inter.noise
HAMBURG 2016

Auditory attention modeling within SONORUS ESR 10

Karlo FILIPAN[1]; Michiel BOES[2]; Annelies BOCKSTAEL[3]; Bert DE COENSEL[4]
Hrvoje DOMITROVIĆ[5]; Dick BOTTELDOOREN[6]

[1, 2, 3, 4, 6] Ghent University, Department of Information Technology, Belgium

[5] University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia

ABSTRACT

This contribution presents the outcomes of the research conducted within the scope of the SONORUS ESR 10 project theme. Within this theme, the influence of human auditory attention on the analysis and design of urban soundscapes is investigated. Several important attention mechanisms are evaluated: saliency, object formation and personal beliefs. Firstly, an auditory saliency model inspired by neural processing was created using features based on spectro-temporal modulations. Testing the model with environmental sounds that include salient events demonstrates its ability to correctly predict the saliency of the sound. Furthermore, object formation and resulting sound recognition was improved using an already developed computational auditory attention model based on a recurrent neural network. The output of the trained model shows that the sounds to which the model attends correspond to the reported sounds from the visitors of urban parks. Finally, to test if people's beliefs influence their attention, their viewpoint on tranquility was evaluated. As a result, it is determined that the visitors of the investigated parks who belong to groups that associate tranquility to natural sound sources and silence hear more mechanical sounds. Future urban sound planners could use all proposed models as tools for steering their decisions in evaluating human attention and perception of sounds in present or planned environments.

Keywords: human attention, artificial models, tranquil areas
I-INCE Classification of Subjects Number(s): 56.3, 68.2, 76.9

1. INTRODUCTION

Urban sound planning – the focus of the ITN SONORUS – should in future become an integral part of the planning process. The urban soundscape, defined in (1) as "an acoustic environment as perceived or experienced and/or understood by a person or people, in context", could be seen as the goal of this process. Therefore, within a context of expectations from stakeholders and experts in urbanism, architecture, ecology and other relevant disciplines, an urban (sound) planner should aim to create a matching soundscape.

As the definition of soundscape suggests, human perception is a vital part of soundscape assessment. Sound is seldom the reason of being in a public space (2), therefore users of such space passively listen to their surrounding acoustic environment. In this manner, bottom-up (inward oriented) attention highly influences their perception. Such attention is triggered by the sounds that exhibit salient characteristics, i.e. stand out of the environment, and thus have the implicit ability to attract attention. Therefore, mostly salient sounds will influence people's perception and understanding, in contrast to subliminal (non-salient, and for this reason unnoticeable) sounds.

Methodology for investigating soundscapes today includes measuring the physical environment with collection of responses from people present in or experiencing the environment (3). People's

[1] karlo.filipan@intec.ugent.be
[2] michiel.boes@intec.ugent.be
[3] annelies.bockstael@intec.ugent.be
[4] bert.decoensel@intec.ugent.be
[5] hrvoje.domitrovic@fer.hr
[6] dick.botteldooren@intec.ugent.be

soundscape evaluation is obtained using various methods, whereas most common include interviews and questionnaires. Although already well tested and structured, these methods do not usually consider the importance of human attention when evaluating the results.

A first issue arises when humans are asked about the sonic environment several minutes (hours) after their listening experience. As with any other cognitive process, their answers are at that time highly influenced by attention and memory (4). In particular, sounds that were noticed during the listening experience could simply be forgotten during the interview time. Additionally, sounds that attract attention at the time of the questionnaire could also influence participants' responses.

Another concern could be that the respondents are usually asked to listen to the sonic environment for which they will later give an evaluation. This task therefore changes their listening behavior to be more observant to the sound. Their attention is then more focused while listening in search and noticing the sounds that would not normally attract their attention.

Given that attention plays such an important role, this research focuses on finding the methods for soundscape assessment that include auditory attention evaluation. This is further complicated by the influence of personal beliefs, expectations, sensitivity and other personal traits on the attention processes. This paper reports on our recent findings and advances, presenting the theoretical background and results from saliency modeling, sound recognition and analysis of beliefs and viewpoints in reported perceptual evaluations.

2. ATTENTION MECHANISMS IN URBAN SOUNSCAPE CONTEXT

Human perception of an environment is generally highly influenced by their attention. If a person does not observe (attend to) an event happening in their surroundings, the event will most likely not influence their perceptual response and therefore will eventually pass unnoticed. This trait is common for all senses, and is equally relevant to auditory perception.

When listening to sound, humans perform an auditory scene analysis (5) to evaluate the complex sound environment. Two separate processes take place concurrently – segmentation and grouping of auditory streams. Segmentation incorporates time and frequency separation of a signal that was heard. The segments are then grouped together to represent auditory streams depicting a single environmental source. This forms a basis for auditory stream segregation, of which the most commonly reproduced example is the "cocktail party effect", i.e. hearing a specific person in a room full of talking people. Originally, the researchers did not consider attention as a part of these low-level processes, however some studies have confirmed its influence on stream segregation (6).

Auditory attention enables the listener to select a single auditory stream from multitude of processed events. When attending to the sound, two mechanisms interplay – bottom-up and top-down selection. Bottom-up selection is based on characteristics (features) of the stream listened to, especially the saliency of the sound. Top-down represents the listener's intention on focusing on a specific event. Important cognitive characteristic that also influences attention is inhibition-of-return. This mechanism enables the listener to focus their attention on other novel events and not only the most persistent one.

On a higher cognitive level, multisensory attention can also impact the perception since the neural pathways and cortex locations are partly overlapping (7). For instance, perceiving the object visually usually provides a better chance of attracting auditory attention to the sound corresponding with the same object. On the other hand, if a certain object is not visible, but its sound is salient enough to attract attention, an even stronger noticing might occur. This interaction between the senses requires multiple interdisciplinary studies, therefore its effects are outside the scope of the research work presented here.

2.1 Saliency models

For modeling how the sonic environment attracts attention, the salience of the sound has to be evaluated. Different models for saliency extraction have been previously proposed (8, 9, 10). All of the reported models base their calculations on a specific feature extractor (spectral, Gammatone, MFCC, etc.) and summarize the obtained features to a single time-varying indicator using an adaptation process.

In earlier models, we used the sound features proposed by Kayser et al. (11). These features capture variations in intensity, time and frequency domain on different scales. However, they somewhat lack in

capturing the transitions not found in the spectrograms. Such transitions are best represented in amplitude and frequency modulations that can often occur in the acoustic environment. Correspondingly, it was previously observed that the human brain contains regions very sensitive to such modulations (12). To include this knowledge in a measurement software, a saliency model based on spectro-temporal modulations (ripple sounds) was constructed.

As an input to the model, the sound is band-filtered and demodulated with the procedure mimicking the processing in human auditory pathways. The demodulated band signals are then cross-correlated with the corresponding modulators from ripple sounds. As a result, saliency features are extracted from the summation and maximization over multiple bands. The second layer of the model comprises time integration of the feature signals. In particular, a leaky integrator with different rise and fall time constants is applied, simulating activation and inhibition processes occurring in the neural pathways. In the last layer, rectified differences of the integrated signals are summed together with the option of including learned weights from logistic regression training. Therefore, the output of the model represents a single-number time-changing saliency evaluation of the input sound.

2.2 Auditory object formation and sound recognition

Features extracted from the input sound – which are likely to be the same as the ones that determine saliency – form an input to a complex neural system of adaptation, voluntary attention, inhibition of return and memory. Such complex system can be represented by a multi-layered neural network computational algorithm. Although the basic implementation of an artificial neural network mimics the neural signal transmission in a human brain, without specific adjustments, these cognitive processes are not simulated accurately. Therefore, all of these mechanisms are implemented specifically in our machine listening model based on a recurrent neural network.

The network consists of three layers connected through weights learned from a long-term processing of environmental sounds (13). The model implicitly implements processes such as memory and inhibition of return by calculating the activation of the neurons based on time-varying thresholds. Finally, the output of the last layer represents the learned concepts that are translated to auditory objects and eventually recognized sounds.

The model is trained through an unsupervised procedure, i.e. connections between the concurrently firing neurons are strengthened based on the similarity of the input features. On the other hand, learning is also accomplished using a top-down (supervised) procedure, during which already labeled concepts are translated to input features and respective connections are enhanced.

2.3 Influence of people's beliefs and viewpoints

Users of a public space have their beliefs and viewpoints on how such space should sound. This could influence what they focus their attention on, and thus what they remember. Considering that there is preexisting knowledge involved, sounds that are familiar are usually easier to recognize. However, unexpected or incongruent sounds could also be more salient because they disturb the overall expectation based on an earlier experience of the environment. These sounds will in turn provoke a response and might additionally hinder the experience by limiting the choices for behavioral response (14).

Investigation of such influences needs to be connected with the evaluated attention response. One approach is to quantify a perceptual concept, for instance a tranquil space. Such method was proposed by Lavandier and Delaitre (15) who evaluated people's viewpoints on tranquility by their ratings of several linguistic statements. The ratings were arranged in order to group the participants based on their beliefs what the concept of tranquility represents. We applied the same methodology by using some of the reported prototypical statements. Therefore, viewpoints were extracted from the questionnaire responses and the connection of beliefs to attended (heard) sound sources analyzed.

3. EXPERIMENTAL RESULTS

3.1 Saliency evaluation of environmental sound recordings

Using the created saliency model multiple environmental sound recordings were evaluated. For two of the recordings the saliency output with sound level and input labeling is displayed in Figure 1. The

first sound was recorded in the one of the streets of Ghent, Belgium, with honks from a car passing by the recording station. The second sound was mixed artificially from recordings of highway traffic noise and an emergency vehicle siren.

The model was trained on the data features and labeled events from the same sounds. Time windowing was chosen to be 10 seconds with one second step (90% overlap). The output shows an excellent correspondence between the events labeled by a human being and those based on the extracted saliency. Finally, the clear distinction between saliency extraction and sound level (bottom portion of the figure), shows that for detection of the salient events, relevant features from the input have to be derived.

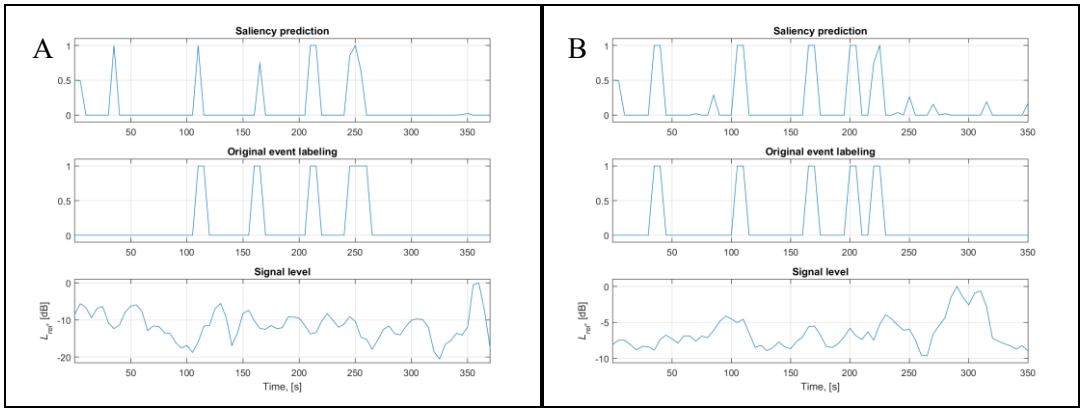

Figure 1. Output of the saliency model with input labels and signal level for two sounds: A) traffic noise background with honk sounds, B) traffic noise from a highway with an artificially mixed emergency siren.

3.2 Automatic detection of attended sounds in a park

The developed human auditory attention model was tested on the data gathered in a large-scale measurement campaign in Antwerp parks (16). During 22 days, the acoustic environment of the parks was recorded using mobile recorders carried inside backpacks. At the same time, visitors of the parks were questioned about their perception of the sonic environment.

Firstly, the attention model was trained in an unsupervised way on the complete dataset of recordings. Tuning of the model was carried out using labeled bird sounds, marked by an expert listener (17). Afterwards, the model output was checked and each of the neurons in the last layer was assigned with a label from one of the sound categories corresponding to people's responses (human, mechanical, natural).

The relationship between the output of the model and the reported sound categories is presented in Figure 2. The model's attended output is characterized by the number of sound events detected per hour of the data recorded in each of the parks. Correspondingly, the responses of the participants are accumulated with the average and denoted standard error. For the two categories of attended human and natural sounds, a strong correlation between model output and the visitors' responses was found.

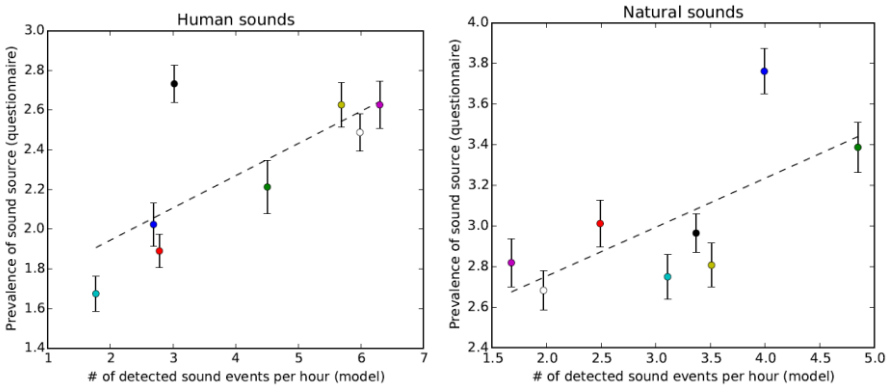

Figure 2. Relationship between reported sound source categories and computational auditory attention model output for human sounds ($r = 0.72$, $p < 0.05$) and natural sounds ($r = 0.68$, $p < 0.1$).

3.3 Influence of beliefs on noticing sounds in a park

The responses from questionnaires administered in Antwerp parks were also used in the analysis on how people's beliefs influence their attention. During the survey, the visitors of the parks were asked about their agreement on 13 statements related to tranquility beliefs. Afterwards, all of their responses were assigned to the calculated tranquility viewpoint group with a relative membership (18).

The results from the responses on the heard (attended) types of sounds and the relative membership of tranquility viewpoint groups are shown in Figure 3. As it can be seen, people who hear a lot of natural sounds generally do not belong to the groups associating tranquility to natural sound sources or to silence. On the other hand, for people belonging to the same tranquility belief groups, there is a pronounced increase in responses of hearing more mechanical sounds. Correspondingly, it can be argued that the mechanical sounds (often characterized as unwanted) are noticed more by those people associating tranquility to silence and to natural sound sources. Therefore, such people hear these antagonizing sounds more than the sounds that they actually want and expect to hear in the tranquil environment.

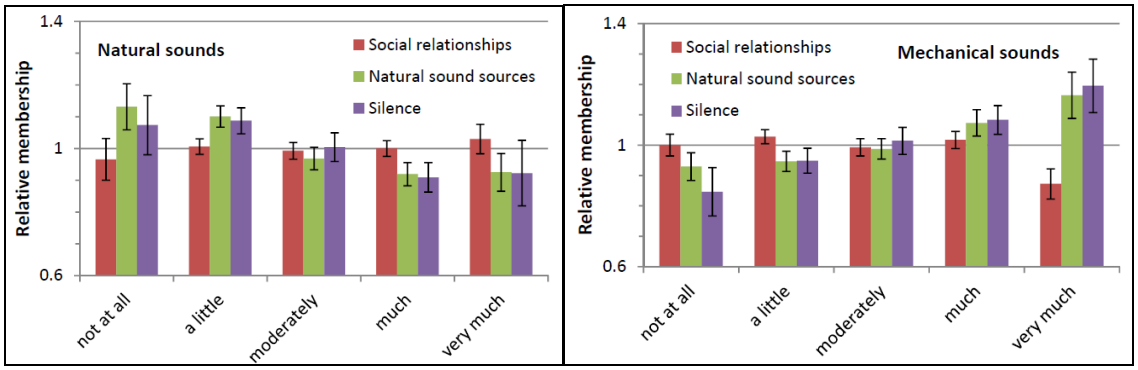

Figure 3. Relative memberships of tranquility viewpoint groups related to the heard sounds reported by the visitors of Antwerp parks.

4. CONCLUSIONS

In this paper, we emphasized the role of attention in urban soundscape assessment. To include this knowledge into measurements analysis, accurate models for saliency – the capacity of sound to attract attention – are needed. Therefore, we constructed a new, biologically-inspired auditory saliency prediction model and illustrated its effectiveness. Ongoing investigation will relate the attention to salient sounds directly to EEG signals in an ecologically valid context. Finally, the saliency extraction will be incorporated in a larger MIMO system trained on explicit human attention response measured by EEG signals.

Object formation and psychological effects such as inhibition of return and memory have been introduced in a computational model for human auditory attention based on a recurrent neural network. Unsupervised and supervised learning constitute essential parts of the model, thus mimicking the plasticity of the human brain. We used the created model for the analysis of data measured in eight urban parks in Antwerp. With extensive training, noticed sounds as reported by visitors can be predicted reasonably well. Accordingly, such use provides the ability for automatic soundscape evaluation.

We also argued that personal factors influence attention, and that the assessment of soundscapes can be shaped by them. In particular, meaning of tranquility was investigated, and the results from parks demonstrated that noticing specific categories of sounds depends on the personal viewpoint on tranquility.

All of the presented statistical and computational models can be used when assessing soundscapes of urban environments. Urban sound planners are thus presented with multiple tools for analysis of human attention and people's response to existing or proposed sonic environments.

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".

Michiel Boes is a doctoral fellow and Annelies Bockstael is a postdoctoral fellow of the Research Foundation Flanders (FWO Vlaanderen); the support of this organization is gratefully acknowledged.

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. Botteldooren D, Andringa T, Aspuru I, Brown AL, Dubois D, Guastavino C, Kang J, Lavandier C, Nilsson M, Preis A, Schulte-Fortkamp B. From Sonic Environment to Soundscape. In: Kang, J, Schulte-Fortkamp, B, editors. Soundscape and the Built Environment. CRC Press; 2015. p. 17-43.
3. Aletta F, Kang J, Axelsson Ö. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. Landscape and Urban Planning. 2016;149:65-74.
4. Cowan N. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. Psychological bulletin. 1988;104(2):163-91.
5. Bregman AS. Auditory scene analysis: The perceptual organization of sound. MIT press; 1994.
6. Shamma SA, Elhilali M, Micheyl C. Temporal coherence and attention in auditory scene analysis. Trends in neurosciences. 2011;34(3):114-23.
7. Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG. The multifaceted interplay between attention and multisensory integration. Trends in cognitive sciences. 2010;14(9):400-10.
8. Tsuchida T, Cottrell GW. Auditory saliency using natural statistics. Proc Annual Meeting of the Cognitive Science (CogSci); 2012. p. 1048-53.
9. Kaya EM, Elhilali M. Investigating bottom-up auditory attention. Frontiers in human neuroscience. 2014;8(327).
10. Wang J, Zhang K, Madani K, Sabourin C. Salient environmental sound detection framework for machine awareness. Neurocomputing. 2015;152:444-54.
11. Kayser C, Petkov CI, Lippert M, Logothetis NK. Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology. 2005;15(21):1943-7.
12. 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.
13. 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.
14. Andringa TC, Van Den Bosch KA. Core affect and soundscape assessment: fore-and background soundscape design for quality of life. Proc INTER-NOISE 13; 15-18 September 2013; Innsbruck, Austria 2013. p. 2273-82.
15. Lavandier C, Delaitre P. Individual and shared representations on "zones calmes" ("quiet areas") among the French population in urban context. Applied Acoustics. 2015;99:135-44.
16. Filipan K, Boes M, Oldoni D, De Coensel B, Botteldooren D. Soundscape quality indicators for city parks, the Antwerp case study. Proc Forum Acusticum 14; 7-12 September 2014; Krakow, Poland 2014. pp. 1-5.
17. Boes M, Filipan K, De Coensel B, Botteldooren D. Machine Listening for Park Soundscape Quality Assessment, Submitted to Acta Acustica
18. Botteldooren D, Filipan K, Boes M, De Coensel B. How the meaning a person gives to tranquility could affect the appraisal of the urban park soundscape. Proc INTER-NOISE 14; 16-19 November 2014; Melbourne, Australia 2014. pp. 1-6.