The role of saliency, attention and source identification in soundscape research

Dick Botteldooren\textsuperscript{a}
Bert De Coensel\textsuperscript{b}

Acoustics group
Department of Information Technology
Ghent University
St. Pietersnieuwstraat 41, B-9000 Ghent, Belgium

ABSTRACT
Most researchers and designers using the concept of soundscape in their work now agree that this concept involves both the sonic environment and the person. The central role of the person can be stressed in research aiming at understanding the soundscape by interpreting the soundscape as the mental representation of the sonic part of the environment. Thus the non-conscious part of neural activity is mostly excluded. The attention of a person enveloped in a sonic environment, engaging in certain activities and pursuing particular goals, could be drawn to the sound by its increased saliency. Whether or not this influences the appreciation of the environment depends on sustained attention and sound source identification, both top-down processes. The prevailing mental image—which contains evoked memories—of the environment determines whether the sensory input is appraised as disturbing and annoying or rather pleasing and completing the mental representation. Computational models mimicking human perception and cognition can be very helpful in understanding the above sketched mechanisms. This paper reports on our ongoing research in this area, with main focus on the bottom-up mechanisms of saliency and attention.

1. INTRODUCTION
When the soundscape concept was first applied to environmental sound\textsuperscript{1}, the focus on active participation of the listener in the urban sonic environment was clear. The interaction of the urban soundscape with the listener is approached with the same mindset as the interaction with music or the landscape. Soundscape research should therefore include personal and social meanings\textsuperscript{2} of environmental sounds, including the historical or aesthetic contexts\textsuperscript{3}.

\textsuperscript{a} Email address: dick.botteldooren@intec.ugent.be
\textsuperscript{b} Email address: bert.decoensel@intec.ugent.be, currently on leave at University of California at Berkeley, USA
Given the central role of the listener in soundscape research, it is useful to remember the basic functions of hearing environmental sound from a bio-acoustical point of view:

- Sound is a warning signal for potential danger. This function relies heavily on detecting salient parts of the soundscape, e.g. (unexpected) changes. Effects such as an increase in blood pressure or heart rate may occur in a pre-attentive phase if saliency is high.
- Sound helps many biological systems to localize prey. Although this ability may be of little use for modern man, it is still used in some forms of advertisement trying to attract our attention.
- Sound helps to create the mental image of the world around us. New environments trigger explorative behavior that may include more attentive listening to the sounds. Source recognition is an essential aspect. A familiar environment can thus be expected to lower attentive listening as long as basic evaluation of the sonic environment matches our expectations. Cognitive resources are spent on other aspects of daily life or the organism just saves energy by slowing down and relaxing.
- Sound is used for communication. Sound initially recognized as being produced by the own species receives more attention than other sound because it may carry important information.

The concept of soundscape, as used in the scientific community, focuses mostly on the third function of sound: constructing a mental image of the living environment. From this point of view, understanding how the soundscape approach might work comes down to understanding how the sonic environment helps creating a mental image of the environment in the mind of the typical users and for the typical use of this environment. This approach puts the listener in the center of the picture and, as a consequence, automatically has to address several of the points of concern raised by soundscape researchers earlier.

It still remains to understand how saliency, attention, and source identification all work together in relating the physical sonic environment to the mental image of the world. For this reason we have been working on a simulation model that allows gaining insight in these mechanisms and drawing conclusions on methods for soundscape design. When attempting to incorporate knowledge on sound perception from recent developments in psychophysics and physiology, including neural research, it is imperative to keep in mind that the typical time scales of soundscape analyses are long and the number of people-environment combinations that need to be studied is high. Moreover, potential implementation in noise measurement equipment must be kept in mind. Translating this laboratory knowledge to simulation models must therefore consider the limited capabilities of computer systems when compared to the human mind. This paper reports on recent developments in our model.

2. CONSTRUCTING THE SIMULATION MODEL

A. Overall model structure
Recent findings in psychophysics and neurophysiology strongly emphasize the role of attention in understanding the everyday complex sonic environment in which we live⁴. Although most of the experimental evidence underlining the importance of what we more specifically would call outward attention or top-down attention is based on speech perception, it can be argued that it also holds for environmental sounds. Based on the analyses of the biological functions of sound, it can be argued that this form of attention is particularly important in building a mental image of the living environment and thus for soundscape perception.

The onset of auditory stream segregation and object formation in audition nevertheless seems very strongly related to the properties of the sound itself. This is the case especially when
attention was not pointed to a particular sound at the time it occurs, for example by visual cues. Saliency of the auditory stream picked up by the ear makes it more susceptible to attracting attention\textsuperscript{5-6}. Saliency based attention focusing may biologically be most important when considering sound as a warning signal. It can also draw mental resources away from more interesting parts of the sonic environment and thus perceptually mask parts of the auditory stream. As a consequence, saliency detection plays an important role in a computational model for soundscape perception.

The mechanisms described above compete for attention and additional mental resources. Attention is expected to rapidly switch (time constants of a hundred to a few hundred milliseconds) between different tasks but it is probably not dividable at any instant in time. Focusing attention implies forming auditory streams\textsuperscript{7}. In prior versions of our soundscape perception model\textsuperscript{8} we assumed auditory streams to be rather static objects. Stream formation in reality is a gradual and dynamic process guided by attention\textsuperscript{9}. Including this dynamic character may lead to better understanding of soundscapes and thus we are gradually introducing it into our model. The more interesting a sound is, the longer attention will be focused on it and the more detailed the auditory stream will be analyzed. Although attention may be initially drawn to the stream by saliency, it could eventually stay there due to the high information content (related to communication or not) of the stream.

Is focusing attention on particular auditory streams important for soundscape evaluation? The answer to this question is related to the definition of soundscape. If one accepts that the soundscape is mainly related to forming a mental image of the living environment, including culture and personal past experience, the components of the sonic environment need to be identified and meaning must be attributed to them. Attention is needed to perform this mentally demanding task. From this point of view, the amount of attention given to “positive” sounds could be an indication of quality. However, if overall health impact is of primary importance, the warning function of the environmental sound may be more significant, a function that needs limited attention. The authors follow the first definition.

B. Saliency

Several computational models for auditory saliency have been proposed in the literature recently\textsuperscript{5-6}, usually having a structure largely based on analogous models for visual saliency. Most of these are designed with applications for speech processing in mind, and therefore consider a fine resolution in time and frequency, at the cost of computational speed. However, this makes these models less suitable for direct application to environmental sound perception, where calculation speed is important. Therefore, a simplified model was developed\textsuperscript{8} which nevertheless implements the key elements present in the calculation of complex auditory saliency maps.

The model starts from the 1/3-octave band spectrograms of the different streams that constitute the sonic environment. A time step of 1 s is considered. In a first stage, a simplified cochleagram is calculated from the 1/3-octave band spectrograms, using the Zwicker model for specific loudness. Subsequently, a set of multi-scale feature maps are extracted in parallel from the specific loudness vs. time maps. These features mimic the information processing stages in the central auditory system. In particular, the human auditory system is, next to absolute intensity, also sensitive to spectro-temporal irregularities (i.e. contrast on the frequency scale, and changes in time)\textsuperscript{10}. The intensity feature maps are calculated by convolving the specific loudness vs. time maps with gaussian filters with varying width. The spectral and temporal contrast feature maps are calculated by convolving the loudness vs. time maps with difference-
of-gaussians filters with varying width, and thus encode the spectral and temporal gradient of the loudness vs. time maps, calculated at varying scales. These feature maps are then normalized using a biologically inspired iterative normalization algorithm\textsuperscript{11-12}, which strongly promotes maps that contain a small number of (conspicuous) peaks, and strongly suppresses maps that contain a large number of comparable peaks. The feature maps are then combined across scales and again normalized into conspicuity maps, which are finally added to yield the (time-varying) spectral saliency of each source. The total saliency value can be calculated by summing over all frequencies, assuming that saliency combines additively across frequency channels\textsuperscript{12}.

Calculated saliency is not sufficient to explain the subjective rating of soundscape quality. An analysis of an extensive questionnaire study in several parks in Stockholm\textsuperscript{13} makes this clear. Average saliency of the (recorded) soundscape at the location of the interviewee was calculated over the 10-minute period during which he/she was asked questions on soundscape quality and perceived presence of sounds. In Figure 1, quality rating is compared with calculated average saliency. Both soundscapes with high and low average saliency give rise to somewhat better quality ratings. When plotting the sources heard by the respondents, it becomes clear that the origin of saliency is quite important: the very salient soundscapes contain a high amount of natural sounds and thus attracting attention to the sonic environment may be appreciated by the listener.

![Figure 1: Rated soundscape quality and perceived presence of sounds generated by different sources, as a function of average saliency (data from the Stockholm study\textsuperscript{15}).](image)

C. Attention

The central mechanisms in the attention model are a combination of activation and inhibition-of-return, and a winner-takes-all competition. Because of the high degree of parallelism in the human system, it is difficult to show the mechanism in a classical flow diagram; in Figure 2 we nevertheless made an attempt. Distinct features in the sound that enters the ear, such as sudden changes over time or changes in frequency content, increase saliency and activate the bottom-up attention mechanism. The auditory system then enters the competition for attention, taking into
account the current state of top-down attention for environmental sound (which depends on context and activity of the listener). If attention is not received, nothing happens. If, on the other hand, attention is received, a process of stream segregation and identification starts. This process may fail or succeed. If a new stream is created, it is taxed on its level of interest for the organism. Interesting streams activate top-down attention. Now inhibition of bottom-up attention (inhibition-of-return) kicks in. Until inhibition exceeds overall activation, the stream re-enters the competition for further attention.

A few examples illustrate this mechanism. As a first example, assume someone starts talking. Initially saliency and bottom-up attention will attract attention. If the listener is not actively involved in another listening task, he/she will immediately turn attention to this speech signal. A stream will be created and the sound will be identified as words. Because of the interesting information embedded in speech, top-down attention will rise. As long as nothing else happens, attention will stay with the speech signal, even when bottom-up inhibition eventually overwhelms bottom-up activation. As a second example, consider a person not particularly interested in car sounds. The sound of a first car passing will probably activate bottom-up attention to the extent that attention is drawn. Now the sound is less interesting and top-down activation will not be activated strongly. Inhibition will increase and, after some time, will become higher than bottom-up activation. Attention is released from the car noise stream. Nevertheless, activation is still there as long as the sound is there and the stream continuous to exist. As a final example, consider a person trying to hear a particular environmental sound, for example the sound of a bird. In this case, top-down attention is high from the start. It could be expected that the non-existing stream is generated from memory even before the sound actually occurs. As soon as the sound actually occurs, it has a good chance to win the competition for attention.

Figure 2: Flow chart of part of the attention model.
This rather complex mechanism can quite easily be reduced to a form suitable for a simulation, mimicking human attention for environmental sound in the context of soundscape evaluation. Activation of attention (for an existing stream) is modeled as an exponential integral of saliency over time with different integration constants for increase and decrease. Inhibition also behaves exponentially with different constants for increase and decrease. It starts to increase as soon as attention is switched to the stream and decreases as soon as attention is drawn away from the stream. The bottom-up request for attention finally is modeled as the difference between activation and inhibition. The model depends on four time constants for which only rough estimates can be extracted from psychophysics literature. Fine tuning is based on the model behaving rational, but sensitivity to tuning is low.

Mechanisms pointing top-down attention to particular streams are more difficult to model. For the time being, we use a base value that is assumed to be relatively constant over time and which depends on intentions and activities of the modeled individual, combined with a function of the total amount of saliency accumulated during the time interval that attention actually is focused on this particular stream. The latter part is a very simplified representation of the interesting features embedded in the sound.

Figure 3 illustrates how the model works during a 10-minute sound excerpt for one particular set of model parameters (one modeled individual). In the upper graph, traffic intensity is very low, but traffic is close and thus noise and saliency peaks are high. Each peak activates bottom-up attention requests very strongly and most of the times these requests are granted. Thus attention goes to traffic noise every time a car passes, but it does not necessarily stay there very long. In the lower graph, traffic intensity is high, but the road is located at some distance. Saliency stays quite high all the time, but now the mechanism behaves differently. Due to the interplay of activation and inhibition, attention switches quasi-periodically to the traffic noise. Neither the occurrence of the periods of attention nor the periodicity strongly depends on car passages or even the number of cars. These quantities now depend more strongly on the time constants chosen for the calculation of activation and inhibition. We could interpret these results as if the modeled individual periodically pays attention to the sound that is continuously present, in order to verify its origin, but he/she does not spend much resource on this task. From a bio-acoustical point of view, this would be a rational strategy.

D. Source identification and meaning
At least for the function of constructing a mental image of the surrounding world, the meaning of the observed sounds seems important. Meaning given to a sound can be seen as the collection of associations the presence of this sound triggers in the listener's mind. For this the source of the sound must be identified. Because human memory and thoughts are highly context dependent, the associations that are triggered depend not only on the sound itself, but also on other observations or even thoughts occurring prior to the observation of that sound. Constructing a computational model that could give some idea about the meaning of a particular sound for a particular person at one instant in time is beyond reach at the moment. However, if one assumes that associations are constructed in people’s minds either by personal experience, by hearing or reading about the experience of others, or by logic deduction, there must be at least some common or cultural part in it. At the moment we are unable to accurately model even this cultural fraction.

Having some insight in the way meaning is associated to a sound nevertheless gives us an indication that attention is important. Recognizing a sound and associating it with other concepts is a demanding cognitive task that probably needs some attention focused on it. Even if the exact
meaning of a sound is not known, the fact that it receives attention can be seen as a prerequisite for attaching meaning and thus influencing the holistic evaluation of the soundscape. In a park environment, visitors prefer natural, human and then mechanical sounds. When categories of sounds that are generally believed to have a negative connotation such as road traffic noise are found in a park, one can safely assume that the time that attention is drawn to these sounds gives an indication on how strongly park visitors will take them into account and thus come to a negative evaluation. For this reason, the time that attention is given to a sound is considered as a model output.

Figure 3: 10-minute excerpt of the temporal variation of various quantities used in the model. Upper panel: park noise combined with traffic noise from a road at 20m carrying 20 vehicles/hour. Lower panel: park noise combined with traffic noise from a road at 200m carrying 300 vehicles/hour.
In this paper, the importance of saliency, attention focusing and sound source identification in soundscape perception is discussed. The main hypotheses of this work state that environmental sounds are used as cues for creating a mental image of the world, and that soundscape perception arises from the match or mismatch of sensory input with this mental representation. Based on these hypotheses, the structure of a model for assessing human perception of environmental sound is outlined. Simulation results of simplified models for saliency and attention focusing are discussed, and it is shown how such results could allow to gain insight in the mechanisms behind soundscape perception, and to draw conclusions on methods for soundscape design.

The modeling reported on in this proceedings paper concerns ongoing research. Extensive examples of applications have been omitted and will be given during the talk using an up-to-date version of the model.

ACKNOWLEDGMENTS

The authors would like to thank Birgitta Berglund and Mats E. Nilsson (Stockholm University & Karolinska Institutet) for the use of their data and useful discussion regarding the model.

REFERENCES

3. B. Schulte-Fortkamp, D. Dubois (editors), Recent advances in Soundscape research (Special Issue), Acta Acustica united with Acustica 92(6) (2006).