ICA 2016 BUENOS AIRES

Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Soundscape, Psychoacoustics and Urban Environment: Paper ICA2016-791

A novel auditory saliency prediction model based on spectrotemporal modulations

Karlo Filipan[a], **Annelies Bockstael**[b], **Bert De Coensel**[c], **Marc Schönwiesner**[d], **Dick Botteldooren**[e]

[a]Ghent University, Belgium, karlo.filipan@intec.ugent.be
[b]Ghent University, Belgium, annelies.bockstael@intec.ugent.be
[c]Ghent University, Belgium, bert.decoensel@intec.ugent.be
[d]Université de Montréal, Canada, marc.schoenwiesner@umontreal.ca
[e]Ghent University, Belgium, dick.botteldooren@intec.ugent.be

Abstract

Previous studies indicate that soundscape perception and appraisal are influenced by the sounds that people hear and pay attention to. Hence, a model that evaluates instantaneous human attention to environmental sounds would be very useful in soundscape research. Attention is triggered by the saliency of a sound within its context. Therefore, we propose a model for predicting saliency of sounds based on dynamic modulation ripples – simultaneous modulations in the frequency and time domain. These ripples exhibit direct response in the auditory cortex of the human brain. Our model contains three stages. In the first stage, the incoming sound signal is demodulated similarly to the early stages of auditory processing, and afterwards it is correlated with each of the modulation functions of the ripples. The obtained ripple features enable the model to detect salient changes that are not accompanied by changes in more commonly used spectrogram features. We demonstrate this by comparing the model output for sound signals with the same amplitude but randomized phase spectrum. The second stage of the model integrates ripple features over time to simulate excitation and inhibition processes happening along neural pathways. In the final stage, spectral saliency is aggregated to an overall saliency using supervised training on sound environments with embedded salient sounds. We evaluate the model with a collection of natural sound fragments previously used in an EEG experiment on attention and illustrate its application in complex environmental sound scenes.

Keywords: attention, saliency, modeling, ripples

A novel auditory saliency prediction model based on spectrotemporal modulations

1 Introduction

Soundscape, as defined in [1] investigates how people perceive or experience and understand an acoustic environment in context. Therefore, when evaluating soundscape of a specific sonic environment, human perception and its characteristics have to be considered. Since the sound of a space is seldom the reason of being there, human perception is generally formed by attention to specific sounds sources found in the environment [2].

In order to separate the specific sources, people listening to the sound perform an auditory stream analysis [3]. This analysis is formed by two coexisting processes – stream segregation and grouping. Segregation enables time and frequency separation and is established by the physical characteristics of human hearing. On the other hand, higher cognitive stages are responsible for grouping the segments into streams representing sound sources.

Although attention partly influences the stream formation [5], its main role is in enabling the listener to focus on a single sound source from multitude of processed events. Auditory attention itself is shaped by two interplaying processes – bottom-up and top-down attention. While top-down is influenced by person's voluntary needs, bottom-up attention is shaped by the saliency of the sounds.

Sounds that are salient exhibit characteristic features that distinguish them from other sounds and enable them to stand out from the background. However, saliency features and corresponding models are not apparent, therefore different saliency predictions have been proposed [5, 6, 7, 8, 9].

In this contribution, we outline the new brain-response-inspired features based on ripple sounds (spectrotemporal modulations). To compare such ripple features to the commonly used features based on spectrograms, feature space representation from sounds with the same amplitude but scrambled phase are shown. Furthermore, we present a layout of the new auditory saliency model. Environmental sound fragments are used to illustrate the capability of the model to correctly predict saliency of events in complex sonic environments.

2 Spectrotemporal modulations features and their strength

2.1 Features based on spectrotemporal modulations

Saliency of a sound can be calculated as a weighted sum of the excitation strength of properly selected features. In turn, such features should account for the response of the auditory system to the relevant characteristics of the sound.

Some saliency models [4, 6, 7] use features that are constructed on the basis of sound spectrogram. Other models, however, explicitly include known contributors to saliency [5]. In our proposed model, we include features based on spectrotemporal modulations that were observed to relate well to spatially located brain response [10].

The model's feature extraction starts by filtering the input audio signal in N_f narrow-band filters with central frequency $f_{c,i}$. The selected frequencies are logarithmically spaced over the complete auditory frequency range (10 octaves) and separated in $1/x$ (x being either 3, 6 or 12) octave bands. The output signals of each of these filters are subsequently demodulated by squaring and low-pass filtering of the result with a cut-off frequency of 20 Hz.

The demodulated signal is finally compared to a set of N_r orthonormal ripple basis functions:

$$S(t,x) = 1 + \Delta m \cdot \sin(\omega t + \Omega x) \tag{1}$$

where ω is the amplitude modulation in rad/s, Ω is the frequency modulation in rad/oct, while Δm represents a modulation depth. Amplitude modulation frequencies are logarithmically spaced between 0.1 and 10 Hz and -10 to -0.1 Hz. Correspondingly, frequency modulation rates are linearly spaced in the range of 0.1 and 10 cyc/oct.

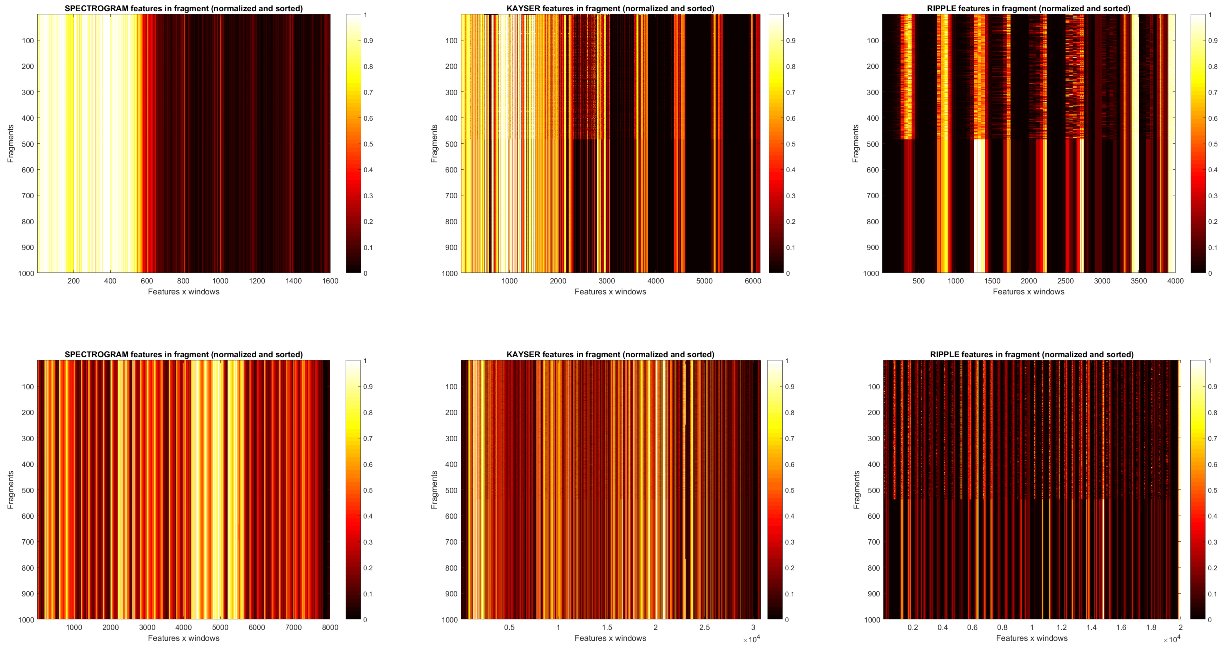
Comparing the demodulated signal with the prototype ripples is done by cross correlation. The resulting signals are summed over the filter bands and maximum values over a sliding window of one octave are used as an output. Features are determined in the overlapping time windows of 10 s (corresponding to the lowest amplitude modulation rate) and with 50% time overlap.

2.2 Predictive strength of spectrotemporal modulation features

To explore the strength of different feature extractors, an inconspicuous form of saliency should be investigated. A well known sound (e.g. music) was transformed with Fast Fourier transform algorithm and its phase scrambled. After inverse transform, the sound was distinctively different from the original for the human listener and phase-scrambled epochs were marked as salient. However, the power spectrum was unchanged and thus the spectrogram stayed exactly the same for both the randomized phase and the original sound. Hence, any features attempting to detect salience starting from a spectrogram extraction will be unsuccessful.

In Figure 1 the results from different feature extractors are shown. Multiple repeated original and modified fragments of two sounds – car honk and guitar music – are used. To obtain a complete comparability of the feature extractions, all of the used procedures were done on the same time windows while modified and original fragments were randomly selected as inputs.

The upper half of each graph in Figure 1 represents phase-scrambled fragments while in the lower part features from original fragments are shown. As it can be seen, the spectrotemporal modulation features allow to clearly distinguish between the fragments while the spectrogram does not show any difference. On the other hand, Kayser features [4] only partly allow to distinguish both groups for the strongly transient car honk sound.



Source: (Author, 2016)

Figure 1: **Comparison of different feature extractors on original and random phase repetitions of two separate sound recording fragments. Sound fragments used (respectively top and bottom row): car honk, guitar music. Feature space outputs (columns from left to right): spectrogram, Kayser features, ripple features.**

3 The saliency model

Based on demodulation, $N_r \times N_f$ features are extracted at each time interval. As these features represent the sensitivities of the human auditory system, the strength of their excitation is proportional to the saliency of the sound. Additionally, our model adds the feature signals integration using different rise and fall time constants representing excitation and activation in neural pathways. For extracting the information that is relevant for soundscape – namely the likelihood of a sound to be noticed by the user of a space within a complex sound environment – a single number indicator is preferred. Therefore, the integration responses are summed ending up with a single number saliency measure.

Until now, the model does not account for the frequency dependent response of the human ear. Detailed and complex models for middle and inner ear dynamics are available that could be used to weigh the input signal. However, for simplicity a straight forward A-weighting followed by summation over frequencies is implemented. Additionally, for weighing the importance of different spectrotemporal modulations, the reported thresholds by Chi et al. [11] are used.

4 Extracting saliency of environmental sounds

To demonstrate the performance of the new saliency model on environmental sounds, two five minute sound fragments containing salient sounds were evaluated. The first one consisted of a highway noise recording on which a sound of passing emergency vehicle with a siren was added five times. The second sound was a five minute recording of urban traffic noise in which one of the vehicles used its horn during four pass byes. In addition to analyzing these environmental sounds on their own, the voice of a person talking in the foreground was mixed by using two different speech fragments.

Figure 2 shows the output of the saliency model and compares it to the sound level and the labeled parts of the salient events. For highway sound, saliency peaks at the instances where an emergency vehicle passage was included. However, other spikes emerge due to the specific trucks and motorcycles that pass by. Once speech is added, the peaks at the locations where the emergency vehicle was inserted seem slightly more pronounced. In comparison, no such distinctions are visible in the simple signal levels at the times of salient events.

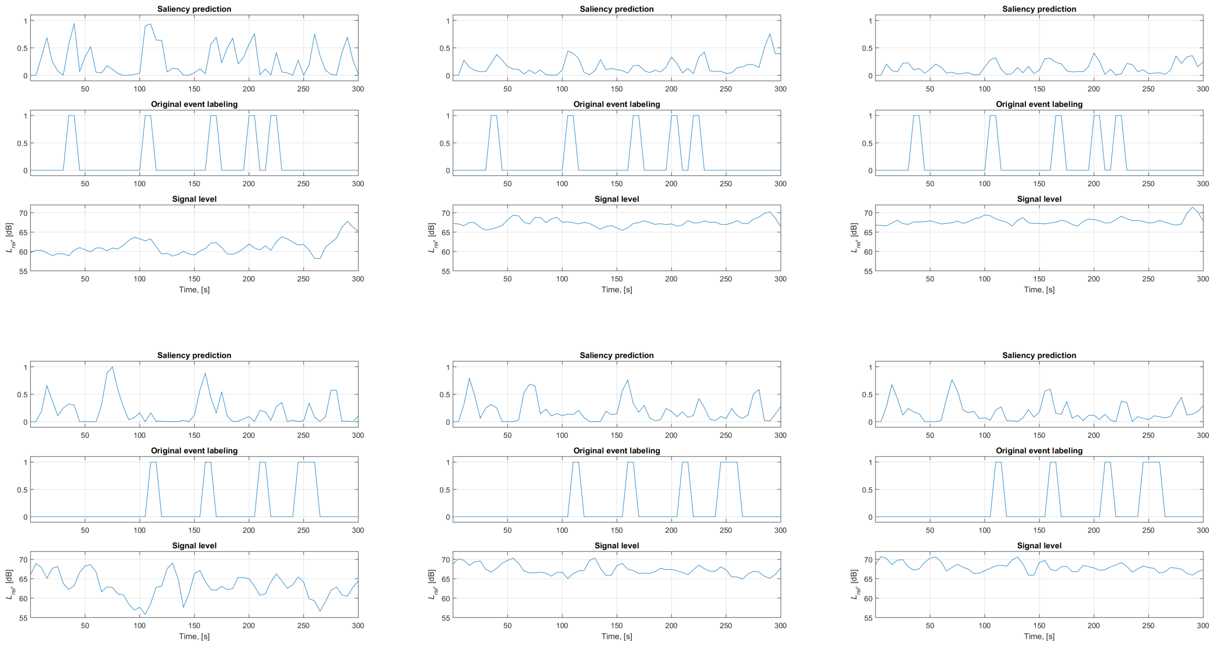
For traffic with car horn sound events, moments where the honking occurs do not emerge in the saliency indication. The strong variation in traffic sounds seems to include more saliency. Therefore, the habituation to the sounds may need to be included in the model to allow for intermittent traffic to receive less saliency. When speech is added, saliency spikes are less pronounced. While speech is a typical sound with significant saliency, having such high levels in comparison to the original recording, determines the detection of only the most prominent salient events.

5 Conclusions

In soundscape analysis and design, noticing particular sounds in a complex sonic environment plays an important role. Indeed, sounds that are noticed are expected to have much stronger influence on perception and understanding of the sonic environment than subliminal sounds. Whether a sound is paid attention to by the average user of a space depends on how much the sound stands out of its environment, i.e. its saliency. To predict the salience of a sound, an accurate and agile model for auditory saliency is needed. In this paper, we presented such model inspired by an observed brain response to spectrotemporal modulations.

It is shown that the model's output predicts purposely incorporated salient sounds reasonably well depending on the content of environmental sound. Furthermore, when foreground speech is added, the model adapts its saliency prediction and predicts only the prominent events in the resulting sound.

The sounds presented in this paper were used in an experiment with participants instructed to attend to the speech, listen for the salient sounds or simply not to attend to the sound stimuli. During this listening test, brain response with EEG signals was recorded. In a follow up study, the calculated saliency will be compared to a direct observation of EEG response in an attempt to find more solid criterion on the algorithms underlying saliency calculation.



Source: (Author, 2016)

Figure 2: **Saliency prediction on two environmental sound recordings and corresponding signals with added speech content. Each figure has three representations: saliency prediction, original labels from embedded salient event and relative sound level. Environmental sounds used (respectively top and bottom row): highway noise with emergency siren, traffic noise with honk sounds. First column contains results from basic recordings, second and third the same recordings with added different speech contents.**

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme *FP7/2007-2013/* under REA grant agreement n° 290110, SONORUS "Urban Sound Planner", and from the Research Foundation Flanders (FWO-Vlaanderen) under grant n° G0D5215N, ERC Runner-up project MAESTRO. Furthermore, Annelies Bockstael is a postdoctoral fellow of the Research Foundation Flanders.

References

[1] ISO T. 43/SC 1/WG 54, 12913-1 Acoustics – Soundscape – Part 1: "Definition and conceptual framework”. International Organization for Standardization. 2014.

[2] Filipan K, Boes M, De Coensel B, Domitrović H, Botteldooren D. Identifying and recognizing noticeable sounds from physical measurements and their effect on soundscape. Proc

European Congress and Exposition on Noise Control Engineering (Euronoise); 2015. p. 1559-64.

[3] Bregman AS. Auditory scene analysis: The perceptual organization of sound. MIT press; 1994.

[4] Kayser C, Petkov CI, Lippert M, Logothetis NK. Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology. 2005;15(21):1943-7.

[5] Shamma SA, Elhilali M, Micheyl C. Temporal coherence and attention in auditory scene analysis. Trends in neurosciences. 2011;34(3):114-23.

[6] Tsuchida T, Cottrell GW. Auditory saliency using natural statistics. Proc Annual Meeting of the Cognitive Science (CogSci); 2012. p. 1048-53.

[7] Boes M, Oldoni D, De Coensel B, Botteldooren D. A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention. Proc International Joint Conference on Neural Networks (IJCNN); 2013. pp. 1-8.

[8] Kaya EM, Elhilali M. Investigating bottom-up auditory attention. Frontiers in human neuroscience. 2014;8(327).

[9] Wang J, Zhang K, Madani K, Sabourin C. Salient environmental sound detection framework for machine awareness. Neurocomputing. 2015;152:444-54.

[10] Schönwiesner M, Zatorre RJ. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proceedings of the National Academy of Sciences. 2009;106(34):14611-6.

[11] Chi T, Gao Y, Guyton MC, Ru P, Shamma S. Spectro-temporal modulation transfer functions and speech intelligibility. The Journal of the Acoustical Society of America. 1999 Nov 1;106(5):2719-32.

22nd International Congress on Acoustics ICA 2016 : Proceedings / Federico Miyara ... [et al.] ; compilado por Federico Miyara ; Ernesto Accolti. - 1a ed . - Gonnet : Asociación de Acústicos Argentinos, 2016.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-987-24713-6-1

1. Acústica. 2. Acústica Arquitectónica. 3. Electroacústica. I. Miyara, Federico II. Miyara, Federico, comp. III. Accolti, Ernesto, comp.
CDD 690.22

ISSN 2415-1599
ISBN 978-987-24713-6-1

Disclaimer: *The material, information, results, opinions, and/or views in this publication, as well as the claim for authorship and originality, are the sole responsibility of the respective author(s) of each paper, not the International Commission for Acoustics, the Federación Iberoamericana de Acústica, the Asociación de Acústicos Argentinos or any of their employees, members, authorities, or editors. Except for the cases in which it is expressly stated, the papers have not been subject to peer review. The editors have attempted to accomplish a uniform presentation for all papers and the authors have been given the opportunity to correct detected formatting non-compliances.*

Hecho en Argentina
Made in Argentina
Asociación de Acústicos Argentinos, AdAA
Camino Centenario y 5006, Gonnet, Buenos Aires, Argentina
http://www.adaa.org.ar

ISBN 978-987-24713-6-1
9 789872 471361

22nd International Congress on Acoustics

PROCEEDINGS

5-9 September, 2016 - Catholic University of Argentina

X Congreso Iberoamericano de Acústica

XIV Congreso Argentino de Acústica

XXVI Encontro da Sociedade Brasileira de Acústica

Editors:

. Federico Miyara | . Vivian Pasch
. Ernesto Accolti | . Nilda Vechiatti

PROCEEDINGS OF THE ICA CONGRESS (online) - ISSN 2415-1599

ASOCIACIÓN de ACÚSTICOS ARGENTINOS
FIA FEDERACIÓN FEDERAÇÃO IBEROAMERICANA DE ACÚSTICA
ICA INTERNATIONAL COMMISSION FOR ACOUSTICS
SOBRAC Sociedade Brasileira de Acústica www.acustica.org.br