet$ is defined (1) as:

$$d_\bullet = t_\bullet - p_\bullet \tag{1}$$

These distances are used as inputs to control the sonification. This method consists of:
Loudness: *X - Axis* - encoded in stereo channels of the speaker. The left $l(t)$ and the right $r(t)$ (2) channels are encoded as:

$$\begin{aligned} l(t) &= \tilde{l}(d) \cdot o(t, d_y, d_z) \\ r(t) &= \tilde{r}(d) \cdot o(t, d_y, d_z) \\ \tilde{l}(d_x) &= \begin{cases} d_x \leq 0, & 1 \\ d_x > 0, & 1 - \frac{d_x}{d_{max}} \end{cases} \\ \tilde{r}(d_x) &= \begin{cases} d_x \geq 0, & 1 \\ d_x < 0, & 1 + \frac{d_x}{d_{max}} \end{cases} \\ d_x &= t_x - p_x \end{aligned} \tag{2}$$

Where d_{max} represent the maximum distance of surgical tip from the target point.

Pitch: *Y - Axis* - represented by two sine tones base and alternate frequency. As long as the user keeps pointer at the target, only one tone can be heard (440Hz). As soon as user start moving away from the target along the *Y* axis (either up or down), variations in frequency of one sine tone happens.
$S(t)$ signal (3) is represented as combination of base and alternate signals:

$$\begin{aligned} S(t) &= \text{base(t)} + \text{alternate(t)} \\ \text{base(t)} &= \sin(t \cdot f) \\ \text{alternate(t)} &= \sin(t \cdot f_{act}(d_y)) \\ f_{act}(d_y) &= \begin{cases} f, & |d_y| < d_{min} \\ f + sgn(d_y) \cdot f_{max}, & |d_y| > d_{max} \\ f + sgn(d_y) \cdot f_d(d_y), & d_{min} \leq |d_y| \leq d_{max} \end{cases} \\ f_d(d_y) &= (f_{max} - f_{min}) \cdot \frac{|d_y| - d_{min}}{d_{max} - d_{min}} + f_{min} \end{aligned} \tag{3}$$

Where f_{act} represents actual frequency for a given distance, and f_{max} and f_{min} represent maximum (500Hz) and minimum (380Hz) frequency. d_{min} and d_{max} represent minimum and maximum distance of surgical tip, from the target along *Y* axis.

Duration: *Z - Axis* - is encoded with pulsed tone (duration of the sound) - Sine (440Hz) tone that pulses gradually faster as the tip of the surgical tool is closer to the target.

$S(t)$ pulsed tone signal (4) is represented as:

$$\begin{aligned} S(t) &= \text{Step(t)} \cdot x(t) \\ f_{act}(d_z) &= f_{max} - (f_{max} - f_{min}) r_{act}(d_z) \\ r_{act}(d_z) &= \frac{d_z - d_{min}}{d_{max} - d_{min}} \\ d_z(\tilde{d}) &= \begin{cases} d_{min}, & \tilde{d} < d_{min} \\ \tilde{d}, & d_{min} \leq \tilde{d} \leq d_{max} \\ d_{max}, & \tilde{d} > d_{max} \end{cases} \\ \text{Step(t)} &= \left\lfloor t \cdot f_{act}(d(\tilde{d}(t))) \right\rfloor \, mod \, 2 \end{aligned} \tag{4}$$

$\lfloor \cdot \rfloor$ represents the number truncated to the closest integer.

2.2. Signal to Noise Sonification(SNR)

This audio guidance, which showed best results in the work of J. Plazak [6] consists of two sounds, white noise and pure sine tone at 440Hz, with the volume mixture of the sounds being controlled (linearly) by distance information:

- Distance $\geq$ 600mm - Results with 100% white noise
- Distance = 300mm - Results with 50% white noise and 50% sine tone
- Distance = 000mm - Results with 100% sine tone

The signal $S_{snr}(t)$ (5) is given by:

$$\begin{aligned} S_{snr}(t) &= r_{act}(d) \cdot n(t) + (1 - r_{act}(d)) \cdot x(t) \\ n(t) &\sim \mathcal{N}(0, 1) \cdot A \\ t &\in \mathbb{R} \\ A &\in \mathbb{R} \\ x(t) &= \sin(f \cdot t) \cdot A(f) \\ f &\in \mathbb{R} \end{aligned} \tag{5}$$

White noise $n(t)$ is generated from standard normal distribution $\mathcal{N}(0, 1)$. A represents the amplitude multiplier and the f denotes the frequency of the tone.

2.3. Pulsed Tone Sonification(PT)

This audio guidance consists of a short (t = 0.1s) sine tone (440Hz) that pulses continuously faster rates as the user approaches the target, up to a point at which pulses linked to form a continuous tone (on target). This type of audio guidance can be found in variety of other commercial applications (e.g. car parking system), and users are familiar with this type, which was the main reason why this method was included in this research. The rate at which the tone pulses is controlled by the distance information, having a range from a slow pulse at 1Hz (60 beats/min) until it reaches a continuous sine tone 20Hz(1200beats/min) on target. The pulsed tone

signal PT(t) (6) is given by:

$$\begin{aligned} \text{PT(t)} &= \text{Step(t)} \cdot x(t) \\ f_{act}(d) &= f_{max} - (f_{max} - f_{min}) r_{act}(d) \\ r_{act}(d) &= \frac{d - d_{min}}{d_{max} - d_{min}} \\ d(\tilde{d}) &= \begin{cases} d_{min}, & \tilde{d} < d_{min} \\ \tilde{d}, & d_{min} \leq \tilde{d} \leq d_{max} \\ d_{max}, & \tilde{d} > d_{max} \end{cases} \end{aligned} \tag{6}$$

Step function Step(t) is defined in (4).

3. EVALUATION OF METHODS

An experimental study has been designed with 20 planned participants in controlled experiment, 10 participants with IGS experience (surgeons and researchers), and 10 participants without experience with IGS. There are total of 7 experimental conditions and each condition should be repeated 10 times by each participant (a sum of 1400 trials):

- 3 audio only conditions (3 methods listed above)
- IGS only
- 3 audio and IGS combined conditions

In each trial, the main goal is to navigate the surgical pointer to randomly placed target points within the CT and/or MRI volume as quickly as possible. Each trial should last for 20 seconds. Following metrics will be used to evaluate results. Experiment setup is presented in Fig. 1. Several metrics are evaluated on the trajectories followed by the participants during the navigation sessions.

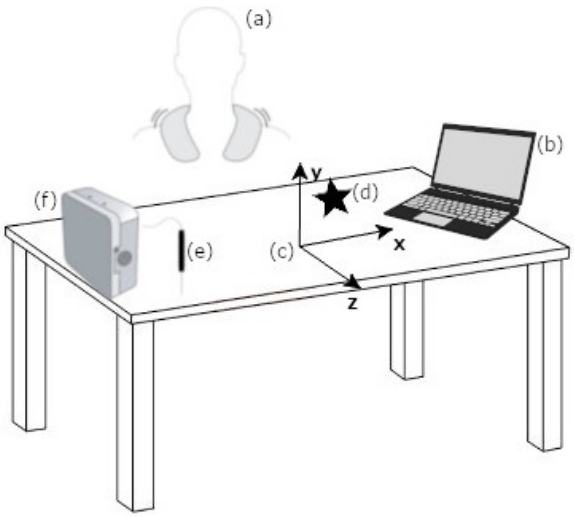


Figure 1: Experiment setup (a) Participant wearing a BOSE audio hardware; (b) Computer containing IGS and audio guidance plugin; (c) Euclidean coordinate system of the table ; (d) Target point in 3D space, which need to be found by participant; (e) Surgical probe tool, which is tracked by magnetic navigation; (f) Field emitter of the NDI Aurora magnetic tracker

3.1. Length of Trajectory

Main hypothesis states that adding audio information to visual display would improve performance within 3D navigation task by shortening the length of trajectory between starting and ending point. For each 20s trial, we can calculate the total length of trajectory by summing the length values between the surgical pointer and the randomly placed 3D target location, and then dividing by the number of samples recorded within the trial. These trial lengths of trajectories were then used as data points to investigate the average trial efficiency as a function of both stimulus modality (audio, visual, audio-visual) and sonification type. Trial lengths of trajectories for the different sonification types that are used in the study are also analyzed. This metric is useful only for cases when user find the target point in given time, otherwise, length of the unsuccessful trajectories will not be considered. Formally, trajectory length $l(x(t))$, for trajectory $x(t)$ is given by (7):

$$\begin{aligned} l(x(t)) &= \sum_{t=0}^{T-1} \|x(t+1) - x(t)\|^2 \\ \widehat{x}(t) &= \frac{1}{2k-1} \sum_{s=t-k}^{t+k} x(s) \\ l(\widehat{x}(t)) &= \sum_{t=0}^{T-1} \|\widehat{x}(t+1) - \widehat{x}(t)\|^2 \end{aligned} \tag{7}$$

3.2. Questionnaire responses

With the questionnaire, we would like to see personal opinion of the participants, and how do they rate the difficulty of each condition, and do they find it useful for specific audio methods. After the experiment, each participant completes a short questionnaire regarding the types of feedback that they find to be most useful. 7-point Likert scale is used to rate difficulty of the three conditions, and also the utility of the three different types of sonification.

3.3. Forward/Backward steps

Main hypothesis states that by adding audio cues, user will need less time to achieve the goal, and by that, most of the time using audio guidance, user will move in the direction of the target. Using this metric, we measure either participant understood audio cues and moves probe towards the target point, or participant misunderstood cues and moves probe in wrong/opposite direction. Logging the coordinates every second, from the beginning until the end of the experiment, information about the tendency (whether the participant is getting closer or further from the target over time) will be provided. The forward fwr(x(t)) and backward bwr(x(t)) steps for trajectory $x(t)$ are represented as (8):

$$\begin{aligned} \text{bwr(x(t))} &= \frac{\sum_T \text{bw(t)}}{T} \\ \text{fwr(x(t))} &= \frac{\sum_T \text{fw(t)}}{T} \end{aligned} \tag{8}$$

T represents overall length of the trajectory.

$$\begin{aligned} d(t) &= \|x(t) - t*\|^2 \\ \text{bw(t)} &= \begin{cases} d(t+1) > d(t), & 1 \\ d(t+1) \leq d(t), & 0 \end{cases} \\ \text{fw(t)} &= \begin{cases} d(t+1) > d(t), & 0 \\ d(t+1) \leq d(t), & 1 \end{cases} \\ \text{fwr(x(t))} &= 1 - \text{bwr(x(t))} \end{aligned} \tag{9}$$

Where $t*$ represents the target in 3D space. $d(t)$ represents Euclidean distance between trajectory at time t and target point $t*$ (9).

3.4. Eye Tracking

Main hypothesis states that by using audio combined with IGS, user will spend less time looking in to the screen, and more time focused *in situ*. By setting up web camera that will record participants during experiment, we can measure time that participants spend looking at the screen or looking in to the virtual patient (using only audio guidance). This metric will be used on Audio-Visual combined condition only. Where e (10) represents experiment and $_i$ stands for iteration, $_r$ for repetition.

$$\begin{aligned} \text{e}_{i,r}(t) &= \begin{cases} 0, & \text{participant looking at the screen} \\ 1, & \text{participant looking at the model} \end{cases} \\ \text{mr}_{i,r} &= \frac{1}{T_{i,r}} \int_0^{T_{i,r}} \text{e}_{i,r}(t)dt \end{aligned} \tag{10}$$

mr stands for model ratio of experiment. T - overall time of the experiment (20s).

4. CONCLUSION

The experiment is ongoing and results are expected to be presented. Metrics mentioned above, will be used for evaluation of results.

5. FUNDING

This project is funded by Austrian Reseach Funding Agency (FFG). Project: NavABI. Project number: 855783.

6. REFERENCES

[1] "Mi1000 CONCERTO ABI, Mi1000 CONCERTO PIN ABI surgical guide," MED-EL GmbH.

[2] D. Black, C. Hansen, A. Nabavi, and *ET AL*, "A survey of auditory display in image-guided interventions," *Int. J. Comput. Assist. Radiol. Surg.*, pp. 1–12, 2017.

[3] Z. Bardosi, Y. Özbek, C. Plattner, and W. Freysinger, "Auf dem Weg zum Heiligen Gral der 3D-Navigation: submillimetrische Anwendungsgenauigkeit im Felsenbein," *CURAC*, pp. 155–158, 2013.

[4] "Aurora V2 User Guide," NDI Europe GmbH.

[5] H. Fletcher and W. Munson, "Loudness, Its Definition, Measurement, and Calculation," *B.T.S.J.*, vol. 12, no. 5, pp. 377–430, October 1933.

[6] J. Plazak, S. Drouin, L. Collins, and M. Kersten-Oertel, "Distance sonification in image-guided surgery," *Healthcare Technology Letters*, vol. 4, pp. 199–203, 2017.

INTERACTIVE REAL-TIME CONCATENATIVE SYNTHESIS IN VIRTUAL REALITY

Carl Moore and William Brent

American University
Audio Technology Program
4400 Massachusetts Ave NW
Washington DC, USA
carlmoore256@gmail.com, w@williambrent.com

ABSTRACT

This paper presents a new platform for interactive concatenative synthesis designed for virtual reality and proposes further applications for immersive audio tools and instruments. TimbreSpace VR is an extension of William Brent's TimbreSpace software using the timbreID library for Pure Data. Design and implementation of the application are discussed, as well as its live performance aspects. Finally, future work is laid out for the project, proposing versatile audio manipulation software specifically for XR platforms.

1. INTRODUCTION

Concatenative synthesis techniques typically use a large database of sounds that are segmented into smaller units or grains for synthesis of new sounds. These grains are analyzed to obtain descriptors or attributes pertaining to their timbre, which allows the grains to be reorganized into a new sound known as a "target." In "Free synthesis," the user manually selects audio grains for real-time playback rather than using an automated system [1]. In short, concatenative synthesis is a platform for creating dynamic soundscapes, unconventional musical performances, and novel sound effects. These ends have been achieved through software implementations such as CataRT, and previous iterations of TimbreSpace by William Brent [2]. This paper introduces a new platform for concatenative granular synthesis and audio analysis, implemented in virtual reality.

Virtual and augmented reality (XR) mediums provide exciting new platforms to experience sonic information in the visual domain that has previously been confined to the two dimensions of a screen. Although virtual reality is rapidly becoming a popular platform for unique types of creation and interaction, few interactive synthesis tools have emerged for platforms such as the Oculus Rift or HTC Vive. Mux, a modular synthesizer available Steam VR, is a notable example; however, as of this paper, it still remains in an early access stage of development.

TimbreSpace VR is a tool that provides a palette of timbres encoded as discrete sonic events that can be easily located, patched together, and rearranged in new ways. The primary goal is to create a versatile sonic workspace that provides visual and haptic feedback for sonic exploration and creation.

2. DESIGN

TimbreSpace is a synthesizer which relies largely on bark-frequency cepstral coefficients (BFCCs) for distribution of audio grains in 3D space. BFCCs are a subset of cepstral analysis, a general process of reducing the complexity of spectral analysis results. BFCCs perform well in timbral analysis because they are based on a frequency scale that closely coincides with human frequency perception relative to critical bands [3].

2.1. Interface

The application presents a cloud of grains derived from a given pre-analyzed audio source and provides the user with two wands as the primary means of interaction. Grains are positioned in the 3D space according to any three of the descriptors derived from analysis in PureData using timbreID. Users can specify the X, Y, and Z spatial ordering of grains upon initialization, as well as set the world scale and grain sphere size multiplier. This flexibility is necessary for different use case scenarios.

Grains appear as spheres which are scaled individually according to amplitude. By default, grains are spaced within the scene according to their 1st, 2nd, and 3rd BFCCs, which are meaningful representations of timbre similarity. Each grain is colored according to their 4rd, 5th, and 6th BFCCs, providing yet another dimension of timbral visualization. Therefore, grains of similar color and location will sound similar.

The aesthetic decisions for the interface are largely carried over from William Brent's original implementation of TimbreSpace. The location and color of grains are simple, easy to understand indicators of timbre, and several first-time users expressed a strong understanding of these tonal-visual correlations. Additional features of grain scaling which were not present in the original version of TimbreSpace also help to further indicate the status of a grain before hearing it play.

The overall aim of the visual aesthetics is to communicate tonal characteristics of regions within the grain cloud so that the user can focus on conceptualizing a tone in their mind and find it quickly, rather than searching for the sound they want by audition.

2.2. Controls

In previous real time concatenative synthesis applications, user interaction was very much constrained within the spatial dimensions. William Brent's open-air fingertip navigation was an improvement to existing one-dimensional gestures available in concatenative synthesizer applications such as CataRT, utilizing IR

https://doi.org/10.21785/icad2019.068

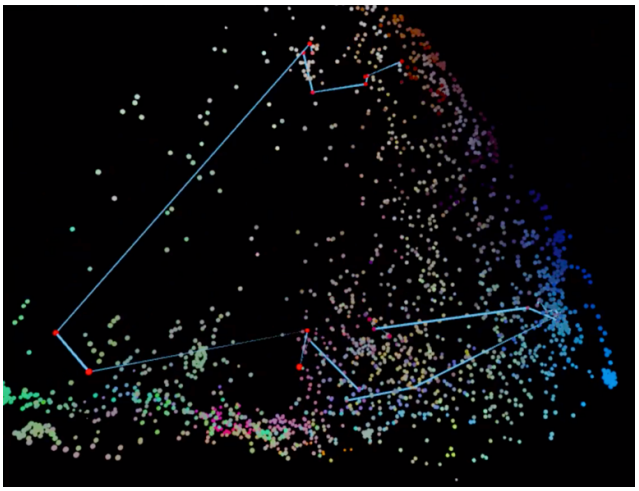


Figure 1: A grain cloud comprised of around 2000 electronic drum samples

fingertip tracking as a means of interaction with audio grains [4]. In addition, gestural combinations such as pinch and rotation of the hands enabled the modulation of effects parameters such as delay and transposition [4].

TimbreSpace VR extends the concept of wrist rotation as an effects modulation parameter, however, gestural tracking is now gathered from the Oculus Touch controller's built in gyroscope and accelerometer.

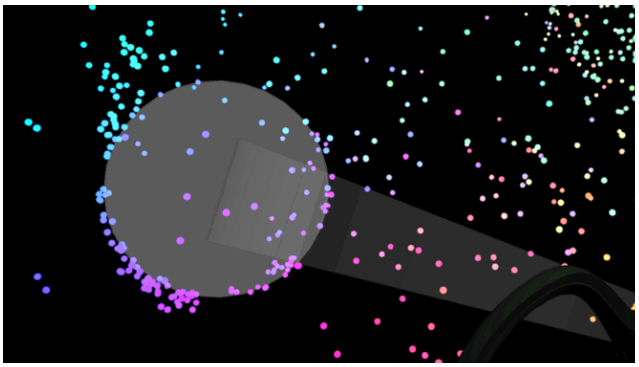


Figure 2: Grains colliding against the bubble cursor wand

Two separate "wands" with spherical tips are provided for playing and sequencing the grains. They use the various features of the Oculus Touch controllers to accomplish all the actions available within TimbreSpace VR. The main goal is to allow the user to reach any grain visible within the scene through the use of natural, intuitive controls, while maintaining precise sonic control.

The spherical wand tip is a three-dimensional implementation of a bubble cursor, a GUI selection tool which possesses benefits from accuracy by dynamically resizing [5]. The diameter can be resized via the combination of a control button and a "doorknob twist" style action, allowing for quick and accurate moves which are useful for live performances. Both wand tips can also be repositioned along the Z-axis independently, allowing the user to reach closer or further into the grain cloud.

2.3. Workflow

Audio scenes or "soundscapes" are currently loaded into the Unity editor as a collection of text files (containing attribute data,) and mono .wav files, which can then be read by TimbreSpace VR. To generate these assets, preliminary steps involving normalization and silence removal, and then audio analysis for feature extraction in PureData are required to import new soundscapes. The process is not particularly user-friendly at this stage but can easily be integrated into a single package with further work.

Soundscapes are typically generated with an artistic concept in mind, or out of curiosity. Collections of thousands of flute samples, bird calls, or rain storms are just a few examples of content that has been experimented with. These samples are often organized first in a handful of ways. Dynamic scenes can be created by combining sonic elements that evoke certain imagery into a DAW session. The result is a nonlinear audio scene in which a performer, much like a Foley artist, acts out the sonic scene within TimbreSpace. Because the process of timbre analysis removes the temporal component entirely, clips can be sewn together in random orders. TimbreSpace also works well with collections of instrument samples and loops, which can be navigated and played as an instrument.

Samples of a percussive nature are often put through a custom PureData patch before analysis, in order to optimize feature extraction in timbreID. This stage utilizes the Bark~ object within timbreID to detect transients and concatenate thousands of samples into a specified window for further analysis.

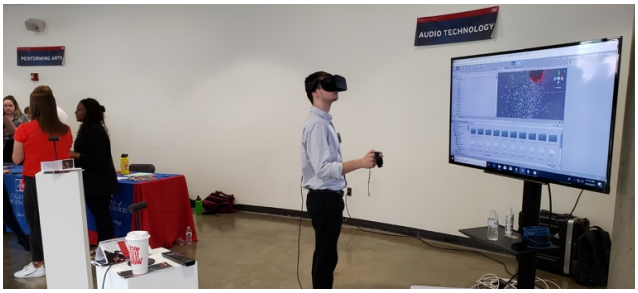


Figure 3: A user exploring a grain cloud in TimbreSpace VR

Figure 4: A large constellation seen from a distance in the scene

2.4. Constellations

Sequencing of audio events is one the primary new innovations of TimbreSpace VR. Grain sequences (referred to as Constellations) are dynamic groups of audio segments that are looped and played back in a specified order. They are displayed as lines running from one grain in the loop to the next, graphically indicating the trajectory of the sound through the scene. Using the constellation editor, the user can form musical ideas in live and non-real-time scenarios. The ordering of grains within the 3D space based on audio features allows for sounds to blend together in novel, timbrally coherent arrangements.

Constellations allow for the creation of musically discrete patterns and rhythms, making TimbreSpace VR a space for dynamic musical composition. They take a completely different interactive and visual approach to musical sequencing when compared to traditional musical sequencers, leading to abstract and unexpected phrases, loops, and textures. The spatial workflow shows the potential of XR applications to bring new tools to artists and contribute to new forms of creative expression.

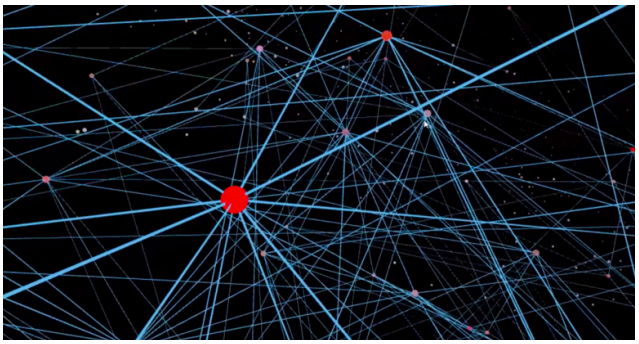


Figure 5: A constellation up close

2.5. Physics

Simulated physical response of audio grains provides a basis for many interactions unique to TimbreSpace. The kinetic response of grains attempts to introduce dynamics to a performance that provide feedback to the performer and audience in the form of action-sound relationship cues. Each unit is pinned to a given position in 3D space with an elastic bond. Interactions via the two bubble wands can stretch these links upon intersection with a grain, displacing all grains within the wand's diameter. Grains bounce about the surface of the wand causing new collisions with the wand as well as with other grains, activating those sonic events in random orders. The result is a controlled method of exciting grain clusters, creating unpredictable concatenation patterns that avoid audibly looping.

3. TECHNOLOGY

The core application is built using the Unity game engine and written in C#. TimbreSpace is being developed for the Oculus Rift headset and controllers. Audio descriptor metadata is currently generated externally using William Brent's timbreID library in PureData, and exported as a text file containing each of the 26 attributes (descriptors) for each grain event.

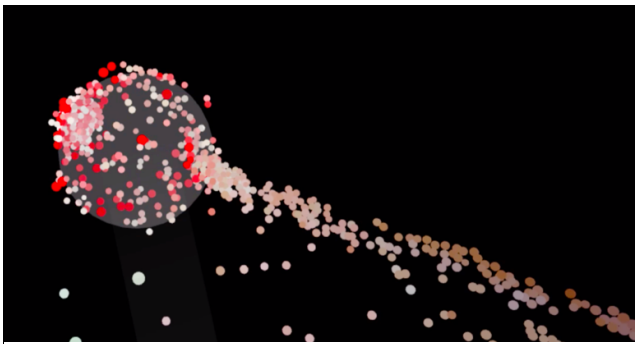


Figure 6: The bubble cursor's simulated physical interaction with audio grains

In this implementation of timbreID, the patch segments a given audio file and measures various audio attributes within each frame. These attributes include frame number, pitch, amplitude, harmonicity, spectral centroid, and BFCCs. Window size has a significant impact on the sonic results of concatenative synthesis, therefore timbreID employs several different window sizes for its analyses, ranging from 2048 to 16384 samples in length. This provides flexibility in TimbreSpace VR, which has the ability to load a scene at a variety of different grain sizes. Because the exported database contains a list of attributes for each grain, any of this information can be referenced in the scene.

Running on current graphics hardware (Nvidia GTX 1070), TimbreSpace VR can render scenes with grain clouds up to around 5000 grains without significant performance reduction or framerate stutter. This equates to around 15 minutes of audio content if played back at a grain size of 8192 samples, an ideal length for musical exploration.

4. FUTURE WORK

The full realization of TimbreSpace is a fully modular sonic sandbox for artists. A collection of tools will extend the capabilities of the control set so that any audio event or cluster of audio events can be added, removed, modified, networked and sequenced. These events will include audio grains, filters, and logical operators which will be exposed to a patching network. Alongside this, greater functionality will be added to the live performance controls to allow precise real-time synthesis.

Additional GUI controls currently being implemented will greatly expand the functionality of TimbreSpace VR. Constellations will be able to be duplicated, saved, and recalled, as well as played polyphonically alongside any number of other active sequences. Information about each grain (determined during analysis) will be accessible in the interface, making TimbreSpace VR a useful platform for visual audio analysis. The ultimate goal is to provide a space for sonic exploration and experimentation that feels natural and provides new tools for artistic expression.

5. REFERENCES

[1] Schwarz, Diemo. "Concatenative Sound Synthesis: The Early Years" Journal of New Music Research35, no. 1 (2006): 3–22. doi:10.1080/09298210600696857.

[2] Diemo Schwarz, Gregory Beller, Bruno Verbrugghe, Sam Britton. “Real-Time Corpus- Based Concatenative Synthesis with Catart”, *Expanded version 1.1 of submission to the 9 Int. Conference on Digital Audio Effects*

[3] Brent, William. “Perceptually Based Pitch Scales in Cepstral Techniques for Percussive Timbre Identification”.

[4] Brent, William. “Physical Navigation of Virtual Timbre Spaces with TimbreID and DILib”, Proceedings of the 18th International Conference on Auditory Display, 2012.

[5] Grossman, Tovi, and Ravin Balakrishnan. “The Bubble Cursor.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems — CHI 05, 2005. doi:10.1145/1054972.1055012.

BREATHING SPACE: BIOFEEDBACK SONIFICATION FOR MEDITATION IN AUTONOMOUS VEHICLES

Yota Morimoto

mdoos,
Bezuidenhoutseweg 65-11,
The Hague, 2594AC, Netherlands
yota@tehis.net

Beer van Geer

Calmspaces,
Wagenstraat 125,
The Hague, 2512AT, Netherlands
beer@calmspaces.nl

ABSTRACT

The collective, c*almspaces,* sets out to create spaces for relaxation and contemplation through traditional architectural approach combined with modern digital technology.

The ongoing project of the collective, *breathing space* (*ademruimte* in Dutch), uses unobtrusive sensing technology to monitor one's breathing, and through designed light and sonic guides, the project tries to enhance the breathing exercise beneficial to regulating one's emotion.

The paper illustrates the project and its relevance to and potential for in-vehicle development. We then discuss the details of our implementations, along with video documentations of the early prototype, and a recently completed installation work.

1. BACKGROUND

Calmspaces gathered individual specialists for the project *breathing space* – a composer designing ambient public sonic spaces, interaction designers working with attention disorder children, a psychologist, meditation therapists, an architect, and an urban planner – those who share concerns regarding the increasing anxiety issues of present day life.

In 2018 the project received the Dutch national grant for creative industry and the collective is currently working on prototypes of space design for augmented breathing exercise as a possible intervention to counter such issues.

2. BREATHING AS MEDITATION

The effects and its underlying neural mechanism of attention-to-breath as a basic mindfulness practice has been studied through fMRI viewing [1] and self-regulation of breathing is proposed as first-line treatments for stress, anxiety, depression, and some emotional disorders [2][3]. As the practice of attentive breathing merely requires one's conscious action, it can be performed inside a fully autonomous automobile without much hassle and may become a fulfilling in-vehicle experience connected to one's well-being.

3. IN-VEHICLE BREATHING SPACE

As sonification for emotion regulation/meditation of drivers in autonomous vehicles, we propose an in-vehicle adaptation of our breathing space project. Such breathing exercise does not require extensive duration (typically 10-15 minutes with measurable beneficial effect [1]) and can be performed by the driver of an autonomous vehicle as well as by other passengers on a daily routine (e.g., during commuting).

To avoid safety risks we propose the use of auditory display as the primal guide for the exercise (perhaps with optional ambient light feedback for an enhanced cue). The exercise can be immediately terminated upon unexpected traffic situations or takeover requests because it does not require any manual or feet operation, posture change or the involvement of sight. In the following paragraphs we detail the implementations of the realized installation work which we propose for the adaptation for autonomous vehicles.

3.1. Sensing

Ballistocardiography (BCG), the measure of ballistic forces on the heart, allows for noninvasive cardiac monitoring without direct contact to the skin [4]. BCG sensor technology thus integrates well into car seats. By adding a frequency analysis algorithm we create a user feedback parameter useful for realtime ambient sound synthesis and lighting control.

In the current standalone installation version completed in early May this year, the user is provided with a knob to adjust and set the pace of the breathing exercise, guided with sound, image projection and lighting [5]. The BCG sensor has also been installed at the bench (the sheet seen shortly at around 8 seconds into the video).We use frequency analysis algorithm to observe if the user's breathing (measured through BCG) matches that of the guiding pace. When the user's breathing pace matches closely to that of the guiding pace, the system gives feedback with brightened image and light as well as with a more resonant sound.

3.2. Sonification

The author has developed a library for in-vehicle auditory display using the sound synthesis programming environment SuperCollider [6]. The early prototype of realtime sonification of BCG is demonstrated in the video [7].

For the sake of demonstration only, we used a GUI slider to show how the overall sound structure follows the difference in speed setting (Fig. 1). The expanding and contracting circle corresponds to the breathing pace, which, in case of the installation work, is adjusted by the participant, but can also be directly controlled by the BCG data for sonification purpose.

https://doi.org/10.21785/icad2019.031

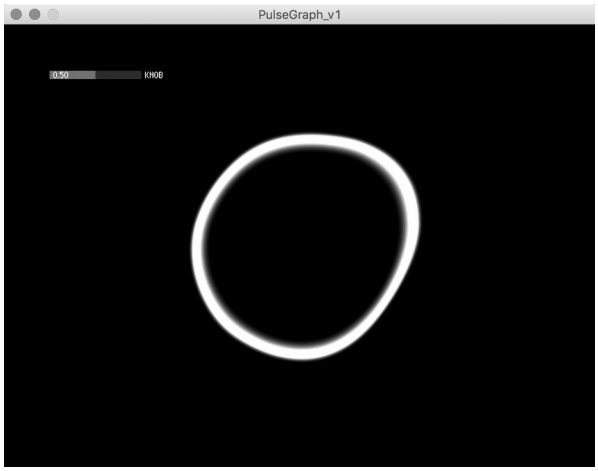


Figure 1: prototype demonstration GUI and visualization.

The overall sonic contour represents the inhale/exhale of breathing with the aim to support the exercise and provide a relaxing and meditative atmosphere. There are other bell-like sounds which tick different number of breathing cycles.

4. DISCUSSION

The exercise demonstrated in the current installation work can well be adapted for fully autonomous cars. As is seen in demonstration video, the BCG sensor system can relatively easily be installed in the car seats. The lighting and image can also be adapted as an ambient light effect to complement the primal auditory guide of the exercise. We are therefore in search of partners of researchers and automobile industry to develop prototypes of the in-vehicle breathing exercise.

The current installation work only provides instantaneous feedback to the participant (whether his/her breathing pace matches with the guide). For future work we consider implementing analysis of heart rate variability to follow a longer period of change the user causes through the exercise. We also plan to create more variations of sound so that the users can choose what they favor.

5. ACKNOWLEDGMENT

The authors thank the Dutch Creative Industries Fund for their support.

6. REFERENCES

[1] Mindful Attention to Breath Regulates Emotions via Increased Amygdala-Prefrontal Cortex Connectivity. Anselm Doll in NeuroImage, Vol. 134, pages 305–313; July 1, 2016.

[2] Self-Regulation of Breathing as a Primary Treatment for Anxiety. Ravinder Jerath et al. in Applied Psychophysiology and Biofeedback, Vol. 40, No. 2, pages 107–115; June 2015.

[3] Efficacy of Paced Breathing for Insomnia: Enhances Vagal Activity and Improves Sleep Quality. H. J. Tsai et al. in Psychophysiology, Vol. 52, No. 3; pages 388–396; March 2015.

[4] Simplified real-time heartbeat detection in ballistocardiography using a dispersion-maximum method. Sun-Taag Choe, We-Duke Cho in Biomedical Research 2017; 28 (9): 3974-3985

[5] https://youtu.be/3_nSyOc3LzM

[6] A Software Library for In-vehicle Auditory Display. Yota Morimoto in Proceedings of the "In-Vehicle Auditory Interactions" Workshop at the 21st International Conference on Auditory Display (ICAD-2015) Graz, Austria: TU Graz.

[7] https://youtu.be/vbS49LJCKhk

PRELIMINARY GUIDELINES ON THE SONIFICATION OF VISUAL ARTWORKS: LINKING MUSIC, SONIFICATION & VISUAL ARTS

Chihab Nadri, Chairunisa Anaya, Shan Yuan, and Myounghoon Jeon

Mind Music Machine Lab
Virginia Polytechnic Institute and State University,
Department of Industrial and Systems Engineering,
1185 Perry Street Blacksburg, VA 24061 USA
{cnadri, danaya14, shany9, myounghoonjeon}@vt.edu

ABSTRACT

Sonification and data processing algorithms have advanced over the years to reach practical applications in our everyday life. Similarly, image processing techniques have improved over time. While a number of image sonification methods have already been developed, few have delved into potential synergies through the combined use of multiple data and image processing techniques. Additionally, little has been done on the use of image sonification for artworks, as most research has been focused on the transcription of visual data for people with visual impairments. Our goal is to sonify paintings reflecting their art style and genre to improve the experience of both sighted and visually impaired individuals. To this end, we have designed initial sonifications for paintings of abstractionism and realism, and conducted interviews with visual and auditory experts to improve our mappings. We believe the recommendations and design directions we have received will help develop a multidimensional sonification algorithm that can better transcribe visual art into appropriate music.

1. INTRODUCTION

The relationship between visual art, music, and technology has frequently been close, as their combined use resulted in masterpieces in the past [1]. With the increasing amount and interest in sonification techniques [2] and image processing techniques using machine learning, transcribing visual experiences into appropriate auditory experiences is now a real possibility. However, the use of these advances in technology has been concentrated on the sonification of images in general and increasing accessibility through short sound feedback. Based on that, we decided to design a sonification algorithm tailored to transcribing visual artworks and appropriately conveying its wider cognitive and emotional message, with the objective of establishing guidelines for the sonification of artworks based on their characteristics (art style, mood, elements…etc.). Additionally, we have conducted interviews with experts from the fields of sonification, visual arts, and music in order to ascertain the main objectives our sonification algorithm should accomplish, as well as learn from their expertise. We expect this study contributes to the development of sonification algorithms in providing the most effective image and data processing techniques suited for the sonification of different artworks.

1.1. Related Works

Research on visual graphics sonification has focused on providing individuals with visual impairments more ways and tools to experience visuals [3]–[5], and highlight useful data processing techniques used to accurately aid the transcription process, such as shape and edge detection machine learning algorithms. In fact, machine learning algorithms have played an increasing role in image processing tasks [6]. Much research has been done concerning saliency detection and salient region cropping in images [7], [8], as saliency provides a good indicator for the importance and relevance of specific areas of the image. Additionally, saliency is an image parameter that plays an important role on eye fixations and visual perception [9], [10], which machine learning algorithms take into account when going through training datasets.

Machine learning has also been used in the arts to classify artworks and identify artistic styles [11], [12]. Such algorithms can play an important role in streamlining classification tasks for galleries and art directories with large collections of artworks and prove practical when implementing sonification techniques. On the other hand, machine learning can also play a role for the composition of music, adapting and learning from performers' musical genre and style to identify and create similar pieces [13].

Aesthetic research on visitor experience at art galleries [14] has revealed visitor patterns of short yet often repetitive viewing of the same paintings and artworks, which can serve as performance criteria for the effectiveness of different visual artwork design experiences. Related research on sound and color mappings has shown that strong associations between music and appearance can exist for nonsynesthetic individuals [15]. Strong associations were also found between music-color mappings and emotions, as different emotions can be evoked through them [16], and the study also found other musical parameters like tempo and mode can play an important yet unclear role.

Prior research on photographic sonification has also yielded several sonification methods and algorithms, adhering to a musical approach focused on musical structure [17] or a naturalistic one following viewing tendencies [18]. Additionally, research on visual saliency in paintings has yielded several algorithms that can imitate human gaze perception and fixations, regardless of art movement the visual artwork belongs to [19]. Indeed, psychological experiments and other saliency-based algorithms have highlighted how the key to understanding art is the

https://doi.org/10.21785/icad2019.074

identification of the perceptual process, and how salient regions can correlate with art movements [20].

1.2. Initial Plans and Mappings

Our sonification would apply to visual artworks from different art styles, with abstract paintings yielding a different sound output than that produced from realistic paintings (Figure 1). The sonification algorithm would use mapping strategies between a variety of visual and musical parameters in order to compose and differentiate sound outputs for different artworks (Table 1). While a variety of mapping strategies already exist [2], we planned on conducting experiments to evaluate the significance of each mapping pair.

Table 1: Initial mapping pairing strategies considered between visual and auditory parameters

Visual Parameters \ *Auditory Variables*	*Pitch*	*Tempo*	*Mode*	*Timbre*	*Loudness*	*Musical Composition*
Hue	X					
Brightness		X			X	
Saliency	X		X			
Size	X					
Art Style				X		X

Following the proposed experiments, data processing techniques, including shape detection and sentiment analysis machine learning algorithms, would supplement the sonification program with further accuracy. Later trials and implementation for art galleries would be tested using the completed algorithm, which should accomplish goals set at the start of the project.

Performance criteria currently considered for the success of the algorithm are split between musical criteria and semantic ones, listed below:

1. Sonification sound organization: the ability of the sound output to be arranged in a musical way, as opposed to a random succession of notes that are off-key.
2. Sound quality: how pleasant the sonification sounds to participants, and how likely they would want to listen to the music produced.
3. Mood and emotion: feeling conveyed by the sonification. Learning about the emotional meaning of each piece as identified by participants can allow us to validate our experimental mappings and uncover unexpected pairings.
4. Matching quality: the ability of the sonification to match the painting's cognitive and emotional message and reflect the scene presented by the painting. Measuring this parameter provides a clear, albeit subjective, scale on the sonification's success at one of its primary goals (carrying similar meaning to the artwork).

Due to the inherent subjectivity in art appreciation, performance criteria were also dependent on the user. However, through the expert interview process, more parameters would be identified and used for the evaluation of the sonification algorithm and its successive iterations.

Figure 1: Composition VII, 1913 - Wassily Kandinsky (left), The Raging One, c.1830 - Camille Corot (right), two sample artworks belonging to abstractionism and realism respectively, the two art styles considered for initial experimental tests.

2. METHODS

2.1. Technical Details

Our program is an application to transcribe digital images of different artworks into MIDI files. We have used JythonMusic[21], a python-based environment for music creation that can also use Java libraries. Given the image provided, our software creates midi files that map image parameters to musical characteristics, creating the music piece.

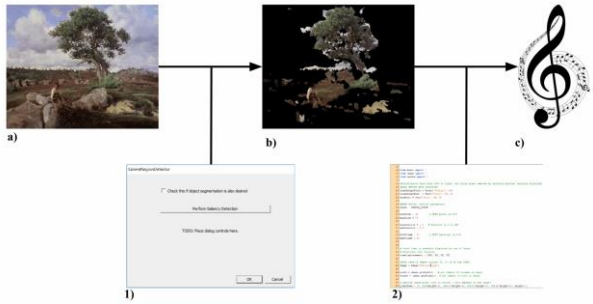


Figure 2: Sample sonification process using saliency mapping, running an image a) through the saliency segmentation software 1) before using JythonMusic code 2) to turn it into music c). The Raging One, c.1830 - Camille Corot.

Over the course of the expert interviews, the saliency and segmentation model created by Achanta et al. was used [7] as an example of a saliency algorithm.

2.2. Participants

As the interview process is still ongoing, five experts, from fields of interest in our study (psychoacoustics, visual arts, sonification), participated in our interview separately. Professors in their field, all had a lengthy experience in their field, with an average experience of 19.4 years.

2.3. Procedure

Experts were first asked questions relevant to their field, such as their previous work and projects. Then, the interviewer asked questions to find out which visual, musical, or sound parameter each expert uses to convey meaning and

accomplish their work objectives, as well as performance measures used in their respective fields. Questions regarding the potential implementation of artwork sonification at art venues were asked next, aiming to discover similarities and differences in perspective between each field. Finally, experts provided advice and observations based on the experimental design, paintings, sonification outputs, or machine learning techniques considered by the team, such as saliency detection.

3. RESULTS

Expert responses varied according to fields of study, with intragroup similarities being found, as well as general trends concerning performance measures shared between all experts (Table 2). Visual arts experts alongside a psychoacoustics expert expressed a preference for the sonification output to carry limited meaning at first. From the perspective of psychoacoustics, following a gradually more complex sound methodology and protocol is essential to reach a complete algorithm. While visual arts experts shared comparable views, not overshadowing the experience of viewing the painting was their main reason. Simply capturing the general feeling of the painting was deemed enough, although greater information would prove beneficial depending on the target audience. Indeed, every visual art expert independently concluded that multiple sonification outputs for any single painting, different in terms of breadth of cognitive and emotional information carried over, would offer the most flexibility for subjective art appreciation at art venues and potential accessibility issues for people with visual impairments.

Expert views on challenges facing the project were less uniform and heavily shaped by their experience. While sonification and psychoacoustics experts emphasized sound quality and methodology respectively, visual arts experts' main priority was to respect the individuality involved in art interpretation. This manifested itself in concerns over matching the painting and not limiting the artwork to a single interpretation.

Important parameters used by experts in their work included brush movement, color choice, geometric shapes, and depicted objects and scenes present in the artwork for visual arts experts. Pitch, timbre cutoff frequency, tempo, and stereo pan were some of the musical parameters mentioned by musical experts, which helped verify mapping strategies used in earlier sonification projects [2].

Concerning relevant data processing techniques, experts held mixed views on saliency detection. For sound experts (psychoacoustics and sonification), saliency detection would constitute a novel way to determine the temporal flow of the artwork. Using the most salient regions detected first, the sonification algorithm could imitate naturalistic viewing patterns and focus on composing salient elements in detail. For visual arts experts, the importance of the background in setting the mood of a painting meant that saliency detection should be used as an additional parameter when sonifying the entire painting. Additionally, sentiment analysis of an artwork's description (as present in art venues) was perceived as an efficient way to capture information on the painting's background, which can then be implemented into the sonification output (e.g. location of artist at time, state of unrest at location… etc.). Since artworks descriptions usually include the art style present in the artwork, using machine learning algorithms to classify paintings was deemed unnecessary.

Experts agreed researchers should make use of visitor viewing patterns as an objective performance measure. Specifically, visual arts experts considered the number of times a participant reviews a visual artwork more important than viewing length. Another suggestion was to test participants' understanding of the artworks' intended message as a measure of success.

Lastly, experts' opinions on differences that should be present between the sonification of realistic art and abstract art followed similar reasoning, although each expert provided different guidelines (Table 3). According to experts, the sonification of realist art needs to incorporate or convey the scene presented in the artwork by adding iconic sounds for objects in the painting, although sound selection would become a challenge. Also, saliency in realism should be limited, as backgrounds carry important meaning necessary for interpretation. As for abstract art, brush movements, geometric shapes used, and color choice should have a large impact on the sonification output. The use of saliency in abstract art could include the sonification of salient regions only, although the preferred approach would be to alternate sound output between different salient areas. Common amongst both styles was the need for relevant instruments to

Table 2: Initial Expert Feedback during Sonification Project Interview

Experts	Experience	Parameters to convey meaning	Art venue visitor approaches	Challenges
Sonification	23yrs	Pitch, amplitude, timbre cutoff frequency, performance frequency, stereo pan, counterpoint. Respecting painting shape and viewing pattern	Consistent sound quality for painting, i.e. a musical masterpiece for a great work of visual artwork	Sound quality
Psychoacoustics	11yrs	Speech intelligibility, tempo, color contrast for abstract art. Painting elements for realist art	Simple sonification output that can become more complex later, focusing on carrying over a few meanings while letting visitors complete understanding with visual data	Sound protocol and methodology
Visual Arts	20yrs	Brush movement and complementary color choices for abstract art. Objects for realistic paintings	Offering multiple sonification results. Output can be different from painting, focus on offering different modality without minding all musical qualities	Respecting differences in interpretation
Visual Arts	13yrs	Appropriate choice of music and art, choice remains arbitrary, especially for VR projects	Capturing general idea of painting would be fine, emotional ellicitation and clues on painting elements	Matching painting
Visual Arts	30yrs	Color brightness, lines and shapes (geometric or real objects)	Offering multiple sonification results. The more text, information, and meaning provided to visitors the better. Amount of times one views artwork more impactful than length of time viewed	Music not overpowering visual experience

be used according to the artwork's background.

Table 3: Expert consensus on sonification differences between realism and abstractionism

Realism	Abstractionism
Proper representation of painting elements through iconic sounds (e.g. wind blowing)	Color choice, brush movement, and shapes used as most important visual parameters that express artist's mood
Limited scope of saliency (used for object identification, not impacting breadth of painting elements)	Saliency can be used to reduce amount of visual data sonified, but preference on alternating between salient regions in a general sonification of the work better
Instruments/timbre used relevant for time period/geographical location/state of the artist	**Instruments/timbre used relevant for time period/geographical location/state of the artist**

4. CONCLUSION & FUTURE WORK

In this exploratory study, we interviewed experts from the fields of music, sound, visual arts, and sonification to establish design directions for a sonification algorithm that can appropriately transcribe visual artworks into music.
Through the interviews, initial mappings and feedback were gathered, and preliminary guidelines on the sonification of realistic art and abstract art were determined.
Additional expert interviews will be conducted, gathering more data and validating early findings. Future experiments will take the findings of this study into account, adjust the sonification algorithm for recommended techniques and changes experts pointed out, and include additional performance measures found in the study.

5. REFERENCES

[1] B. N. Walker and C. M. Bruce, "ICaD 2013 of biocybernetics A case study of the Accessible Aquarium Project," pp. 39–44, 2013.

[2] G. Dubus and R. Bresin, "A systematic review of mapping strategies for the sonification of physical quantities," *PLoS One*, vol. 8, no. 12, 2013.

[3] M. Jeon, R. J. Winton, J.-B. Yim, C. M. Bruce, and B. N. Walker, "Aquarium fugue: interactive sonification for children and visually impaired audience in informal learning environments," *Proc. 18th Int. Conf. Audit. Disp. (ICAD 2012)*, pp. 246–247, 2012.

[4] S. Cavaco, J. T. Henriques, M. Mengucci, N. Correia, and F. Medeiros, "Color Sonification for the Visually Impaired," *Procedia Technol.*, vol. 9, pp. 1048–1057, 2013.

[5] T. Yoshida, K. M. Kitani, S. Belongie, and K. Schlei, "EdgeSonic : Image Feature Sonification for the Visually Impaired Categories and Subject Descriptors," *Image Rochester NY*, pp. 1–4, 2011.

[6] S. van der Walt *et al.*, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 2014.

[7] R. Achantay, S. Hemamiz, F. Estraday, and S. Susstrunky, "Frequency-tuned salient region detection," *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, no. Ic, p. 1597 { 1604, 2009.

[8] L. Zhang, Y. Gao, R. Ji, Y. Xia, Q. Dai, and X. Li, "Actively learning human gaze shifting paths for semantics-aware photo cropping," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2235–2245, 2014.

[9] I. Fuchs, U. Ansorge, C. Redies, and H. Leder, "Salience in Paintings: Bottom-Up Influences on Eye Fixations," *Cognit. Comput.*, vol. 3, no. 1, pp. 25–36, 2011.

[10] U. Leonards and W. Singer, "Conjunctions of colour, luminance and orientation: The role of colour and luminance contrast on saliency and proximity grouping in texture segregation," *Spat. Vis.*, vol. 13, no. 1, pp. 87–105, 2000.

[11] Y. Bar, N. Levy, and L. Wolf, "Classification of artistic styles using binarized features derived from a deep neural network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8925, pp. 71–84, 2015.

[12] B. Saleh and A. Elgammal, "Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature," 2015.

[13] R. De Prisco, D. Malandrino, G. Zaccagnino, R. Zaccagnino, and R. Zizza, "A Kind of Bio-inspired Learning of mUsic stylE," in *Computational Intelligence in Music, Sound, Art and Design*, 2017, pp. 97–113.

[14] C. C. Carbon, "Art perception in the museum: How we spend time and space in art exhibitions," *Iperception.*, vol. 8, no. 1, 2017.

[15] S. E. Palmer, K. B. Schloss, Z. Xu, and L. R. Prado-Leon, "Music-color associations are mediated by emotion," *Proc. Natl. Acad. Sci.*, vol. 110, no. 22, pp. 8836–8841, 2013.

[16] T. Tsang and K. B. Schloss, "Associations between Color and Music are Mediated by Emotion and Influenced by Tempo," *Yale Rev. Undergrad. Res. Psychol.*, pp. 82–93, 2010.

[17] N. Rönnberg and J. Löwgren, "Photone: Exploring Modal Synergy in Photographic Images and Music," no. Icad, pp. 73–79, 2018.

[18] A. Polo and X. Sevillano, "Musical Vision: an interactive bio-inspired sonification tool to convert images into music," *J. Multimodal User Interfaces*, no. i, 2018.

[19] R. G. Condorovici, R. Vrânceanu, and C. Vertan, "Saliency Map Retrieval for Artistic Paintings Inspired from Human Understanding," *Proc. SPAMEC*, pp. 101–104, 2011.

[20] B. Stoica, L. Florea, A. Badeanu, A. Racoviteanu, I. Felea, and C. Florea, "Visual saliency analysis in paintings," *ISSCS 2017 - Int. Symp. Signals, Circuits Syst.*, 2017.

[21] B. Manaris, B. Stevens, and A. R. Brown, "JythonMusic: An environment for teaching algorithmic music composition, dynamic coding and musical performativity," *J. Music. Technol. Educ.*, vol. 9, no. 1, pp. 33–56, 2016.

IS SONIFICATION DOOMED TO FAIL?

John G. Neuhoff

The College of Wooster,
Wooster, OH 44691, USA
jneuhoff@wooster.edu

ABSTRACT

Despite persistent research and design efforts over the last twenty years, widespread adoption of sonification to display complex data has largely failed to materialize, and many of the challenges to successful sonification identified in the past persist. Major impediments to the widespread adoption sonification include fundamental perceptual differences between vision and audition, large individual differences in auditory perception, musical biases of sonification researchers, and the interdisciplinary nature of sonification research and design. The historical and often indiscriminate mingling of art and science in sonification design may be a root cause of some of these challenges. Future sonification design efforts that explicitly strive to meet either artistic or scientific goals may lead to greater clarity and success in the field and more widespread adoption of useful sonification techniques.

1. INTRODUCTION

This year marks the 20th anniversary of the publication of *The Sonification Report* [1]. An international panel of sonification researchers produced the report which identified the state of the field at that time and a research agenda going forward. In the time since the report, sonification has grown slowly but steadily. A March 2019 search of all *Web of Science* databases showed a nearly four-fold increase in the appearance of the term "sonification" in literature over the previous 20 years. However, despite some innovative one-off successes, the widespread adoption of sonification to present complex data has largely failed to materialize. In fact, most sonifications in widespread use today are simple binary messages (e.g., *ding! your seatbelt is unlatched*).

However, researchers have been anticipating a tipping point in the field for some time. The authors of *The Sonification Report* in 1999 wrote that *"Sonification will gain significant momentum once several specific applications become widely used. However, until there are intuitive, efficacious applications, skeptics will adhere to current display solutions."* Twenty years later the quest for success and the "intuitive, efficacious application" or the *killer app*, as it came to be known, continues [2, 3].

2. HOW SHOULD SUCCESS BE DEFINED?

Is the killer app the appropriate metric by which we should measure the success of sonification? Should researchers continue to strive to make data sonification as ubiquitous a means of data representation as the bar graph? Some would say no. Nees [2] for example, argues that sonification is simply one kind of tool that can be used to display data. He cites several successful (if not ubiquitous) examples where sonification "works." Nees argues that if sonification in the appropriate context conveys the intended information, then the field as a whole can be considered successful. However, even those who promote most strongly the viability of widespread sonification and argue that a killer app is *not* required for success acknowledge that many of the roadblocks to successful sonification identified in *The Sonification Report* by Kramer et al. in 1999 are still prevalent today [2, 4-6].

3. PERSISTENT CHALLENGES TO SONIFICATION

The quest for sonification success has yielded several different approaches to representing data with sound. *Audification, auditory icons, earcons, parameter mapping,* and *model-based sonification* all have strengths and weaknesses. Their collective promise has led to somewhat of a public fascination with the idea of sonification and a relentless sense of optimism within the sonification research community [7, 8]. Unfortunately, sonification is more often viewed by the public as an entertaining curiosity than as a scientific tool for understanding data [9]. Some of the reasons for this include fundamental perceptual differences between vision and audition, large individual differences in auditory perception, perceptual crosstalk in audition, inherent musical biases of sonification researchers, and the interdisciplinary nature of the field.

The precision of vision versus audition. In humans, there are approximately ten times as many cortical neurons devoted to vision as there are to hearing. It should come as no surprise then that in all but the perception of time, perceptual judgments made with the eyes are usually more precise than those made with the ears. For example, the most common representational dimension used in visual graphs is length. The most commonly used dimension in auditory graphs is pitch [10]. If we examine the just noticeable difference (the minimum amount that a stimulus needs to change in order for the observer to notice the change) in each dimension, we find the percentage of pitch change required to notice a change is about twice the percentage of line length change required [11, 12]. If we examine the spatial resolution of the two modalities we find that the auditory system has a resolution or *Minimum Audible Angle* of between one and two degrees azimuth [13]. The corresponding visual measure, the *Minimum Angle of Resolution* is about 60 times more precise with a resolution of 1-2 minutes of arc [14]. In almost all dimensions but time, the precision with which we can perceptualize data is greater in vision. This disparity obviously presents some difficulty for making a sonification that is on par with a typical visualization.

Individual differences in audition. In addition to differences in precision, the *polarity* of mapping data to an auditory representation is more unreliable than mapping data to visual representation. For example, in a visual graph, "up" almost always represents "more." However, the same cannot be said for sonification. When data variables such as physical size or number of dollars are mapped to pitch, listeners are almost

https://doi.org/10.21785/icad2019.069

evenly split on the question as to whether increasing pitch should represent increasing or decreasing values of the variable in question. Other variables show similar individual differences [5, 15, 16]. Some listeners with little to no musical experience even show a poor grasp of what the words "up" and "down" *mean* in the context of pitch change [17].

Widespread individual differences in music cognition further compound these problems. For example, musicians have lower thresholds for pitch discrimination than non-musicians [18], show enhanced attentive processing of non-speech sounds [19], and demonstrate better acuity in pitch and time [20]. Perhaps the most critical difference between musicians and non-musicians in extracting information from sonification lies in the ability to segregate auditory streams. Extracting information about any one variable from a display requires selectively attending to the variable of interest and suppressing attention to the other simultaneously sounding streams, a task at which musically experienced listeners excel and novices struggle [21]. Thus, variability in music cognition leads to variability in the comprehension of most sonifications.

Auditory perceptual interaction and asymmetry. Compounding the problem of individual differences is the finding that many auditory perceptual dimensions that are used to represent multidimensional data have been shown to interact perceptually [22]. Changes in loudness can influence perceived changes in other dimensions such as pitch or timbre [23]. This type of interaction can distort the underlying relationships between the data variables. Complicating matters even further are findings that show increases in acoustic dimensions such as pitch, loudness, and tempo are perceived as changing more than identical decreases in those dimensions [24-26]. Thus, a data variable mapped to one of these dimensions that exhibits an increase of ten units would be heard as changing more than if the same variable decreased by ten units.

The musical nature of sonification researchers. There are 1,103 conference papers in the *ICAD* proceedings from the years 1994-2018. The word *music* appears in 74% of these papers, and musical terminology is used widely [3]. Over 30% of the authors listed on the 1,103 conference papers have an institutional or departmental affiliation related to music (e.g., *School of Music*). In addition to those whose primary employment is in the field of music, a large percentage of sonification researchers in other fields also have some background in music. There is typically a higher proportion of musicians among those who do research in audition as can be evidenced by both the programmatic and impromptu "jam sessions" that occur at among attendees at professional conferences such as the *Meeting of the Acoustical Society of America*, *The Society for Music Perception & Cognition*, and *ICAD*. Among psychologists who study music cognition, over 97% report having a musical background [27].

The overrepresentation of musicians in the sonification community coupled with the dramatic differences between the brains and perceptual abilities of musicians and non-musicians has the potential to skew sonification design in a way that is not aligned with the listening practices and abilities of the general public [28]. Musicians employ analytical listening strategies that can be beyond the immediate grasp of non-musicians [29]. Importantly, the analytical listening advantage that musicians have is present even when listening to non-musical audio [30].

While some researchers have stressed the importance of taking individual differences like musical background into account from the start when designing a sonification [31], others have suggested that sonification designers "use their own introspection and intuition" in sonification design before moving to more formal usability testing [32]. Still, others have eschewed musically naïve listeners entirely and focused exclusively on those with domain expertise [33]. Thus, a major stumbling block to effective sonification design for the masses is a failure of designers to take the perspective of musically naïve and "non-attentive listeners" [34].

Interdisciplinarity. Sonification is an inherently interdisciplinary field. Economist George Steigler once said, *"The main insight learned from interdisciplinary studies is the return to specialization."* Challenges to interdisciplinary work include differences in the underlying assumptions of the various disciplines, differences vocabulary, methods, and in values among many others. Perhaps nowhere is this more apparent in sonification work than when it comes to the evaluation of sonification. Should the sonification be evaluated simply by the designer? By process of iterative participatory design? Or by tests of statistical significance with appropriate sample size? The answer generally depends on the discipline of the person answering the question. Is sonification art, design, science, or a mixture of all three?

4. THE BIFURCATION OF SONIFICATION

It may be that bifurcating sonification into well-defined paths of art and science would lead to greater success. The paths need not be mutually exclusive and would be most effective if pursued simultaneously. There are advantages to both.

Shift Toward Artistic Sonification. Given the challenges to sonifying data in a manner that stays empirically faithful to the underlying data, perhaps some researchers should abandon this pursuit altogether. Instead, "artistic sonification" would embrace the more aesthetic aspects of sonic representation, giving listeners a "sense" of the underlying data while perhaps not always perfectly preserving the underlying data relations. Barrass [35] has suggested this approach as prioritizing "usefulness" in design even if it means sacrificing a veridical understanding of the underlying data.

This technique might be considered analogous to a courtroom sketch artist who makes drawings of the key figures at a trial. The representation is certainly not a "precise" representation of the courtroom scene, yet it does convey information to the viewer in a way that is "useful." In fact, the creativity of the artist might even provide a *better* representation of the mood of the courtroom than a still photograph. This artistic approach would facilitate multiple interpretations of the same data [36]. The shift in focus might also enhance the role that sonification plays in generating enthusiasm for science both with the public. For example, Ballora [7] has suggested that despite concerted efforts to sonify data empirically, *"sonification's potential value, like much of the scientific visualisation content, probably lies less in hard facts and more in how it may serve as a stimulant for curiosity."*

Shift Toward Empirical Sonification. Others in the field might pivot more toward the empirical. If nothing else, the last twenty years of sonification research have clarified what does not work [4]. If sonification is to be considered a scientifically legitimate way of representing data, we should heed the lessons of the past. Specifically, the following points should be emphasized:

1. Design efforts should be focused in a perceptual space where audition performs well and individual differences are smallest.

2. To avoid perceptual interactions, parameter mapping that uses simple acoustic dimensions like pitch and loudness should be largely abandoned.
3. Empirical sonification researchers should evaluate their designs with a focus on the poorest rather than the best analytical listeners in their target user population.

Leveraging audition's temporal advantage would likely be a more fruitful approach than concentrating on other perceptual dimensions (e.g., pitch and loudness). Similarly, although the spatial resolution of the visual system is better than that of the auditory system, we can only see a limited field of vision while we can hear in 360 degrees. Concentrating on design efforts that exploit these kinds of advantages is likely to produce significant advances in the field. Abandoning the use of simple acoustic dimensions in parameter mapping would also be a step in the right direction. It has been known for decades that representing multidimensional data with multiple acoustic dimensions introduces distortions [23, 37, 38]. However, many current attempts to sonify data still take this approach [39-41]. As an alternative, ecological parameter mapping techniques might provide a more effective approach. The well-known work on ecological acoustics by Gaver [42] suggests that people listen to sounding objects and events rather than acoustic dimensions. As such, a better approach to parameter mapping might be assigning data variables to acoustically complex but ecologically simple sounds (e.g., footsteps) that indicate changes in sounding objects or events [43].

Finally, the importance of perspective taking by musically experienced sonification designers cannot be overstated. It is well known that musicians hear, think about, and speak about sound differently than non-musicians. Musicologist Sarah Cassie Provost gives her music students an assignment entitled *"Communicating with Non-Musicians"* [44]. Others provide "translations" for musicians who may find themselves working with non-musicians in a professional production environment [45]. A sonification designed by someone with a musical background could be largely lost on someone without one. An empirical approach to sonification would benefit from a design process that seeks input at the start from non-musicians and is evaluated empirically with a representative target population.

Avoid the "muddled middle." The line that separates art and science in sonification design is, in fact, not a line at all. Integrating art and science in sonification work has resulted in a continuum. Unfortunately, the closer a given sonification is to the midpoint, the more frequently it fails to live up to the goals of either art or science. Shifting sonification closer to the endpoints of this continuum would result in moving away from the muddled middle ground. Artistic sonification would have the goal of aesthetically enhancing user experience, capturing attention, and stimulating curiosity. It would be data-based without requiring an isomorphic tie between data and sound. Scientific sonification would have the goal of reliably representing the underlying data across listening conditions and listeners. It would not ignore aesthetics but would hold reliable representation in priority above aesthetics. There would certainly still be crosstalk. Art and science would continue to influence each other in design. However, a clear delineation of the goals, methods, and evaluation of the sonification would avoid design efforts that try to be both art and science and end up being neither.

5. CONCLUSIONS

Many of the challenges that faced early sonification researchers persist to the present day. Clearly outlining the goals of a given sonification, whether scientific or artistic, and holding fast to design principles and evaluations that best serve those goals may help us overcome some of these challenges. In a keynote address at ICAD in 2017, Carla Scaletti suggested that sonification may be near the tipping point of "scientific legitimacy" [34]. Sonification may also be at a tipping point of "artistic legitimacy." Decoupling these approaches may facilitate tipping points for sonification in both domains.

6. REFERENCES

[1] G. Kramer *et al.*, "Sonification report: Status of the field and research agenda.," Palo Alto, CA:, 1999.

[2] M. A. Nees, "Auditory Graphs Are Not the "Killer App" of Sonification, But They Work," *Ergonomics in Design*, vol. 26, no. 4, pp. 25-28, Oct 2018, doi: 10.1177/1064804618773563.

[3] A. Supper, "The Search for the "Killer Application": Drawing the Boundaries around the Sonification of Scientific Data," in *The Oxford Handbook of Sound Studies*, T. Pinch and K. Bijsterveld Eds. New York: Oxford University Press, 2012, pp. 249-270.

[4] J. H. Flowers, "Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions," presented at the International Conference on Auditory Display, Limerick, Ireland, 2005.

[5] B. N. Walker and M. A. Nees, "Theory of Sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff Eds. Berlin: Logos Publishing House, 2011.

[6] T. Hermann, A. Hunt, and J. G. Neuhoff, "Introduction," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff Eds. Berlin: Logos Publishing House, 2011, pp. 1-6.

[7] M. Ballora, "Sonification, Science and Popular Music: In search of the 'wow'," *Organised Sound*, vol. 19, no. 1, pp. 30-40, Apr 2014, doi: 10.1017/s1355771813000381.

[8] A. Supper, "Lobbying for the Ear: The Public Fascination with and Academic Legitimacy of the Sonification of Scientific Data," Ph. D., Maastricht University, 2012.

[9] A. Supper, "Sublime frequencies: The construction of sublime listening experiences in the sonification of scientific data," *Social Studies of Science*, vol. 44, no. 1, pp. 34-58, Feb 2014, doi: 10.1177/0306312713496875.

[10] G. Dubus and R. Bresin, "A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities,", *Plos One*, Article vol. 8, no. 12, p. 28, Dec 2013, Art no. e82491, doi: 10.1371/journal.pone.0082491.

[11] H. Ono, "Difference threshold for stimulus length under simultaneous and nonsimultaneous viewing conditions," *Perception & Psychophysics*, vol. 2, no. 5, pp. 201-207, 1967, doi: 10.3758/bf03213050.

[12] R. Teghtsoonian, "Exponents in stevens law and constant in ekmans law," *Psychological Review*, Article vol. 78, no. 1, pp. 71-+, 1971, doi: 10.1037/h0030300.

[13] D. R. Perrott and K. Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1728-1731, Apr 1990, doi: 10.1121/1.399421.

[14] A. Shaqiri *et al.*, "Sex-related differences in vision are heterogeneous," *Scientific Reports*, vol. 8, May 2018, Art no. 7521, doi: 10.1038/s41598-018-25298-8.

[15] B. N. Walker, "Magnitude estimation of conceptual data dimensions for use in sonification," *Journal of Experimental Psychology-Applied*, Article; Proceedings

Paper vol. 8, no. 4, pp. 211-221, Dec 2002, doi: 10.1037/1076-898x/8.4.211.

[16] L. Axon, M. Goldsmith, and S. Creese, "Sonification Mappings: Estimating Effectiveness, Polarities, and Scaling in an Online Experiment," *Journal of the Audio Engineering Society*, Article vol. 66, no. 12, pp. 1016-1032, Dec 2018, doi: 10.17743/jaes.2018.00057.

[17] J. G. Neuhoff, R. Knight, and J. Wayand, "Pitch change, sonification, and musical expertise: Which way is up?," presented at the International Conference on Auditory Display, Kyoto, Japan, 2002.

[18] L. Kishon-Rabin, O. Amir, Y. Vexler, and Y. Zaltz, "Pitch discrimination: Are professional musicians better than non-musicians?," *Journal of Basic and Clinical Physiology and Pharmacology*, Article vol. 12, no. 2, pp. 125-143, 2001.

[19] C. Marie, T. Kujala, and M. Besson, "Musical and linguistic expertise influence pre-attentive and attentive processing of non-speech sounds," *Cortex*, Article vol. 48, no. 4, pp. 447-457, Apr 2012, doi: 10.1016/j.cortex.2010.11.006.

[20] P. Janata and K. Paroo, "Acuity of auditory images in pitch and time," *Perception & Psychophysics*, vol. 68, no. 5, pp. 829-844, Jul 2006, doi: 10.3758/bf03193705.

[21] B. R. Zendel and C. Alain, "Concurrent Sound Segregation Is Enhanced in Musicians," *Journal of Cognitive Neuroscience*, Article vol. 21, no. 8, pp. 1488-1498, Aug 2009, doi: 10.1162/jocn.2009.21140.

[22] J. G. Neuhoff, "Interacting Perceptual Dimensions," in *Ecological Psychoacoustics*, J. G. Neuhoff Ed. New York: Academic Press, 2004.

[23] R. D. Melara and L. E. Marks, "Interaction among auditory dimensions - timbre, pitch, and loudness," *Perception & Psychophysics*, Article vol. 48, no. 2, pp. 169-178, Aug 1990, doi: 10.3758/bf03207084.

[24] J. G. Neuhoff, "Perceptual bias for rising tones," *Nature*, Letter vol. 395, no. 6698, pp. 123-124, SEP 10 1998 1998, doi: 10.1038/25862.

[25] C. C. Wang, "Effects of some aspects of rhythm on tempo perception," *Journal of Research in Music Education*, vol. 32, no. 3, pp. 169-176, 1984, doi: 10.2307/3344836.

[26] P. G. Vos, M. vanAssen, and M. Franek, "Perceived tempo change is dependent on base tempo and direction of change: Evidence for a generalized version of Schulze's (1978) internal beat model," *Psychological Research-Psychologische Forschung*, Article vol. 59, no. 4, pp. 240-247, Feb 1997, doi: 10.1007/bf00439301.

[27] C. Wollner, J. Ginsborg, and A. Williamon, "Music researchers' musical engagement," *Psychology of Music*, Article vol. 39, no. 3, pp. 364-382, Jul 2011, doi: 10.1177/0305735610381592.

[28] S. E. Pitts, "Amateurs as Audiences: Reciprocal Relationships between Playing and Listening to Music," in *Audience Experience: a Critical Analysis of Audiences in the Performing Arts*, J. Radbourne, H. Glow, and K. Johanson Eds. Oxford: Intellect Ltd, 2013, pp. 83-+.

[29] B. P. Gold, M. J. Frank, B. Bogert, and E. Brattico, "Pleasurable music affects reinforcement learning according to the listener," *Frontiers in Psychology*, Article vol. 4, p. 19, Aug 2013, Art no. 541, doi: 10.3389/fpsyg.2013.00541.

[30] A. Harris and E. Flynn, "Medical education of attention: A qualitative study of learning to listen to sound," *Medical Teacher*, vol. 39, no. 1, pp. 79-84, Jan 2017, doi: 10.1080/0142159x.2016.1231916.

[31] L. M. Mauney and B. N. Walker, "Individual Differences and the Field of Auditory Display: Past Research, A Present Study, and an Agenda for the Future," presented at the International Conference on Auditory Display, Montréal, 2007.

[32] T. L. Bonebright and J. H. Flowers, "Evaluation of auditory display," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff Eds., 2011, pp. 111-144.

[33] S. Landry and M. Jeon, "Participatory Design Research Methodologies: A Case Study In Dancer Sonification," presented at the The 23rd International Conference on Auditory Display, Pennsylvania State University, 2017.

[34] C. Scaletti, "Why sonification is a joke. Keynote address delivered at the 23rd International Conference on Auditory Display,," ed. University Park, PA., 2017, p. https://www.youtube.com/watch?v=T0qdKXwRsyM.

[35] S. Barrass, "The aesthetic turn in sonification towards a social and cultural medium," *AI & Society*, vol. 27, pp. 177-181, 2012, doi: 10.1007/s00146-011-0335-5.

[36] S. Barrass, M. Whitelaw, and F. Bailes, "Listening to the mind listening: An analysis of sonification reviews, designs and correspondences," *Leonardo Music Journal*, Article vol. 16, pp. 13-19, 2006, doi: 10.1162/lmj.2006.16.13.

[37] J. Neuhoff and M. McBeath, "The Doppler illusion: The influence of dynamic intensity change on perceived pitch," *Journal of Experimental Psychology-Human Perception and Performance*, vol. 22, no. 4, pp. 970-985, AUG 1996 1996, doi: 10.1037/0096-1523.22.4.970.

[38] M. G. Boltz, "Illusory tempo changes due to musical characteristics," *Music Perception*, vol. 28, no. 4, pp. 367-386, Apr 2011, doi: 10.1525/mp.2011.28.4.367.

[39] D. E. MacDonald, T. Natarajan, and R. C. Windeyer, "Data-driven sonification of cfd aneurysm models," presented at the International Conference on Auditory Display, Michigan Technological University, 2018.

[40] M. Ballora, C. Roman, R. Pockalny, and K. Wishner, "Sonification and science pedagogy: Preliminary experiences and assessments of earth science data presented in an undergraduate general education course," presented at the International Conference on Auditory Display, Michigan Technological University, 2018.

[41] R. Wheeler and D. Worrall, "Representing twitter users' engagement by sonification," presented at the International Conference on Auditory Display, Michigan Technological University, 2018.

[42] W. W. Gaver, "What in the world do we hear? An ecological approach to auditory event perception," *Ecological Psychology, 5(1), 1-29.* 1993, pp. 1-29.

[43] J. G. Neuhoff and L. M. Heller, "One small step: Sound sources and events as the basis for auditory graphs," in *11th International Conference on Auditory Display*, Limerick, Ireland, 2005.

[44] S. C. Provost. "Communicating with Non-Musicians." https://sarahcprovost.domains.unf.edu/communicating-with-non-musicians/ (accessed April 2, 2019, 2019).

[45] M. Gallant. "Collaborating and communicating with non-musicians." https://blog.discmakers.com/2018/07/collaborating-and-communicating-with-non-musicians/ (accessed April 2, 2019, 2019).

DESIGNING AUDITORY COLOR SPACE FOR COLOR SONIFICATION SYSTEMS

Dominik Osinski

Norwegian University of Science and Technology
Department of Electronic Systems,
Trondheim, 7491, Norway
dominik.osinski@ntnu.no

Patrycja Bizon

Jagiellonian University
Institute of Psychology
Ingardena 6,
Kraków, 30-060, Poland
patrycja.bizon@gmail.com

Helene Midtfjord
Norwegian University of Science and Technology
Department of Computer Science
Gjøvik, 2815, Norway
helene.midtfjord@ntnu.no

Michał Wierzchoń
Jagiellonian University
Institute of Psychology
Ingardena 6,
Kraków, 30-060, Poland
michal.wierzchon@uj.edu.pl

Dag Roar Hjelme
Norwegian University of Science and Technology
Department of Electronic Systems
Trondheim, 7491, Norway
dag.hjelme@ntnu.no

1. ABSTRACT

Designing of color sonification systems provides a possibility of contribution to various fields ranging from rehabilitation of visually impaired through color perception, multisensory art experience to consciousness studies. The design process itself requires understanding and integrating knowledge from many difficult and inherently different branches of science and the resulting sonification method will be highly dependent on the purpose of the system. We present work in progress on designing and experimental verification of color sonification method that will be implemented in Colorophone – a wearable assistive device for the visually impaired, which enables perception of the information about color through sound. Although our system shows promising results in color and object recognition, we would like to enhance the existing color sonification method by designing a framework for experimental verification of our color sonification algorithm. The goal of this paper is therefore to briefly describe our way of thinking in order to provide the basis for the discussion.

2. INTRODUCTION

Our interest in designing intuitive color sonification algorithms is directly related to development of Colorophone – a visual-to-auditory sensory substitution device (SSD) [1]. The main goal of the Colorophone project is to develop an affordable, wearable SSD which will enhance cognitive capabilities of visually impaired by providing auditory information about color and distance. Color sonification systems proved to enhance object recognition and orientation of visually impaired as documented in [2],[3],[4]. Although speaking color monitors are commercially available, coding colors as sound provides much faster and language independent way of delivering the information to users. It also enables active user engagement in the process of scanning of the environment, and development of new sensorimotor contingencies by integration of movement and sound-coded visual information into one multisensory experience. If we looked closer on necessary elements for building a color sonification system, we would conclude that current developments in consumer electronics such as mobile phones, camera technology, and bone conductive headphones are at the level which enables designing SSDs that provide real-time color to sound conversion. The missing element is an intuitive color sonification method.

3. COLOR SONIFICATION

Since the goal of the color sonification is to convert information from visual to auditory channel, which are inherently different, we believe that the necessary preliminary step is to specify the function of such a conversion system. In SSDs used for visual rehabilitation of the visually impaired the function of color sonification algorithms is to provide an intuitive information about color by sound. Such systems should therefore be focused on the usability and at the same time provide continuity between different sensory modalities while avoiding sensory overload and limiting interference with other perceptual functions [5].

3.1. What can we learn from existing systems?

The existing color sonification methods used in SSDs can be divided into two categories: the first category contains systems which use direct association between color category and

https://doi.org/10.21785/icad2019.076

presented sound [2],[6],[7], for example red color is coded by the sound of choir. In other words, every color is coded by an associated sound, which imposes strict color categorization and step transition between sounds corresponding to colors. The second category of systems uses basic color components associated with sound components [3],[4],[8]. In such systems, auditory color representation is constructed from many sound components, which are merged into one auditory stimulus. The latter category gives significantly better results in experiments related to color recognition, topping out at 98% of correct answers on the task of identifying 14 colors. Therefore we decided to use the approach described in the second category while designing new color sonification methods. Overview of color sonification methods together with corresponding experimental results are summarized in [3],[4].

3.2. Mapping sensory components

In order to associate one sensory modality (color) with another (sound) we have to take into consideration many factors like what the relationship between color and sound components should be, number of used color components, differences in perceptual characteristic of each sensory channel, cross-modal correspondences and finally, which sounds should we use.

3.2.1. Number of color components

While thinking about the number of basic color components we should remember that if this number will be too large it could be difficult for the naïve user of the system to remember and recognize all the sounds associated with color components. If the number of basic color components will be too low the user will not have necessary variety in the auditory signal to be able to recognize color change. Our preliminary experiments indicate that although 4 color components (red, green, blue and white) allowed very good auditory color recognition for 14 tested colors (black, white, red, pale red, green, pale green, blue, pale blue, yellow, pale yellow, violet, pale violet, cyan, pale cyan), the recognition of colors in vicinity of yellow (orange, olive green) remains challenging. Since the yellow component is central in opponent process theory [9] and yellow-blue axis is present in many advanced color spaces we consider the yellow channel to be necessary in our color sonification design. Black remains a special color component, because the information about this color, which effectively means lack of any light is conveyed by silence – lack of any sound. Definition of five color components plus black strongly reminds of color component definition from Natural Color System (NCS) [10].

3.2.2. Psychophysics

Since senses of sight and hearing show different psychophysical characteristics, we implemented inverted Stevens's power law [11] for auditory channel in order to compensate for non-linear response of the human auditory system. The information about the color intensity is pre-processed by the inverted Stevens's power law function which then is annulated by the influence of the human auditory system.

3.2.3. Cross-modal correspondences

Cross-modal correspondences are natural associations between different sensory modalities. Although finding an universal mapping remains ambiguous, we can utilize existing research results as a guideline in designing color sonification method. The first intuitive mapping between a color component and a sound component would be the mapping of the intensity of the color stimuli to the intensity of the sound stimuli. The more intensive color will be associated with the sound of higher volume. We chose to associate color components to corresponding sound frequencies on basis of pitch-croma relationship described in [8].

3.2.4. Sounds associated with color components

While choosing sound components corresponding to color components, we used the following guidelines: the sounds should be calibrated in amplitude corresponding to maximal color intensity in order to provide equal loudness for every sound component, perceptually equally spaced in the frequency domain and be associated with colors on basis of chosen cross-modal correspondences i.e. blue – low pitch, green – middle pitch, yellow – high pitch, red – high pitch, white – white noise. Since we know which sound pairs will be presented at the same time we can decide if we will present consonant or dissonant pairs of sounds at the same time. For the first version of the system all sounds besides white noise are pure sine tones.

3.3. Color spaces

There are numerous color spaces which define the conventions of coding information about color by numerical values. CIELAB and CIELUV are often used, uniform color spaces based on opponent process theory. However CIELAB does not have focal red, blue or green any close to the corresponding color axis, and CIELUV has the biggest deviation from axes for green, yellow and red color components [9]. Focal colors are the best example of a given color category [3]. While designing auditory color space on basis of previous considerations, we need to use a color space based on opponent process theory, where color axes are as close as possible to focal red, yellow, green and blue. We propose to call the color space equipped with the features described above as RYGBW. Possible candidates for being a prototype for developing RYGBW, which meet our requirements, are YCiCii [9] and oRGB [12] color spaces.

3.4. Auditory color space

Since the suggested RYGBW color space will be the base for experimental evaluation of color sonification method it does not have to be calibrated in terms of perceptual color distance. Non-linearities in color perception will be mapped by the experiments and could be minimized by an iterative calibration process. We have to remember that the iterative mapping of sound components to color components will compensate for non-linearities in both visual and auditory channels, which is positive for enhancing color sonification algorithms for SSDs. However this compensation process makes research of color perception limited to relative comparisons between participants or participant groups.

4. EXPERIMENTAL VERIFICATION

The main technical part of the system is a software framework developed for experimental verification of color sonification method. It allows automated presentation of test data, logging and postprocessing of the results. A short video which presents our software framework can be found at https://s.ntnu.no/sonification

Figure 1: Example of a stripe presenting transition path from red to blue.

4.1. Experimental procedure

In order to assure consistency in representation of color stimuli we use an EIZO monitor equipped in build-in color calibration system. Stimuli are presented on a standardized mid-grey background in a dark room. The mid-gray color fills up the whole background on the test monitor. The experiment consists of showing the participant a colorful stripe (Fig. 1). The stripe shows color for one of the basic color transition paths (for example from red to blue through violet). At the same time the system plays a multicomponent sound, where the sound components correspond to red and blue color components. The task of the participant is then to choose the point on the colorful stripe which participant associates with the presented sound. When the participant clicks on the stripe the chosen color is presented in form of rectangle in the middle of the screen in order to eliminate color illusions (Fig. 2.). The participant has then a possibility to correct the choice by clicking on the rectangle or to go to the next trial by clicking below the rectangle. The next trial contains the same colorful stripe but sound components corresponding to red and blue have different amplitude than previously. After performing the whole experiment for one transition path the participant repeats trials with the rest of the transition paths.

Figure 2: Chosen color presented with grey background to eliminate color illusions.

4.2. Result analysis

The proposed method for evaluation of color sonification algorithms enables comparison of numerical RGB values used for sound generation with RGB values chosen by participants. This allows quantitative evaluation of errors in colors picking on basis of auditory signals, which can be used in iterative method for optimization of color sonification algorithms. However, if we assure proper calibration of amplitudes of auditory signals, even the results from the first experimental round can be potentially used to identify differences in color perception between participants.

5. DISCUSSION AND FUTURE WORK

We believe that the system for evaluation of color sonification method described in this paper will allow enhancing our existing color sonification method. The enhanced method will then be implemented in the Colorophone SSD and utilized in developing and evaluating of a wearable electronic travel aid for visually impaired as well as in consciousness research in the project "Cognitive and Neural Plasticity and the Subjective Experience. Interdisciplinary Analysis of Sensory Substitution", where the sonification method is used in prolonged training so to induce subjective color perception through audition. Although the beta version of our system is functional, there still are some design challenges that need to be addressed. Which RYGBW color space should we use? Which sound frequencies should be associated with color components? Should the sound pairs be consonant or dissonant? How can we ensure repeatability of stimuli for every participant? The proposed system has been designed to improve our sonification method, however the usage of this system is not limited to design of SSDs. Color research, in spite of growing evidence from neuroscientific studies still remains a place for intensive universalist-relativist debate. Using independent sound variable for evaluation of color perception independently from language constraint seems to be a very interesting path to investigate individual differences in color perception in humans. Development of a web-based system similar to the one described here will allow easy verification of cross-cultural differences in color perception.

6. ACKNOWLEDGMENT

This work was supported by the National Science Centre, Poland, grant OPUS (2016/23/B/HS6/00275) given to Michał Wierzchoń.

7. REFERENCES

[1] http://www.colorophone.com/

[2] G. Bologna, B. Deville, and T. Pun, "Sonification of Color and Depth in a Mobility Aid for Blind People," *in Proceedings of the International Conference on Auditory Displays (ICAD)*,Washington, D.C, USA, 2010, pp. 9-13.

[3] G. Hamilton-Fletcher and J. Ward, "Representing colour through hearing and touch in sensory substitution devices," *Multisens Research,* vol. 26, pp. 503- 532, January 2013.

[4] D. Osiński and D. R. Hjelme, "A Sensory Substitution Device Inspired by the Human Visual System," *in Proceedings of the 11th International Conference on Human System Interaction (HSI),* Gdansk, Poland, 2018, pp. 186-192.

[5] A. Kristjansson, A. Moldoveanu, O. I. Johannesson, O. Balan, S. Spagnol, V. V. Valgeirsdottir, and R. Unnthorsson, "Designing sensory-substitution devices: Principles, pitfalls and potential," *Restorative Neurology and Neuroscience*, 34(5), pp.769-787, September 2016.

[6] S. Cavaco, J. T. Henriques, M. Mengucci, N. Correia, and F. Medeiros, "Color Sonification for the Visually Impaired," *Procedia Technology*, vol. 9, pp. 1048–1057, 2013.

[7] S. Abboud, S. Hanassy, S. Levy-Tzedek, S. Maidenbaum, and A. Amedi, "EyeMusic: Introducing a 'visual' colorful experience for the blind using auditory sensory

substitution," *Restor. Neurol. Neurosci*., vol. 32, no. 2, pp. 247–257, January 2014.

[8] G. Hamilton-Fletcher, J. Ward, and T. D. Wright, "Cross-Modal Correspondences Enhance Performance on a Colour-to-Sound Sensory Substitution Device," *Multisens. Research*, pp. 1–27, February 2016.

[9] N. Moroney, "The opposite of green is purple?" *in Proceedings of IS&T/SPIE Electronic Imaging,* San Jose, USA, 2009, Vol. 7241 72410N-(1-7).

[10] https://ncscolour.com/

[11] S. S. Stevens, "On the psychophysical law.," *Psychological Review*, vol. 64, no. 3, p. 153, June 1957.

[12] M. Bratkova, S. Boulos, and P. Shirley, "orgb: A practical opponent color space for computer graphics." *Computer Graphics and Applications, IEEE,* 29(1):42–55, Jan.-Feb. 2009.

DESIGN AND EVALUATION OF A NEW AUDITORY DISPLAY FOR THE PULSE OXIMETER

Estrella Paterson
Penelope Sanderson
Neil Paterson

The University of Queensland
Sir Fred Schonell Drive
St Lucia, Queensland 4051, Australia
estrella.paterson@uqconnect.edu.au

Robert Loeb

Department of Anesthesiology
University of Florida College of Medicine
Gainsville, FL 32610, USA
RLoeb@anest.ufl.edu

ABSTRACT

During surgery the pulse oximeter device provides information about a patient's oxygen saturation (SpO_2) and heart rate via visual and auditory displays. An audible tone is emitted after every detected pulse (indicating heart rate), and the pitch of the tone is mapped to SpO_2. However, clinicians cannot reliably judge SpO_2 using only the current auditory display. In a series of three studies, we compared auditory displays based on current pulse oximeters with displays designed to provide more information about SpO_2 levels using additional acoustic properties. Results from the first two laboratory studies show that the new auditory displays support better identification of specified ranges of SpO_2, and better detection of when saturation transitions a critically relevant threshold. The analysis of a third study in a high-fidelity simulator is currently under way. An auditory display that provides more information about SpO_2 levels and when SpO_2 changes from one range to another may be useful for clinicians when they are engaged in other visually demanding tasks but have to detect and treat patient deterioration, often in time-pressured and stressful situations.

1. INTRODUCTION

Over the past three decades, the pulse oximeter (PO) has become standard equipment in a number of hospital settings including the operating room (OR), recovery room, intensive care unit and patient transport[1]. It provides a visual display (numerical and waveform) and an auditory display (variable pitch plus alarms) of the patient's oxygen saturation (SpO_2), heart rate, and heart rhythm.

The auditory display is especially important when clinicians are engaged in other visually demanding tasks, when the visual display is obscured, or when visual overload occurs [2-4]. During an operation, anaesthetists are usually not looking at the visual display; they look at the display only around 5–30% of intraoperative time [5, 6]. Thus, clinicians depend on the auditory display to provide patient information. However, clinicians cannot reliably identify SpO_2 levels accurately using the current auditory display alone [7-9]. The current auditory display is based on tones of variable pitch, supplemented with alarms set at a clinically relevant threshold. As SpO_2 decreases from a maximum of 100%, the pitch of the tones decreases. Although people find it easy to recognise pitch changes, very few people have absolute pitch [10] making it difficult to identify SpO_2 values using pitch alone.

The problem of identifying SpO_2 from the pulse oximeter tone is exacerbated by a number of factors. First, clinicians have many tasks to perform while monitoring patients [11] and therefore have to divide attention between these tasks and patient monitoring. Second, the OR can be a noisy environment. Research shows that as noise levels increase, anaesthetists' ability to distinguish between SpO_2 levels diminishes [9]. Third, during surgical procedures, anaesthetists are frequently interrupted and distracted [12].

In a series of three studies, devised as incremental design experiments, we evaluated a new auditory display for the pulse oximeter. Our aim was to test whether the new auditory display better supports judgements about SpO_2 range, and when the SpO_2 changes from one range to another, than does a standard display. Visual display of SpO_2 was not provided in any of the three experiments.

In the first laboratory study we compared the ability of non-clinician participants to identify SpO_2 levels using five different auditory pulse oximetry displays, including a standard display similar to those in current use, while they performed a visual distractor task[13]. In the second laboratory study we compared the ability of clinician and non-clinician participants to identify SpO_2 levels using the standard display and the best enhanced display from Study 1, while they also performed visual and auditory distractor tasks. Finally, in our current study, we are comparing anaesthetists' ability to distinguish SpO_2 levels in a high-fidelity simulator using the displays from Study 2.

2. STUDY 1

In former studies, we found that listeners can distinguish SpO_2 levels more accurately when the pulse oximeter's variable pitch tone is enhanced with tremolo and acoustic brightness than when variable pitch alone is used [14, 15]. This may be because listeners can more easily use auditory displays with multiple heterogeneous features indicating state changes than displays with only one such feature [16]. A limitation was that participants' only task was to judge SpO_2 levels, whereas in the OR anaesthetists have many tasks to perform while monitoring patients' states. Furthermore, the experiments were conducted in a quiet room. Noise levels in the OR average 51–75 dB [17] and can reach levels of 120 dB[18] .

In the current Study 1, using a between-subjects design, we measured 100 non-clinician participants' accuracy and latency at detecting transitions into and out of an SpO_2 target range, identifying SpO_2 range (target, low, critical), and identifying the absolute SpO_2 value, using five different auditory displays[13]. We addressed limitations of the above studies by including a secondary distractor task (arithmetic verification) plus background noise.

https://doi.org/10.21785/icad2019.070

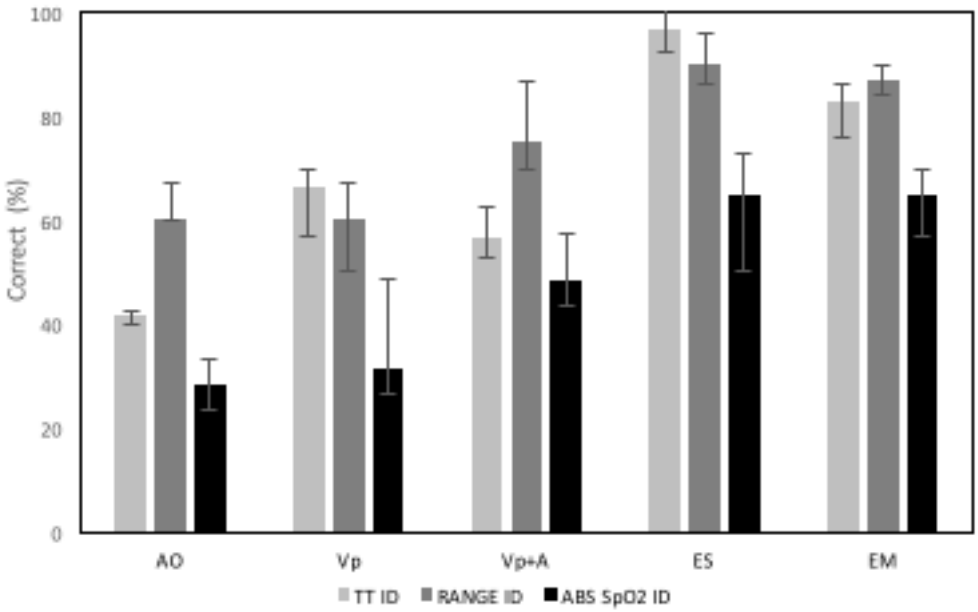


Figure 1. Participants' accuracies of Target Transition identification, Range identification, and Absolute SpO_2 Value identification, for Alarm only (AO), Varying pitch(Vp), Varying pitch plus alam (Vp+A), Enhanced single (ES) and Enhanced Multiple (EM) conditions. (Mean ± CI)

Background noise contained dialogue, OR noises and pop music with vocals.

We tested SpO_2 values from 100%–80% and divided them into three ranges: target (100%–97% SpO_2), low (96%–90% SpO_2) and critical (89%–80% SpO2). The five auditory displays are described in detail below and in this sound file: https://www.dropbox.com/sh/r8xfqiyyeto2r1w/AACgLdH-feMAb6JvJu-9hU69a?dl=0

The control condition, ***Variable pitch plus alarm (Vp+A)***, was based on the auditory display of current pulse oximeters and comprised variable pitch pulse tones with an audible alarm set at 89% SpO_2 [19]. Tones were sine wave functions ranging logarithmically from 150 Hz at 80% SpO_2 to 950 Hz at 100% SpO_2; Each tone lasted for 150 ms with a 10 ms fade-in and 10 ms fade-out to eliminate acoustic artefact.

In the ***Alarm only (AO)*** condition there were no pulse tones; only an alarm [IEC-Medium-General alarm (IEC-60601-1-8)] that sounded when SpO_2 entered the critical range from the low range and every 15 s thereafter that SpO_2 remained in the critical range. This condition represents typical use of the auditory display used in the intensive care unit, with the variable pitch deactivated to reduce noise levels.

The ***Variable pitch (Vp)*** condition was the same as the control condition but without an alarm. This condition corresponds to use of pulse oximetry displays when the alarm is silenced or alarm limits are set very wide.

The first experimental condition for single patient monitoring, ***Enhanced single (ES)***, comprised the same variable pitch mapping as the control condition but with tremolo added to tones in the low and critical ranges (96%–80% SpO_2) and brightness added in the critical range (89%–80% SpO_2). Tremolo was produced by modulating the peak amplitude of the tone: four cycles of tremolo with 90% wet. Brightness was produced by adding odd harmonics of the tone's fundamental frequency (third, fifth and seventh harmonic) to produce a sharper sound.

The second experimental condition, ***Enhanced multiple (EM)*** was the same as Enhanced single except that variable pitch tones were excluded when SpO_2 was in the target range. Instead pulse tones were replaced by a chirp (an "all well" sound) that sounded every 5 s that SpO_2 remained in the target range. The chirp had a duration of 100 ms, started at 1000 Hz that decreased linearly to 500 Hz at the 50 ms midpoint and increased to 1000 Hz at the end of the tone. Volume increased from 0 to 0.3 (on a scale of 0–1) at the midpoint and decreased to 0 at 100 ms. This display represents a prototype that we have developed for monitoring multiple patients. When all patients' SpO_2 remains in target range, only a series of chirps is heard. If SpO_2 for one patient moves from the target range, the "all well" sound changes to the enhanced variable pitch tone.

Participants were trained to identify SpO_2 range and absolute SpO_2, values and to detect when SpO_2 moved into or out of the target range (target transitions). They then completed two blocks of fifteen 60-second trials each.

Results are shown in Figure 1. Participants using either of the two experimental auditory displays enhanced with additional acoustic properties (ES and EM) were more accurate and faster at detecting target transitions, and more accurate at identifying SpO_2 range and absolute SpO_2 values, than participants using the Variable pitch plus alarm condition (Vp+A). Participants in the Alarm only condition were less accurate and slower at detecting target transitions, and less accurate at identifying SpO_2 ranges and absolute SpO_2 values that those in the Vp+A condition. There was no difference for participants in the Variable pitch (Vp) and Vp+A condition for target transition detection accuracy or latency but participants in the Vp+A condition were more accurate than those in the Vp condition for SpO_2 range and absolute SpO_2 identification accuracy.

This study provides evidence that auditory displays comprising variable pitch with additional acoustic properties of tremolo and brightness are more effective for identifying

SpO_2 levels than an auditory display similar to that of current pulse oximeters.

3. STUDY 2

Study 1 established the superiority of displays enhanced with additional acoustic properties over a standard display for SpO_2 parameter identification. Participants performed only one distractor task that was presented visually, and participants were from a non-clinical population. However, many anaesthetic tasks involve verbal communication, some essential for effective team performance and some irrelevant to case management. Verbal processing may interfere with perception of the pulse oximeter's auditory signal. Furthermore, clinicians' greater familiarity with standard pulse oximetry auditory displays might mean they perceive the signal differently from non-clinicians. Thus, in Study 2, we added a distractor task in the same perceptual modality as the monitoring task, and tested non-clinician and clinician participants.

In a laboratory study using a counterbalanced, within and between-subjects, crossover design, non-clinician participants (n=28) and specialist/trainee anaesthetists (n=25) from a tertiary hospital identified SpO_2 levels using the Variable pitch plus alarm (standard) and the Enhanced single (enhanced) displays from Study 1. Participants performed two distractor tasks simultaneously: arithmetic verification task from Study 1 and a new keyword detection task. Each participant performed the experiment over two blocks of 15 trials each: one using the standard display and the other using the enhanced display. Each trial lasted 60 s with heart rate set at 72 bpm. Participants received training before each block. Participants identified SpO_2 target transitions during a trial, and SpO_2 range and absolute SpO_2 value at the end of each trial. Ranges were the same as in Study 1: target, low and critical.

For the keyword detection task, we designed 30 linguistic scenarios, one per trial. In each trial there were seven spoken phrases comprising 0–4 keyword phrases. Participants identified keywords: BLOOD, PATIENT or TABLE. Background noise contained OR noises, and music with vocals played throughout the experiment.

Participants were more accurate and faster at detecting SpO_2 target transitions with the enhanced display (87%, 2.4 s) than with the standard display (57%, 8.7 s), $p<.001$ for each measure. Participants were more accurate at identifying SpO_2 range and absolute SpO_2 value with the enhanced display (86%, 66%) than with the standard display (76%, 46%), $p<.001$ for each measure. Participants reported that they found the monitoring task easier and were more confident of their judgements with the enhanced display than with the standard display. We found no differences between clinicians and non-clinicians for performance accuracies or speeds, or for subjective judgements.

This study provides additional evidence that an auditory display enhanced with tremolo and brightness is more effective for identifying SpO_2 levels than a standard display using only pitch and alarms, even when participants are engaged in an auditory distractor task as well as a visual computational task. There was no difference in performance between clinicians and non-clinicians, which may not be surprising given that the experiment tested only perception and classification performance [20].

4. STUDY 3

Anaesthetists have many tasks to perform while they are continuously monitoring patients' states, and they are subject to numerous distractions and interruptions [11, 21]. High-fidelity simulators are powerful environments for investigating equipment usability in safety critical systems. They let investigators test devices in more challenging and authentic clinical settings, such as the OR. [22]. We designed a study to test whether the enhanced display would help anaesthetists monitor SpO_2 levels more accurately compared with the standard display. We used the simulator suite at a large paediatric hospital, and set it up as an OR. Participants were consultant anaesthetists (N=20) who identified SpO_2 levels using standard and enhanced displays from Study 2. In addition, participants identified changes in heart rate, blood pressure and CO_2. Each participant performed two different experimental scenarios, one for each display and each lasting 20 minutes. Scenarios were counterbalanced across both displays, were deterministic, and were controlled from the simulator control room.

Participants were trained to use the auditory display before each scenario. They performed a cognitively-demanding distractor task during each scenario: categorisation of patient details. Participants were also interrupted during scenarios, both directly and via telephone. All scenarios were video recorded. The video recordings will be coded for verbal responses relating to detection of SpO_2 range transition (target to low and low to critical in both directions) and identification of SpO_2 range. We are currently still analyzing the results of Study 3, but early results are promising.

5. GENERAL DISCUSSON AND CONCLUSION

The purpose of this program of research was to evaluate a new auditory display for the pulse oximeter. In a series of three studies we tested listeners' ability to identify SpO_2 levels using different auditory displays. In Study 1 we established that non-clinician participants detected SpO_2 target transitions and identified SpO_2 ranges and absolute SpO_2 values more accurately using an auditory display enhanced with tremolo and brightness compared with a pitch plus alarm display. Participants performed these tasks while doing a visual distractor task and in the presence of simulated background OR noises. In Study 2 we found superiority of the enhanced auditory display held, even when participants performed the visual task, plus a keyword detection task presented in the same modality as the monitoring task. There was no difference between performance of clinicians and non-clinicians, indicating that the new display has potential for use by novices. In Study 3 we have tested whether the effect holds in the more realistic environment of a simulator.

The experimental display enhanced with tremolo and brightness for non-normal ranges provides more information about SpO_2 levels than does the standard display of variable pitch plus alarm. In the first two experiments, when SpO_2 transitioned the target-low threshold, participants were able to detect transitions using the enhanced display far more accurately and faster than when using the standard display. Such a display may enable clinicians to monitor patients pre-attentively and continuously, allowing attention to be directed to other visually demanding tasks. [23] The additional sound properties may attract auditory attention to pre-set thresholds, thus indicating a change in saturation levels so remedial action

can be taken before a critical threshold is breached and an alarm sounds. This may help reduce the number of audible alarms and decrease annoyance from noise. The display may also help clinicians monitor whether treatment has been effective and detect exactly when SpO_2 levels return to normal once more. These results may have implications for clinical practice. If clinicians can detect changes in SpO_2 more accurately and faster they may be able to make decisions about treatment more effectively.

Our research shows that a PO auditory display enhanced with features such as tremolo and brightness to distinguish clinically important SpO_2 ranges allows for more accurate judgment of SpO_2 levels compared with displays similar to those of current pulse oximeters. If results from Study 3 show that SpO_2 levels can still be distinguished much more effectively with the enhanced display than with the standard display in an environment similar to the OR, further clinical trials could be conducted. Importantly, commercial manufacturers and users would need to be consulted in evaluation of a new PO auditory display for it to be taken up successfully.

6. ACKNOWLEDGMENTS

We thank Dr Birgit Brecknell and Ismail Mohammed for software development, Dr Peter Moran, Princess Alexandra Hospital, Brisbane, and actors Isaac Salisbury, Jelena Zestic, Garry Mann, Felicity Burgmann, Tom Davidson, T-lok Tang and Lachlan Peterson (The University of Queensland, Brisbane).

7. REFERENCES

[1] A. Shah, and K. H. Shelley, “Is pulse oximetry an essential tool or just another distraction? The role of the pulse oximeter in modern anesthesia care,” Journal of Clinical Monitoring and Computing, vol. 27, no. 3, pp. 235-242, 2013.

[2] J. M. Ansermino, "Intelligent patient monitoring and clinical decision making," Monitoring Technologies in Acute Care Environments, M. C. J. M Ehrenfeld, ed., pp. 401-407, New York: Springer, 2014.

[3] R. M. Craven, and A. K. McIndoe, “Continuous auditory monitoring--how much information do we register?,” British journal of anaesthesia, vol. 83, no. 5, pp. 747, 1999.

[4] P. Sanderson, D. Liu, and S. A. Jenkins, “Auditory displays in anesthesiology,” Current Opinion in Anesthesiology, vol. 22, no. 6, pp. 788-795, Dec, 2009.

[5] S. Ford, E. Birmingham, A. King, J. Lim, and J. M. Ansermino, “At-a-Glance Monitoring: Covert Observations of Anesthesiologists in the Operating Room,” Anesthesia & Analgesia, vol. 111, no. 3, pp. 653-658, 2010.

[6] C. M. Schulz, E. Schneider, L. Fritz, J. Vockeroth, A. Hapfelmeier, T. Brandt, E. F. Kochs, and G. Schneider, “Visual attention of anaesthetists during simulated critical incidents,” British Journal of Anaesthesia vol. 106, no. 6, pp. 807-813, 2011.

[7] R. W. Morris, and P. J. Mohacsi, “How well can anaesthetists discriminate pulse oximeter tones?,” Anaesthesia and Intensive Care, vol. 33, no. 4, pp. 497-500, 2005.

[8] G. T. Schulte, and F. E. Block, “Can people hear the pitch change on a variable-pitch pulse oximeter?,” Journal of Clinical Monitoring, vol. 8, no. 3, pp. 198-200, 1992.

[9] R. A. Stevenson, J. J. Schlesinger, and M. T. Wallace, “Effects of divided attention and operating room noise on perception of pulse oximeter pitch changes: A laboratory study,” Anesthesiology, vol. 118, no. 2, pp. 376-381, 2013.

[10] D. J. Levitin, and S. E. Rogers, “Absolute pitch: Perception, coding, and controversies,” Trends in Cognitive Sciences, vol. 9, no. 1, pp. 26-33, 2005.

[11] D. Phipps, G. H. Meakin, P. C. Beatty, C. Nsoedo, and D. Parker, “Human factors in anaesthetic practice: Insights from a task analysis,” British journal of anaesthesia, vol. 100, no. 3, pp. 333-343, 2008.

[12] M. van Pelt, and M. B. Weinger, “Distractions in the Anesthesia Work Environment: Impact on Patient Safety? Report of a Meeting Sponsored by the Anesthesia Patient Safety Foundation,” Anesthesia and analgesia, vol. 85, no. 6, 2017.

[13] E. Paterson, P. M. Sanderson, N. A. B. Paterson, and R. Loeb, “The effectiveness of enhanced pulse oximetry sonifications for conveying oxygen saturation ranges: A laboratory comparison of five auditory displays,” British Journal of Anaesthesia, vol. 119, no. 6, pp. 1224-30, 2017.

[14] K. Hinckfuss, P. Sanderson, R. G. Loeb, H. Liley, and D. Liu, “Novel pulse oximetry sonifications for neonatal oxygen saturation monitoring A laboratory study,” Human Factors, vol. 58, pp. 344-359, 2016.

[15] E. Paterson, P. Sanderson, N. A. B. Paterson, D. Liu, and R. G. Loeb, “The effectiveness of pulse oximetry sonification enhanced with tremolo and brightness for distinguishing clinically important oxygen saturation ranges: A laboratory study,” Anaesthesia, vol. 71, no. 5, pp. 565-572, 2016a.

[16] J. Edworthy, E. Hellier, K. Titchener, A. Naweed, and R. Roels, “Heterogeneity in auditory alarm sets makes them easier to learn,” International Journal of Industrial Ergonomics, vol. 41, no. 2, pp. 136-146, 2011.

[17] D. Hasfeldt, E. Laerkner, and R. Birkelund, “Noise in the operating room—What do we know? A review of the literature,” Journal of PeriAnesthesia Nursing, vol. 25, no. 6, pp. 380-386, 12//, 2010.

[18] J. M. Kracht, I. J. Busch-Vishniac, and J. E. West, “Noise in the operating rooms of Johns Hopkins Hospital,” The Journal of the Acoustical Society of America, vol. 121, no. 5 Pt1, pp. 2673-2680, 2007.

[19] R. G. Loeb, B. Brecknell, and P. Sanderson, “The sounds of desaturation: A survey of commercial pulse oximeter sonifications,” Anesthesia & Analgesia, vol. 122, no. 5, pp. 1395-1403, 2016.

[20] J. Rasmussen, A. M. Pejtersen, and L. P. Goodstein, “Cognitive systems engineering,” 1994.

[21] A. N. Healey, N. Sevdalis, and C. A. Vincent, “Measuring intra-operative interference from distraction and interruption observedin the operating theatre,” Ergonomics, vol. 49, no. 5-6, pp. 589-604, 2006.

[22] D. M. Gaba, “The future vision of simulation in healthcare,” Simulation in Healthcare, vol. 2, no. 2, pp. 126-135, 2007.

[23] D. D. Woods, “The alarm problem and directed attention in dynamic fault management,” Ergonomics, vol. 38, no. 11, pp. 2371-2393, 1995.

Concert Pieces

PLEIN AIR | *Silva Datum Musica*

Tim Collins

Colins & Goto Studio,
Rm 1M, Whisky Bond,
Glasgow, G4 9SS, Scotland
Tim@collinsandgoto.com

Reiko Goto

Colins & Goto Studio,
Rm 1M, Whisky Bond,
Glasgow, G4 9SS, Scotland
Reiko@collinsandgoto.com

ABSTRACT

Selections from PLEIN AIR | Silva Datum Musica
Plant bioacoustics, data-soni cation, computer

Side 1 - PLEIN AIR Live in Glasgow, Scotland , 2017
ALDER 5:32
OAK 8:04
ELDERBERRY 3:36
BIRCH 8:48
Side 1 playing time 26:00

Side 2 - PLEIN AIR Live in Cologne, Germany, 2015
PEAR 24:42

We shaped the PLEIN AIR project by working through a series of iterations with collaborators to refine the form, systems, and the audio to the point that it has become a simple sound instrument that sits between ourselves, one leaf, one tree. Empathy is the concept that drove the design of an experience that intends to initiate ethical consideration of trees using sound to focus attention and imagination. Outcomes include a significant exhibition consisting of the easel hooked up to live trees, processing sound in real time, as well as photographs and video that describe various aspects of the theoretical and practical development of the project. PLEIN AIR is now touring internationally and a new vinyl album – PLEIN AIR | Silva Datum Musica is available in May.

1. Artist's Statement

Collins and Goto worked with a team of scientists, technologists, and musicians to reveal the breath of a tree. Their intention was to explore the empathic interrelationship we may have with trees. PLEIN AIR integrates aesthetics, ethics, and awareness in the pursuit of a better understanding of the limitations of people-plant and culture-nature relationships. The artwork provides an experiential interface to an important but generally invisible aspect of carbon sequestration. The experience produced by PLEIN AIR is metaphoric; through the mediation of sensors and software, we hear a sound of one leaf – one tree breathing. Does our sense of moral duty change as we listen? A tree is commonly understood as property, as a utilitarian resource, and as a non-sentient thing. Yet the presence of trees in our daily lives and their bio-chemical agency, their carbon dioxide / oxygen exchange, can be construed as an essential condition of the public realm.

The idea of PLEIN AIR began in Duke Forest, while visiting the Duke University Teaching and Research Laboratory, in North Carolina. The scientists had wired the forest to test the reaction of the trees to future levels of carbon dioxide. Collins and Goto were invited to climb a forty-foot structure built among pine trees to measure photosynthesis, transpiration, and sap rise. As the scientists set up their sensors, the sun rose but was covered in cloud. When the sun emerged, the sap began to rise and the photosynthesis rate went up immediately. One scientist asked Goto to put her hand on the leaves to block the sunlight. The meter went down immediately. The response of the tree astonished the artists.

Reflecting on the experience, Goto recognized this as an 'epiphany', a moment when the essential nature of a thing was being revealed to us. Later she began to understand this as an indication of 'the phenomena of life', Edith Stein's concept describing a sense of lived connectedness and awareness of the relationship between body, mind, and environment. Stein writes, "[The phenomena of life] includes growth, development and aging, health and sickness, vigour and sluggishness" (Stein 2002 p.69). Stein confirms the phenomena of life can be observed amongst all living things including plants.

This experience and this concept have been central to the development and testing of the PLEIN AIR instrument over a series of iterations. Collins and Goto's collaboration with a plant physiologist established the authenticity of their method of gathering data and its quality, as well as giving support to their pursuit of programming that would let them hear the tree through its physiological responses. To 'hear the tree', the artists had to focus on sound as the key idea and push the technology into the background. The sound developed through years of collaboration with Chris Malcolm, with input from Georg Dietzler and many, many others.

2. Producers Statement

Visiting Glasgow in 2014, I had a chance to see and hear PLEIN AIR in the Collins & Goto Studio. I was impressed by its audio-visual richness. I then decided to present this trans-disciplinary project within a 2015 sound-art series called VISUAL SOUNDS – BIOACOUSTIC MUSIC. A short residency was provided for the artists, funded by ON – Neue Musik, Cologne. Reiko, Tim and I visited local tree nurseries that offered native regional trees and bushes. We decided to buy a mix of large potted plants for the exhibition: a butterfly bush, elderberry and hazel, and a German heritage pear tree. After the project presentation, all have been replanted in private gardens.

PLEIN AIR was presented in a square-shaped music room. Collins and Goto titled the room the 'Tree Study Sound

Chamber' (Baumklang-Studien- Zimmer), a sort of intimate music room, making reference to historical chamber music. I worked with them often during the exhibition and found myself thinking about the changing timbre and pitch of the music in relation to changes within and outside the room. Light intensity and carbon dioxide from visitors' breathing would change each tree's responses and the sound quality. At the end of the day,too, changes to the light quality would affect the trees and the sound was very different to when the trees were under full light at mid-day.

We would listen to PLEIN AIR for hours while we made the long Cologne recordings. I found an impressive richness of sounds, comparable with minimal music: steady pulses slowly changing, gradual transformations, phase shifting, consonant harmony – music that was easy to listen to. One could hear plants getting tired, stressed, hear the difference of tones in the morning, noon, afternoon, evening. Reiko, Tim and I met and talked each day, often eating together, discussing the range of sounds and the public reaction to the work. After a few days, I asked them if they ever had considered an artist edition vinyl. They were very interested. We discussed a new plan – an exhibition and more recordings at the Kibble Palace, a historic glasshouse in the Glasgow Botanic Gardens in Scotland, in 2017.

The different venues have an impact on what we heard onsite and in the experience of the recordings. Architectural scale, shape, and building materials create specific spatial acoustics. Qualities of intensity and frequency, temporal effects, and tonal attributes – all are contributing to the differing sound experiences across the two sides of the PLEIN AIR live recording. Cologne was more of a sheltered, quiet room, a reverberating hard cube with one window. Outside we had bright blue skies, warm days, and, at night, no clouds. The recording device used was a ZOOM H2n. In Glasgow, PLEIN AIR was presented in a curved, Victorian-era glasshouse. The plants chosen were all native deciduous trees of Scotland. We could only record when the public left the building at the end of the day. The weather conditions were dramatic: a mix of sunny and rainy days, very intense sunlight interrupted by fast moving clouds. The glass made for fast-changing temperatures. The recording device was a Zoom H4n using two external microphones facing towards the half-dome shape end of the glasshouse.

You will notice the differences in the trees, the venues and the sound between the two sides of the vinyl recording. The plants and context in Cologne produced minimal music, slowly changing, a steady pulse, soothing sounds. Glasgow was Much more dramatic, an extreme and dynamic range of sounds not at all like minimal music. The cities of Cologne and Glasgow are as different as the sound we hear on the vinyl. PLEIN AIR is a touching and impressive artwork, a sound piece embedded into a carefully crafted wooden painting easel. What Reiko and Tim have accomplished with PLEIN AIR opens up rigorous sensor data to an immediate and intuitive experience through sound. Over time the changes of light, humidity and carbon dioxide are all revealed in relationship to the tree and to the venue it is presented in, to the spatial positioning and to the work's relationship to the audience. This is very promising area for more generative artworks that bring us face-to-face with the sound of the breathing of a tree.

3. PAGE TITLE SECTION

The paper title (on the first page) should begin 1 inch (25.4 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information.

4. Biographies

Collins & Goto Studio: Tim Collins and Reiko Goto have developed long-term, socially engaged environmental research (SEER) that examines the cultural meaning of semi-naturalancient forest: Future Forest (2013-present); Sylva Caledonia (2015); Caledonian Decoy (2017); PLEIN AIR: The Ethical Aesthetic Impulse (2010); CO2 Edinburgh (2013); Sound of a Tree: Cologne (2016); PLEIN AIR Live at Glasgow Botanics (2017); Nine Mile Run (1997-2000); and 3 Rivers 2nd Nature (2000-2005). Outputs include artworks, exhibitions, seminars, workshops, and publications that embrace an arts-led dialogue method of research-and theory-informed public practice. They have worked with other artists, musicians, planners, communities, scientists, and technologists as well as historians and philosophers to realize work for over twenty years.

5.1 Sound Programmer: Chris Malcolm is a Scottish computer programmer and software developer with over twenty years of experience writing computer code with vector graphics, sound and interactive systems for industry clients. He also has an extensive background in experimental music developing innovative tools and instruments for studio and live performance. Malcolm is recognized within the electronic music scene for his use of retro-computers and consoles to generate unexpected interactive audio and visual experiences. The work with Collins & Goto Studio is driven by a curiosity about human relationships to technology as a tool and as an interface to bio-events. The PLEIN AIR system and software opens up new programming challenges and levels of expression not available with traditional electronic instruments and methods.

5.2 Producer: Georg Dietzler is a Cologne-based artist, author, curator, and consultant. He is recognized as an active producer of cross-disciplinary cultural projects, exhibitions, seminars and conferences, audio-visual concerts, media, dance, improv-theatre, and more. As a socio-political and conceptual artist with an international reputation, he works on ecological future visions linked to social and political change. His latest art work is a concept for an inner-city citizens' heirloom orchard, introduced at 'Ecovention Europe' at De Domijnen, in Sittard Netherlands in 2017.

5.3 Mastering/Postproduction: Dirk Specht is a sound artist, electronic and electroacoustic musician, sound recordist, and curator. From 2011 to 2016 he has been assistant professor for sound at the Academy of Media Arts Cologne, Germany. His works include electroacoustic compositions, field recording and soundscape-compositions, ars acustica/radio drama, sound art pieces and spatial installations, music for dance and choreography, and soundtracks for films and video. He also works on projects focusing on sound archives, audio restoration, sound post-production, and mastering. Having studied architecture (Berlin) and media art (Cologne), he

shares a great interest in the relations between sound and space(s), intermediality and experimental approaches to sound, music, and spatial arts. He is a founding member of Therapeutische Hörgruppe Köln and Frequenzwechsel, two media/sound-artist collectives. Specht lives and works in Cologne, Germany.

5.4 Artist: Reiko Goto Collins was born in Japan and has lived in both the US and UK. She is a principal in the Collins & Goto Studio. She has been a research fellow at the Institute for Advanced Studies in the Humanities at the University of Edinburgh. She participates in an international climate change network, Council on the Uncertain Human Future, and is currently involved in the working group 'Living Organisms and Their Choices' at the University of Edinburgh. She is a distinguished research fellow at the STUDIO for Creative Inquiry at Carnegie Mellon University in Pittsburgh, Pennsylvania.

5.5 Artist: Tim Collins is from the US, an artist, author, and planner; a principal in the Collins & Goto Studio; and an honorary research fellow in the School of Social Science at the University of Aberdeen. He works across science, technology, and philosophy to develop projects related to nature, culture, and to changing ideas about ethical duty and public space. In 2017, he was on the development committee for the 'Art and Artists in Landscape Environment Research Today' seminar at the National Gallery in London. He currently serves on the board of directors for the Landscape Research Group and Glasgow Sculpture Studios.

To download digital sound files for review

https://collinsandgoto.com/PleinAirLP/

Password: Pl31n@1r2019.

5. ILLUSTRATIONS, GRAPHS, AND PHOTOGRAPHS

Photo below: The team prepares the instrument and the recording devices at the Kibble Palace in Glasgow as an oak, and aspen and an elderberry await their performance

7. REFERENCES

Stein, E. (2002) On the Problem of Empathy. (W. Stein, trans.). Washington D.C.: ICS Publications. (Original work published 1917)

8. ACKNOWLEDGMENTS

Primary funding for exhibition planning, final development of PLEIN AIR, and the recording and editing of the sound files was provided by Creative Scotland in 2017. Final production was completed at the Glasgow Sculpture Studios. The album was produced by for the Sound Art Series by Gruenrecorder, Frankfurt, Germany.

WEATHERSYSTEMS

Stuart Duncan Haffenden Cornejo

Lancaster University
s.haffenden1@lancaster.ac.uk

https://orrest.bandcamp.com/track/weathersystems-icad-2019

ABSTRACT

WeatherSystems is a generative composition that explores the boundaries between data sonification, electronic music and citizen science. Through the use of inexpensive and open source hardware such as Arduino and different sensors, weather data was captured and then processed using different sonification techniques such as, parameter mapping and auditory icons, to provide listeners with a sonic and musical experience of weather data.

1. BACKGROUND

This project arose from research into experimental approaches to composition, with a focus on the use of data as a source for musical material. Key influences on the conceptualization of this project can be found in composers such as Robert Alexander II [3]; Byrne [2]; Zell [4], and Ryoji Ikeda 4/18/2019 10:40:00 PM, whose works present an exploratory boundary between scientific data sonification, installation art and music composition.

The use of sonification for the creation of artistic work presents several important areas for exploration, namely the consideration of aesthetics and musicality of sonification, as well as the role of the composer/sound designer.

> The craft of composition is important to auditory display design. For example, a composer's skills can contribute to making auditory displays more pleasant and sonically integrated and so contribute significantly to the acceptance of such displays. There are clear parallels between the composer's role in AD and the graphic artist's role in data visualization. Improved aesthetics will likely reduce display fatigue. Similar conclusions can be reached about the benefits of a composer's skills to making displays more integrated, varied, defined, and less prone to rhythmic or melodic irritants.
> —Gregory Kramer, Auditory Display,1994 [49, pp. 52–53] [1]

Additionally, the use of data to create compelling musical experiences, allows listeners to feel more connected to the realm of science, which can feel impenetrable to most. By allowing people to experience the intangible systems around them, interest, curiosity and embodiment are fomented, leading to a more democratic and inclusive discussion of data.

2. WEATHERSYSTEMS

A custom-made data collection system was developed using environmental sensors for: temperature, humidity, dew point, light intensity, carbon dioxide, carbon monoxide and methane gas. This was then placed in the woods in Windermere, Lake District and set to record 24 hours of weather data.

As this piece was not intended to provide any sort of scientific contribution, a conceptual and artistic approach was employed to determining how to scale and map the collected data. Thus, it was determined that to provide a structural framework for the piece the light intensity data stream was to be used. This was done with the intention of creating an A B structure, where A represents the night (ie. low light intensity) and B represents the day time (high light intensity). This structure can be observed as the envelope of the graph in Fig 1.

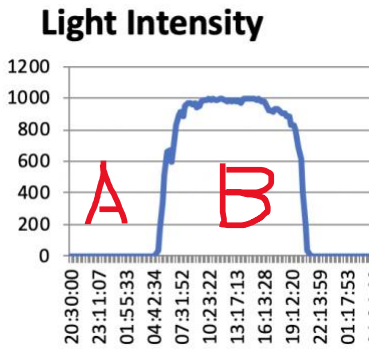


Figure 1:

Each of the two sections presents a contrasting timbre and instrumentation to give the impression of night (darkness) and day (light). The A section features the sound of a bell, which demarcates every passing hour, its pitch is determined by temperature. Humidity and dew point are mapped to trigger high-pitched metallic sounds whose pitch and rhythm is determined by stochastically chosen thresholds being crossed by the data streams.

As the light intensity increases, the bells and metallic sounds give way to three synthesizer sounds whose pitch is controlled by humidity, temperature and dew point. This signals the transition into day and the B section of the piece and creates and emotive crescendo corresponding to the sun rising. Here the demarcation of time is represented with each pulse, thus giving the piece an even rhythm throughout that corresponds to the time of day. As the light fades away the nighttime instrumentation fades in, signaling the end of the day.

Further to the melodic elements that indicate temperature, humidity and dewpoint there are a number of background sounds that correspond to weather features that occur when these three parameters meet certain thresholds. Periods where humidity was over 90% and dew point equal to temperature (the point at which precipitation occurs), were demarcated by the sound of a field recording of rain. The pollutant gas data streams where used to trigger granular synthesis bursts and textures, whenever they crossed stochastically chosen thresholds determined by the envelope of the graphs of the corresponding data streams. This was done to represent pollutant gasses in a metaphorical way with sounds that are perceived as noisy and gritty.

While these parameter mapping methods resulted from compositional nous and the usual trial and error, rather than through scientific rigor, they still yielded a clearly audible and communicable contrast in the data between day and night. By using light intensity to dictate the structure of the piece, a distinct difference between night and day periods can be experienced in a musical way. Through this methodology WeatherSystems hopefully presents a compelling experience to the listener, showcasing an alternative compositional approach that has data at its core.

3. REFERENCES

[1] S. Barrass and P. Vickers (2011) 'Chapter 7: Sonification Design and Aesthetics' in *The Sonification Handbook* (T. Hermann, A. Hunt, and J.G. Neuhoff, eds), Berlin: Logos-Verlag.

[2] M. Byrne (2013). Spaced Out: "The Space Composer" is Making Music With the Sun. Motherboard. https://motherboard.vice.com/en_us/article/wnnyym/spaced-out-making-music-with-the-sun (accessed 4.18.19).

[3] R. Alexander II (2014). Two turn tables and a satellite: Robert Alexander at TEDxUofM.

[4] H. Zell (2014). More Than Meets the Eye: NASA Scientists Listen to Data [WWW Document]. NASA. http://www.nasa.gov/content/goddard/more-than-meets-the-eye-nasa-scientists-listen-to-data (accessed 4.18.19).

LIGHT-CURVE-DRIVEN SOUNDSCAPES

Adrián García Riber

Image and Sound Art,
Francesc Martí i Mora 1-B 22-3,
Palma de Mallorca, 07011, Spain
adrian@imageandsoundart.com

ABSTRACT

The darkness, in some of its forms, has traditionally been one of the sources of inspiration for almost any composer regardless of the moment, place or musical genre under analysis. Far from this beautiful, necessary and natural romantic orientation, this composition tries to use light not only as inspiration but also as an endless source of musical information that can be layered and structured to conform experimental electroacoustic music pieces. Under this composition paradigm, selected fragments of the publicly available light curves from the Mikulski Archive for Space Telescopes (MAST) [1,2], the simulated light curves of the Planet Hunters project from the Transiting Exoplanet Survey Satellite (TESS) [3,4], the curves generated with the *Lightkurve* software package for Kepler & TESS time series analysis in Python [5] and the Catalog and Atlas of Eclipsing Binaries (CALEB) [6], have been used to create a multimodal composition that uses *Sonifigrapher* prototype as sound and visual engine. This *ad hoc* developed sonified graph's synthesizer, implemented in CSound plus Cabbage environment [7], [8], converts light curves fragments into audio spectra through software additive synthesis .

1. REFERENCE WORKS AND INSPIRATIONS

The wide range of influences and references used in this multidisciplinary work that proposes a way of listening to the stars' light, can be simplified in two main axes of inspiration: Experimental musical instruments and Astronomical data Sonifications. From the musical perspective, Alexander N. Scriabin had already represented at the beginning of the twentieth century the opposite idea of converting sound into light with his '*Clavier a Lumières*', created to be used in '*Prometheus: The Poem of Fire*' [9]. On the other hand, in the last decades of the century and attending to the idea of transducing graphic information into sound, the Iannis Xenakis' Unité Polyagogique Informatique du CEMAMu (UPIC) represented a new paradigm in the use of experimental music devices [10], which underlies almost every current project focused on graphic to sound conversion. Centered on the Sonification field, worth a mention the multi-purpose auditory graph tool *Sonification Sandbox* by Walker & Cothran [11] -based on previous projects *MUSE* and *MUSEART*-, the Bell3D audio-based Astronomy Education System [12] by Jaime Ferguson and the xSonify astronomical data sonification software [13] from Diaz-Merced et al. [14], these last, both designed to improve accessibility in Data-Driven Astronomy.

2. ABOUT LIGHT CURVES AND TRANSITS

Light curves are graphic representations of the brightness flux variations observed in celestial objects along time. If an object passes between the observer and a star, it generates a partial eclipse that produces a flux decrement in its light curve which can be measured in terms of time and depth.

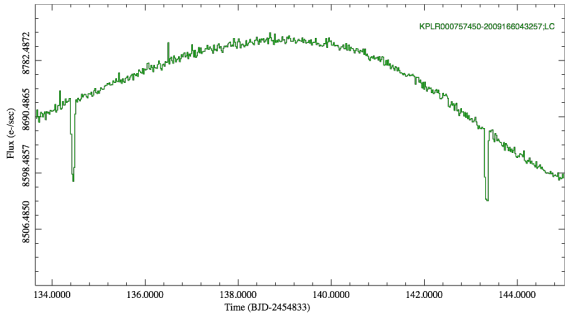


Figure 1: Enlarged KPLR000757450-2009166043257 sequence showing two planet transits [15]. Sap Bright Flux (Simple Aperture Photometry, electrons per second) [16] vs time expressed in BJD-2454833 (Barycentric Julian Date, 2454833.0 offset) [17].

As described by Seager & Mallén-Ornelas [18] in the field of exoplanet exploration, a planet's orbit can be characterized by analyzing the decrease of energy presented in the light curve of its star. The orbital period of the planet might be obtained measuring time between these light decrements, called transits, and the planet's temperature and atmospheric properties can be determined through transmission spectroscopy methods consisting on the observation of transit light curves at different wavelengths [19].

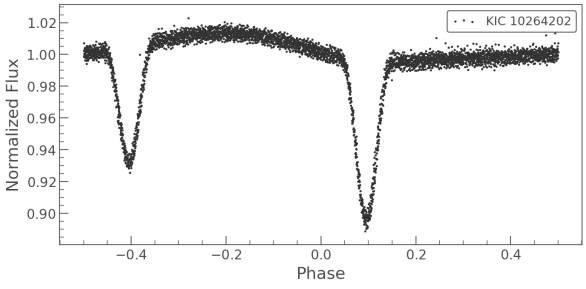


Figure 2: X-ray light curve for KIC 10264202 eclipsing binary star system. Normalized flux vs phase [5].

In a similar way, analyzing the X-ray emission of an eclipsing binary star system, it is possible to estimate the relative size of the two stars that orbit their common center of mass. In *Figure 2*, the X-ray intensity is highest when both stars are completely visible, and lowest when the X-ray

emitting star is eclipsed by the central star [20]. This kind of light curves often present two different sized dips while planet transits tend to have a similar loss of flux. This feature results crucial in light curve's classification and can be notice if comparing *Figures 1* and *2*. On the other hand, this kind of representation can also be used to characterize supernovas, allowing to generate graphic descriptions of the star's extinction massive explosions. Under a cross-discipline correlation perspective, supernova's light curves could keep a certain similarity to echograms representing the decrease of sound energy inside a room.

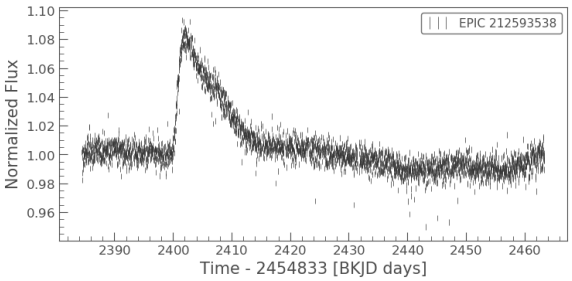


Figure 3: Supernova light curve. EPIC 212593538. Normalized flux vs phase [5].

3. ABOUT SONIFIGRAPHER

Sonifigrapher is a quadraphonic virtual instrument prototype designed for sonifying light curves. It works in a similar way to a wavetable sampler, generating filter-controlled audio spectra through an additive synthesizer which control variables have been mapped from the graphic representation input. *Sonifigrapher* can be downloaded for testing as a packed 'ZIP' file from:

https://archive.org/details/SonifigrapherMacOSX

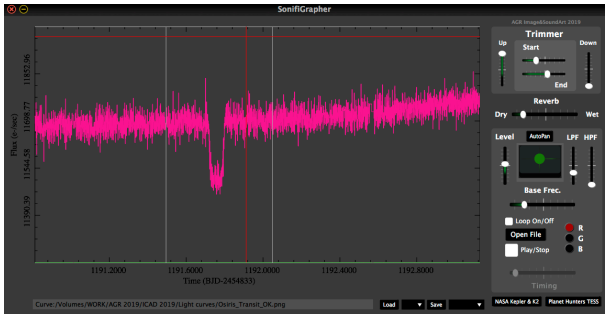


Figure 4: *Sonifigrapher* interface capture during a single planet transit sonification [15]. PDCsap Bright Flux [16] vs time expressed in BJD-2454833 [17].

Based on Marilungo's examples of CSound's image processing opcodes [21], [22], and McCurdy's Cabbage and CSound examples [23], *Sonifigrapher* uses additive synthesis with a non-quantified frequency scale generated from a user-defined base frequency. This approach makes it possible to create tonal sweeps and microtonal sounds or chords and improves accuracy in light curves' pitch tracking. As the final sonified spectrum relies on the mentioned user-defined base frequency, it is possible to adapt the graphic changes in the curves to different frequency ranges for a better perception of the sonification, or fine tuning in creative applications. To maintain coherence with the transit detection method, lowest flux values in the curves correspond to highest frequencies in the sonification. In this way, when a transit is produced, a high frequency sine is reproduced facilitating its detection.

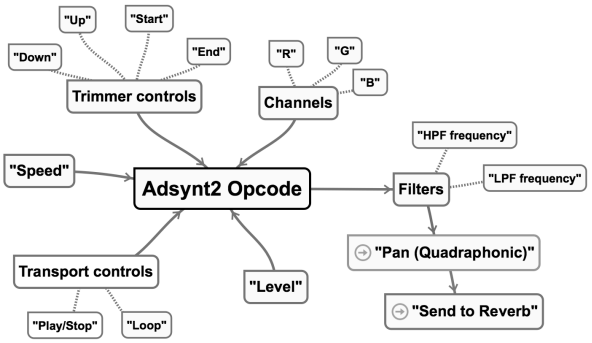


Figure 5: *Sonifigrapher* design map showing control variables and signal flow.

The core of the prototype is the CSound's a*dsynt2* opcode [21]. This opcode also provides interpolation to lightly soften the most pronounced graphic transitions. *Figure 5* describes the design implementation map with all the variables used to control the sonification process. The R, G and B values of the loaded image are extracted using CSound's image processing opcodes [21] to work as input arguments for *adsynt2*. Its monophonic output is low- and high-pass filtered and sent in parallel to a quadraphonic matrix and reverberation processor to be used in creative applications. End users can select the sonified R, B or G channel input to reduce noise and focus attention. The synthesizer provides trimmer controls to adjust the 'start' and 'end' points of reproduction as well as the 'up' and 'down' graphic limits in order to avoid the sonification of non-relevant information printed in the sampled image. The speed control allows both detailed analysis and fast monitoring, if the loop playback is not enabled. The loop reproduction works on an 8 seconds Mellotron-based time scale [24] and disables timing control. All changes made to the 'trimmer', 'speed' and 'loop' controls are applied once the current reproduction is completed. High- and low-pass filters frequencies are controlled before the signal is sent to the reverberation processor to improve sound quality. A user controlled "x-y" matrix is also provided for sound allocation in a quadraphonic reproduction system. Default auto-panning configuration follows the graphic timeline bar with stereo compatibility. The 'level' fader acts over both the 'dry' and 'wet' signals by minimizing the number of controls required.

4. ABOUT THE PIECE

Light-curve-driven soundscapes is an unreleased multimodal sonification composition framed within the electroacoustic experimental music context. Built from selected samplers of the Mikulski Archive for Space Telescopes (MAST) [1,2], the simulated light curves of the Planet Hunters project from the Transiting Exoplanet Survey Satellite (TESS) [3,4], the curves generated with the *Lightkurve* software package for Kepler & TESS time series analysis in Python [5] and the Catalog and Atlas of Eclipsing Binaries (CALEB) [6] public archives and catalogues, the piece has been created to be reproduced in a quadraphonic environment using a laptop, an audio interface and Cabbage's *Patcher* capabilities. With a total duration of around eight minutes, this composition proposes a way to explore the universe through the possibilities offered by sonified light curves, adding a sound dimension to their inherent information. A promotional video with some fragments of the piece is available at:

https://vimeo.com/331259492

5. REFERENCES

[1] https://www.nasa.gov/kepler/education/getlightcurves

[2] http://archive.stsci.edu/kepler/data_search/search.php

[3] https://www.zooniverse.org/projects/nora-dot-eisner/planet-hunters-tess

[4] https://tess.mit.edu

[5] https://docs.lightkurve.org/

[6] http://caleb.eastern.edu/

[7] CSound software, accessed March 2019: http://www.csounds.com

[8] Cabbage software accessed March 2019: http://cabbageaudio.com

[9] Triarhou, L. (2015). *Scriabin for Neuroscientists: A Study in Syn-Aesthetics*. CreateSpace Independent Publishing Platform.

[10] Xenakis, I. (1992). *Formalized Music*. *Thought and Mathematics in Music*. Hillsdale, NY: Pendragon Press.

[11] Walker, B. N. and Cothran, J. T. (July 2003). *Proceedings of the 2003 International Conference on Auditory Display,* Boston, USA.

[12] Ferguson, J. (2016). Bell3D: An Audio-based Astronomy Education System for Visually-impared Students. *CAPjournal*, No.20, pp35.

[13] Diaz Merced, W. L. (2013). *Sound for the exploration of space physics data.* (Doctoral Thesis). University of Glasgow.

[14] Diaz-Merced, W. L., Candey, R.M., Brickhouse,N., Schneps, M., Mannone, J.C., Brewster, S. and Kolenberg, K. (2012). Sonification of Astronomical Data. *New Horizons in Time-Domain Astronomy Proceedings IAU Symposium No. 285, 2011. R.E.M. Griffin, R.J. Hanisch & R. Seaman, eds.*

[15] http://archive.stsci.edu/kepler/condition_flag.html

[16] https://keplergo.arc.nasa.gov/PyKEprimerLCs.shtml

[17] http://archive.stsci.edu/kepler/manuals/archive_manual.pdf

[18] Seager, S. & Mallén-Ornelas, G. (2002). A Unique Solution of Planet and Star Parameters from an Extrasolar Planet Transit Light Curve. Retrieved from: https://iopscience.iop.org/article/10.1086/346105/fulltext /

[19] Alapini Odunlade, A. E. P. (2010) *Transiting exoplanets: characterization in the presence of stellar activity.* Doctoral Thesis. University of Exeter.

[20] https://imagine.gsfc.nasa.gov/educators/hera_college/binary-model.html

[21] Vercoe, B. MIT Media Lab et al. *The Canonical Csound Reference Manual*. Retrieved from: http://www.csounds.com/manual/html/

[22] Boulanger, R. (Ed.) (2000). *The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming.* Cambridge, MA, USA: MIT Press.

[23] McCurdy, I. accessed March 2019, http://iainmccurdy.org/

[24] *Mellotron service manual,* https://web.archive.org/web/20111218174345/http://www.cem3374.com/docs/Manuals/Misc/Mellotron_MkII_SM.pdf

6. ACKNOWLEDGMENT

Special thanks to Ruth Capó Mesa for listening to the stars.

WE INTERACT

Daniel Grayvold

Composer, Sound Designer, Musician
daniel@dgrayvold.com

1. INTRODUCTION

There is a simple beauty in the concept of a text message. With just a limited set of characters we can tell someone our favorite joke, or perhaps give a biting argument. We can wax poetic about our love or just say "hello." Never in the course of history has humanity had such a swift and effortless method of communication as it does now with cell phones. Text messaging is a nearly universally shared experience, and these little bits of text and icons have a story to tell about the people who send them. When brought together, they have a voice that can tell us much about the interactions we have with others and the ways we use these devices in our everyday lives. *We Interact* is a composition that tells that story through my own personal data.

2. FOUNDATION

Over the course of 2018, I sent and received more than 8,000 text messages. The content of those messages brought me happiness, sadness, laughter, love, and so much more. Each single message could have been part of a long conversation or a simple question and answer. They were tiny interactions with the world around me. A few may not tell much of a story, but hearing them all together, I knew, would paint another picture entirely.

We Interact is a story told through the sonification of digital interactions with my friends, family, and more over the course of the year 2018. Each measure of the piece contains a single day's worth of interactions; the arrival of each day is marked with an arpeggio played softly by a piano. At each moment when I would receive a text message from someone, a note would play. If it was the first message within twelve hours, the pitch of the note would be the root note of the song's key's scale—D3 in this case. If it was following a message sent or received by the same person within twelve hours, the note's pitch would climb further up the scale. Each time more than twelve hours would pass between messages, the pitch would reset to D3.

With just this simple algorithm, a pattern instantly emerged. By listening to the number of notes with the base pitch, I could hear how many times a new conversation started. Some conversations would only increase in pitch once or twice. These were quick interactions that I had with people:

> "What time does the show start?"
> "Made it home okay."
> "Looks like rain."

Some of these conversations, however, emerged as clusters of notes played in rapid succession. The pitches shot up to the ceiling of the scale and remained there for a time, marking conversations that lasted hours or days. These were perhaps the most memorable of the interactions that I made—the ones in which secrets were shared or heated arguments were waged.

By hearing these messages played out over time, I was able to learn a little bit more about myself and the way I used technology to interact with the people in my life. I was able to remember specific conversations I had based on clusters of notes on specific days; I could hear when I was particularly busy and could not check my phone as often. When the notes would play very early in the measure, those were the nights in which I stayed up a little too late talking with someone who was perhaps hundreds or thousands of miles away.

These tiny insights not only showed me who I was in a unique way but also endued me with a greater appreciation for a technological marvel that I had been taking for granted. Without putting any thought towards it, I was able to stay close to people I cared about even when they were all over the world.

3. CREATION

To create *We Interact*, I had to have a source of message data that I could search through and play back. The operating system my cell phone uses relies on a database file named chat.db for its storage of messages. Within this file are all of the conversations I've ever had since I bought the phone. With a little programming using SQL and cleanup with basic spreadsheet software, I converted the database into a comma separated value file (CSV) containing the date the message was delivered, the person for which the message was sent or received, and whether or not I sent the message. Since the composition relied on timing that was accurate to the minute, I calculated each message's delivery date in minutes from midnight on January 1, 2018. From there, I was able to start building the piece. A short excerpt of this data is found on the table below.

I used a graphical media programming application called Pure Data to create a patch that would perform *We Interact*. With the data being in a CSV format, I was able to load it into a text object and search through it as needed for playback. The patch counted each day by minute using a metro (metronome) object. When playback began, the patch counted up at a rate of 360 ticks per second or 1,440 per 4/4 measure. This corresponded to the 1,440 minutes that each day contained. The algorithm for determining melody note pitches was programmed in using a text object storing the person's id, last message date, and last note pitch. This was compared to the current message playing at the moment to determine the pitch. Note pitches were stored using additional text objects.

Logic Pro X connected to Pure Data through MIDI provided the actual sonification with various virtual instruments and synths.

The composition features piano front and center since it has both a relatively sharp attack and a long sustain; it also is an important instrument in my life and further reflects this piece's personal connection to me. Reverbs and sustained notes in the piano and synths provide a more ethereal timbre in order to convey a sense of the days and weeks blurring together. As each message needs to be a distinct point in time, the piano/synth melody notes are layered with pizzicato strings to sharpen the attack just enough to make each note easily distinguishable. I also randomized the bass arpeggios slightly to add interest.

4. PERFORMANCE

Live performance of *We Interact* varies slightly from the computer-generated recording. Instead of playing a note for every single message, I instead focus on the beginnings and endings of each new conversation; a moderate degree of accuracy is traded off for a more easily playable composition.

Date	Date (Minutes)	Unique Person ID	Is From Me
2018-01-06 22:30:51	8551	7	FALSE
2018-01-07 09:48:11	9228	7	FALSE
2018-01-07 21:36:04	9936	8	FALSE
2018-01-07 21:36:08	9936	8	TRUE
2018-01-11 19:27:04	15567	12	TRUE
2018-01-12 03:24:25	16044	12	TRUE
2018-01-14 20:55:03	19975	15	FALSE
2018-01-14 21:07:20	19987	15	FALSE

Table 1: A selection of data from my personal messages database

Only a computer and a MIDI keyboard is needed to play the composition. The addition of an optional external display can give context to the composition through visualization of the underlying data during performance. The performance of this composition including its optional supplemental material lasts less than ten minutes.

5. CONCLUSION

Creating *We Interact* showed me a part of myself that I had not considered before. I was able to use the technology that I interact with every day to demonstrate how I interact with people every day. This was using data that had been collected automatically by my phone with no prompting from me. In a time when the guarding of personal data is something on our minds and a cause of fear, it is encouraging to know that there is an opportunity for positive use of this sensitive information. The possibilities for the evolution of this composition are endless, too. The addition of the message data of others or the use of other personal information such as purchases and picture data provides exciting possibilities for evolution of *We Interact*.

Each person with a cell phone carries with them stories that are waiting to be found. We leave a trail that can reveal much about ourselves when we interact with others through this technology. All it takes to hear that story is some data manipulation and sonification.

5.1. Media

A sample recording of *We Interact* can be found at `dgrayvold.com/features/weinteract`.

LISTENING BACK

Jasmine Guffond

School of Art and Design, University of N.S.W. Sydney, Australia
j.guffond@student.unsw.edu.au

ABSTRACT

Listening Back is a plug-in for the Chrome browser that sonifies internet cookies in real time as one browses online. Utilising digital waveform synthesis, the *Listening Back* browser add-on provides an audible presence for hidden infrastructures that collect personal and identifying data by storing a file on one's computer. By sonifying Internet cookies this browser add-on functions to expose real-time digital surveillance and consequently the ways in which our everyday relationships to being surveilled have become normalised. Online surveillance equates to the extraction, management, selling, and ultimately control of personal data. Veiled by the browser interface, these algorithmic processes remain largely obscured from users of online platforms and digital services. *Listening Back* provides tangible access to the real-time relationship between monitoring infrastructures and Web browsing experience by enabling users to go beyond the event on the screen and engage with complex phenomena behind its graphical interface. Sonification is therefore employed as a compositional tool to reveal asymmetrical relationships of power inherent to online surveillance cultures. For ICAD 2019 I present a live concert in the form of a lecture performance. The *Browser Duo* utilises the *Listening Back* add-on to create a composition generated by real-time cookie activity.

1. INTRODUCTION

Listening Back is a plug-in for the Chrome[1] browser that sonifies internet cookies in real time as one browses online. Utilising digital waveform synthesis, *Listening Back* provides an audible presence for hidden infrastructures that collect personal and identifying data by storing a file on one's computer. Addressing the proliferation of ubiquitous online surveillance and the methods by which our information flows are intercepted by mechanisms of automated data collection, *Listening Back* functions as a poetic exposition of real-time digital surveillance and consequently the ways in which our everyday relationships to being surveilled have become normalised. Acknowledging the contemporary 'surveillant assemblage'[2] as increasingly based upon a consistent flow of immaterial data tracking, sound as a similarly immaterial medium is particularly suited to represent algorithmic online surveillance processes of data collection. The aim is to provide critical awareness and engagement with the proliferation of ubiquitous online tracking technologies by providing a sonic experiential platform for the real-time activity of Internet cookies.

Within the online tracking ecosystem, multiple surveillance technologies are implemented via the World Wide Web. Our personal data is thereby collected, aggregated, compiled and sold. Such data can include our IP address, type of computer or mobile phone, operating system, the plugins we have installed, our searches, our likes, the websites we visit, what we buy, watch, and how long our cursor lingers on a page. Some of the lesser known online surveillance technologies include web bugs, audio beacons, Web RTC IP discovery, third-party HTTP requests, device fingerprinting, canvas fingerprinting, browser fingerprinting, font fingerprinting, audio context fingerprinting and battery API fingerprinting. A recent measurement study of online privacy across one million of the most visited websites reveals how each and every characteristic of our devices, that is, every technical component and property across hardware and software can be, and will be, repurposed to identify and track us.[3] Even with the enforcement of the European General Data Privacy Regulation (GDPR)[4] in May 2018, by which websites are mandated to inform visitors of the tracking technologies embedded on their website, the majority of Web users within the European Union and beyond remain unaware of the plethora of surveillance technologies monitoring their every online move. However everyone has come across the tracking technology known as the cookie. By co-opting the Internet cookie, users of the *Listening Back* browser add-on already have some concept of, and thereby immediate means of engaging with, the sonified data. Furthermore, the cookie represents a significant case study for online surveillance cultures. As I outline in my lecture performance, the invention of cookies in 1994 is historically situated at the origins of online automated data collection and the commercialisation of the World Wide Web.

2. THE CHROME API

The data set I have access to for this project is determined by the Chrome API (application programming interface) and therefore which data Google gives third party developers access to. A Web API is a programming interface for a Web server or browser

[1] https://chrome.google.com/webstore/detail/listening-back/gdkmphlncmoloepkpifnhneogcliiiah

[2] Haggerty, K.D., and Ericson, R.V. 2000. "The Surveillant Assemblage." British Journal of Sociology, 51(4), pp. 605–622.

[3] Englehardt, S and Narayanan, A. 2016, "Online Tracking: A 1-Million-Site Measurement and Analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security – CCS'16*, the 2016 ACM SIGSAC Conference, Vienna, Austria: ACM Press, 1388–1401, https://doi.org/10.1145/2976749.2978313.

[4] https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en (accessed 10/06/2019).

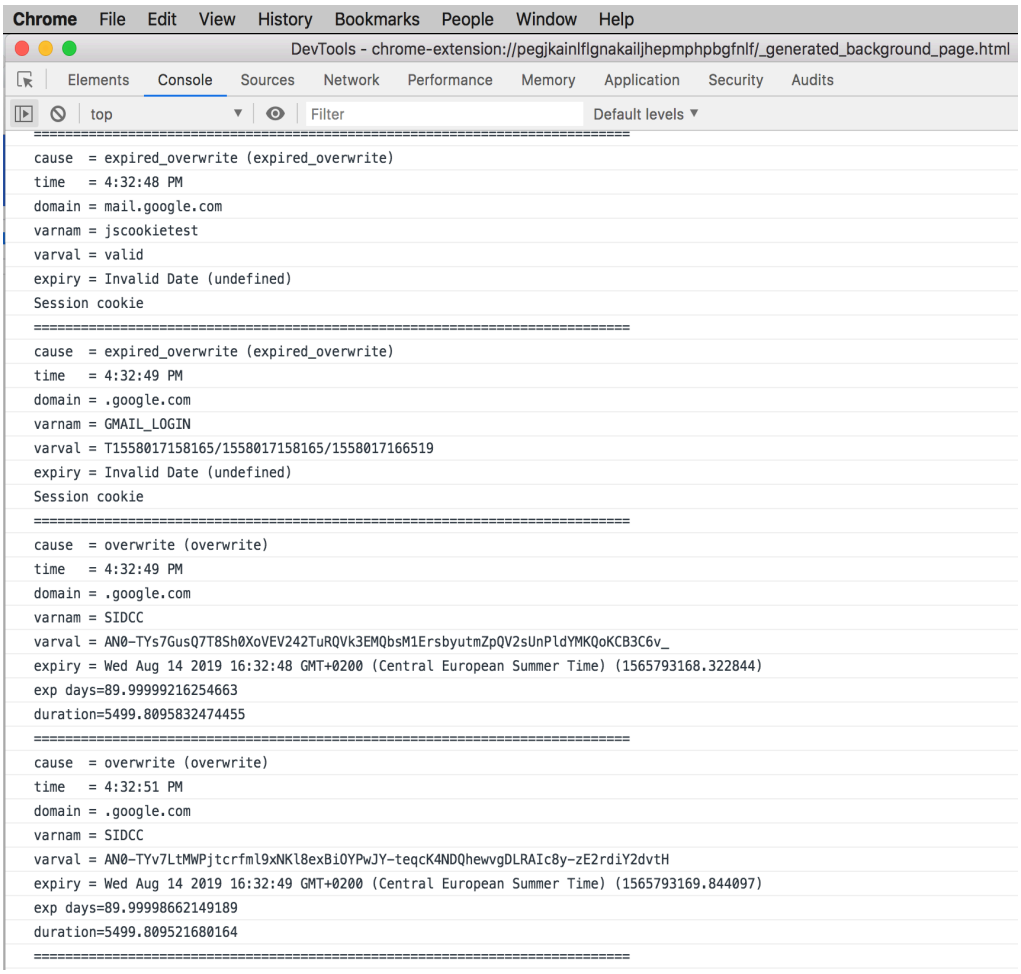


Figure 1: Chrome browser cookie log—logging into my gmail account

which predetermines the objects, actions or protocol the developer may need to access in order to develop a third party application. For the Chrome API the data set includes each time a cookie is inserted onto the user's computer, deleted from the user's computer or overwritten. Other information such as each time a cookie is read by the Web browser and server is not included. A cookie log is provided by the Chrome API and this is the text I often referred to in trying to understand and decipher the cookiesphere (see Fig. 1). This log is provided for troubleshooting purposes and ranges from difficult to impossible to interpret. Its inherent indecipherability drew attention to the fact that certain technical processes are hidden or obscured even from tech savvy programmers, as they are in fact well kept business secrets for major tech corporations and the online data broker industry.

In Fig. 1, 'varnam' = variable name and 'varval' = variable value. These values are determined by the programmer behind the Web server hosting the website and are often difficult if not impossible to interpret. The varnam 'GMAIL_LOGIN' is the most clearly named variable. One could probably assume that the 'ID' in the 'varnam = SIDCC' refers to ID, since cookies are essentially identifying tokens placed on the user's computer so that the browser and Web server can identify the user. What information is collected regarding the user would probably be revealed by deciphering the 'varval'. However this is untranslatable to anyone but the programmer of the cookie.

Other information that one can glean from the cookie log is the time, date and domain name of the cookie. The domain name determines if it is a first or third party cookie. A first party cookie has the same domain name as the website one is currently visiting. A third party cookie is any cookie with a different domain name to the website one is currently visiting. The expiry date indicates the duration for which the cookie is programmed to be on the user's computer. If the date is defined as 'invalid' it indicates that the cookie is a session cookie. A session cookie is deleted from the user's computer when the user quits the Web browser. A persistent cookie is inserted onto the user's computer for as long as it has been programmed to be there for. This can be anything from minutes, to hours, to days, to months, to years. It is however possible for the user to clear the cache via the browser settings and therefore remove cookies until they are again reinserted by the browser.

3. AESTHETIC CONSIDERATIONS

By engaging sound as a means of inquiry that exploits 'the ever-openness of the ear'[5] it is my aim that the user can listen either attentively or peripherally to the Internet cookie continuum while simultaneously engaging in their daily browsing routines — checking emails, shopping, posting, communicating, liking, commenting etc. I'm particularly interested in our auditory ability to peripherally monitor or listen in parallell[6] for extended periods of time. The *Listening Back* browser add-on operates across two modalities — (un)private use on personal computers, and for live performance and installation primarily within art and music contexts. Aesthetic decisions were partly determined by parameters and limitations imposed by the technology itself. Browser add-ons were never intended to process large amounts of sound. During the development process certain websites were crashing the browser due to the sheer amount of cookie activity so that I had to limit the amount of cookies that can trigger sound at any one time to forty three. A Web browser however, is capable of sending a lot more cookies from the Web server to the user. This is indicated by the Internet Engineering Task Force (IETF), the de-facto internet standardisation body's cookie implementation considerations"[7]

> **6. Implementation Considerations**
>
> **6.1 Limits**
>
> Practical user agent implementations have limits on the number and size of cookies that they can store. General-use user agents SHOULD provide each of the following minimum capabilities:
>
> - At least 4096 bytes per cookie (as measured by the sum of the length of the cookie's name, value, and attributes).
> - At least 50 cookies per domain.
> - At least 3000 cookies total.

For *Listening Back*, the sonification of data is employed and integrated via a Web browser add-on. Max Breedon, the tech saavy programmer I worked with on this project, suggested we use the timbre.js library[8] as the most practical method for generating web audio. I therefore designed sounds via digital waveform synthesis. I use sine, saw or triangle waves, white noise and a range of effects — EQ, delay, phasor, flanger, reverb. With live performance in mind, I chose four scales from which notes are randomly generated — D minor, F major, G minor, Bb major. The default scale is D minor and live performers have the option to select between different scales via the user interface. I deliberately chose scales that are in tune with each other as my previous experience with earlier sonification projects[9] has indicated that dissonance can have the effect of interrupting, making people disengage or simply turn the sound off. Moreover, when working with real-time data the results tend to occur unpredictably and in an ever changing state of flux, as an indeterminate composition derived from real-time data unfolds. A simple harmonic structure can help listeners to decipher from the complexity of simultaneous layers of sonic information.

I designed specific signature sounds for major online platforms — Google, Facebook, Amazon, YouTube, Expedia and some of the third party advertising cookies that are prevalent across many websites, such as krxd.net. From within this group the Google analytics cookie is the most prevalent cookie embedded across the Web and our personal computers, and Facebook is the second largest data collector in the online tracking ecosphere.[10] I chose one sound from the timbre.js library called 'pluck' that plays outside of the tuning of the four scales. I assigned the 'pluck' sound to be the generic cookie sound, because I found it to be a pleasant sound and therefore suited to being continually played. The note of each generic cookie is generated from a number produced by a hash of the domain name of the cookie. The generic cookies are divided into two data sets, first party cookies and third party cookies. For the third party cookie sound I added distortion to the pluck sound as a means of differentiating between the two. *Listening Back* incorporates an interface that allows listeners to modify the volume of cookies according to domain name as well as two specific volume sliders — one for first party cookies and the other for third party cookies. The intention is to allow listeners to be able to decipher between first and third party cookies. This is significant because first party cookies are mostly functional and third party cookies are implemented for tracking browsing behaviour.

The duration that a cookie is programmed to be on a user's computer is mapped to the duration of the signature sounds on an exponential scale from one hundred milliseconds to seven seconds. Due to CPU limitations it was not possible to program individual sounds to play back for any longer.

4. THE WORLD WIDE WEB AS A MUSICAL INSTRUMENT

Anyone can download the *Listening Back* browser add-on from the Chrome store. Additionally, I have used it for live performance with the Browser Duo, Trio or Ensemble, for lecture performances, and for immersive sound installation. In these situations it is ideal to have a high fidelity full range sound system — that is a four speaker system that includes subwoofers. In order to experience the ubiquitousness nature of online surveillance and the way by which it is woven into the fabric of our everyday lives, I've explored immersive sound environments and the potential to experience sound in an embodied way. Philosopher Jean-Luc Nancy in his book *Listening* notes the acoustic functioning of sound to propagate "throughout the entire body something of its effects, which could not be said to occur in the same way with the visual signal."[11] What Nancy is referring to is the fact that our bones conduct sound and so in effect we listen with our entire body. Sound engages us

[5] Kim-Cohen, S. 2009, *In the Blink of an Ear, Toward a Non-Cochlear Sonic Art*, New York & London: Continuum, XVIII.

[6] Brown, M.L., Newsome, S.L., & Gilbert, E.P. 1989 "An experiment into the use of auditory cues to reduce visual workload." Proceedings oft he ACM CHI 89 Human Factors in Computering Systems Conference CHI 89: pp. 339–346. Fitch, W.T., & Kramer, G. 1994. "Sonifying the body electric: Superiority of an auditory over a visual display in a complex, multivariate system." In Kramer, G (Ed.), Auditory Display: Sonification, Audification, and Auditory Interfaces, pp. 307–326. Vickers, P. 2011. "Sonification and Process Monitoring." In Hermann, T., Hunt, A. and Neuhoff, J. G. (eds), The Sonification Handbook. Berlin: Logos Publishing House, pp. 455–92.

[7] Barth, A. 2011, "HTTP State Management Mechanism", *Internet Engineering Task Force, Requests for Comments 6265*, `https://tools.ietf.org/html/rfc6265` (accessed 28.03.2019).

[8] `http://mohayonao.github.io/timbre.js/` (accessed 05.03.2017).

[9] Guffond, J, `http://jasmineguffond.com/?path=art/Anywhere+All+The+Time` (accessed 12.04.2019).

[10] Englehardt and Narayanan, "Online Tracking", p. 8.

[11] Nancy, J. L. 2007, *Listening*, trans. Charlotte Mandell, New York: Fordham University Press, p. 14.

physically as well as mentally and therefore provides for an analytic means through embodied experience.

With live performances I'm interested in exploring the add-on as a musical instrument, and consequently the World Wide Web as a musical instrument. The *Listening Back* interface (see Fig. 2) allows performers to shift between the four scales as well as transpose the pitch up or down one or two octaves. In addition the interface enables the performer to select a domain name and then individually change the volume and octave for the cookies generated by that domain. By employing octaves a relatively simple harmonic composition is maintained despite the fact that a lot of sound is continually generated in real time, especially in situations where there is more than a single performer. As previously mentioned, there are additional volume sliders for first and third party cookies so that it is possible to discern between the two and reduce the amount of cookie generated sounds.

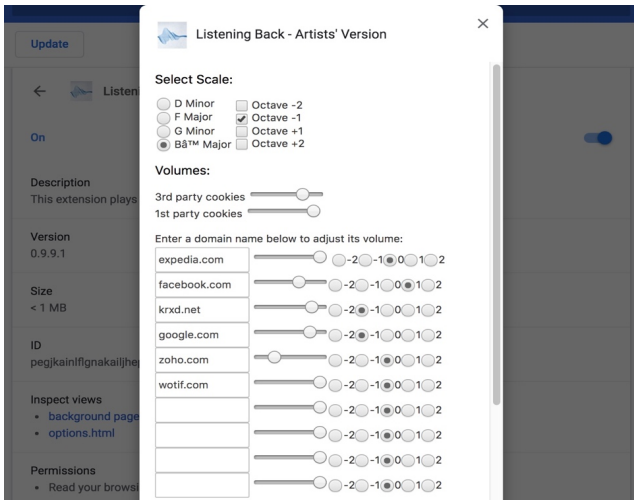


Figure 2: The *Listening Back* User Interface

5. BEYOND THE EVENT ON THE SCREEN

I situate this work as artistically engaging sonification as a 'mediator of the invisible[12] and adopt composer Barry Truax's notion of 'acoustic communication'. By acoustic communication Truax is referring to a creative sound practice that maps audio to real world data in order to 'direct the listener's attention back to an understanding of some facet of that world'[13] *Listening Back* directs the listener's attention to hidden processes of online data collection, specifically cookies. In practice online surveillance equates to the extraction, management, selling and ultimately, control of data. These algorithmic processes remain largely obscured from users of online platforms and digital services. Our access to the World Wide Web is mediated by screen devices and *Listening Back* enables users to go beyond the event on the screen and experience some of the algorithmic surveillance processes that underlie our Web experience. By directly intervening with the World Wide Web as a technological, social and political platform, the *Listening Back* browser add-on explores how sound can help us engage with complex phenomena beyond apparent materiality and therefore functions to highlight a disconnect between the graphical interface of the World Wide Web and the socio-political implications of background algorithmic processes of data capture. The Web browser as an interface marks the point where technology becomes apparent to the user. Built upon a distributed infrastructure of cables, servers, satellites and coded protocol that dates back to the invention of packet switching during the 1950s cold war era,[14] the Web browser is crucial as a site of engagement that conceals the algorithmic processes intrinsic to the functionality of the World Wide Web. Networked power manifests invisibly through algorithmic surveillance as monitoring technologies such as cookies are largely obscured from the user, making these systems difficult to approach, analyse and understand. Sonification is employed as a compositional tool to reveal asymmetrical relationships of power inherent to surveillance societies by enabling the opportunity to listen back to some of the imperceivable surveillance infrastructures that monitor and control our habitual online browsing. By providing a tangible sonic experience of real time surveillance data for examination, reflection and discussion this project asks, what is the potential of sonification as a means of addressing contemporary political questions?

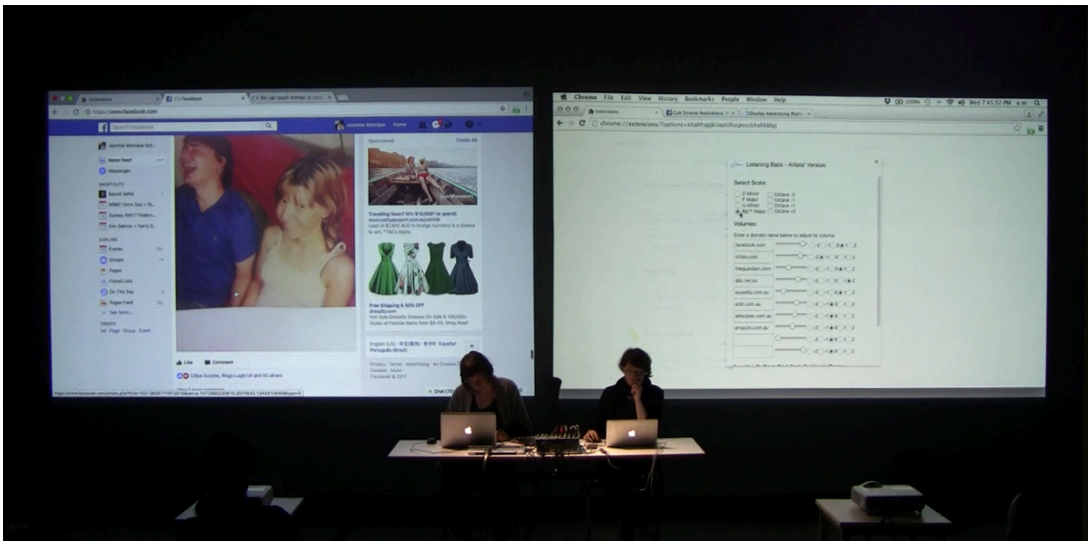


Figure 3: Live Performance with the Browser Duo at Black Box, UNSWAD, 2017

[12]Wolfe, K. 2014, "Sonification and the Mysticism of Negation." Organised Sound, 19(3), pp. 304–309. Doi:10.1017/S1355771814000296.

[13]Truax, B. 2012, "Sound, Listening and Place: The Aesthetic Dilemma", Organised Sound 17(3), p. 196.

[14]Galloway, A.R. 2004, *Protocol: How Control Exists after Decentralization*, Cambridge, Mass.: MIT Press, p. 5.

SONIFICATION AS ACTIVISM: A SPATIAL SONIFICATION OF SCHOOL SHOOTINGS SINCE COLUMBINE

Justin Kuhn

justintkuhn@gmail.com

ABSTRACT

Sonification possess the ability to engage a broad audience using online platforms. This work explores how data sonification can be used to accurately and effectively portray a political worriment for social activism purposes. Modeled as a compliment of the Washington Post's data visualization work on school shooting data, I use ambisonics to create a soundfield of gunshots that provide a time lapse of shootings in the United States. This work examines how individual data points can be placed into a soundfield using ambisonics for geographical, yet emotional effect. The hope is that audience members will feel motivated to act on gun safety activism within the United States, and have a full picture of the violence within this country. Link to 360 video: https://www.youtube.com/watch?v=78rFWVSDffo&list=PLGaph72tiQAka7czFEyzdiEtxN8_gVvNP&index=6&t=2s

1. DESCRIPTION

A *Spatial Sonification* explores how data sonification can be used to accurately and effectively portray a political worriment. School shootings are a crisis within the United States, with 1,300 school shooting incidents since 1970, with about 250 since the infamous Columbine shooting in 1999 (Campus Safety Magazine, 2019). The issue has still not been fully addressed or resolved, as the year 2018 had the greatest number of incidents since 1970, with 82 recorded incidents. After seeing the data visualization work on this topic by The Washington Post [1], I hoped to provide a compliment to this organization's efforts. Where visualization provides geometric and holistic information, sonification (and especially immersive sonification) engages its audience with an emotional experience, without sacrificing integrity.

I modeled this piece to highlight the affordances of sonification through an abstract, yet accurate presentation of data. As noted in the Sonification Handbook, sonic information is processed in a different way than visual information [2]. I aimed to utilize those advantages to put the listener into an experience that would be impactful to their position on the issue. One of the most shocking qualitative findings in the dataset was the frequency of assault rifles used by school shooters. I used a recording of a gunshot of each type of gun to represent a single shooting to accurately represent this. The location of the shooting was captured with spatial position of the sound in the soundfield. I wanted to treat each tragedy with equal weight, regardless of how many fell victim. In a single minute, I can bring the listener through the entire history of school shootings since Columbine.

A Spatial Sonification is different than other sonification pieces that I've seen in that it uses immersive technology for a political call-to-action. The Climate Symphony, led by directors Leah Borromeo and Catherine Round, and composed by Jamie Perera, sought to raise awareness of climate change through symphonic sonification of climate data [5]. Borromeo signifies the Climate Symphony's performance as a warning bell for the rapid imbalance of the earth's condition. In a different piece, Scott Lindroth, a computer music composer, addressed the news of Osama Bin Laden's death with a sonification of tweets, [3] which happened to be the first source of the news. *A Spatial Sonification* shares resemblance in that Lindroth's work keeps the timing of tweet sounds to the real-timing of when the tweets were sent. Nevertheless, *A Spatial Sonification* goes further than political kairos with the use of gunshots for political action, and it doesn't require hundreds of people and a large budget like the Climate Symphony to commission.

2. REFERENCES

[1] J. W. Cox, S. Rich, A. Chiu, J. Muyskens, and M. Ulmanu, (2019, April 8). "Analysis — More than 210,000 students have experienced gun violence at school since Columbine". Retrieved April 13, 2019, from https://www.washingtonpost.com/graphics/2018/local/school-shootings-database/?utm_term=.a11ab6974f4e.

[2] T. Hermann, A. Hunt, and J.G. Neuhoff, (2011). The Sonification Handbook. Retrieved March 23, 2019, from https://sonification.de/handbook/download/TheSonificationHandbook-HermannHuntNeuhoff-2011.pdf

[3] S. Lindroth (2012, July 27). Retrieved April 13, 2019, from https://www.youtube.com/watch?v=MUsBeJoBRxw

[4] T. Lossius and J. Anderson, (2014). ATK Reaper: The Ambisonic Toolkit as JSFX plugins. Proceedings of the joint 40th International Computer Music Conference & 11th Sound and Music Computing Conference, Athens, pp. 1338–1345.

[5] A. Simon-Lewis (2017, July 05). Climate change data is being transformed into beautiful, haunting symphonies. Retrieved April 13, 2019, from https://www.wired.co.uk/article/climate-symphony-data-sonification Staff, C. (2019, January 23).

[6] The K–12 School Shooting Statistics Everyone Should Know. Retrieved April 13, 2019, from https://www.campussafetymagazine.com/safety/k-12-school-shooting-statistics-everyone-should-know/

56FE

Falk Morawitz

Novars Research Centre, Manchester, UK
falk.morawitz@gmail.com

ABSTRACT

56Fe is a nine-minute acousmatic composition concerned with the recontextualization of everyday noise. The composition presents the sound of modern machinery such as trains or car engines and devolves them slowly into their archaic steam-driven counterparts. By presenting sounds removed from their visual source, the piece aims to de-normalize the noise of our daily lives. The composition uses sound recordings of steam vents, struck metal, and working machinery to invite the listener to ponder their relationship with their industrialised environment and to trace their relationship with technology from a different angle.

The piece can be reviewed under: https://soundcloud.com/falk-morawitz/56fe-2018

1. DATA SOURCES AND SONIFICATION METHODOLOGY

The piece is based on sound material related to iron and water, using recordings of struck metal and steam vents. Complementing this sound library are sonifications of iron and water orbital energy level data.

1.1. Water orbital data sonification

Water is a covalently bound molecule and the orbitals of oxygen and the two hydrogen atoms merge to create an array of new orbitals at discrete energy levels (Figs 1 and 2). These binding energies were translated to frequencies* and scaled. The scaled frequency values and their relative intensities were set to control a series of bandpass filter frequencies and their relative gains, respectively. Each filter was fed with white noise and the outputs of all filters were summed to form the sonification result.

1.2. Iron orbital data sonification

In metallic iron, orbitals mix just as in water molecules, but because iron metal consists of a high number of atoms, the energy gaps between individual orbitals of the same type become negligible and continuous energy bands are formed (Fig. 3), resulting in a broad set of electron energy distribution (Fig. 4).

[1]C.G. Ning and others, "High Resolution Electron Momentum Spectroscopy Of The Valence Orbitals Of Water", Chemical Physics, 343.1 (2008), 19–30.

[2]Ibid.

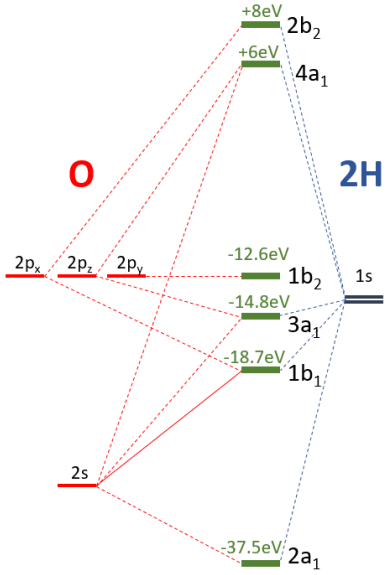


Figure 1: Orbitals of oxygen (red) and hydrogen (blue) re-combine when forming H_2O, leading to new hybrid orbitals (green). Orbital energies in electron volt (eV), computed by Ning et al.[1]

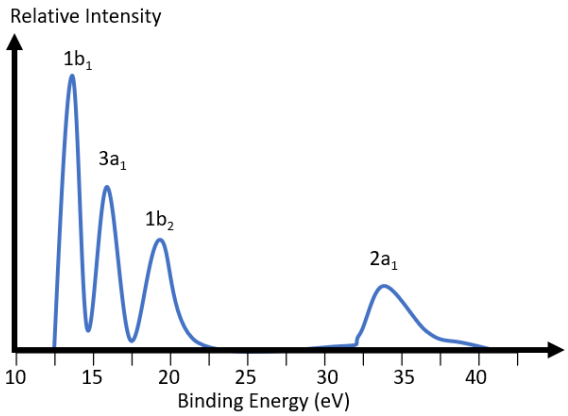


Figure 2: Ionisation spectrum of water measured at an impact energy level of 1200 eV, computed by Ning et al.[2]

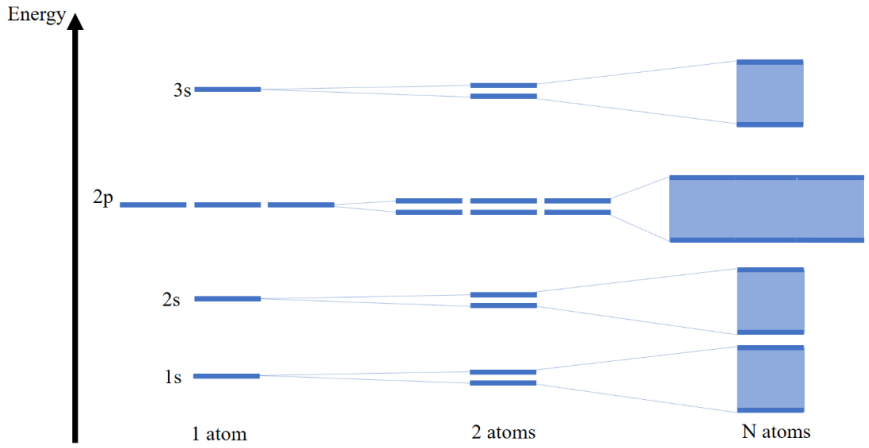


Figure 3: When two or more atoms interact, their orbitals will form new, mixed orbitals above and below the energy levels of the original orbitals. If enough atoms interact, the gaps in energy between these orbitals become negligible and energy bands are formed.

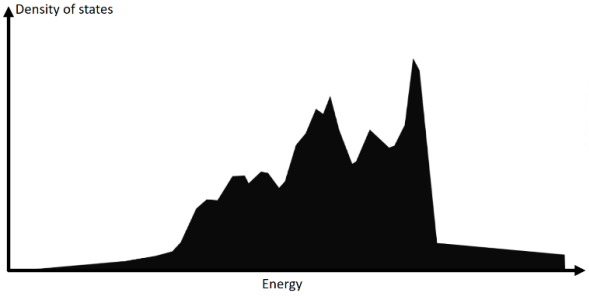


Figure 4: Energy distribution of electrons in metallic iron, data from Dronskowski[2].[3],[4]

The energy distribution data of electrons in metallic iron (figure 4) were sonified exactly like the orbital energy data of water, by scaling and assigning each energy peak to control the frequency of a series of bandpass filters. White noise was run through each filter separately and their outputs were summed according to their relative energy peak amplitudes.

1.3. Sound characteristics and sonification procedure

56Fe explores the idea of using reference sounds to help contextualise the sonification sounds. In addition to using real-world reference sound material (e.g. the sound of a passing train), *56Fe* employs auditory icons. Auditory icons mimic real-world sounds to encode data or represent processes, e.g. the sound of water filling a glass representing a copying process on a computer, or the sound of crumpling paper when moving computer files into an operating system's trash bin.[5]

In the sonification of water's and iron's electron energy distributions (Figs 2 and 4) the bandwidth (Q) of each bandpass filter can be adjusted, as well as its volume envelope. Narrow bandwidths and exponential decay envelopes result in bell-type sounds, whereas large bandwidths and linear decay will result in sounds similar to steam released from steam vents (Fig. 5).

2. STRUCTURE

Like other acousmatic pieces in the portfolio, *56Fe* explores the combination of different sets of sound material in each of its sections. The sound arrangement within a section and the arrangement of the music structure follow aesthetic and compositional concerns. No data-driven structure was imposed. Two main aspects that were explored in this piece are the placement of sonification sounds in space and in time, with sound material being crafted to sound far away or very close, and sections that either have highly dense sound material or contain almost no sounds at all.

0:00 – 1:50 minutes

The section starts with the introduction of sound motifs that will recur throughout the piece: a metallic bell made via sonification of water binding energy data, a drone based on the sonification of iron binding energy data (see Fig. 5), and the sound of passing trains. This section features iterations of mixed gestures in which the arrival and departure parts of the gestures are each based on one of the three sound motifs, but not necessarily the same one.

1:50 – 3:10 minutes

This part of the composition is an exploration of the positions of sounds in space, with the bell-type sound moving in and between panoramic, proximate, and distal composition space.

[3]"SFB 761 – Glossar", *Abinitio.Iehk.Rwth-Aachen.De* `http://abinitio.iehk.rwth-aachen.de/glossar/?text_id=109&division=Array&scale=Array` [Accessed 3 January 2019].

[4]Richard Dronskowski, *Computational Chemistry Of Solid State Materials* (Weinheim: Wiley-VCH, 2007).

[5]Eoin Brazil and Mikael Fernstrom, "Auditory Icons", in *The Sonification Handbook* (Berlin: Logos Verlag, 2011), pp. 325–338.

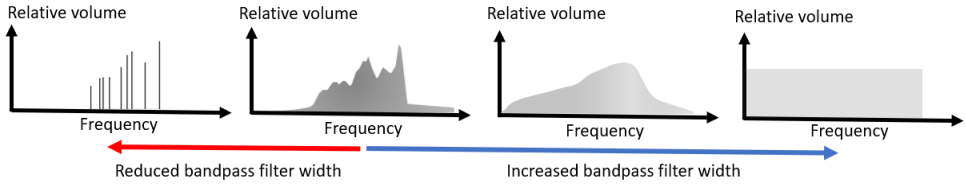


Figure 5: Spectrograms of sonifications of iron electron energy level data for various bandpass filter widths.

3:10 – 4:10 minutes

The gesture of the passing train is re-introduced again as departure or arrival halves of multiple mixed gestures. The gestural material in the first half of this section is combined with a low-frequency drone. The second half features contact microphone recordings of vibrating train tracks.

4:10 – 6:15 minutes

This section explores the fluidity between gestural and textural features of orbital energy level data of iron that is sonified and sculpted to resemble steam escaping from steam vents.

6:15 – 9:00 minutes

The last section of the piece reiterates and elaborates on previously introduced sound material. Sounds from the composition, such as the bell-type sound, steam vents, passing trains, and the sonification drone are explored via a variety of sound transformations including recursive effect loops.

PHOTONE - SONIFICATION CONCERT PROPOSAL

Niklas Rönnberg

Linköping University
Media and Information Technology
SE-581 83 Linköping, Sweden
niklas.ronnberg@liu.se

Jonas Löwgren

Linköping University
Media and Information Technology
SE-581 83 Linköping, Sweden
jonas.lowgren@liu.se

1. INTRODUCTION

Photone is an interactive installation combining color images with musical sonification. The musical expression is generated based on the syntactic (as opposed to semantic) features of an image as it is explored by the user's pointing device. The intention is to catalyze a holistic user experience that we refer to as modal synergy, where visual and auditory modalities multiply rather than add. In a scenic performance Photone is played by a performer who uses the features in the image, such as gradients, textures, different shades of color, and contrasts to perform a musical piece. The projector mirrors the image that is used, showing the mouse cursor and the interaction path as the performer brings the audience on a sonic journey with drone-like chords and harmonies, with different melodic movements, and rhythmic components.

In most cases where music is used as a complement to an image, the music is composed to images based on their denotative and connotative meaning; film music is one obvious example of this. However, our artistic intention in Photone is another: Music is not used as a complement, but by building the sonification upon pixel values of hue and brightness, that is, syntactic rather than semantic properties of an image, we aim to cut through conventional ways of seeing to a more foundational level.

2. GOALS AND AESTHETICS OF PHOTONE

The design and composition rationale for Photone was to compose the fundamental musical elements in a style inspired by electronic drone music, where the musical elements would be changed by the user's interaction and exploration. The initial goal was to explore an image with musical sounds, but we soon discovered that the image is also used to explore the musical sounds. The interaction with image and music has holistic qualities that combine into what we call modal synergy, creating an experience that is larger than its individual components.

The musical expression differs in terms of harmonics and melodies between an image's overall hue. The intention is to create different impressions of, for example, a whiter image compared to a more green image. Depending on the pixel value (i.e. the color) under the mouse cursor the musical expression varies, harmonies change, melodic movements change direction, and rhythmic instruments are attenuated or amplified. This, in a way, makes Photone into a musical instrument, but with predefined intervals, tonal scales, and tempo.

When a performer uses Photone to perform music for an audience, Photone is then like any other musical instrument performing a written musical score where the score (i.e. the image) is displayed to the audience. And, this interplay between musical elements and image pixel values transfers the unique qualities of Photone to the audience.

3. COMPOSITIONAL CHOICES

Based on psychology of colors the composition in Photone, as well as the synthesis of the sounds, are adapted to better fit, even if not mimic, the impression of the overall color in the image. A number of musical elements are adjusted according to the selected overall color (see Table 1). The harmony of the harmonic ambience is changed due to the color, where major chords are used for the warmer colors while minor chords are used for the colder colors. Furthermore, the complexity of the chords is chosen to represent the energy and the complexity in the colors. The dissonance of each tone, i.e. the spread in frequency of the pitches creating each tone used in the harmonic ambience, varies in relation to the impression of the colors. Colors with more energy have a greater dissonance, creating tones with more energy, while colors with less energy have more unison and relaxing tones. The timbre of the harmonic ambience and the melodic components is changed by altering the pulse-width of the square wave forms. Colors associated as more positive with more energy have pulse-widths creating more harmonics. All musical sounds pass through a low-pass filter and the cutoff frequency is adjusted according to the overall color, where the sonification for the more positive colors has more high frequency content while the less positive colors have their high frequencies attenuated. And finally, the tempo of the rhythmic composition is also generally faster and the rhythm is more complex for the colors with more energy.

The composition in Photone consists of seven musical elements (see Figure 1): 1) the overall harmonic ambience, 2) melodic components, 3) two low bass tones, 4) a high light intensity chord, 5) a bell-like sound for white, 6) a low frequency sweep for black, and 7) rhythmic instruments to emphasize contrasts.

The *harmonic ambience* consists of two-tone intervals multiplied over five octaves, creating a harmonic ambience with ten tones for each color channel (red, green, and blue). Depending on the pixel value the harmonic ambience varies from a two-tone interval to a complex chord. The harmonic ambience also becomes louder and with more high frequency content in bright areas in the image compared to darker areas.

Table 1: The table shows the colors used in Photone, the impression of the color, and the musical elements affected by these colors.

Color impression	Happiness & vitality	Energy & joyfulness	Passion & intensity	Purity & cleanliness	Creativity & loveliness	Nature & serenity	Calm & sadness
Harmony	Major ninth	Major minor seventh	Major	Suspended 2nd & 4th	Major major seventh	Minor	Minor minor seventh
Dissonance	+/- 30 cents	+/- 22.5 cents	+/- 15 cents	+/- 1 cent	+/- 10 cents	+/- 10 cents	+/- 2.5 cents
Timbre (PW)	80%	70%	65%	50%	60%	60%	55%
LPF cutoff	14kHz	12kHz	10kHz	8kHz	6kHz	4kHz	3kHz
Tempo (BPM)	98	96	94	96	94	92	90
Rhythm	Most complex and dense rhythmic pattern.	Slightly less complex and dense rhythmic pattern.	Less complex and dense rhythmic pattern.	Less complex and dense rhythmic pattern.	Not that complex and dense rhythmic pattern.	Not complex but sparse rhythmic pattern.	Least complex and dense rhythmic pattern.

The *melodic components* contain ten tones for each color channel, and these tones are played one tone at a time. The intensity in each color channel is divided into ten steps and one of the tones is used accordingly. This creates an upwards going melodic movement when intensity in that specific color channel increases, and a downwards going melody when intensity decreases. The melodic components also vary in amplitude and in band-pass filter cut-off frequency according to the intensity level.

The *bass tones* consist of two low pitched tones that are only present when the overall intensity level in the image is low (i.e. the image is dark), to emphasize the impression of darker colors. The *high light intensity chord* is composed with three tones and is only present when the overall light level is high to create an airy and high-intensity feeling. The short *bell-like sound* is used to further accentuate the dazzling intensity of white, and the *low frequency downwards sweeping sound* is used to emphasize the change in intensity from different shades of color to darkness. The *rhythmic instruments* are synthesized to mimic congas, triangle, and hi-hat sounds, and are used to rhythmically emphasize the amount of contrasts in the images.

4. SCENIC PERFORMANCE OF PHOTONE

Our goal with the concert is to present Photone for the ICAD audience, and demonstrate how a carefully planned path over an image can be used to create a harmonically pleasing and musically interesting musical piece. A video with a few examples of Photone can be found here: `dropbox-file`.

The video demonstrates how Photone can be explored and interacted with, rather than a carefully planned musical performance. The images used in the video examples comes from the public exhibition in the science center in connection to Linköping University that displayed and explained visualization and computer graphics to the visitiors. However, for the ICAD sonification concert we plan to use photos from the conference surroundings.

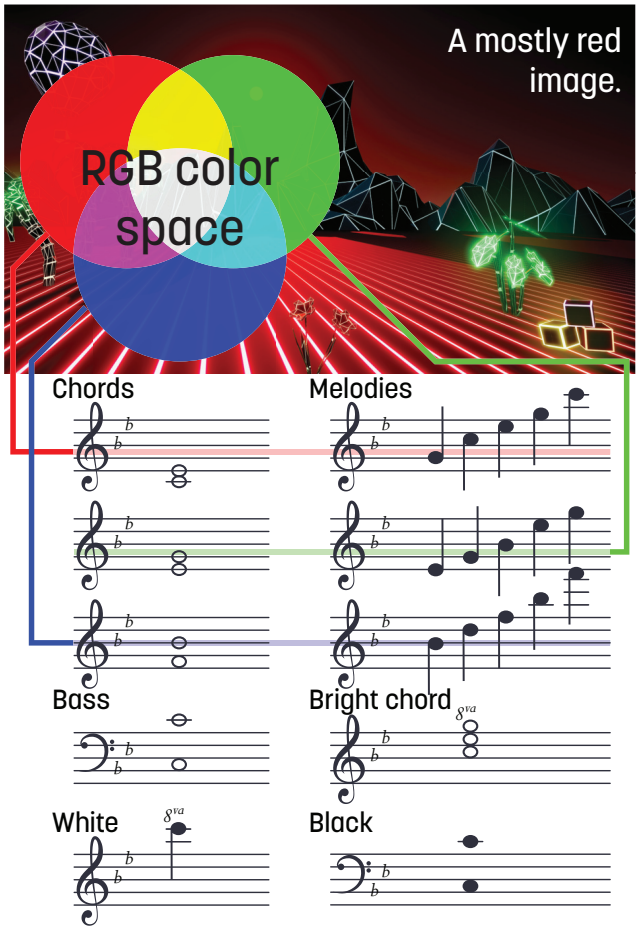


Figure 1: The values in the color channels build up the musical expression in Photone.

Installations

FOGETFULNESS

Ivica Bukvic, Zachary Duer, and Meaghan Dee

Virginia Tech
Blacksburg, VA, USA
{ico,zachduer,meaghand}@vt.edu

Forgetfulness is an interactive VR installations where multiple observers can navigate a mobius poem by Denise Duhamel that is projected onto a figure 8 mobius strip seemingly suspended in the middle of the virtual space (originally Virginia Tech Cube). Using wearable VR packs and the motion capture infrastructure such observers are free to navigate the space and study the shape and its content from multiple vantage points. Using abstract representations (avatars) to highlight their location within the virtual environment participants are also made aware of each other. The poem's words projected onto a virtual Mobius strip phase in and out of existence, reflecting the mental state of the poem's sole character, an alzheimer's patient who engages in a series of activities that without prior knowledge of their condition seem outright unusual. Despite strip's seemingly transparent appearance, the two sides of the strip contain different parts of the poem, thus subtly defying the limitations of the physical world. Specific keywords and thoughts are punctuated by the location-aware spatial soundscapes and events.

As observers navigate the virtual space and are experiencing the poem and supporting spatially-aware soundscape, they experience varying levels of stress and excitation. These are reinforced through the poem, as well as musical elements, some that offer self-standing locations where music can last up to several minutes without repeating. By wearing biofeedback wristbands that capture heart rate, heart rate variance, and skin conductance (level of skin sweat or galvanic skin response), the installation monitors for a relative change in observers' emotional states and projects them as virtual location-aware cloud-like trails. Such trails linger long after their owners have left the virtual environment, resulting in nebula-like structures around the poem with specific punchline spots offering denser clouds of colors indicative of a more dramatic shift in the observed biometric data. The ensuing visual hotspots of emotional shifts serve as beacons, attracting others to explore them in search of the poem's punchline, and thereby feed their luminescence.

Demo video available at: http://ico.bukvic.net/Video/ForgetfulnessVR.mp4

VISUAL ART SONIFICATION: COMBINING IMAGE AND DATA PROCESSING FOR ENHANCING ART APPRECIATION

Chihab Nadri, Chairunisa Anaya, Shan Yuan, Hongrui Hu, and Myounghoon Jeon

Mind Music Machine Lab
Virginia Polytechnic Institute and State University,
Department of Industrial and Systems Engineering,
1185 Perry Street Blacksburg, VA 24061 USA
{cnadri, danaya14, shany9, hongruih, myounghoonjeon}@vt.edu

1. INTRODUCTION

1.1. Goal

Much research has been conducted on sonification algorithms and image and data processing techniques. We aim to make use of all these advances simultaneously for sonifying visual artworks. Even though multiple image sonification algorithms and software programs have already been developed, only a few have been specifically made with the sonification of paintings and visual art in mind. Therefore, to make more appropriate painting sonification according to its art style and genre, we have used a JythonMusic platform to create multidimensional sonification algorithms able to make full use of both image and data processing methods.

1.2. Background

Machine learning has been used in the arts to classify artworks and identify artistic styles [1], [2]. Such algorithms can play an important role in streamlining classification tasks for galleries and art directories with large collections of artworks. On the other hand, machine learning can also play a role for the composition of music, adapting and learning from performers' musical genre and style to identify and create similar pieces [3]. Aesthetic research on visitor experience at art galleries [4] has revealed visitor patterns of short yet often repetitive viewing of the same paintings and artworks. Additionally, various sonification mapping strategies have been devised to transcribe physical quantities into sounds [5], providing us with a plethora of alternatives, as well as highlighting the continuous search for more appropriate designs. Coupled with advances in image and data processing techniques, a set of guidelines and approaches can be developed for the sonification of paintings and other visual works of art.

1.3. Motivation

Our sonification would apply to visual artworks from different art styles, with paintings from artists such as Monet. Our sonification would output different musical pieces depending on the artwork being used as a basis, taking into account parameters such as saliency and contrast.

1.4. Related Works

Research on visual graphics sonification has focused on providing individuals with visual impairments more ways and tools to experience visuals [6]–[8], and highlight useful data processing techniques used to accurately aid the transcription process, such as shape and edge detection machine learning algorithms. In fact, much research has been done concerning saliency detection and salient region cropping in images [9], [10], as saliency provides a good indicator for the importance and relevance of specific areas of the image. Research on photographic sonification has also yielded several sonification methods and algorithms, adhering to a musical approach focused on musical structure [11] or a naturalistic one following viewing tendencies [12]. Despite these earlier works, little has been done on the use of sonification for transcribing paintings and visual artworks specifically.

2. TECHNICAL DETAILS

2.1. Technical Details

Our program is an application to transcribe digital images of different artworks into MIDI files. We have used JythonMusic[13], a python-based environment for music creation that can also use Java libraries. Given the image provided, our software creates midi files that map image parameters to musical characteristics, creating the music piece.

2.2. Visual Artwork Sonification algorithm, Version 1

Our first design iteration included code for the transcription of image hue characteristics into musical parameters such as pitch and volume. Additionally, we made our first attempt at implementing visual saliency into our painting sonification algorithm, making use of Achanta et al's saliency and segmentation software [9]. Then, we worked on increasing the number of musical chords used, combining different instruments, adjusting pitch mappings, and exploring ways to improve the sonified output as well as to differentiate the

sounds between different art styles (e.g., Realism vs. Abstractionism).

3. CURRENT WORK

We are currently conducting interviews with experts in visual arts, music, and sonification to help indicate parameters to consider when sonifying the artworks. The development of new algorithms that can further enhance user experience following subsequent user studies will alter the final design we expect to demonstrate at the conference. The basic directions for future improvement include improvements in sound quality, cognitive and emotional meaning provided, and overall matching to the painting each musical track aims to complement.

4. INSTALLATION AND FUTURE PLAN

This installation would consist of our laptop to visualize and sonify the visual artworks. We will use our headsets for sonification. The audience will use our laptop to listen to the sonification of paintings from different art styles set up on small easels. For this installation, we will only need a small table for our laptop and easels. After this installation at ICAD, we plan on updating our design and sonification algorithm from the user feedback received.

5. REFERENCES

[1] Y. Bar, N. Levy, and L. Wolf, "Classification of artistic styles using binarized features derived from a deep neural network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8925, pp. 71–84, 2015.

[2] B. Saleh and A. Elgammal, "Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature," 2015.

[3] R. De Prisco, D. Malandrino, G. Zaccagnino, R. Zaccagnino, and R. Zizza, "A Kind of Bio-inspired Learning of mUsic stylE," in *Computational Intelligence in Music, Sound, Art and Design*, 2017, pp. 97–113.

[4] C. C. Carbon, "Art perception in the museum: How we spend time and space in art exhibitions," *Iperception.*, vol. 8, no. 1, 2017.

[5] G. Dubus and R. Bresin, "A systematic review of mapping strategies for the sonification of physical quantities," *PLoS One*, vol. 8, no. 12, 2013.

[6] S. Cavaco, J. T. Henriques, M. Mengucci, N. Correia, and F. Medeiros, "Color Sonification for the Visually Impaired," *Procedia Technol.*, vol. 9, pp. 1048–1057, 2013.

[7] T. Wörtwein, B. Schauerte, K. Müller, and R. Stiefelhagen, "Mobile Interactive Image Sonification for the Blind," in *Computers Helping People with Special Needs*, 2016, pp. 212–219.

[8] M. Jeon, R. J. Winton, J.-B. Yim, C. M. Bruce, and B. N. Walker, "Aquarium fugue: interactive sonification for children and visually impaired audience in informal learning environments," *Proc. 18th Int. Conf. Audit. Disp. (ICAD 2012)*, pp. 246–247, 2012.

[9] R. Achantay, S. Hemamiz, F. Estraday, and S. Susstrunky, "Frequency-tuned salient region detection," *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, no. Ic, p. 1597 { 1604, 2009.

[10] L. Zhang, Y. Gao, R. Ji, Y. Xia, Q. Dai, and X. Li, "Actively learning human gaze shifting paths for semantics-aware photo cropping," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2235–2245, 2014.

[11] N. Rönnberg and J. Löwgren, "Photone: Exploring Modal Synergy in Photographic Images and Music," no. Icad, pp. 73–79, 2018.

[12] A. Polo and X. Sevillano, "Musical Vision: an interactive bio-inspired sonification tool to convert images into music," *J. Multimodal User Interfaces*, no. i, 2018.

[13] B. Manaris, B. Stevens, and A. R. Brown, "JythonMusic: An environment for teaching algorithmic music composition, dynamic coding and musical performativity," *J. Music. Technol. Educ.*, vol. 9, no. 1, pp. 33–56, 2016.

Index of Authors

ICAD2019

Sonification for Everyday Life

ISBN: 0-9670904-6-6

@ICAD2019 #ICAD2019

https://icad2019.icad.org