Towards Sonification in Multimodal and User-friendly Explainable Artificial Intelligence

Björn W. Schuller[*]
Chair of Embedded Intelligence for Health Care & Wellbeing,
University of Augsburg
Augsburg, Germany

Tuomas Virtanen
Audio Research Group, Tampere University
Tampere, Finland

Maria Riveiro
Department of Computing, Jönköping University
Jönköping, Sweden

Georgios Rizos
GLAM – the Group on Language, Audio, & Music,
Imperial College London
London, UK

Jing Han
Department of Computer Science and Technology,
University of Cambridge
Cambridge, UK

Annamaria Mesaros
Audio Research Group, Tampere University
Tampere, Finland

Konstantinos Drossos
Audio Research Group, Tampere University
Tampere, Finland

ABSTRACT

We are largely used to hearing explanations. For example, if someone thinks you are sad today, they might reply to your "why?" with "because you were so Hmmmmm-mmm-mmm". Today's Artificial Intelligence (AI), however, is – if at all – largely providing explanations of decisions in a visual or textual manner. While such approaches are good for communication via visual media such as in research papers or screens of intelligent devices, they may not always be the best way to explain; especially when the end user is not an expert. In particular, when the AI's task is about Audio Intelligence, visual explanations appear less intuitive than audible, sonified ones. Sonification has also great potential for explainable AI (XAI) in systems that deal with non-audio data – for example, because it does not require visual contact or active attention of a user. Hence, sonified explanations of AI decisions face a challenging, yet highly promising and pioneering task. That involves incorporating innovative XAI algorithms to allow pointing back at the learning data responsible for decisions made by an AI, and to include decomposition of the data to identify salient aspects. It further aims to identify the components of the preprocessing, feature representation, and learnt attention patterns that are responsible for the decisions. Finally, it targets decision-making at the model-level, to provide a holistic explanation of the chain of processing in typical pattern recognition problems from end-to-end. Sonified AI explanations will need to unite methods for sonification of the identified aspects that benefit decisions, decomposition and recomposition of audio to sonify which parts in the audio were responsible for the decision, and rendering attention patterns and salient feature representations audible. Benchmarking sonified XAI is challenging, as it will require a comparison against a backdrop of existing, state-of-the-art visual and textual alternatives, as well as synergistic complementation of all modalities in user evaluations. Sonified AI explanations will need to target different user groups to allow personalisation of the sonification experience for different user needs, to lead to a major breakthrough in comprehensibility of AI via hearing how decisions are made, hence supporting tomorrow's humane AI's trustability. Here, we introduce and motivate the general idea, and provide accompanying considerations including milestones of realisation of sonifed XAI and foreseeable risks.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Model development and analysis*; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Explainable artificial intelligence, sonification, human computer interaction, multimodality, trustworthy artificial intelligence

ACM Reference Format:
Björn W. Schuller, Tuomas Virtanen, Maria Riveiro, Georgios Rizos, Jing Han, Annamaria Mesaros, and Konstantinos Drossos. 2021. Towards Sonification in Multimodal and User-friendly Explainable Artificial Intelligence. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3462244.3479879

[*] Corresponding author: schuller@ieee.org. The author is further affiliated with GLAM – the Group on Language, Audio, & Music, Imperial College London, London, UK.

ICMI '21, October 18–22, 2021, Montréal, QC, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479879

1 INTRODUCTION

In the future, AI systems will be ubiquitous in our everyday life, for example in self-driving cars, robots, and intelligent home assistants. They will be constantly making critical decisions that have a severe impact on many aspects of human lives. The behaviour of such systems needs to be explainable for non-expert users, and sonification will be a powerful tool for this purpose. In comparison to visualisations, which are the standard way for explaining AI, sound has several clear benefits: for example, sonification will allow grasping the attention of a user, for example the passenger of an autonomous car, or presenting information to a user whose vision is focused on another task. Furthermore, humans are highly capable of listening to and interpreting complex polyphonic signals that involve different rhythms, harmonies, etc., allowing the usage of sonification to present complex sequential data. Sonified explanations may play a crucial role supporting visually impaired people, children with disabilities, the elderly, drivers in autonomous vehicles, and medical personnel, e. g., during complex surgeries.

The number of methods for explaining AI black-box models has boomed in recent years; even if many challenges remain, multiple solutions have been proposed in the literature [9]. Several of these approaches rely on identifying the group of input data points that affect or alter the decisions of the utilised classifier or regressor. There are gradient-based approaches, where the explanations of the decisions are given based on the effect that each input data point has to the gradient [21], or approaches that examine the performance of the machine learning model by substituting different groups of input data points with noise [3]. The typical output of methods that explain the decisions of a model is a salience map that identifies the salient input data points. For example, in an image processing task, a salience map would identify data point regions of the input image that affect a particular decision made by the classifier; however, this approach conveys limited information in a limited manner.

Such salience maps have a clear meaning for 2-D data such as images, but are not an intuitive way to explain decisions on multidimensional data (e. g., banking transactions, or medical signals from multiple sensors). Additionally, a visual cue requires users to focus towards the illustration of the explanation, making it impractical in situations where visual attention of the user has to be focused on crucial matters (e. g., textual explanations when driving a vehicle [15]). Furthermore, current approaches consider mostly static information, while in many cases, decision making is a dynamic process; tasks like speech emotion recognition, sound event detection, dialogue systems, or source separation are time-evolving, sequence processing tasks. Currently, there are no XAI approaches tailored for audio data: while existing approaches indicate data points that affect the decisions in image processing tasks, they fail to identify entities that affected that decision, being unable to explain the causal relationship between input and output. An overview of a potential sonified explanation system is depicted in Figure 1.

1.1 Vision and ambition

Sonified explanations will focus on providing intelligent algorithms that will advance the state-of-the-art in XAI. The particularities of the audio modality should be exploited in order to identify factors in the input audio data that explain the decision made by the AI, which will then be sonified. Further opportunities lie in the sonification of data other than audio (e. g., medical, images, multimodal) – either standalone, or in a multimodal manner alongside visual or other ways of explanation aiming at maximal informativeness and usability. Furthermore, in any attempt at evaluation, sonifications of AI should be considered with respect to different types of users (highlighted as an open challenge in XAI for Natural Language Processing (NLP) [6]), e. g., represented by gender-balanced user groups, young individuals and elderly age groups, and high and low tech-affinity. The objectives of sonified explanations are:

- Intuitive and trustworthy explainability by sonification in Audio Intelligence, as well as general AI tasks.
- Personalised sonified explanations, by taking into account diversity of demographics, tech-affinity, and AI expertise.
- Benchmarking sonified explanations against visual and textual state-of-the-art alternatives.
- Best multimodal embedding of sonified explanations and combination with explanations given in other, non-audio ways, where going beyond a mono-modal approach appears more informative, usable, and efficient.

2 HOW TO USE XAI WITH AUDIO?

One goal of sonified explanations is to produce audio-based, human-like artificial explanations for decisions of Machine Learning (ML) models. The backward path to produce such explanations highly depends on the data itself. Hence, it is important to investigate innovative explainable methodologies, such as pointing back at which data drive the decisions, retrieving similar audible samples from the training set, and generating more appealing samples with emphasis on particular patterns via generative models (e. g., Generative Adversarial Networks (GANs) [8] and Variational Auto-Encoders (VAEs) [16]). In particular, by deploying generative models, unlimited realistic and salient audio data can be generated towards more fitting explanations and justifications. This opens pathways to personalised explanations that can be more understandable to diverse non-technical audiences such as end-users and other stakeholders. Also note that, since the above-mentioned example-based explanation approaches focus on the data itself, sonified explanations can also be applied in a model- and task-agnostic manner.

Existing advanced example-based interpretability methods can mainly be divided into two groups [1]: a) Prototypes and criticisms, and b) Counterfactual explanations. In the first group, representative instances are normally selected either to represent each category as prototypes, or outliers are selected as critics. In the sonified explanations domain, algorithms need to be developed to select/generate audible instances of such prototypes and critics, aiming at characterising the dataset and explaining corresponding decisions as a whole. In counterfactual explanations (as in [10] for textual explanations), however, instead of explaining explicitly how a decision is derived, adversarial examples are computed to explain implicitly what minimum conditions can lead to other decisions – an approach that can be adopted for the case of audio.

2.1 Source separation for sonified explanations

Realistic data used as an input to AI systems are often composed of multiple factors that interact together. In audio processing, source

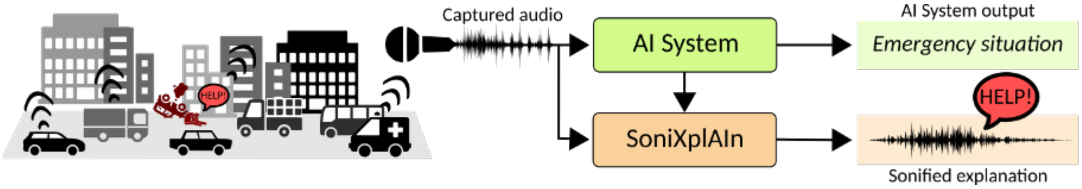

Figure 1: Sonic explanation will explain AI by sonifying data factors that drive its decision. Here, we illustrate an example of a sonifying explanation module that provides explanations for an audio modelling AI decision-making system. That being said, sonified explanations can be provided for non-audio modalities, or as complementary to multimodal XAI systems.

separation is a thoroughly studied field, with significant results for speech and speaker separation [11, 17] or music source separation [7]. We believe that source separation methods for XAI are needed, that are able to identify the factors in the data that are important in the decision making, to separate and enhance them for sonification. Potential solutions can build upon the state-of-the-art source separation methods, e. g., deep clustering [11], auto-encoders [7], or non-negative matrix factorisation [24], such that any factor that is involved in the decision making of an AI system can be isolated.

2.2 Rethinking evaluation of XAI

Incorporating XAI evaluation metrics as a standard in ML research has been proposed recently, in the context of XAI for Natural Language Processing (NLP) [6]. That being said, the evaluation of explainable AI is a research challenge on its own. Despite rapid advances in AI that we are experiencing in multiple application areas, less progress is seen in how to evaluate users interacting with AI-systems [20]. Conventional evaluations of humans interacting with such systems are carried out using traditional methods and metrics either from the machine learning community (algorithm-centred evaluation) or the Human-Computer-Interaction (HCI) community (human-centred evaluation), as indicated in related surveys [12, 19]. This has led to a paucity of holistic and integrative methods that assess the overall collaboration over particular components [2, 23].

ML often uses performance evaluation metrics that focus on the algorithms employed, such as accuracy, precision, recall, squared error, posterior probability, information gain, etc. In the meantime, interactive machine learning builds on these metrics by typically combining them with some form of usability assessments (see e. g., [5, 22]), i. e., the extent to which a product or system can be used by specified users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specified context of use. In turn, usability evaluations can be categorised in exploratory, formative and summative evaluations. Exploratory evaluations assess the current usage of a system, and, typically, use interviews, observations, surveys and logging. Formative evaluations help improve the system during the design process through heuristics and thinking aloud methods. Summative evaluation assesses the overall quality of a system once it is more or less finished, by collecting bottom-line data and quantitative measurements of performance, e. g., how long did users take, how many errors did they make, were they successful, number of commands/features used, etc. To capture and evaluate the interactions between the users of sonified explanations, case studies and the interfaces developed during an according sonification for XAI process, we need to look not only into traditional HCI and ML evaluation methods and metrics, but also include theoretical principles from cognitive and social sciences that account for human preconceptions about systems' inner workings and behaviour, which can also explain other expectations, fears, and trust issues towards AI-systems. To consider insights from social sciences is a current trend in AI evaluation [18] that applies here as well.

Evaluating explanatory sonifications would focus on assessing sound explanations that either complement visual, textual, or other ones or are the only ones provided to the user. Principles from Theory of Mind can be investigated in order to measure user mental models [14] of the inner workings and behaviour of AI-systems; a connection between those theories with principles from psychoacoustics (scientific study of sound perception) is desirable.

3 WHY RESEARCH SONIFIED XAI?

We believe that sonifying explanations is expected to have a profound impact on the advancement of XAI:

- **sonifying explanations**: Development of technology for integrating explainability into black-box models based on diverse sonification approaches. This potential innovation is largely different from other existing works such as natural language explanations and various visualisation techniques. It refers to explaining the data as well as the decision processes via sound, as a person has a better ability to perceive and understand massive information audibly than visually. If attained, this technology could provide an efficient and understandable audio-based XAI system; furthermore, it could provide comprehensive explanations when combined with visual and text explanations, leading towards more reliable and trustworthy multimodal XAI systems.
- **user-centric XAI**: development of technology that incorporates users in the explanation loop, to generate human-like and user-friendly explanations. This would enhance the usability and efficacy of XAI systems for the stakeholders through personalised and user-centric interfaces.
- **identification and elimination of dataset biases**: application of sonifying XAI on stratifying biases imposed onto datasets, which affect AI decisions in a non-inclusive way for under-represented cases. While using visual stimuli to identify such biases might require strong visual indications (e. g., colour, or significant difference in physiology), employing audio is likely to result in more easily and intuitively noticeable patterns that clearly indicate biases in the dataset. Furthermore, sonification tools can be used to develop strategies that will promote inclusiveness in the data collection.

- **evaluating explanations**: production of appropriate evaluation benchmarks and metrics to compare, validate, quantify and evaluate the explainability of sonification XAI methods in general, as well as for evaluating the effectiveness of audible explanations compared to visual and textual ones.

4 SONIFIED XAI MILESTONES

We envision methodological milestones and other great opportunities that should be covered in this research direction.

4.1 XAI core

Advanced XAI techniques need to be developed and utilised to mitigate the black-box nature of deep models and thus make their decisions more traceable, transparent, and trustworthy. For that, a holistic XAI approach should be explored on the full AI pipeline, covering explanations of sample-based v.s. feature-based manner, global v.s. local explanations, and counterfactual decision explanation v.s. querying internal state, with emphasis on audio-specific data algorithms. It is important to identify which algorithms and audible explanations are more appropriate for which end-user cases.

4.2 Sonification

Recent studies have proposed methods for sonifying visual input, targeted towards people with visual impairment [4, 13]. However, extended development of novel algorithms and techniques is required, to intuitively communicate the explanation and reasoning of the decisions using audio as a communication channel. Different types of information could be leveraged, ranging from raw input data (e. g., audio) to multi-dimensional tensors (e. g., gradients, weights, and learning signals).

4.3 Personalisation

A great challenge lies on defining the concrete requirements of particular end users (speech vs general audio, children vs elderly); specifying the modalities and types of explanation required in each case, taking into account sound requirements and needs.

4.4 Evaluation

The aim of this module is to assess the proof-of-concept prototypes developed during sonified explanations. For that, we need to select or develop evaluation methods and metrics that support our users interacting with intelligent systems. For instance, it would be necessary to analyse what objective and/or subjective measures should be developed to assess the results in terms of design guidelines for the use of sound coupled to explanations from AI-systems, particularly for complex human-system interactions.

4.5 Other possibilities

There is potential value in moving away from bespoke XAI, towards an XAI module that extracts explanations from an internal AI core module for sonification, in a model-agnostic manner.

Apart from sonifying salient audio factors, there is value in explanations in realistic speech, using advanced language generation and speech synthesis techniques, potentially including socioemotional competency. With this approach, audible explanations can be an alternative option for end-users with visual impairment or provide complementary information when incorporated with visible/textual explanations, for both audio and non-audio data applications, e. g., related to relate to biomedical signals or capturing motion activity.

5 FORESEEABLE RISKS

Researching sonified explanations would advance the state of the art on XAI by adding audio as a communication channel. As such, it can significantly extend current visual and textual means to provide explanations, reaching out to users using AI-systems where reading text or processing visuals are either not recommended or simply not preferred: for instance, surgeons carrying out complex tasks. It also reaches out to groups that can more easily process audio and sound, like children, the elderly, or visually impaired people, etc. However, the use of sound for providing explanations is totally unexplored in XAI, and there is a risk that it is challenging to find optimal or effective combinations of these modalities for a different variety of end-users. Therefore, we believe that the innovation potential of sonified explanations is very high. Particularly, a lot of efforts are required to address the named and further risks, including:

- design a case study and experimental set up for sonified XAI's proof-of-concept is a novel problem.
- what kind of data are required for such a given study case? There are no data available at hand to work with for this purpose. What to do to maintain the data security and privacy?
- evaluating the effectiveness of an XAI sonification solution is a novel problem, with no existing benchmark for comparing sonification with, e. g., a visual explanation method.
- how to expand the findings in the case study or multiple studies to a wider range of users and be applied at scale?

6 CONCLUSION

We introduced the concept of Sonified XAI, which aims to provide sonification solutions in order to satisfy the 'right to an explanation' for AI-generated decisions of end-users of a gamut of social and tech-affinity backgrounds. This ambitious, from the ground-up concept of sonification algorithmic methodology, is not only a much under-explored explanation avenue compared to explanation vectors such as visualisation and text, but an altogether crucial one in applications where visual, textual, or explanations given in any other modality or combination thereof are not an option, as well as holding promise if utilised along with the more established approaches in existing applications. Specifically, we believe that the entire AI pipeline should be addressed, and model-agnostic XAI sonification approaches explored, in order to design transparent, reproducible XAI methodologies that foster trust in AI. A crucial issue will be the careful evaluation thereof, taking into account state-of-the-art approaches to XAI, as well as specific end-user cases for which sonification based XAI holds promise. Further, we look forward to multimodal explanations in AI in a personalised manner for best individual explanation provision in tomorrow's AI.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton. 2018. Evaluation of interactive machine learning systems. In *Human and Machine Learning*. Springer, 341–360.

[3] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2018. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024* (2018).

[4] Angela Constantinescu, Karin Müller, Monica Haurilet, Vanessa Petrausch, and Rainer Stiefelhagen. 2020. Bring the Environment to Life: A Sonification Module for People with Visual Impairments to Improve Situation Awareness. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 50–59.

[5] Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Vittorio Papadopoulos, and Roberto Turrin. 2011. Looking for "good" recommendations: A comparative evaluation of recommender systems. In *IFIP Conference on Human-Computer Interaction*. Springer, 152–168.

[6] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 447–459.

[7] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. 2019. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174* (2019).

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[10] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating Counterfactual Explanations with Natural Language. In *ICML Workshop on Human Interpretability in Machine Learning*. 95–98.

[11] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 31–35.

[12] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[13] Zihao Ji, Weijian Hu, Ze Wang, Kailun Yang, and Kaiwei Wang. 2021. Seeing through Events: Real-Time Moving Object Sonification for Visually Impaired People Using Event-Based Camera. *Sensors* 21, 10 (2021), 3558.

[14] Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness.* Number 6. Harvard University Press.

[15] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*. 563–578.

[16] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 27, 8 (2019), 1256–1266.

[18] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[19] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv preprint arXiv:1811.11839* (2018).

[20] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592* (2019).

[21] Badri N Patro, Mayank Lunayach, Shivansh Patel, and Vinay P Namboodiri. 2019. U-cam: Visual explanation using uncertainty based class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7444–7453.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[23] Serge Thill and Maria Riveiro. 2019. Memento hominibus: on the fundamental role of end users in real-world interactions with neuromorphic systems. In *Workshop on Robust Artificial Intelligence for Neurorobotics (RAI-NR) 2019, 26–28 August, University of Edinburgh, United Kingdom*.

[24] Tuomas Virtanen. 2007. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing* 15, 3 (2007), 1066–1074.