Identifying and recognizing noticeable sounds from physical measurements and their effect on soundscape

Karlo Filipan, Michiel Boes, Bert De Coensel
Department of Information Technology, Ghent University, Ghent, Belgium

Hrvoje Domitrović
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia

Dick Botteldooren
Department of Information Technology, Ghent University, Ghent, Belgium

Summary
In contrast to the classical noise control, the soundscape approach analyzes the person-environment interaction in more detail including positive as well as negative effects. Environmental sound is often a by-product of the environment and listening to it is rarely the purpose of being in a place. Therefore, noticing and inhibition-of-return play an important role in the theoretical model for people's perception. The proposed model extends from an initial physiological response to environmental sound over noticing, identifying, and recognizing to appraisal within a context of personal beliefs and expectations. Consequently, it attempts to encompass the whole interaction of the person and the environment from sensory inputs to actions related to the response on the environment. During the recent years, environmental monitoring and sound monitoring as its part have experienced a technology driven growth to which various governing bodies have shown a significant interest. However, the challenge now presents itself in the analysis of the acquired big data especially when it comes to perception. Several aspects of the above mentioned theoretical model for perception of environmental sound have been implemented in the computational models for this purpose. The models are based on the artificial neural network structure that mimics many of the low level neural processes occurring in the human brain. However, the models do not attempt to make a simulation of a complete brain, which is still well out of reach even for the most advanced computer architectures. This contribution will focus in particular on the object formation and attention processes in an attempt to predict which sounds would be noticed by the user of a space and how this will affect the soundscape. Examples from urban parks and residential areas will be shown to illustrate how accurately the model based on physical inputs solely can match the human response.

PACS no. 43.50.Rq, 43.50.Qp, 43.50.Yw, 43.66.Lj

1. Introduction

The environment and the environmental sound that people live in could be an influential factor for their health, wellbeing and overall living satisfaction [1, 2, 3]. Sound or rather noise is in most cases byproduct of the environment and rarely the purpose of being in a place. Although previous regulations as well as the measurement techniques have considered the negative effects of noise, a lot of current research in environmental sound is directed towards the soundscape approach, thus considering sound as a resource rather than a waste [4, 5].

During the recent years, noise monitoring has become one of the areas to which various governing bodies have shown a significant interest. Present technology allows monitoring cities with high spatial resolution thus gathering and storage of large amounts of sensor data. However, the challenge then becomes to analyze those data especially when it comes to perception. State-of-the-art research in the field of neural network modeling is still not being able to replicate even the simplest models of animal brains [6], nevertheless borrowing the principles and making them more suitable for human sound perception, computa-

tional auditory scene analysis [7] could become one of the tools for future soundscape monitoring.

This contribution will encompass a theoretical model for people's perception extending from initial reaction to environmental sound, to appraisal and coping mechanisms. The theoretical model accounts for complex personal factors such as expectations and personal beliefs on the sound environment. In comparison to this model, a machine listening model based on artificial neural network that implements some of the building blocks will be presented.

2. Human sound perception model

Humans that occupy any environment in the world today are constantly affected by the different sounds present in these environments. Whether it is indoors, for instance in homes and at work places, or outdoors (urban public spaces as well as in the nature), the sound environment differs greatly.

Sound perception is a process that could be described as the whole auditory scene analysis and its concurrent interpretation by the person. While environmental sound is a rather complex conglomeration of various sounds originating from different sources, humans tend to dissolve this mixture into the individual auditory streams using auditory, but also visual as well as other cues [7].

Environmental sounds can be regarded as any sound that does not posses a communication value to a specific listener considered, as opposed to speech or other informational sounds (which examples, depending on the occasion, could include horn for traffic participants, telephone ring, fire alarm, etc.). Therefore, initially there is no particularly strong attention focus on any of the specific sounds, and the person is listening in readiness or rather, listening in a holistic way to the sonic environment. Consequently, most of the environmental sounds that humans are exposed to are not being regularly noticed, but form a background mix or hum. However, from these sounds the listener's attention, using the concepts of perception, occasionally selects and fine tunes the auditory streams.

Sound saliency forms a part of the bottom-up process, and is one of the properties that should be accounted for when considering human perception of environmental sounds. Although salient sounds do attract attention, they alone cannot explain all of the attention components, thus a multisensory part and a voluntary aspect in the final selection for attention focusing have to be considered as well. Accordingly, attention provided from the saliency of the sound may change with the gating coming from top-down personal consciousness. In addition, an important factor in attention switching is the process of inhibition-of-return which prevents the attention from permanently staying focused on one single item [8]. Therefore, it reduces the focusing on the single auditory source or stream and gives the possibility to perceive other than the currently most salient ones.

Additionally, appraisal of and the response to the attended sound are guided by the meaning that the person assigns to it. Provided that the sound is perceived as negative, behavioral actions that will occur can be observed as focusing, denial and active coping [9]. On the other hand, for sounds with positive connotation, paying attention to them is something that would be considered as an improvement of the sonic environment. Accordingly, in soundscape research, it was observed that by adding positively contextualized sounds, such as bird sound or the sound of water streams, their traits improve the overall appreciation of the sonic environment [2, 10].

Furthermore, the home environment assumes that almost every environmental sound coming from the outside will be regarded as an intrusion and in turn have a negative appraisal. Therefore, noticing any sound from outside would almost definitely lead to annoyance. However, for the public space, the sonic environment is part of the experience of the area or the place, and listening in search emerges as a visitor's natural listening state. Therefore, attention that a person gives to the noticed sounds could have a significant importance for soundscape description and (artificial) determination of these sounds might prove to be a reasonable method for comparing sound environments.

Nevertheless, it should be noted that different people also assign different meaning on soundscape properties based on their previous experiences and their own personal beliefs. As it was shown in [11] for tranquility of the public spaces, most of the people could be confined to three main groups based on their preferences. Thus, finding the preference of the listeners and assigning the specific sounds to their preference might prove a very promising tool for sonic environment reshaping or design.

Furthermore, one should also consider the core affect that is evoked by the sonic environment. As it was previously shown, this outcome can be evaluated with the soundscape perceptual dimensions on an eight-scale space [4]. Effects of the sounds and noticed effects take part in the affect forming, however part of it is evoked from background sound, consequently inducing a "musical" listening experience [12]. Additionally, a person then experiences an emotion that is evoked by the core affect.

In particular, during listening it becomes apparent that the novelty in the sound might evoke a pleasant or unpleasant emotion. Events that are correctly predicted by the brain cause an aesthetic emotion – the reward given to the system for correctly predicting the immediate future predictability of sound. Negative emotion appears however when non-matching expectations of the sound occur. Additionally, surprise can also be welcomed and not trigger the negative context

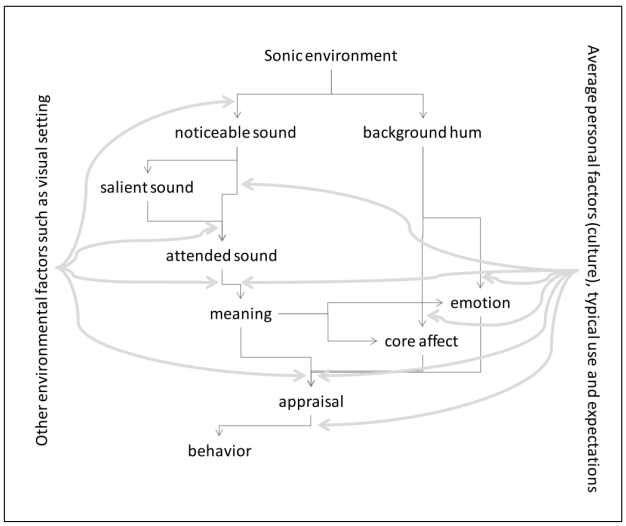

Figure 1. Human sound perception shaped from sonic environment and resulting in appraisal and behaviour. Feedback and sensors gating, as well as learning, are part of the process and shape the process continuously.

when only a contrastive valence introduced. Possibility of learning the sound environment sequences can also shape appraisal. For instance, extremely unpredictable sequences of sound allow only reduced opportunity for learning. Correspondingly, extremely predictable sequences do not offer much challenge to the brain and the learning is not existent. However, the sequences with moderate degree of expectation violation are perceived as pleasurable [13].

Described model of auditory perception proceeding from the sonic environment and shaping towards subjective behavior and appraisal is given in Figure 1. Moreover, represented stages are gated with inhibition-of-return coming from multisensory perception, while feedback is provided from attention focusing which sharpens the sensors coping action. Learning itself occurs on all stages, however coming to the later stage personal factors give way for cultural and social ones (i.e. person's social environment).

3. Artificial sound perception model

The artificial (machine) sound perception model relates directly to the previously presented model for human sound perception. It implements two basic two different listening styles. The first machine listening model focuses on the characteristic of holistic, non-focused "musical" listening experience. The second machine listening model implements analytic listening and considers the person's attention to and the noticing of sounds. It uses current advances in machine learning and artificial intelligence research.

With the first alternative which characterizes the holistic, background listening experience, the artificial sound perception model directly extends the currently widely used classical A-weighted levels. Correspondingly, in order to better represent the response of the peripheral auditory system, stationary or time-dependent loudness [14] is used. In addition, several other spectral cues, such as sharpness and tonality, which are also used in assessing products' sound quality [15], can serve as an indicator for background hum quality. Somewhat different measures that also illustrate spectral content are the center of gravity and the difference between C-weighted and A-weighted level (solely because of the simplicity of using widely available A and C-weighting filters). Furthermore, this spectral information enables indirect indication of saliency and the actual sound source (e.g. birds sounds and human voices tend to contain more higher frequencies than the distant traffic sounds).

Temporal dynamics of the sound are also useful representation of a holistic listening experience directly similar to music [16]. In addition to the more generally applied peak counts and amplitude dynamic measures such as $L_{10} - L_{90}$, the slope of the spectrum of both amplitude and pitch serves as the balance representation between predictability and change (novelty) in the sound environment. Additionally, for temporal dynamics an $1/f$ slope in these spectra appears to be more closely related to the aesthetic of natural sound and music. At the same time, time variations are also an indirect indicator of sound saliency – the stronger the variations the higher the sound saliency – and therefore the noticeability of individual sounds in the mixture.

To mimic the part of the human model regarding the noticed sounds (left side of Figure 1), the machine listening should correctly identify the sounds within the sonic environment that a person would likely pay attention to. With the relatively recent development of advanced machine learning techniques and more importantly the increase in computational power, it is now becoming possible to partially simulate the essential human perceptual dynamics involved in this process. Even though the current computational power and the system architecture are still years behind in managing to approach even the most simplest brains, the possibility of machine listening with attention models emerge from two concepts: deep neural networks [17] and adaptive resonance theory [18].

Artificial neural networks emerged from an effort to mimic biological characteristics of a single neuron's signal transmitting properties and the connections that the (100 billion) neurons in a human brain form with each other. Nevertheless, although the artificial networks were initially designed to shape the complexity of a human brain, training an immense number of connection weights was nearly impossible given the proportionally small number of typically available training samples. As a consequence, the networks were reduced to usually three layers with a number of hidden neurons, the output of which results in complex nonlinear functions, which are however very far from

the functionality of a human brain. Nevertheless these simplified artificial neural networks have been used for many different purposes, one of which also includes sound recognition and classification [19]. For the purpose of the current work this classical simple artificial neural network structure needs to be extended and refined.

Apart from the supervised training based upon an already labeled set of data, used as a main principle in many machine learning problems, some neural network architectures can be extended with an unsupervised learning phase. This unsupervised training of complex features is based on co-occurrence and works effectively for environmental sound object creation as shown in [20]. Moreover, including long unsupervised training periods is much closer to how a biological brain tries to organize the world around it.

Unsupervised training on long sound sequences – i.e. the continuous recording of an outdoor microphone – has some consequences. Firstly, careful parameter setting should prevent that well tuned neurons get detuned over and over again, or in other words, the artificial neural network catastrophically forgets. Secondly, unsupervised training on a stream of environmental sound may lead to overspecialization on non-informative but frequently occurring background sound. Therefore, a selection based on the saliency of the sounds [22] has to be included. If included properly in the model, the most prominent sounds that the person would likely attend to are trained more carefully.

Even at the lowest level of sound identification or sound source recognition, context plays an important role when interpreting the acoustic features. For this reason, sounds with similar characteristics, such as traffic noise or sea waves breaking at a distance, could be perceived as either, depending on whether the person is inside a city park or at the sea shore. Moreover, sounds that have emerged in specific environment recently are more likely to occur in the same environment again. Therefore the model needs to account for this as well. With that in mind, a short term memory which enables accounting for recent history has been included in the machine sound perception model.

Consequently, by assessing the sounds that are often present at the location on which the machine listener has been trained, another aspect from the unsupervised training approach appears: specialization. Nevertheless, context awareness also assumes that the listener possesses a general knowledge about the environment. A human listener expects to hear certain sounds at a particular location based on a previous sensory experience or based on the knowledge obtained from social or cultural factors. However, this spatial awareness if not incorporated explicitly cannot be expected to emerge in machine listeners, so the only option remaining is to provide this information separately. Context is established by increasing a priori probability for detection. The feedback included in its implementation automatically increases the machine model's attention focusing onto specific sounds.

4. Applications in environmental sound monitoring

The artificial sound perception model that incorporates the characteristics of human sound perception and simulates the attention to sounds, as discussed in the previous two sections, is currently used for environmental sound monitoring. The core of the model lays within a four-layer recurrent neural network which structure incorporates the attention mechanisms such as gating, inhibition-of-return and saliency, as well as the short term memory [20].

As input to the model, environmental sound is captured by 1/3-octave bands every 125 milliseconds and stored in the database [21]. Furthermore, features based on the perceptual characteristics of low-level human hearing characteristics [22], are extracted from these data and fed to the artificial neural network as input layer neurons.

The neural network is then trained on an extended sequence of environmental sound. Firstly, an unsupervised training without labeled data is conducted, i.e. no teacher top-down input of labels is provided. This approach however results in learning based only on co-occurence. In addition, the model is also trained at regular intervals in a supervised way with a set of labeled sounds. This process however requires human labeling, thus the Noiseplay game for human sound labeling [23] has been used to create a competitive environment for performing the rather tedious process of labeling the sounds.

Finally, after the network has been sufficiently trained, the expected attention to different sounds is gathered in the evaluation run. The output of the network now consists of activations that characterize the attention to the specific sounds. Furthermore, for clearer representation, attended sounds are categorized into three main categories: human, mechanical and natural sounds.

The dataset, that is used for attention model in this paper, was obtained within the measurement and survey campaign carried out for 22 days in eight urban parks in Antwerp, Belgium, during August and September 2013 [24]. The data was gathered with mobile sensor nodes placed in the researchers' backpacks which were carried on all the paths inside the parks. The recorded data consisted of 1/3-octave bands and audio signal together with the GPS position, which in turn enabled the spatial representation of all measurements. Additionally, questionnaire surveys were taken from the parks' visitors. Noticed sounds within the sonic environment, core affect, general soundscape appraisal and tranquility viewpoints were captured.

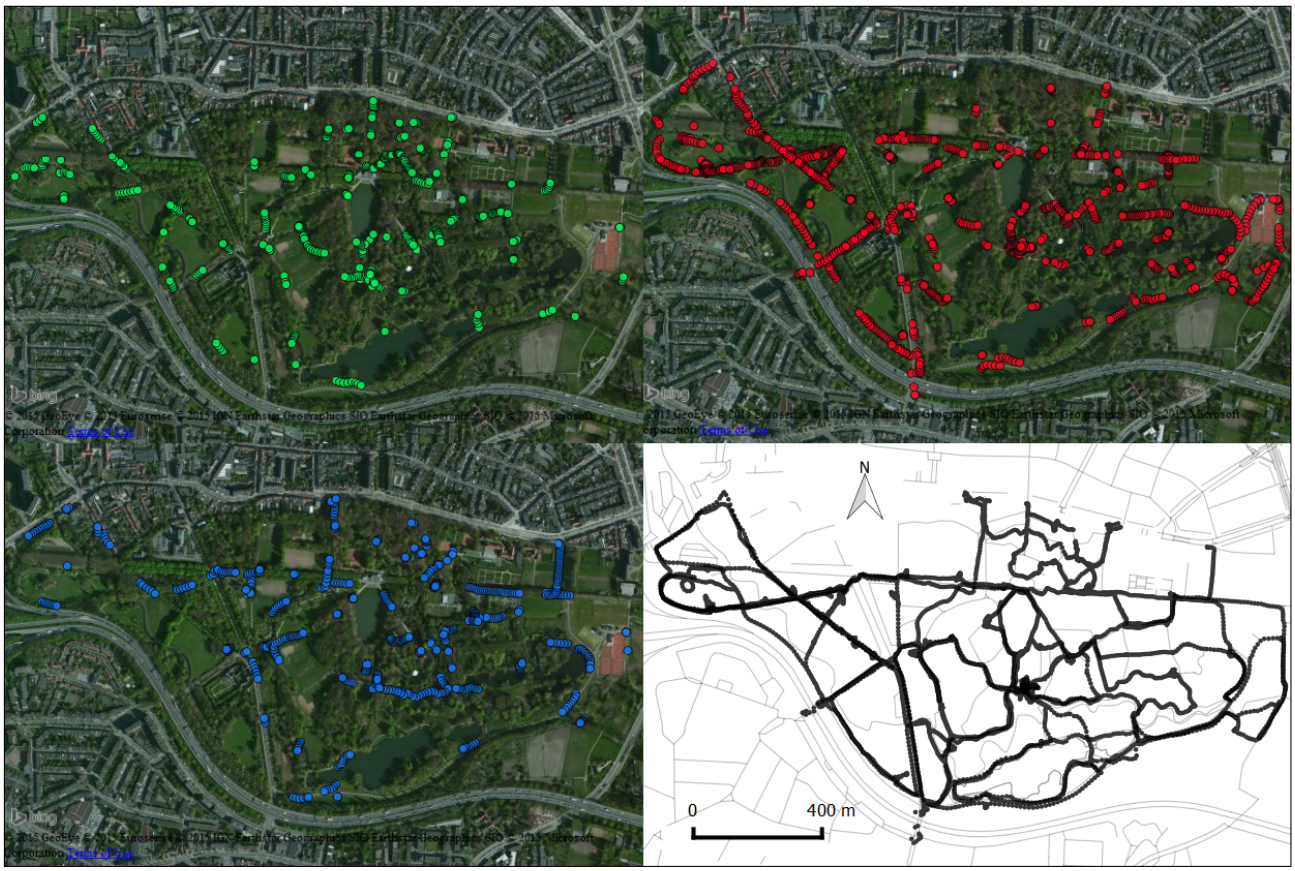

Figure 2. Artificial sound attention model output for Rivierenhof park in Antwerp, Belgium. Categories of natural, mechanical and human sounds are given in green, red and blue color respectively (background source: Bing maps). All possible walked paths are shown in black color on a simplified map.

Figure 2 shows an example of the output created by the sound attention model for Rivierenhof park in Antwerp on 6th of August 2013. Sound categories of natural, mechanical and human sounds to which the model paid attention to are represented based on their spatial positions. Although the model cannot be expected to find the exact sounds that each individual person would pay attention to, the spatial separation of the categories should give information that statistically resembles the individual perception. For instance, mechanical sounds were almost constantly noticed next to the busy roads (west part of the park) even though the output also gave a lot of attended mechanical sounds inside the park. On the other hand, attended human and natural sounds appear predominantly in the center of the park. Finally, natural sounds were more often noticed in the north of the park which is positioned in a fairly isolated area without many people.

However, the output also shows a significant portion of human sounds activation in the areas which did not accommodate many people during the measurement campaign (central forested area). In contrast, human sounds were correctly noticed by the model on the busy main west-east walking connection in the park, and, together with natural sounds, around the main pond with the fountain. Note that at some locations the model predicts that no attention will be paid to either of the sounds. This is a consequence of the implementation of attention processes that allow the model not to listen attentively just like a human park visitor would.

5. Conclusions and future work

In this paper a model for human perception of environmental sound and its translation to an artificial model were presented. Both models were based on the psychological and physiological characteristics of human perception and sound appraisal. Furthermore, it was shown how the artificial model for attention, based on the recurrent neural network, can be used in the evaluation of an urban sound environment. Even though the results are promising, the comparison of the model outcome with the perceptual questionnaire data requires a further non-trivial validation step.

The presented artificial sound perception model finds its natural place as an integral part of a sound monitoring sensor network. As a result, real-time output, as well as the gathered historical data, would enable interested stake holders to assess and consequently understand the overall sonic environment and

its evolution over time. In turn, this could prove as an important tool for soundscape assessment with further extension to the future urban sound planning.

Acknowledgement

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 Bert De Coensel is a postdoctoral fellow of the Research Foundation Flanders (FWO Vlaanderen); the support of this organization is gratefully acknowledged.

References

[1] P. Lercher, D. Botteldooren, U. Widmann, U. Uhrner and E. Kammeringer. Cardiovascular effects of environmental noise: Research in Austria. Noise and Health 13:234-250, 2011.

[2] B. De Coensel, S. Vanwetswinkel and D. Botteldooren. Effects of natural sounds on the perception of road traffic noise. J. Acoust. Soc. Am. 129(4):EL148-EL153, 2011.

[3] J. Terroir, B. De Coensel, D. Botteldooren and C. Lavandier. Activity interference caused by traffic noise: experimental determination and modelling of the number of noticed sound events. Acta Acust. Acust. 99(3):389-398, 2013.

[4] O. Axelsson, M.E. Nilsson and B. Berglund. A principal components model of soundscape perception. J. Acoust. Soc. Am. 128(5):2836-2846, 2010.

[5] A.L. Brown, J. Kang and T. Gjestland. Towards standardization in soundscape preference assessment. Appl. Acoust. 72:387-392, 2011.

[6] P.J. Werbos. From ADP to the Brain: Foundations, Roadmap, Challenges and Research Priorities. In Proc. Int. J. Conf. Neur. Netw. (IJCNN), 107-111, 2014.

[7] A.S. Bregman. Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, Cambridge, Massachusetts, USA, 1994.

[8] B. De Coensel and D. Botteldooren. A model of saliency-based auditory attention to environmental sound. In Proc. Int. Cong. Acoust., 2010.

[9] D. Botteldooren and P. Lercher. Soft-computing base analyses of the relationship between annoyance and coping with noise and odor. J. Acoust. Soc. Am. 115(6):2974-2985, 2004.

[10] 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. J. Acoust. Soc. Am. 127:1357-1366, 2010.

[11] P. Delaitre, C. Lavandier, R. Dedieu and N. Gey. Meaning of quiet areas in urban context through people viewpoints. In Proc. Acoust., 2012.

[12] D. Huron. Sweet anticipation: Music and the psychology of expectation. The MIT Press, Cambridge, Massachusetts, USA, 2006.

[13] M.T. Pearce and G.A. Wiggins. Auditory Expectation: The Information Dynamics of Music Perception and Cognition. Topics Cogn. Sci. 4:625-652, 2012.

[14] H. Fastl and E. Zwicker. Psychoacoustics: Facts and models. Springer, Berlin, 2007.

[15] K. Genuit. The sound quality of vehicle interior noise: a challenge for the NVH-engineers. Int. J. Veh. Noise Vib. 1(1):158–168, 2004.

[16] D. Botteldooren, B. De Coensel and T. De Muer. The temporal structure of urban soundscapes. J. Sound Vib. 292:105-123, 2006.

[17] Y. Bengio. Learning deep architectures for AI. Found. Trends Mach. Learn. 2(1):1–127, 2009.

[18] G.A. Carpenter and S. Grossberg. Adaptive Resonance Theory, In: M.A. Arbib (eds.) The Handbook of Brain Theory and Neural Networks, 2nd ed., The MIT Press, Cambridge, Massachusetts, USA, 2003.

[19] B.W. Schuller. Intelligent audio analysis. Springer, Berlin, 2013.

[20] M. Boes, D. Oldoni, B. De Coensel and D. Botteldooren. A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention. In Proc. Int. J. Conf. Neur. Netw. (IJCNN), 596-603, 2013.

[21] B. De Coensel and D. Botteldooren. Smart sound monitoring for sound event detection and characterization. In Proc. INTER-NOISE & NOISE-CON, 3442-3451, 2014.

[22] D. Oldoni, B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem and D. Botteldooren. A computational model of auditory attention for use in soundscape research. J. Acoust. Soc. Am. 134(1):852-861, 2013.

[23] http://www.noiseplay.org/, accessed on 27.02.2014.

[24] K. Filipan, M. Boes, D. Oldoni, B. De Coensel and D. Botteldooren. Soundscape quality indicators for city parks, the Antwerp case study. In Proc. Forum Acust., 2014.