Informational masking and attention focussing on environmental sound

D. Botteldooren[1], B. De Coensel[2]

[1] *Ghent University, Belgium, Email: dick.botteldooren@intec.ugent.be*

[2] *Ghent University, Belgium (on leave at UC Berkeley, USA), Email: bert.decoensel@intec.ugent.be*

Introduction

It has been established that the appreciation of a soundscape depends on the cognitive and emotional evaluation of the sonic environment, with particular emphasis on the meaning associated with the perceived sound [1]. Before meaning can be attributed to a sound, it nevertheless has to be detected and recognised. Energetic masking due to the working of the inner ear has been mentioned as a reason for not hearing an environmental sound and thus proposed as a potential solution for improving negative soundscapes [2][3][4]. The central nervous system however also plays a crucial role. Informational masking may prohibit recognising or understanding the sound. In this context attention focussing is always around the corner. As attention gets focussed on an auditory stream, informational masking could be reduced even if the signal to noise ratio drops considerably after initial detection. Making practical use of this knowledge in the acoustic design of (urban) outdoor spaces is one of the long-term goals of soundscape research.

In this paper, a computational model for simulating (parts of) the perception of environmental sound is described. In its current form, the model mainly consists of two steps. Firstly, conspicuous (salient) sounds are discerned by computing an auditory saliency map [5], using an algorithm optimized for use with environmental sound. Secondly, consciously noticed sounds are discerned using a model for attention focussing and shifting, which implements a balance between top-down and bottom-up focussing to sounds with high saliency [6]. As an illustration, the model is used to study the ability of typical urban parks to mask road traffic noise.

Environmental noise perception and soundscape evaluation model

Auditory saliency of environmental noise

The human sensory systems are continuously exposed to huge amounts of stimuli. Because the brain can not fully process all these stimuli at once, neural mechanisms exist to select possibly important (i.e. salient) but small subsets of these stimuli for further processing [7][8]. Several computational models for auditory saliency have been proposed in the literature recently [5][9][10], usually having a structure largely based on analogous models for visual saliency [7]. 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. The timescales involved in soundscape evaluation, as well as the strong effect of personal factors, make it necessary to consider longer stretches of time (typically minutes to hours), and to average results over vast numbers of simulated individuals (typically hundreds to thousands). Therefore, a simplified model is developed based on the models given in [5][9][10], which nevertheless implements the key elements present in the calculation of complex auditory saliency maps.

Figure 1 shows the general layout of the saliency model. The model starts from the 1/3-octave band spectrograms of the different sounds that constitute the sonic environment. When spectrograms for different types of sounds are simulated (or possibly extracted from sound recordings), as is considered in this paper, the non-trivial problems of modeling auditory scene analysis [8] and sound source recognition are bypassed. A timestep of 1s is considered, which is much larger than the typical timesteps used in the models cited earlier (usually 10ms).

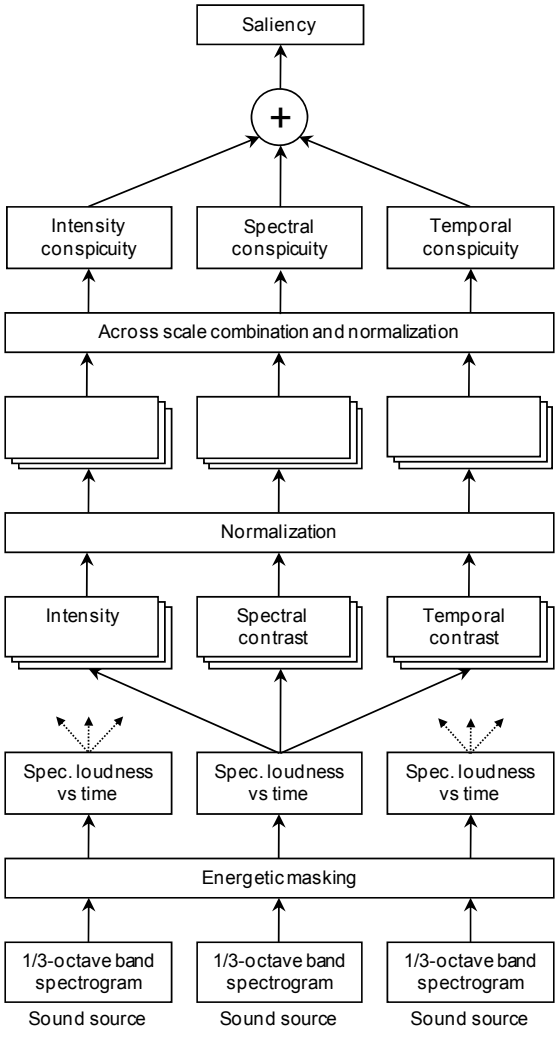

Figure 1: Structure of the saliency model for (simulated) environmental noise sources (adapted from [5][9][10]).

In a first stage, a simplified cochleagram is calculated from the 1/3-octave band spectrograms, using the Zwicker model for specific loudness [11]. Energetic masking is accounted for by considering, for each source, all other sources as the background. The specific loudness vs. time maps thus only contain non-zero values for those spectro-temporal parts of each source that are not energetically masked by the sum of all other sources.

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) [12]. 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 (subtraction between a "center" fine scale and a "surround" coarser scale), 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 [9][13], which strongly promotes maps that contain a small number of (conspicuous) peaks, and strongly supresses 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 [9].

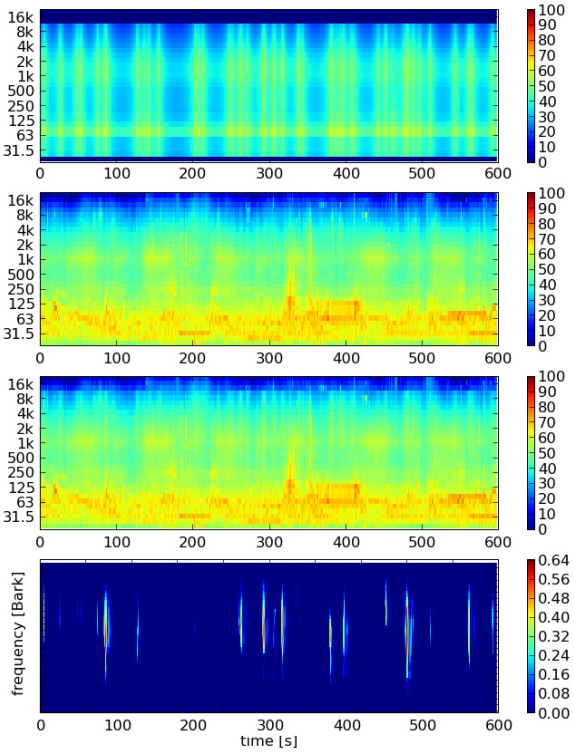

Figure 2: Example calculation of auditory saliency, with panels (from top to bottom): simulated traffic, park background, superposed spectrogram, saliency of traffic.

Figure 2 illustrates the saliency calculation algorithm. The upper panel shows a simulated 1/3-octave band spectrogram of 10 minutes of traffic noise, originating from a road at 30m, carrying a flow of 300 vehicles/hour, with cars travelling at an average speed of 70 km/h. The example uses the Harmonoise road traffic noise source model [14], and a simple geometric divergence propagation model. The second panel shows the spectrogram of a 10-minute recording of sound in an urban park. The third panel shows the superposition of both, as if the road was added next to the park. The fourth panel finally shows the saliency of the parts of the road traffic that are not masked by the sounds already present in the park.

Attention focussing and shifting

It is widely acknowledged that attention focussing is directed using both bottom-up, saliency based cues, and top-down, activity dependent cues [6][15]. The latter is guided by higher level cognitive processing, in which meaning is attached to the sounds within an experienced and expected context, and may depend on the current activity, personal states and traits (e.g. noise sensitivity), and the information content of the sound. To date, the basic knowledge needed to model this part of the complex problem is still lacking. To provide the essential higher-level cognitive information on top-down attention, a simplified feedback mechanism is used. The general layout of the attention focussing model is sketched in Figure 3.

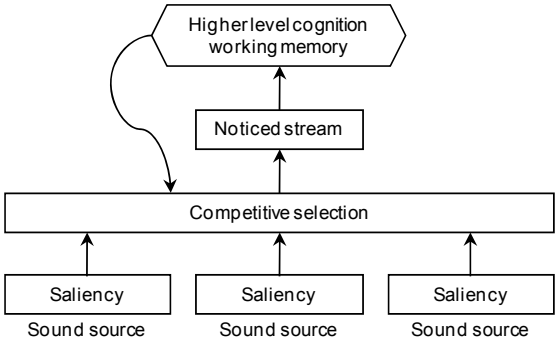

Figure 3: Structure of the attention focussing model for (simulated) environmental noise sources.

The calculated saliency triggers bottom-up attention. The competitive selection mechanism selects the current attended source, if any, using an activation and inhibition-of-return process. Activation is increased by saliency, while inhibition-of-return is increased by received attention (integrated over time). Both are increased in (saturating) steps, and decreased by an exponential decay, with inhibition slower (τ = 1000s) than activation (τ = 100s). The switch then selects at each timestep the attended source using a winner-takes-all mechanism. The top-down attention sets the threshold. Bottom-up requests for attention introduce steps that fade away with a relatively small time constant (τ = 10s). Non-sound attention takes part in this competition. The switch can be considered as a gate to further cognitive evaluation.

To illustrate the proposed model for attention focusing, let us again consider the example of the previous section. Figure 4 shows the calculation of the attended sources. The upper

panel shows the time-varying saliency for traffic and ambient noise (note that this time series does not exactly correspond to the saliency map of Figure 2, because of the stochastic nature of the road traffic noise simulation). The third and fourth panel show both bottom-up and top-down attention. Note that for top-down attention, a "thought" source is introduced, which aggregates all attention directed towards other modalities, and which now and then distracts the attention from sound. This seemed to be an essential requirement in making the model work. The bottom panel shows which sound is receiving the attention. Initially, the simulated individual attends the ambient sound, until a sufficient loud car passes, which receives attention. At about 50s, no sound sources are attended anymore, until at about 140s, a sufficiently salient ambient event attracts attention. A similar example is shown in Figure 5, but this time a park with a relatively quiet ambient sound is used. Clearly, the traffic noise is more salient in this case, and thus attracts more attention.

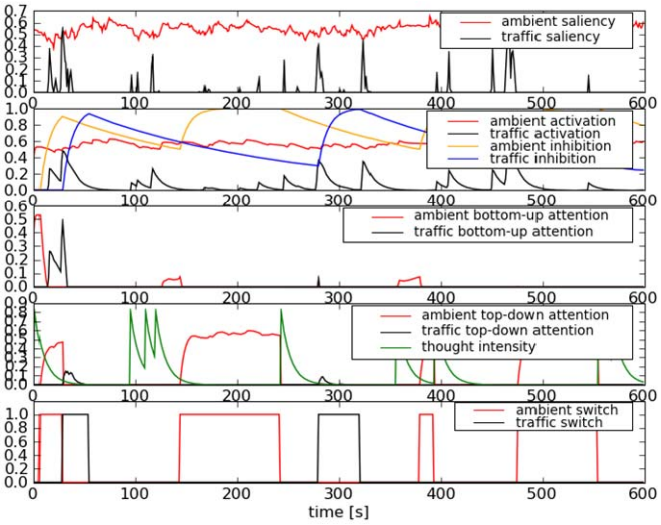

Figure 4: Example calculation of attention focussing, based on a park with loud ambient sound.

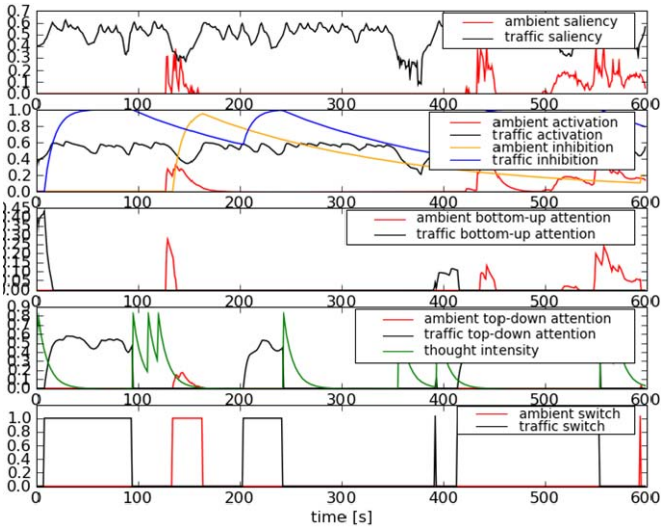

Figure 5: Example calculation of attention focussing, based on a park with quiet ambient sound.

In the model proposed thus far, it is assumed that saliency triggers attention for each stream separately. One could nevertheless argue that saliency attracts auditory attention before the source is recognised. Thus it should be based on the overall sound rather than on a particular stream. In Figure 6 the effect of calculating overall saliency and attributing it to the dominant source is shown. Although saliency takes different values, the strength of the activation and inhibition model avoids the whole system to suffer strongly from this difference in approach.

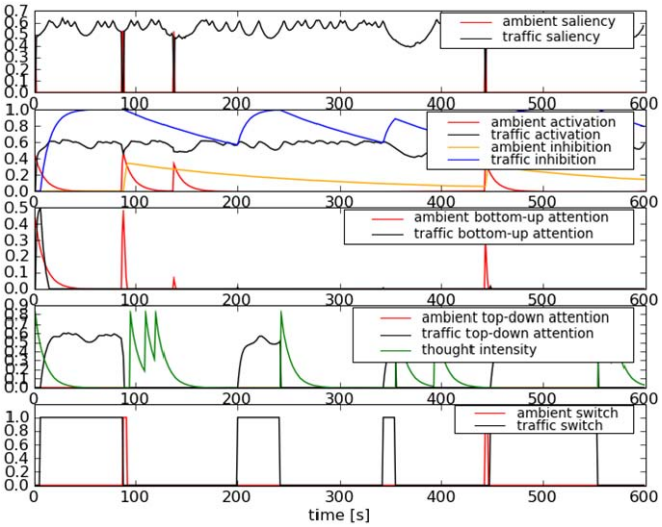

Figure 6: Example calculation of attention focussing, based on a park with quiet ambient sound (saliency before stream segregation).

Traffic masking ability of typical park sonic environments

To investigate how the model discussed above could be used to quantify the ability of typical park sounds to mask traffic noise, a virtual experiment is conducted. First the sonic environment in 16 parks is recorded, resulting in over 1000 sound fragments of 10 minutes [16]. To simulate the effect on the soundscape of additional traffic noise, several traffic noise situations (Table 1) are modelled and added to the recorded sounds. The result is fed to the model.

Parameter	Range
traffic intensity	10 – 1000 veh/h
traffic speed	25 – 75 km/h
distance	10 – 500 m

Table 1: Modelled road traffic noise situations.

Figure 7 summarizes the time that attention is paid to traffic noise over all traffic situations and all park ambient sounds. Results are shown as a function of signal to noise ratio of traffic noise vs. park ambient, expressed as 10-minute L_{Aeq}. Once the traffic noise level is less than 10 dBA below the ambient level, the traffic noise gradually gets attention more frequently. The gradual increase seems more realistic than a crisp masking / no masking decision. However, to fully prove that the proposed model can deliver more information than a fuzzy model based on L_{Aeq} differences, different park

sounds have to be compared (Figure 8). It is observed that the masking capability is different for different "quiet park" sounds: in park 1 and 9 traffic sound more easily gets attention.

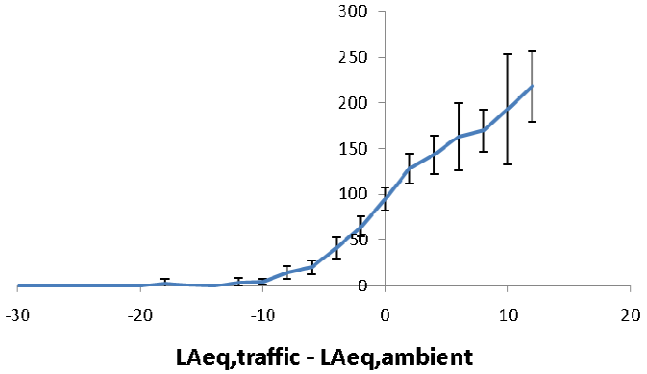

Figure 7: Time that traffic noise receives attention as a function of signal to noise ratio in L_{Aeq}.

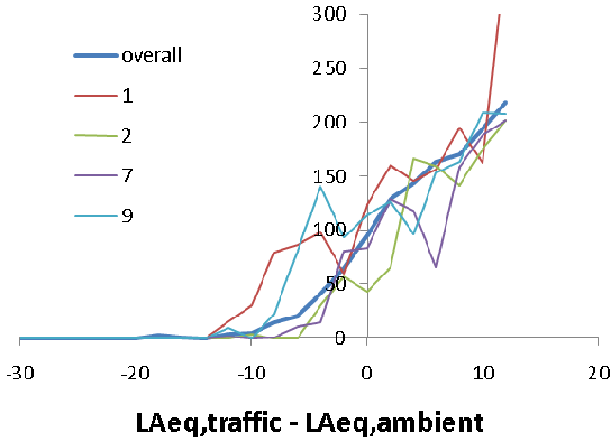

Figure 8: Time that traffic noise receives attention as a function of signal to noise ratio in L_{Aeq} for a selection of quiet parks.

Conclusions

Simplified models of human processing of combined environmental sound exposure are used to simulate the appreciation of a variety of sonic environments. These simulations allow us to better understand how soundscapes are constructed in the head of the human listener. In particular we showed how perceptual masking could work in addition to energetic or physiological masking to improve the mental image of a sonic environment. This paper also illustrates a potential use of the models that we have been building as a tool to predict the perception and attention focussing on (unwanted) sounds that are added to existing park environments.

References

[1] D. Dubois, C. Guastavino, M. Raimbault, "A cognitive approach to urban soundscapes: using verbal data to access everyday life auditory categories", *Acta Acustica united with Acustica* **92** (2006), 865-874.

[2] G. W. Siebein, Y. Kwon, P. Smitthakorn, M. A. Gold, "An acoustical palette for urban design", *Proceedings of ICSV13*, Vienna, Austria (2006).

[3] J. Kang, "A systematic approach towards intentionally planning and designing soundscape in urban open public spaces", *Proceedings of Internoise'07*, Istanbul, Turkey (2007).

[4] A. L. Brown, A. Muhar, "An approach to the acoustic design of outdoor space", *Journal of Environmental Planning and Management* **47** (2004), 827-842.

[5] C. Kayser, C. I. Petkov, M. Lippert, N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map", *Current Biology* **15** (2005), 1943-1947.

[6] J. B. Fritz, M. Elhilali, S. V. David, S. A. Shamma, "Auditory attention – focusing the searchlight on sound", *Current Opinion in Neurobiology* **17** (2007), 437-455.

[7] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", *IEEE Tranactions on Pattern Analysis and Machine Intelligence* **20** (1998), 1254-1259.

[8] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sounds*, The MIT Press, London (1990).

[9] O. Kalinli, S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech", *Proceedings of Interspeech '07*, Antwerp, Belgium (2007).

[10] V. Duangudom, D. V. Anderson, "Using auditory saliency to understand complex auditory scenes", *Proceedings of EUSIPCO '07*, Poznan, Poland (2007).

[11] E. Zwicker, H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin, Germany (1999).

[12] S. Shamma, "On the role of space and time in auditory processing", *Trends in Cognitive Sciences* **5** (2001), 340-348.

[13] L. Itti, C. Koch, "Feature combination strategies for saliency-based visual attention systems", *Journal of Electronic Imaging* **10** (2001), 161-169.

[14] H. G. Jonasson, "Acoustical source modelling of road vehicles", *Acta Acustica united with Acustica* **93** (2007), 173-184.

[15] E. I. Knudsen, "Fundamental components of attention", *Annual Reviews Neuroscience* **30** (2007), 57-78.

[16] M. E. Nilsson, D. Botteldooren, B. De Coensel, "Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas", *Proceedings of ICA'07*, Madrid, Spain (2007).