Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

A model of saliency-based auditory attention to environmental sound

Bert De Coensel and Dick Botteldooren

Department of Information Technology, Ghent University, Ghent, Belgium

PACS: 43.50.Qp, 43.50.Rq, 43.66.Ba

ABSTRACT

A computational model of auditory attention to environmental sound, inspired by the structure of the human auditory system, is presented. The model simulates how listeners switch their attention over time between different auditory streams, based on bottom-up and top-down cues. The bottom-up cues are determined by the time-dependent saliency of each stream. The latter is calculated on the basis of an auditory saliency map, which encodes the intensity and the amount of spectral and temporal irregularities of the sound, and binary spectro-temporal masks for the different streams. The top-down cues are determined by the amount of volitional focusing on particular auditory streams. A competitive winner-takes-all mechanism, which balances bottom-up and top-down attention for each stream, determines which stream is selected for entry into the working memory. Consequently, the model is able to delimit the time periods during which particular streams are paid attention to. Although the main ideas could be applied to all types of sound, the implementation of the model was targeted at environmental sound in particular. As an illustration, the model is used to reproduce results from a detailed field experiment on the perception of transportation noise. Finally, it is shown how this model could be a valuable tool, complementing auralization, in the design of outdoor soundscapes.

INTRODUCTION

During recent years, there has been a growing awareness that sound forms an integral part of the urban environment, and that it should be considered at the same level of importance as visual aesthetics in the urban planning and design process [1–5]. Consequently, there is a need for models and techniques for acoustic design, that specifically account for aspects of spatial scale, and for the nature of sound sources in the urban environment. Including auditory aspects and knowledge on human perception of environmental sound in the urban planning and design process, an approach often referred to as *soundscape design*, has great potential. The key idea is that sounds can be considered as resources for improving the quality of the acoustic environment, and not just as a waste to be managed. An example of this approach is the use of natural sounds to mask unwanted sounds, such as those from surface transportation, in the design of urban parks or squares [3, 6].

Recent findings in psychophysics and neurophysiology strongly emphasize the important role of selective *auditory attention* in perceiving the complex acoustic environment to which we are exposed [7–9]. Humans have a great proficiency in *auditory scene analysis* (ASA)—decomposing the mixture of incoming sounds from different sources into individual auditory streams, based on a combination of auditory and visual cues [10]. Auditory attention allows us to focus our mental resources on a particular stream of interest while ignoring others, differentiating foreground from background. This stream is then analyzed in detail in working memory, and its information may be used for making decisions and taking actions [11]. On a longer time scale, the sounds to which we pay attention will contribute to the creation of a mental image of our acoustic environment, and ultimately will shape our perception of its quality.

Given the importance of attention in the perception of environmental sound, it seems like a natural step to consider the possibilities of using specific sounds for *informational masking* of unwanted sounds [12, 13]. While it could be unfeasible to energetically mask all unwanted sound, particular sounds could still be used to distract attention from unwanted sound as much as possible. The goal of this paper is to present a computational model of auditory attention, which can be used in the design of acoustic environments. The model simulates how listeners switch their attention over time between different sounds, and can be applied to complement auralization by replacing the listener. Because of the large time scales involved in the perception of environmental sound, and because of the huge variation between listeners, compromises between biological accuracy and computational efficiency are inevitable. The model is aimed to be valid on a statistical basis, rather than on an individual basis, and will be of functional rather than of neurobiological nature. Nevertheless, the model will still be firmly rooted in available knowledge.

In the next section, a short overview of the literature on auditory attention is given, summarizing the theoretical and empirical foundation for the model, without going into much detail on the neurobiological basis (more information can be found in the cited references). Subsequently, the model is presented, and parameters are estimated on the basis of a listening experiment on environmental sound. The model in this paper builds upon different ideas presented in earlier work [14–16].

AUDITORY ATTENTION

General overview

Auditory attention can be defined as "the cognitive process underlying our ability to focus on specific aspects of the acoustic environment, while ignoring others" [9]. More in particular, auditory attention is responsible for selecting the information that is to be processed in more detail. Central in most theories on attention (visual as well as auditory) is the interplay of *bottom-up* (saliency-based) and *top-down* (voluntary) mechanisms in a *competitive selection* process [7, 11].

The bottom-up mechanism selectively gates incoming auditory information, enhancing responses to stimuli that are conspicuous. This is accomplished by a sophisticated novelty detection system, that continuously monitors the acoustic environment for changes in frequency, intensity, duration or spatial location of stimuli [17]. This pre-attentive mechanism operates rapidly and independent of the nature of the particular task which the listener may be performing. The top-down mechanism focuses processing resources on the auditory information that is most relevant for the current goal-directed behavior of the listener. This mechanism is guided by information already held in working memory, through sensitivity control, in which the relative strengths (signal-to-noise ratios) of different information channels that compete for access to working memory are regulated [11]. Examples are directing eye movement (for visual attention), changing the orientation of the head, or even modulating the sensitivity of the neural circuits that process the information. The selection of information for entry into working memory is found to be a highly competitive, hierarchically structured process [11, 18]. At low hierarchical levels, competition occurs within neural representations of basic sound parameters, such as frequency or temporal structure; at higher levels, competition occurs between different auditory streams. Finally, at the interface with working memory, competition occurs between information from the different senses. At each level, the stimulus with the highest relative strength is selected (combining the effects of bottom-up saliency and top-down bias), in a winner-takes-all fashion. The process of voluntary selective attention involves working memory, sensitivity control and competitive selection operating in a recurrent loop [11], and may prohibit involuntary switching of attention to task-irrelevant distractor sounds [19], leading to the cognitive benefits that are associated with attention.

In a simplifying manner, ASA is often regarded as a two-stage analysis-synthesis process [10, 20]. In the first stage (segmentation), the acoustic signal is decomposed into a collection of time-frequency (T-F) segments. In the second stage (grouping), segments that are likely to have arisen from the same environmental source are combined into auditory streams. Traditionally, it has been assumed that the perceptual mechanisms behind this process are largely pre-attentive: only after auditory streams are formed, they can become an object of attention [8, 10]. Although this view is appealing because of its conceptual simplicity, recent findings suggest that attention also plays a role in the formation of auditory streams [21]. Overall, it can be stated that ASA draws on low-level principles for segmentation and grouping, but is fine-tuned by selective attention [7]. Nevertheless, the interplay between the processes of ASA and attention remains the focus of intensive research.

Models of auditory attention

Several computational models of auditory attention have been proposed in the literature recently [17, 22–25]. Most of these models focus on bottom-up attention, and have a structure that is largely based on similar models for bottom-up visual attention, of which the one by Itti and Koch [26] is probably the most well-known. Central to most models of auditory attention is the calculation of an *auditory saliency map* [27]. This map provides a weighted representation of the acoustic environment, emphasizing elements that are conspicuous and thus likely to be detected. The calculation of this map follows a general structure (see Figure 1, yellow) that is common to most models.

First, a T-F representation or *spectrogram* of the (usually monaural) acoustic input is calculated, from which a number of low-level features are extracted in parallel, mimicking the information processing stages in the central auditory system. Different sets of receptive filters at varying spectral and temporal scales are applied to quantify intensity, spectral and temporal contrast (some models also consider pitch [24] or spectro-temporal orientation [23, 24]). Subsequently, center-surround differences across scales within each feature are calculated, mimicking the properties of local cortical inhibition [17]. The resulting feature maps have to be normalized, because they represent non-comparable modalities, having different dynamic ranges. Most models also apply a nonlinear amplification step as part of this normalization, before combining maps. Because many maps have to be combined, salient elements in one map risk being masked by noise or other less salient elements in other maps. The nonlinear amplification algorithm simulates competition between neighboring salient locations in each map, promoting peaks while supressing background noise [28]. The normalized feature maps are then combined (added) across scales within each single feature, and are again normalized into so called conspicuity maps. Finally, the auditory saliency map is computed by combining the conspicuity maps.

Selective visual attention is often compared to a stagelight [29], sequentially illuminating different parts of the visual scene for further analysis. Computational models for bottom-up visual attention apply the saliency map as a base for the selection of locations for successive attentional focus. An important factor in this dynamic process is *inhibition-of-return* (IOR), which prevents attention from permanently focusing on the most salient location in the map. This can be modelled in a biologically plausible way by feeding the saliency map into a 2D layer of leaky integrate-and-fire neurons [26]. The potential of neurons coupled with a salient location in the map will increase faster, and when one neuron reaches its threshold charge, it fires and its accumulated charge is shunted to zero. This 2D layer of neurons can in turn be coupled to a 2D winner-takes-all neural network, implementing a neurally distributed maximum detector. The combination of both networks will naturally generate an attentional scanpath over time. The existence of a similar IOR effect in auditory processing has received experimental support (see e.g. [30, 31]). However, as far as the authors are aware, IOR has not been implemented in any computational model of auditory attention.

As with auditory attention, ASA has also been studied extensively by computational means (see [20] for an overview). The goal of computational models for ASA can be defined as the estimation of a T-F mask: a weighting of a T-F representation of the acoustic environment, such that T-F units that are dominated by a particular (target) stream are emphasized, and units that are dominated by other streams are suppressed [20]. Often, binary masks are used, motivated by the phenomenon of masking in auditory perception. In essence, this T-F mask does not contain any information about which stream will be paid attention to—most models have been designed with applications for speech processing in mind, and are as such only interested in separating speech (the target stream) from background noise. Note the complementarity with the saliency map, which emphasizes the T-F units that are most likely to be the subject of auditory attention, without containing any information about the attribution of units to auditory streams.

COMPUTATIONAL MODEL

In the following paragraphs, the above ideas are worked out mathematically, with the main goal of obtaining a model that can be used in soundscape design. Formally, the model takes as input the sound signals $x_i(t)$ present at the location of the listener, originating from N sound sources, which are considered to be the objects of the design process. How these sound signals are obtained/rendered, is not a subject of this work; when the design of outdoor acoustic environments is considered, this will probably involve simulating the dynamic behavior of a series of

sound sources (vehicles passing by, sound from nature, music etc.), coupled with detailed modelling (auralization) of sound propagation from source to listener location [32]. The model has as output an auditory attention switching function $X_i(t)$, which returns 1 if source i is paid attention to at time t, and 0 otherwise. When integrated over time, $X_i(t)$ may be regarded as a measure for the potential of source i to attract attention.

For reasons of computational efficiency, the model will be of functional rather than of neurobiological nature. Considering the field of application and the large timescales involved in the perception of environmental sound (e.g., when distraction by train passages is assessed, it may be necessary to consider several hours of sound), the use of detailed auditory processing models is not feasible. Next to this, a number of conceptual simplifications are made. First, the model only accounts for monaural sound, disregarding the influence of spatial cues on attention. Second, the model does not distinguish between sound sources and auditory streams. More in particular, the T-F content of each auditory stream is assumed to be directly related to the T-F content of the source signal in the auralized mixture (in the following, the terms stream and source are considered equivalent). Because the model does not solve the problem of ASA, it is, in its current form, not readily applicable to arbitrary recordings of environmental sound containing a mixture of sources. A second consequence of this assumption is that the model disregards the influence of attention on the formation of auditory streams, i.e. ASA is considered to be pre-attentive.

The computational model is comprised of four stages, illustrated in Figure 1: peripheral auditory processing, the calculation of a saliency map, the derivation of a time-varying saliency score associated with each auditory stream, and finally, the simulation of auditory attention switching.

Peripheral auditory processing

In the first stage, a T-F representation of the acoustic input is derived (Figure 1, red). The total sound wave $x(t) = \sum x_i(t)$ is filtered with a gammatone filterbank, modelling the frequency selectivity of the basilar membrane:

$$y_f(t) = (x * g_f)(t) = \int_{-\infty}^{+\infty} x(u)\, g_f(t-u)\, \mathrm{d}u, \tag{1}$$

with g_f the gammatone filter of order n with center frequency f (in Hz), given by

$$g_f(t) = t^{n-1} e^{-2\pi b(f)t} \cos(2\pi f t + \phi)\, \mathscr{H}(t). \tag{2}$$

In Eq. (2), $\mathscr{H}(t)$ denotes the Heaviside step function, ϕ denotes the phase and $b(f)$ denotes the equivalent rectangular bandwidth (ERB, in Hz) [33] at center frequency f:

$$b(f) = 1.019\,(24.7 + 0.108 f). \tag{3}$$

For $n = 4$, the gammatone filter has been found to fit well to experimentally derived estimates of human auditory filter shapes. Finally, in order to estimate the auditory nerve response, the output of each filter is rectified, integrated and compressed logarithmically into a spectrogram:

$$s(t,f) = 10 \log_{10} \left[\frac{1}{\Delta t} \int_{t-\Delta t}^{t} y_f^2(u)\, \mathrm{d}u \right], \tag{4}$$

in which a temporal resolution Δt = 10 ms is used. In a similar way, the spectrograms $s_i(t,f)$ of all constituent sounds (auditory streams) are calculated as well.

To perform the convolution in the time domain, an efficient digital implementation of the gammatone filterbank is used [34]. A total of 128 filters are considered, with center frequencies distributed evenly on the ERB scale, defined as [35]

$$f^{\mathrm{ERB}} = 21.4 \log_{10}(0.00437\, f^{\mathrm{Hz}} + 1), \tag{5}$$

spanning the full audible range f^{Hz}= 20 Hz to 20 kHz. This corresponds to a spectral resolution $\Delta f^{\mathrm{ERB}} \approx 0.32$. In the following discussion, all frequencies are expressed on the ERB scale; for readability, the ERB label is dropped from here on.

Auditory saliency map

The calculation of the saliency map largely follows the scheme presented in [24] (see Figure 1, yellow), with the major adjustment that spectro-temporal orientation and pitch are not considered. For reference, we will give a brief overview. In a first stage, sets of raw feature maps are extracted at varying spectral and temporal scales, by convolving the spectrogram with filters that mimic the receptive fields in the auditory cortex:

$$r_{\alpha,\vartheta}^{\delta,\lambda}(t,f) = (s * h_{\alpha,\vartheta}^{\delta,\lambda})(t,f), \tag{6}$$

in which $h_{\alpha,\vartheta}^{\delta,\lambda}$ is the 2D Gabor filter, defined as [36]

$$h_{\alpha,\vartheta}^{\delta,\lambda}(t,f) = \exp\left[-\frac{t'^2 + f'^2}{2\delta^2} \right] \cos^{\alpha}(2\pi t'/\lambda) \tag{7}$$

with

$$\begin{aligned} t' &= \frac{t}{\Delta t}\cos\vartheta + \frac{f}{\Delta f}\sin\vartheta \\ f' &= -\frac{t}{\Delta t}\sin\vartheta + \frac{f}{\Delta f}\cos\vartheta \end{aligned} \tag{8}$$

The intensity filters mimic the receptive fields with only an excitatory phase ($\alpha = 0$, $\vartheta = \pi/2$; shorthand: $\mu = 1$). The spectral contrast filters mimic the receptive fields with an excitatory phase and simultaneous symmetric inhibitory side bands ($\alpha = 1$, $\vartheta = \pi/2$; $\mu = 2$). The temporal contrast filters mimic the receptive fields with an excitatory phase and subsequent inhibitory phase ($\alpha = 1$, $\vartheta = 0$; $\mu = 3$), because only the past is considered. The raw feature maps are extracted using the filters described above on eight scales $\sigma = \{1,\ldots,8\}$, with $\delta = 2^{\sigma-1}$ and $\lambda = 3 \cdot 2^{\sigma-1}$. We will use the shorthand $r_{\mu}^{\sigma}(t,f)$ from here.

Subsequently, center-surround differences are calculated from the raw maps obtained at different T-F scales, followed by rectification:

$$d_{\mu}^{\sigma_c,\Delta\sigma}(t,f) = |r_{\mu}^{\sigma_c}(t,f) - r_{\mu}^{\sigma_c+\Delta\sigma}(t,f)|, \tag{9}$$

in which $\sigma_c \in \{2,3,4\}$ and $\Delta\sigma \in \{3,4\}$. In total, 3 × 6 = 18 feature maps are computed, which are then normalized using an iterative nonlinear algorithm $\mathscr{N}(\cdot)$ that simulates competition between neighboring salient locations [24, 28]. Each feature map is first scaled to the range $[0,1]$ to eliminate the difference in dynamic range between the different modalities and scales. Then, each iteration consists of a self-excitation and inhibition induced by neighbors, implemented by convolving each map with a 2D difference-of-Gaussians (DoG) filter, and clamping the negative values to zero. Formally, a feature map d is transformed in each iteration step as follows:

$$d \leftarrow |d + d * \mathrm{DoG} - 0.02|_{\geq 0} \tag{10}$$

The details of the DoG filter shape can be found in [24, 28]. The normalized feature maps are then combined across scales, into conspicuity maps for each single feature:

$$q_{\mu}(t,f) = \sum_{\sigma_c=2}^{4} \sum_{\Delta\sigma=3}^{4} \mathscr{N}\left(d_{\mu}^{\sigma_c,\Delta\sigma}(t,f) \right). \tag{11}$$

Finally, the auditory saliency map is computed by combining the normalized conspicuity maps:

$$S(t,f) = \frac{1}{3} \sum_{\mu=1}^{3} \mathscr{N}\left(q_{\mu}(t,f) \right). \tag{12}$$

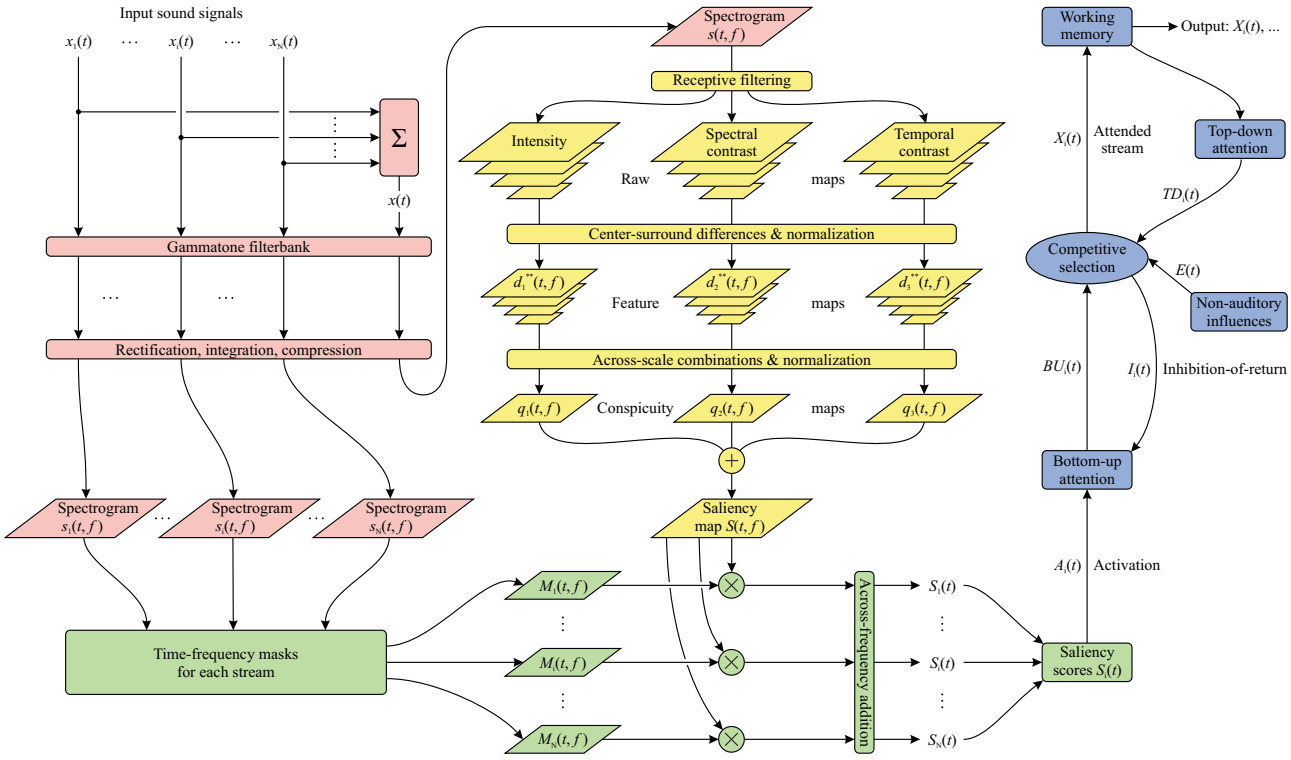

Figure 1: Structure of the model for auditory attention: (red) peripheral auditory processing; (yellow) calculation of a saliency map (adapted from [17, 24]); (green) derivation of time-varying saliency scores; (blue) simulation of auditory attention switching.

Stream-specific saliency scores

T-F masks for all streams can be calculated easily on the basis of their spectrograms $s_i(t,f)$. One option is to use ratio masks [37, 38], which return the ratio between the energy attributed to a particular stream and the energy attributed to all other streams (i.e. a time-varying Wiener filter):

$$M_i^R(t,f) = \frac{10^{s_i(t,f)/10}}{\sum_{j=1}^{N} 10^{s_j(t,f)/10}} \tag{13}$$

Another option is to consider binary T-F masks, which return 1 if the T-F unit centered around t and f is dominated by stream i, and 0 otherwise (i.e. sound i is energetically masked by the mixture of all other sounds):

$$M_i^B(t,f) = \begin{cases} 1 & \text{if } M_i^R(t,f) > T, \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

in which usually $T = 0.5$. Binary masks are a good choice if there is little overlap in spectro-temporal content between the different sound sources, which is often the case for speech and music [37], but less for broadband environmental sounds. Irrespective of the choice of T-F mask type, the saliency score for stream i can be calculated as

$$S_i(t) = \int M_i(t,f)\, S(t,f)\, \mathrm{d}f. \tag{15}$$

This procedure is visualized in Figure 1, green. Eq. 15 assumes that saliency combines additively across frequency channels [24].

Auditory attention switching

The attention submodel is mainly based on the functional model described by Knudsen [11], and implements an interplay between bottom-up and top-down influences in a winner-takes-all competition. The model can best be illustrated using a few examples. As a first example, assume someone starts talking. Initially, salient features of the speech signal will attract attention, and the bottom-up mechanism is activated. If the listener is not actively involved in another listening task, he/she will immediately turn attention to this speech stream, and will indentify the words. Because of the interesting information embedded in speech, the top-down mechanism will be activated. As long as nothing else happens, attention will keep being focused on the speech stream. As a second example, consider a person not particularly interested in car sounds. The sound of a car passing by will probably activate the bottom-up mechanism to the extent that attention is drawn to the passage. Now the sound is less interesting, and the top-down mechanism will not be activated strongly. The IOR mechanism will, after some time, cause attention to be released from the car passage stream. As a third example, consider a person trying to hear a particular environmental sound, for example the sound of a bird. In this case, top-down influence will be high from the start. As soon as the sound actually occurs, it has a good chance to win the competition for attention.

The above described mechanisms can quite easily be reduced to a form suitable for simulation, mimicking human attention switching to environmental sound (see Figure 1, blue). Activation of the bottom-up mechanism for a particular stream may be modelled as a leaky integration of saliency over time:

$$A_i(t) = \int_0^t S_i(u) \exp\left[\frac{u-t}{\tau_A}\right] \mathrm{d}u, \tag{16}$$

with time constant τ_A. As part of the implementation, different time constants for increase and decrease are used ($\tau_A^{\nearrow}$, $\tau_A^{\searrow}$). IOR is also modelled to behave exponentially. It starts to increase as soon as attention is drawn to a particular stream, and decreases from the moment attention is drawn away from the stream:

$$I_i(t) = \int_0^t X_i(u) \exp\left[\frac{u-t}{\tau_I}\right] \mathrm{d}u \tag{17}$$

with $X_i(t)$ as defined in the beginning of this section. Again, different time constants τ_I for increase and decrease may be used. The bottom-up mechanism is then modelled as the difference between activation and IOR:

$$BU_i(t) = |A_i(t) - I_i(t)|_{\geq 0} \tag{18}$$

The top-down mechanism is more difficult to model, as it relies both on the information encoded in the attended stream, and on the intentions and activities of the modelled individual. As a simplified representation of this mechanism, the former may be modelled as a leaky integration of saliency over the time period that the stream is actually paid attention to, while the latter may be modelled as a (potentially time-varying) bias term β_i:

$$TD_i(t) = \int_0^t X_i(u)\, S_i(u) \exp\left[\frac{u-t}{\tau_T}\right] \mathrm{d}u + \beta_i(t) \tag{19}$$

Finally, the attended source at time t is decided according to both bottom-up and top-down influences:

$$X_i(t) = \begin{cases} 1 & \text{if } i = \arg\max_k \left[BU_k(t) + TD_k(t)\right], \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Together, Eqs. 19 and 20 form a positive feedback loop, modelling voluntary selective attention. However, this mechanism may result in attention being focused continuously on a particular stream, as long as other streams do not become too salient. In reality, attention may be drawn away from the sound by non-auditory influences (visual cues, thoughts etc.). Therefore, for the model to behave in a realistic way, a non-auditory stream has to be included in the competitive selection process of Eq. 20. The total attention $E(t)$ of this stream will depend, among other things, on the activity of the modelled individual. When no information regarding non-auditory influences is available, $E(t)$ may be modelled as a sequence of peaks that fade away exponentially with time, and that are distributed over time according to a Poisson or $1/f$ distribution with rate ρ_E.

Figure 2 illustrates how the attention switching submodel works. Two streams are considered: road traffic noise (individual passages are considered part of the same auditory stream) and ambient noise (e.g. birds singing). Figure 2(a) shows the (simulated) saliency scores of both sounds; the attended stream is shown at the bottom. The first two car passages receive attention, but the IOR mechanism causes the third passage to not receive attention. Some attention is spent on listening to the ambient noise, but attention to non-auditory objects now and then kicks in, and even causes the car passage at ca. 475 s to not be noticed.

The submodels for peripheral auditory processing, for the calculation of a saliency map and for calculating stream-specific saliency scores are largely based on work by others. Consequently, the parameters of these submodels have been taken from literature; it is assumed that these parameters do not vary too much between (normal hearing) individuals. The submodel for auditory attention switching contains 5 internal parameters: the time constants τ_A (2×, for increase and decrease), τ_I (idem) and τ_T. For the model to behave in a meaningful way, these parameters are subject to some constraints. First, activation and IOR are modelled as saturation processes, for which $\tau_A^{\nearrow} < \tau_A^{\searrow}$ and $\tau_I^{\nearrow} < \tau_I^{\searrow}$. Second, for the IOR mechanism to work, time constants have to be larger than those for activation, i.e. $\tau_A^{\nearrow} < \tau_I^{\nearrow}$ and $\tau_A^{\searrow} < \tau_I^{\searrow}$. Third, for the top-down mechanism to work in a smooth way, such that events occurring shortly after each other are lumped into a single period of attention, it is necessary that $\tau_T < \tau_I^{\nearrow}$. Next to this, external influences are incorporated in the top-down bias terms β_i for each stream, and the rate ρ_E of

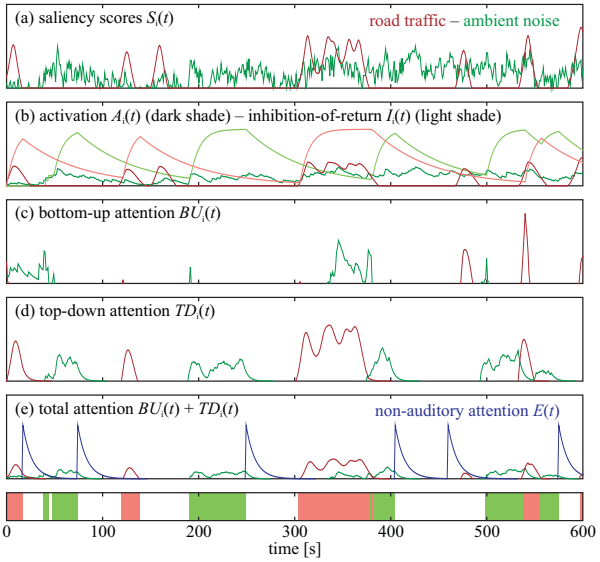

Figure 2: Excerpt of the temporal course of various quantities (arbitrary units) used in the attention switching submodel ($\tau_A^{\nearrow}$ = 1.5 s; $\tau_A^{\searrow}$ = 10 s; $\tau_I^{\nearrow}$ = 10 s; $\tau_I^{\searrow}$ = 60 s; τ_T = 5 s).

non-auditory attention shifts. These parameters are expected to be highly variable between individuals, and will strongly depend on context.

PARAMETER ESTIMATION

The model presented in the previous section (in particular, the attention switching submodel) contains a number of parameters, for which values are not readily available in literature, especially not if the extent of interindividual differences are to be fully considered. Consequently, values will have to be estimated on the basis of carefully designed experiments. One important methodological caveat, as pointed out by Fritz *et al.* [7], is that it is notoriously difficult to precisely measure the selectivity, intensity and/or duration of auditory attention. A commonly accepted, quantifiable measure of attention is still lacking. Nevertheless, most studies infer the presence/absence of auditory attention from performance on well designed tasks, or from neuroimaging data [7]. Here, we will take a first step by applying the model to reproduce the results of a field experiment. The original aim of this experiment was to investigate potential differences in annoyance in an at-home context, caused by the noise of various types of surface transportation: road traffic with different traffic intensities, conventional trains, high-speed trains and trains based on magnetic levitation. First, we will perform a sensitivity analysis of the model, in order to find its most important parameters. Subsequently, we will give a brief description of the experiment; the methodology and results have already been presented in detail in [39]. In particular, we will limit the description to those aspects of the experiment that are relevant to the current discussion. Finally, we will show how our model can be used to replicate the results of this experiment, and how the parameters of the model are estimated in this process.

Sensitivity analysis

In order to find those parameters that are most responsible for the variation in the model output, a sampling-based sensitivity analysis is carried out. A fixed set of two auditory streams is considered (traffic noise and ambient noise; an excerpt is shown in Figure 2), for which a 5-hour timeseries of saliency scores is simulated. Model parameters are randomly sampled from the intervals shown in Table 1, according to a uniform distribution.

The parameter intervals were chosen wide enough, in order to encompass a wide variance in model behavior, from no attention at all to the foreground traffic noise, up to almost continuous attention. In total, 10^4 cases are considered. It is found that the rate ρ_E of non-auditory attention events has the largest influence on the total duration that the foreground traffic noise is paid attention to (58.8 % of variance explained), followed by $\tau_I^{\searrow}$ (7.3 % of variance explained) and τ_T (3.7 % of variance explained). All parameters were found to have a significant influence on the model output, except for $\tau_A^{\searrow}$ ($p > 0.05$). About 30 % of the variance in attention (the remainder from 100 % in Table 1) is due to the stochasticity in the model, in particular the distribution of non-auditory events over time. Although the activation and IOR mechanisms are essential parts of the model, changes in the parameters $\tau_A^{\nearrow}$, $\tau_A^{\searrow}$ and $\tau_I^{\nearrow}$ have only a small influence on the model outcome, and therefore we may safely assign fixed values to these parameters.

Table 1: Ranges for the parameters in the sensitivity analysis, together with the fraction of total variance explained.

Parameter	Interval		Variance
	Min	Max	explained
$\tau_A^{\nearrow}$	Δt	10 s	0.3 %
$\tau_{A_{\searrow}}^{\nearrow}$	$\tau_{A_{\nearrow}}^{\nearrow}$	60 s	0.0 %
$\tau_{I_{\searrow}}^{\nearrow}$	$\tau_{A_{\nearrow}}^{\nearrow}$	10 s	0.1 %
$\tau_I^{\searrow}$	$\max(\tau_I^{\nearrow}, \tau_A^{\searrow})$	300 s	7.3 %
τ_T	Δt	$\tau_I^{\nearrow}$	3.7 %
ρ_E	$0/h$	$300/h$	58.8 %

Field experiment

Subjects. One hundred participants were selected to be representative of the Dutch population, for criteria such as age, gender, education and noise sensitivity. An invitation to participate was sent to 1500 persons, living within short distance of the experiment site, together with a questionnaire. Participants were selected by comparing answers with distributions taken from a recent Dutch nation-wide environment survey that included the same questions [39].

Stimuli. A wide range of traffic noises were used, including passages of conventional high speed trains at approx. 140 and 300 km/h, Dutch intercity trains (approx. 140 km/h) and Maglev trains (approx. 200, 300 and 400 km/h), all passing by at distances of 25, 50, 100, and 200 m. In addition, sounds from a highway and from local roads at the same distances were also included. All experimental sounds were recorded in the field at the stated distances from the source track or road. Subsequently, a series of 10-minute stimuli were composed, consisting of 2 or 4 passages of the same train type at the same distance and speed, or alternatively, of continuous highway/road traffic noise. As these stimuli were played back through loudspeakers placed outdoors (see below), we refer to these as the "outdoor stimuli".

Realistic setting. The experiment was conducted in an ecologically valid setting: participants were seated in the living room of an actual house, and transportation noise was reproduced through loudspeakers placed outdoors, that were not visible from inside the living room. During the experimental sessions, participants were free to engage in light daily activities with varying cognitive demands, such as reading a magazine, having something to drink or holding a conversation. In contrast to typical laboratory experiments on noise annoyance, the participants were not asked specifically to focus attention to the sounds played back, although these sounds could distract them from their activity. The playback equipment was placed outside the experimental house, and was calibrated in such a way that playing back the stimuli would give the same levels and spectral content (full hearable spectrum) at the façade as if the house would be located at the stated distances from the track or road. During the experiment, the sound inside the living room was recorded using a binaural head and torso simulator seated among the participants. These 10-minute "indoor stimuli" exactly match the stimuli played back outside the house as heard indoors, but also include the sounds made by the (activities of the) participants themselves.

Procedure. Four to six participants jointly participated in a session. The overall structure of the experiment was identical for each group of participants: 14 stimuli of 10-minute duration were presented, with a break after the first 7 stimuli (in total 20 sessions were organized, resulting in $14 \times 20 = 280$ unique indoor stimuli). At the end of each stimulus, the participants were asked to write down how annoyed they were by the sound during the past 10 minutes. The method of free-number magnitude estimation was used: participants were asked to use a number on a relative scale (e.g. if one is twice as much annoyed by a subsequent stimulus, one had to use the double of the previous number), with the condition to use zero if they were not annoyed at all by the sound. Before the start of both series of 7 stimuli, a short training session was held, which helped the participants to define their own scaling context, and more importantly allowed every participant to produce individual reference functions to be used for calibrating their annoyance scales [40]. The empirically derived individual reference functions were then used to transform the free-number magnitude estimations for each individual to the corresponding annoyance values in units of a common master scale, making the annoyance values comparable across subjects.

Virtual experiment

For listening tests in which short sound fragments are presented under laboratory conditions, and in which participants are asked to pay attention to the sound itself, noise annoyance is often found to be mainly related to perceptual properties (mainly loudness) of the presented sounds. However, several authors have pointed out that when noise annoyance is assessed in realistic conditions, the interference of the noise with the task at hand, or with the daily activity when an at-home context is considered, is essential [41, 42]. Because of the particular procedure applied in the experiment described above, it can therefore be expected that the participants mainly considered the amount of disturbance that the sound produced in their activity, when scaling their annoyance, especially because the intruding sound (traffic noise) may have a negative connotation. There is clear evidence that the presence of irrelevant information (sounds) degrades selective attention, impairing the performance on the task at hand [43], because this irrelevant information may also take part in the competition for entrance in working memory. Therefore, we could expect a certain amount of correlation between the annoyance values in the above experiment, and the amount (duration) of attention that was paid to the transportation sounds.

Synthetic population of subjects. For each participant in the field experiment, we model an additional "virtual participant", that is subjected to the same exposure (indoor stimuli as recorded by the artificial head and torso) as the original participant. The total duration of attention that these modelled individuals pay to the transportation noise can then be compared to the reported annoyance values of the actual participants. It is assumed that the modelled individuals have no a priori top-down bias to any sound ($\beta_i = 0$). The other parameters of the attention switching submodel ($\tau_I^{\searrow}$, τ_T and ρ_E in particular) are subject to the estimation procedure described below. In order to cope with the variance in output caused by the model stochasticity, results are averaged over 5 simulation runs for each modelled individual.

Auditory saliency of indoor stimuli. The indoor stimuli can be regarded as consisting of the superposition of the sound originating from the outdoor stimuli, and the sound originating from all other ambient sources. The latter will give rise to a multitude of auditory streams, e.g. the speech from different participants will obviously give rise to distinct auditory streams. However, because of practical considerations in separating these sources in the T-F domain, it is only feasible to consider two streams in this virtual experiment: the transportation noise, and all other ambient sound. Although separate sound signals are not available, it is still possible to calculate binary T-F masks, and consequently, saliency scores for both streams, as illustrated in Figure 3. Binary masks are estimated based on the spectrogram of the outdoor stimuli, accounting for the insulation of the experimental house (in particular, the transfer function between the loudspeakers outside of the house, and the right ear of the artificial head). The saliency map is calculated on the basis of the spectrogram of the sound recorded by the artificial head. The sound sources in the example fragment of Figure 3(c), which give rise to peaks in the saliency map, are, in chronological order: turning pages in a magazine, peaks of somebody sniffing his nose (2×), clicking of ballpoint pen, stimulus (train passage), people talking on a subject not related to the passage of the train. The lower panel of Figure 3 shows the time-varying saliency scores for the train passage (the early peak is due to the short rise time) and the other ambient sounds.

Estimation of model parameters. In general, two strategies can be applied to estimate the model parameters. A first strategy is to find those model parameter values that best describe the results of the field experiment. A single set of parameter values, common to all modelled individuals, is optimized to achieve maximum correlation between the attention paid to the transportation noise in the virtual experiment and the annoyance values of the field experiment. This strategy considers all modelled individuals to be identic, representing the "average" participant to the field experiment. The parameters of this average participant can then be used when the model is being applied. A second, more elaborate strategy is to use a separate set of parameters for each participant, and to optimize these parameters to achieve a full correspondence (except for a constant scaling factor) between modelled attention and annoyance values. This way, parameter distributions are obtained, rather than fixed values. Consequently, when the model is to be applied, one has to consider a population consisting of a large number of individuals, and to draw samples from the appropriate distributions for each parameter (see e.g. [16] for an example of this approach). It has to be stressed that, irrespective of the strategy followed, the model will only be valid on a statistical basis, rather than on an individual basis. As an illustration, we will follow the first strategy in this paper.

There is almost no correlation between the measured $L_{\mathrm{Aeq,10min}}$ of the indoor stimuli, and the annoyance values ($r^2 = 0.045$), obviously because the former also includes the sound produced by the participants themselves. On the other hand, the saliency scores only consider those T-F units that originate from the stimuli, and account for spectro-temporal irregularities instead of only sound pressure level. When the total saliency scores of the transportation noise over time $\int S(t)\,\mathrm{d}t$ are considered, the correlation increases ($r^2 = 0.180$). Subsequently, a brute-force search was carried out to find those values for the parameters $\tau_I^{\searrow}$, τ_T and ρ_E of the attention submodel that result in the highest correlation with the annoyance values. The results are given in Table 2. For these parameter values, a correlation $r^2 = 0.383$ between duration of attention and annoyance values is obtained. For comparison, the correlation between the outdoor façade $L_{\mathrm{Aeq,10min}}$ (mainly containing the sound from the stimuli) and the annoyance values, on an individual basis, is $r^2 = 0.267$.

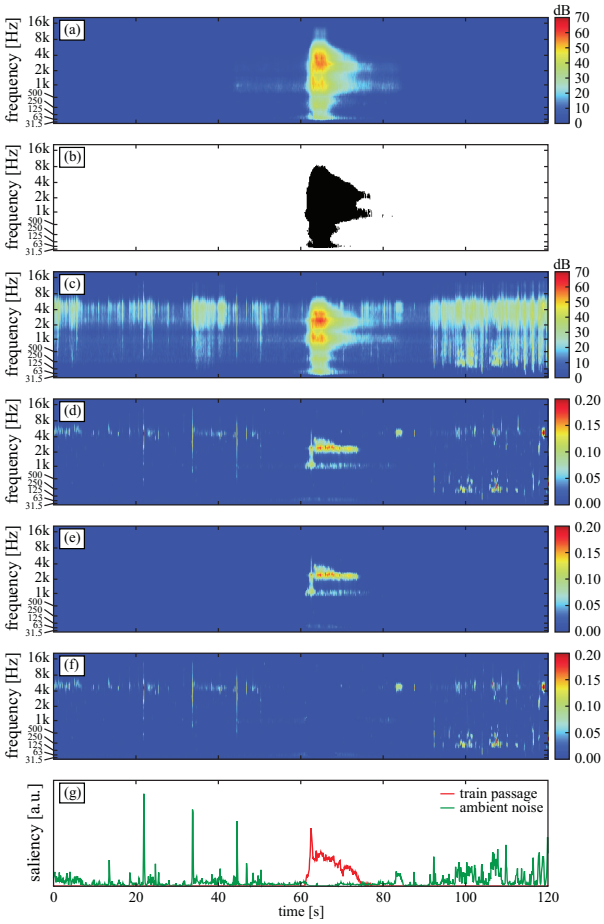

Figure 3: Illustration of the extraction of saliency scores from the sound recorded by the artificial head placed among the participants: (a) spectrogram of the stimulus (passage of a high speed train with a speed of 300 km/h, at a distance of 50 m), as heard inside the house (i.e. corrected for the insulation of the house); (b) binary mask extracted from the stimulus, using a fixed threshold of 15 dB; (c) spectrogram of the sound recorded by the artificial head; (d) saliency map of the head recording; (e) saliency attributed to the train passage; (f) saliency attributed to the other sounds; (g) saliency scores for the train passage and the other sounds.

Table 2: Optimal parameter values.

$\tau_A^{\nearrow}$	$\tau_A^{\searrow}$	$\tau_I^{\nearrow}$	$\tau_I^{\searrow}$	τ_T	ρ_E
1 s*	5 s*	5 s*	40 s	1.5 s	$15/h$

*Fixed values were assigned.

DISCUSSION AND CONCLUSIONS

A computational model of auditory attention was presented in this paper, aimed to be used in soundscape design, as a tool to assess the potential of specific sounds for informational masking of unwanted sounds. The model is based on a set of simulated sound signals at the location of the listener, coming from various modelled sound sources, and delimits the time periods during which particular sounds are paid attention to. The model makes it possible to conduct virtual listening experiments; however, no meaning is attached to the sounds, and it is left to the soundscape designer to decide if particular sounds fit in a given context or not. Furthermore, the model does not consider auditory stream segregation; auditory streams are assumed to be equivalent with the input sound signals, representing the various sound sources in the mixture. As a result, the model

is more suited to be applied in design and simulation, rather than to analyze arbitrary recordings of environmental sound containing a mixture of sources. A second simplification is that the influence of spatial cues on attention is disregarded.

In order to cope with the long timescales associated with the perception of environmental sound—which maks it necessary to consider sound signals with a sufficiently long duration—the model is of functional rather than of neurobiological nature. Because of the huge variance between listeners, the model is only valid on a statistical basis. The computational complexity of the model is, for large part, determined by the temporal and spectral resolution considered, and these settings thus provide some opportunity for simplification. If speech is to be considered, lowering the temporal resolution may not be a good idea, but it could be argued that for general, environmental sounds, a lower resolution (0.1 s or even 1 s, see e.g. [44] for some motivation) could be sufficient for saliency calculations. For lowering the spectral resolution, one could replace the 128-channel gammatone spectrogram by a 1/3-octave band spectrogram, possibly complemented with a time-varying loudness calculation [15].

The model contains a number of parameters, for which (distributions of) values may, for a given context and nature of sounds, be estimated on the basis of carefully designed listening experiments. The field experiment described in this paper forms a first step in validating the model, but more detailed experiments, in which the sound exposure and the tasks/activities of the participants are much more controlled, will be necessary. The presented early results suggest that a meaningful calibration of the model should be possible. It has to be noted that any given set (distribution) of model parameters can only be valid for a single combination of context and sound source nature. For example, the amount of top-down bias for particular sounds will depend on the particular goals and expectations of the listener, tied to its particular location (e.g., people visiting a park may be spending more attention to bird songs, a priori).

ACKNOWLEDGMENTS

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

REFERENCES

[1] B. Hellström, *Noise Design: Architectural Modelling and the Aesthetics of Urban Acoustic Space*. PhD thesis, School of Architecture, Royal Institute of Technology, Stockholm, Sweden, Sept. 2003.
[2] A. L. Brown and A. Muhar, "An approach to the acoustic design of outdoor space," *J. Environ. Plann. Manage.*, vol. 47, no. 6, pp. 827–842, 2004.
[3] J. Kang, "A systematic approach towards intentionally planning and designing soundscape in urban open public spaces," in *Proc. Inter·noise*, (Istanbul, Turkey), Aug. 2007.
[4] M. Adams, B. Davies, and N. Bruce, "Soundscapes: an urban planning process map," in *Proc. Inter·noise*, (Ottawa, Canada), Aug. 2009.
[5] B. De Coensel, A. Bockstael, L. Dekoninck, D. Botteldooren, B. Schulte-Fortkamp, J. Kang, and M. E. Nilsson, "The soundscape approach for early stage urban planning: a case study," in *Proc. Inter·noise*, (Lisbon, Portugal), June 2010.
[6] 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.*, vol. 127, no. 3, pp. 1357–1366, 2010.
[7] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention — focusing the searchlight on sound," *Curr. Opin. Neurobiol.*, vol. 17, no. 4, pp. 437–455, 2007.
[8] E. S. Sussman, J. Horváth, I. Winkler, and M. Orr, "The role of attention in the formation of auditory streams," *Percept. Psychophys.*, vol. 69, no. 1, pp. 136–152, 2007.
[9] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLoS Biol.*, vol. 7, no. 6, p. e1000129, 2009.
[10] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts, USA: The MIT Press, 1994.
[11] E. I. Knudsen, "Fundamental components of attention," *Annu. Rev. Neurosci.*, vol. 30, pp. 57–78, 2007.
[12] C. S. Watson, "Some comments on informational masking," *Acta Acust. Acust.*, vol. 91, no. 3, pp. 502–512, 2005.
[13] N. Durlach, "Auditory masking: Need for improved conceptual structure," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 1787–1790, 2006.
[14] D. Botteldooren and B. De Coensel, "The role of saliency, attention and source identification in soundscape research," in *Proc. Inter·noise*, (Ottawa, Canada), Aug. 2009.
[15] B. De Coensel, D. Botteldooren, B. Berglund, and M. E. Nilsson, "A computational model for auditory saliency of environmental sound," *J. Acoust. Soc. Am.*, vol. 125, no. 4, p. 2528, 2009.
[16] B. De Coensel, D. Botteldooren, T. De Muer, B. Berglund, M. E. Nilsson, and P. Lercher, "A model for the perception of environmental sound based on notice-events," *J. Acoust. Soc. Am.*, vol. 126, no. 2, pp. 656–665, 2009.
[17] C. Kayser, C. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Curr. Biol.*, vol. 15, no. 21, pp. 1943–1947, 2005.
[18] A. Baddeley, "Working memory: looking back and looking forward," *Nat. Rev. Neurosci.*, vol. 4, no. 10, pp. 829–839, 2003.
[19] E. Sussman, I. Winkler, and E. Schröger, "Top-down control over involuntary attention switching in the auditory modality," *Psychon. Bull. Rev.*, vol. 10, no. 3, pp. 630–637, 2003.
[20] D. Wang and G. J. Brown, eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2006.
[21] R. Cusack, J. Decks, G. Aikman, and R. P. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis," *J. Exp. Psychol.–Hum. Percept. Perform.*, vol. 30, no. 4, pp. 643–656, 2004.
[22] S. N. Wrigley and G. J. Brown, "A computational model of auditory selective attention," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1151–1163, 2004.
[23] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," in *Proceedings of the 15th European Signal Processing Conference (EUSIPCO 2007)*, (Poznań, Poland), pp. 1206–1210, Sept. 2007.
[24] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. Interspeech 2007*, (Antwerp, Belgium), pp. 1941–1944, Aug. 2007.
[25] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 5, pp. 1009–1024, 2009.
[26] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
[27] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.
[28] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electron. Imaging*, vol. 10, no. 1, pp. 161–169, 2001.
[29] G. Sperling and E. Weichselgartner, "Episodic theory of the dynamics of spatial attention," *Psychol. Rev.*, vol. 102, no. 3, pp. 503–532, 1995.
[30] C. Spence and J. Driver, "Auditory and audiovisual inhibition of return," *Percept. Psychophys.*, vol. 60, no. 1, pp. 125–139, 1998.
[31] D. J. Prime, M. S. Tata, and L. M. Ward, "Event-related potential evidence for attentional inhibition of return in audition," *NeuroReport*, vol. 14, no. 3, pp. 393–397, 2003.
[32] B. De Coensel, T. De Muer, I. Yperman, and D. Botteldooren, "The influence of traffic flow dynamics on urban soundscapes," *Appl. Acoust.*, vol. 66, no. 2, pp. 175–194, 2005.
[33] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception* (Y. Cazals, L. Demany, and K. Horner, eds.), pp. 429–446, Oxford, UK: Pergamon, 1992.
[34] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Tech. Rep. 35, Apple Computer, Inc., 1993.
[35] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, no. 1–2, pp. 103–138, 1990.
[36] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.
[37] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, vol. 51, no. 3, pp. 230–239, 2009.
[38] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501, 2006.
[39] B. De Coensel, D. Botteldooren, B. Berglund, M. E. Nilsson, T. De Muer, and P. Lercher, "Experimental investigation of noise annoyance caused by high-speed trains," *Acta Acust. Acust.*, vol. 93, no. 4, pp. 589–601, 2007.
[40] B. Berglund, "Quality assurance in environmental psychophysics," in *Ratio Scaling of Psychological Magnitudes — In Honor of the Memory of S. S. Stevens* (S. J. Bolanowski and G. A. Gescheider, eds.), Hillsdale, New Jersey, USA: Erlbaum, 1991.
[41] K. Zimmer, J. Ghani, and W. Ellermeier, "The role of task interference and exposure duration in judging noise annoyance," *J. Sound Vib.*, vol. 311, no. 3–5, pp. 1039–1051, 2008.
[42] D. S. Michaud, S. E. Keith, and D. McMurchy, "Annoyance and disturbance of daily activities from road traffic noise in Canada," *J. Acoust. Soc. Am.*, vol. 123, no. 2, pp. 784–792, 2008.
[43] S. P. Banbury, W. J. Macken, S. Tremblay, and D. M. Jones, "Auditory distraction and short-term memory: phenomena and practical implications," *Hum. Factors*, vol. 43, no. 1, pp. 12–29, 2001.
[44] K. Kaliski and D. Joyce, "Characterizing soundscapes using spectrograms from long-term third octave band data," in *Proc. Inter·noise*, (Ottawa, Canada), Aug. 2009.