A computational model of auditory attention for use in soundscape research

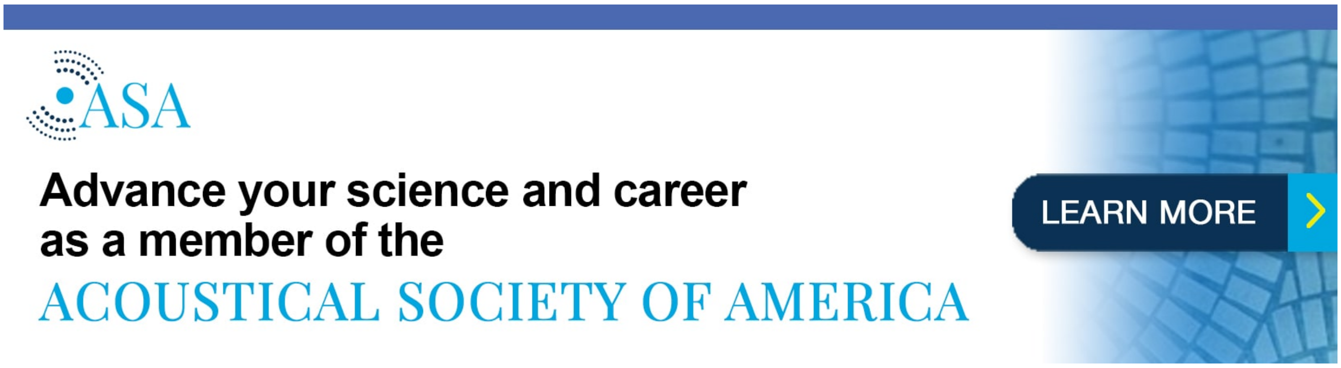
Damiano Oldoni, Bert De Coensel, Michiel Boes, Michaël Rademaker, Bernard De Baets, Timothy Van Renterghem, and Dick Botteldooren

Citation: The Journal of the Acoustical Society of America **134**, 852 (2013); doi: 10.1121/1.4807798

View online: https://doi.org/10.1121/1.4807798

View Table of Contents: https://asa.scitation.org/toc/jas/134/1

Published by the Acoustical Society of America



A computational model of auditory attention for use in soundscape research

Damiano Oldoni, Bert De Coensel,[a] and Michiel Boes
Acoustics Research Group, Department of Information Technology, Ghent University, St.-Pietersnieuwstraat 41, B-9000 Ghent, Belgium

Michaël Rademaker and Bernard De Baets
Research Unit Knowledge-based Systems, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, B-9000 Ghent, Belgium

Timothy Van Renterghem and Dick Botteldooren
Acoustics Research Group, Department of Information Technology, Ghent University, B-9000 Ghent, Belgium

(Received 30 January 2012; revised 17 April 2012; accepted 23 April 2012)

Urban soundscape design involves creating outdoor spaces that are pleasing to the ear. One way to achieve this goal is to add or accentuate sounds that are considered to be desired by most users of the space, such that the desired sounds mask undesired sounds, or at least distract attention away from undesired sounds. In view of removing the need for a listening panel to assess the effectiveness of such soundscape measures, the interest for new models and techniques is growing. In this paper, a model of auditory attention to environmental sound is presented, which balances computational complexity and biological plausibility. Once the model is trained for a particular location, it classifies the sounds that are present in the soundscape and simulates how a typical listener would switch attention over time between different sounds. The model provides an acoustic summary, giving the soundscape designer a quick overview of the typical sounds at a particular location, and allows assessment of the perceptual effect of introducing additional sounds.

© *2013 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4807798]

PACS number(s): 43.50.Rq, 43.60.Np, 43.50.Yw, 43.66.Ba [NX] Pages: 852–861

I. INTRODUCTION

Sound is an integral part of the urban environment, and there is a growing awareness that acoustical aspects should be considered at the same level of importance as architecture and visual esthetics in urban planning and the design of urban outdoor spaces.[1–3] For example, it has been shown that easy access to nearby outdoor (green) spaces for public amenity, such as urban squares and parks, leads to important positive effects on stress restoration[4] and general well-being[5,6] of urban residents. In order to create this kind of urban spaces, environments that are of high acoustic quality, it is essential that auditory aspects and knowledge on human perception of environmental sound are included during the urban planning and design process. The goal of the soundscape designer is to compose acoustic environments that are as much as possible pleasing to the ear. More in particular, this means creating spaces in which the sounds that the listener identifies as desired in that context are often heard, while undesired sounds remain mostly hidden to the human ear, or at least are not noticed by the user of the space. This approach obviously goes beyond noise abatement and the striving for silence, and as such, there is a growing need for new models and techniques for soundscape analysis and design, well grounded in human auditory perception.

In this paper, a human-mimicking computational model for soundscape analysis is presented, which combines a self-organized map of acoustical features with a functional model of auditory attention. The model classifies the sounds that are present in the soundscape over time and simulates how listeners would switch their attention over time between different sounds. As such, it can be used within the soundscape design process to assess the influence of soundscaping measures (e.g., adding desired or removing undesired sounds) in the field. Next to this, the model involves constructing an acoustic summary through extensive training, tuning the model to the typical sounds that are heard at a particular location. The latter could be used to quickly provide an overview of a specific soundscape for the soundscape designer.

Auditory scene analysis has already been studied extensively by computational means (see Wang and Brown[7] for an overview). The ultimate goal of most of these models is to extract clean sound samples for individual components of the auditory scene, e.g., for separating speech from background noise. The ultimate aim of the present model is to mimic human evaluation of the sonic environment. In contrast to these previous models, it does not aim at extracting sounds that are as clean as technically possible, but at analyzing the scene as accurately as a human listener would. However, as the model is aimed to be integrated in equipment for long-term outdoor sound measurement, it presents a compromise between biological accuracy and computational efficiency. Furthermore, because of the huge variation between listeners, the model is aimed to be valid on a

[a] Author to whom correspondence should be addressed. Electronic mail: bert.decoensel@intec.ugent.be

statistical basis, rather than on an individual basis. It has to be noted that the model, in its present form, does not involve the automated labeling of classified sounds; rather, it simulates how a soundscape will be analytically perceived by the listener. This work refrains from designing methodologies for identifying which sounds are desired in a given environment with a particular use.[8,9]

In Sec. II, a short overview of the literature on auditory scene analysis, attention, and masking is given, summarizing the empirical foundation for the model, without going into much detail on the neurobiological basis. In Sec. III, a detailed formulation of the model is presented. In Sec. IV, a case study that illustrates the use of the model as a tool in soundscape design is presented. Finally in Sec. V, conclusions and perspectives for future research follow. The work described in this paper builds upon different ideas presented in earlier works.[10–14]

II. EMPIRICAL BACKGROUND

A. Analyzing the auditory scene

Outdoor acoustic environments are usually composed of a wide range of sounds that often overlap in time or frequency. Humans have a great proficiency in disentangling this mixture of incoming sounds into coherent perceptual representations of objects (called auditory streams), usually related to individual sound sources, based on a combination of auditory and visual cues. In a simplifying manner, this process of auditory scene analysis is often regarded as a two-stage analysis-synthesis process.[15] In the first stage (segmentation), the acoustic signal is decomposed into a collection of time-frequency 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.[16,17] 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.[18,19] Overall, it can be stated that the process of auditory scene analysis draws on low-level principles for segmentation and grouping but is fine-tuned by selective attention.[20]

B. Detecting and identifying a sound

Some sounds, although present in the auditory scene, will not be detected; no matter how hard the listener tries, these sounds remain masked. Masking effects have been widely studied using artificial sounds, such as sequences of tones or broadband noises,[21] or using speech,[22] but basic research on auditory masking of environmental sound is lacking.[23] Two types of masking are generally distinguished:[24,25] energetic and informational masking. Energetic masking concerns competing sounds (maskers) overlapping in time and frequency such that parts of one sound (the target) are rendered inaudible. Informational masking regards difficulties to detect a target sound which cannot be accounted for by interfering energy patterns at the peripheral auditory system but are caused by auditory mechanisms at higher levels of processing. An example of the latter is the inability to separate elements of the target sound from elements of the masker sound, due to similarity between the target and the masker.[26]

At this point it is useful to distinguish between detecting and identifying environmental sounds. Detecting a sound means that the listener can observe that a sound is present. Identifying a sound means that the listener can attach meaning to the sound (such as, but not necessarily, attach a linguistic label to the sound), based on prior knowledge. For simple sounds such as pure tones, detecting is almost equal to identifying, but for speech and environmental sound this is not the case. It has been shown that the meanings attributed to sounds act as a determinant for soundscape quality evaluations,[27,28] and therefore identification of sounds is an important factor in the context of soundscape design; sounds that are not identified are expected to influence overall soundscape appraisal to a lesser degree.

Detectability of a particular target sound within a soundscape is expected to depend on the spectral characteristics of both the target sound and the background sound, as can be concluded based on previous research on energetic masking. However, one should keep in mind that both target and background sound may exhibit considerable temporal variations. For example, the use of water sounds for masking road traffic noise in urban parks has recently gained some scholarly interest.[23,29] Reducing the detectability of road traffic noise to 10% of the time by adding water sound might therefore require water sound with an equivalent level up to 10 dB(A) above the equivalent level of road traffic noise. The model of Glasberg and Moore[30,31] summarizes the knowledge on (partial) loudness due to energetic masking, and may be used to quantify the audibility of time-varying sounds in the presence of background sound.

Leech *et al.*[32] and Gygi and Shafiro[33] investigated the particular characteristics of a sound that allow it to be identified in familiar auditory background scenes. The signal-to-noise ratio between the sound and the background noise was found to be the most important factor in their studies. They also found that contextual congruency between the sound and the background noise plays a role, in the sense that sounds that are not readily expected within a given environment are more easily identified. They could not prove that this was due to potential similarities in acoustic features of background noise and congruent target sounds.

Identifying a sound not only involves the ear but also the brain, and both have their limitations. It can be expected that information content plays a role: the more information is embedded in a sound, the easier it will be to detect it. In the experiment by Gygi and Shafiro,[33] some of the physical components of the sounds that made them more identifiable were standard deviation of the spectrum and the number of bursts or peaks. Both characteristics are related to the information content of a sound, and both make the target sound less likely to be masked completely. More generally, it can be expected that identifying a sound within a complex auditory scene also depends on how many unique features the sound has. For example, broadband noises are less likely to be identified than vocalizations that contain a rich variety of tones and tonal fluctuations.

Furthermore, familiarity of the listener with the sound to be detected makes it easier for the listener to detect it.[34] This mechanism could work for desired as well as for undesired sounds. Sensitivity to particular acoustical features of a sound are learned in early childhood, but new sounds can be learned at all ages.[35] Once sounds become familiar, they are identified more easily. It must be noted that learning effects are not limited to high-level associative memory. Several neurophysiological studies have reported on the capacity for holding memory traces (enduring neural records) in the primary auditory cortex (see Weinberger[36] for an extensive review). In particular, the number of neurons of the representational area of a sound is tuned by its importance[37] and the bigger the area, the stronger the memory effects.[38] Neurophysiological correlates of cognitive processes such as selective attention,[39,40] expectancy,[41] concept formation[42] and cross-modality effects[43] have been found in the primary auditory cortex, suggesting that due to neuronal plasticity, the primary auditory cortex is not merely an acoustic analyzer, but an adaptive auditory problem solver.[36] Another important property of the auditory cortex is tonotopy: neurons next to each other are typically excited by similar stimuli. Tonotopic maps have been observed in the auditory cortex of animal species such as cats[44] and monkeys.[45,46] The human cortex also contains several topologically ordered regions,[47–49] similar to regions observed in the macaque monkey brain.[49] Based on the above, it should be clear that a human-mimicking computational model for soundscape analysis will have to take into account the tonotopic mapping in the auditory cortex and incorporate continuous learning effects.

C. Paying attention to a sound

Although a particular sound within the acoustic environment may be hearable if one listens to it, this does not imply that one actually has to. Users of the space may not notice the sound because they are performing tasks—auditive or not—or are involved in activities that require their attention. On a longer time scale, the sounds that we consciously notice will contribute to the creation of a mental image of the acoustic environment at a location, and ultimately will shape our perception of its quality. As such, not noticing a sound can be positive if the sound is not part of the acoustic design, while it is negative if the sound is considered a unique soundmark[50] of the location.

Auditory attention allows us to focus our mental resources on specific aspects of the acoustic environment, while ignoring all other aspects.[51] More in particular, the auditory attention mechanism is responsible for selecting the information that is to be processed in more detail in working memory, and thus that may be used for making decisions and taking actions.[52] It is an essential mechanism in human input processing, as it avoids sensory overload. Central in most theories on attention (visual as well as auditory) is the interplay of bottom-up (saliency-based, depending on the characteristics of the stimulus) and top-down (voluntary, depending on the state of the listener) mechanisms in a competitive selection process.[20,52]

The bottom-up mechanism selectively enhances responses to sounds that are conspicuous, for example, because they have rare or novel physical features, or are of instinctive biological importance. This is accomplished by a novelty detection system that continuously monitors the acoustic environment for changes in frequency, intensity, duration, or spatial location of stimuli.[53,54] This pre-attentive mechanism operates rapidly and independently of the nature of the particular task that the listener may be performing. In contrast, 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 of different information channels that compete for access to working memory are regulated.[52] Examples are directing eye movement or changing the orientation of the head, or modulating the sensitivity of the neural circuits that process the information. Finally, the selection of information for entry into working memory is found to be a competitive, hierarchically structured process.[55] At low hierarchical levels, competition occurs within neural representations of basic sound parameters; at higher levels, competition occurs between different auditory streams; 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 bottom-up and top-down effects), in a winner-takes-all fashion. This is why selective attention is often compared to a stagelight,[56] sequentially illuminating different parts of the scene for further analysis. An important factor in this process is inhibition-of-return[57,58] (IOR), which prevents attention from permanently focusing on the most salient components of the scene, naturally generating an attentional scanpath over time. The process of voluntary selective attention involves working memory, sensitivity control and competitive selection operating in a recurrent loop,[52] and may prohibit involuntary switching of attention to task-irrelevant distractor sounds.[59]

It is difficult to determine whether a particular sound within the acoustic environment is noticed or not, using psychophysical experiments. Simply asking people about the sound may point their attention towards it and make them notice the sound. In laboratory conditions, biophysical measures such as event-related potentials (ERP) can be used to assess the influence of attention to sounds during the performance of various non-auditory tasks.[60,61] Such research suggests that effective orientation of attention toward particular sounds is influenced by a wide range of top-down, personal factors: the prior experience of the listener with the sound and the significance of the sound to the listener, the listener's intentions and activities, its emotional state[62,63] or even a possible genetic component.[64,65] Next to this, the emotional cues carried by a sound also affect the degree to which it captures attention. Unpleasant sounds are known to attract human attention more than neutral sounds,[66] even when the peak sound amplitudes are similar.[67]

The empirical knowledge on human auditory processing (auditory scene analysis, masking, detection of sounds, tonotopic representation, learning, and auditory selective attention) summarized above, will now form the basis for the construction of a human-mimicking computational model of auditory attention, described in detail in the following paragraphs.

III. COMPUTATIONAL FRAMEWORK

A. General considerations

The proposed computational model for analyzing outdoor soundscapes takes as input the sound signal recorded by a microphone at a particular location and has as output a measure of the potential of various soundscape components (related to sound sources) for attracting attention. In view of long-term deployment of the model in outdoor measurement equipment, and for evaluating simulated soundscape design interventions, computational efficiency (low data communication rates, real-time operation, etc.) is advantageous. Consequently, the use of detailed auditory processing models, such as those existing for loudness,[30] masking,[31] stream segregation,[7] auditory saliency,[54] or auditory attention[68] is not feasible. Instead, simplified models for each step of the soundscape analysis process are proposed. Next to this, the proposed model only accounts for monaural sound, disregarding the influence of spatial cues on attention, and does not perform automated labeling of sounds. The proposed computational model is comprised of three stages, illustrated in Fig. 1: (a) peripheral auditory processing and the calculation of a measure of auditory saliency, (b) mapping of acoustical features based on co-occurrence, and (c) modeling auditory attention. A detailed description of each of the three stages follows.

B. Peripheral auditory processing

In a first stage, a feature vector is extracted, at regular time intervals, from the sound signal measured by the microphone. Instead of calculating a detailed time-frequency representation of the raw sound wave (e.g., using a gammatone filterbank), the model starts from the $\frac{1}{3}$-octave band spectrum (31 bands from 20 Hz to 20 kHz), calculated with a temporal resolution of 1 s. This procedure has the main advantage that off-the-shelf sound measurement equipment can be used as a front-end, which increases the applicability of the model. The limited data rate (31 values per second) makes it possible to implement the model on a large-scale measurement network and to store data for longer periods of time. Furthermore, the choice of time resolution can be justified by noting that a wide range of outdoor environmental sounds have a relatively slowly varying temporal envelope.[69–71] Subsequently, a simplified cochleagram is calculated using the Zwicker loudness model,[21,72] which accounts for energetic masking. Again, the complete hearable frequency range is considered (0 to 24 Bark) with a spectral resolution of 0.5Bark, resulting in 48 spectral values at each timestep.

The mechanism for extracting the feature vector, which characterizes the strength and spectro-temporal variability in

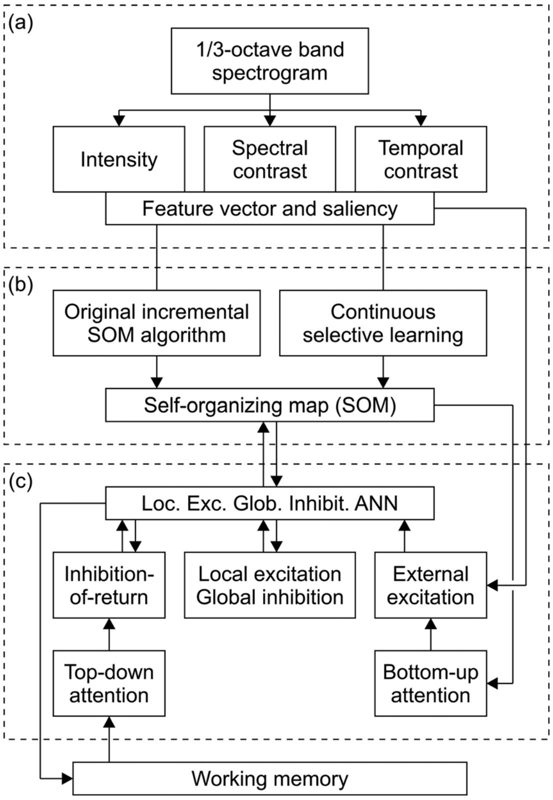

(a)
1/3-octave band spectrogram
Intensity
Spectral contrast
Temporal contrast
Feature vector and saliency
(b)
Original incremental SOM algorithm
Continuous selective learning
Self-organizing map (SOM)
(c)
Loc. Exc. Glob. Inhibit. ANN
Inhibition-of-return
Local excitation Global inhibition
External excitation
Top-down attention
Bottom-up attention
Working memory

FIG. 1. Schematic overview of the proposed computational model: (a) peripheral auditory processing, (b) self-organized map of acoustical features based on co-occurrence, and (c) auditory attention.

the sound signal, is inspired by the way the human auditory system biases its attention towards particularly conspicuous events. Based on existing models for auditory saliency,[54,73] the proposed model calculates measures for intensity, spectral and temporal contrast using a center-surround mechanism, which mimicks the receptive fields in the auditory cortex. In particular, multiscale features are calculated in parallel by convolving the cochleagram with various 2D Gaussian and difference-of-Gaussian filters. The former encode intensity, while the latter encode the spectral and temporal gradient of the cochleagram. In total, 16 scales (4 for intensity, 6 for spectral contrast and 6 for temporal contrast) are considered. Figure 2 shows a section of the filters along the time or frequency axis. Using this procedure, a feature vector is constructed at each timestep, consisting of $16 \times 48 = 768$ values.

Based on the feature vector, a measure for the saliency of the sound at each timestep is calculated. The calculation largely follows the scheme presented by Kalinli and Narayanan,[73] with the major adjustment that the effects of spectro-temporal orientation and pitch are not considered. First, rectified center-surround differences are calculated from the raw features obtained at different scales within the same modality (intensity, spectral or temporal contrast), mimicking the properties of local cortical inhibition.[54] The resulting center-surround differences are then scaled to a common range, in order to eliminate the difference in dynamic range between the different modalities and scales, and normalized using an iterative nonlinear algorithm that

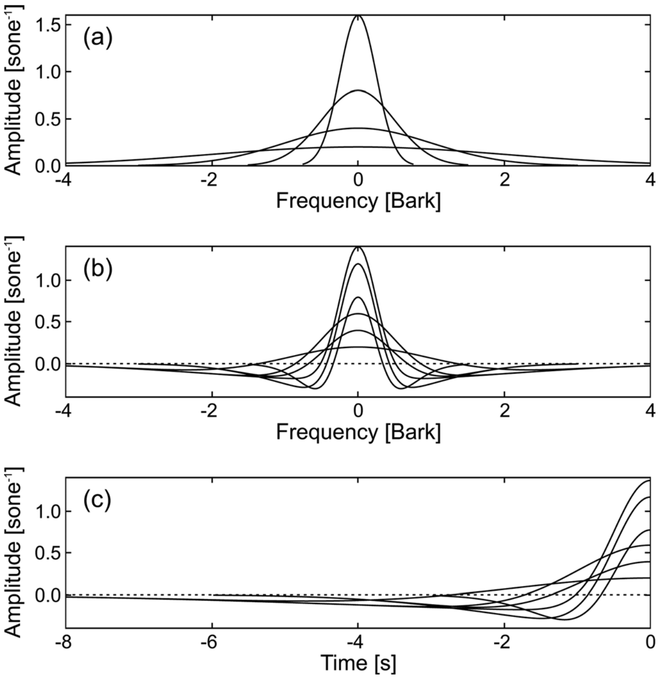


FIG. 2. Cross section of the receptive filters that are used to calculate (a) intensity, (b) spectral contrast, and (c) temporal contrast. For the latter, causality is preserved by only convolving with the past.

simulates competition between neighboring salient locations on the tonotopic scale, promoting peaks while suppressing background noise.[74] The normalized center-surround difference vectors are then combined (added) across scales within each modality, and the resulting vectors are again normalized using the same algorithm, and combined to achieve a single tonotopic vector, encoding the saliency of the sound at each timestep and at each frequency channel. Finally, a single saliency score at each timestep is calculated by summing all values of the saliency vector, hereby assuming that saliency combines additively across frequency channels.[75] A detailed description of the algorithm can be found in De Coensel and Botteldooren.[12]

C. Co-occurrence mapping of features

Biological systems learn which auditory features belong to the same auditory object based on co-occurrence. However, auditory learning as described at the end of Sec. II B is not a straightforward process, and is still far from being fully understood and computationally replicable. Moreover, learning and memory are not observable phenomena; they have to be inferred from behavior.[36] Nevertheless, in the computational framework here presented, an initial unsupervised learning strategy based on feature co-occurrence is used. It is implemented as a Self-Organizing Map (SOM) or Kohonen Map,[76] an abstract model of topographic mapping in the sensory cortex (see Sec. II B).

A SOM is a 2D grid of units, each of which is represented in the high-dimensional feature space through a reference vector. The Original Incremental SOM Algorithm[76] to train the map consists of iterating the following two steps until some stopping criterion is met.

(1) An input feature vector is provided at each time step and the unit corresponding to the closest reference vector, generally called the best-matching unit (BMU), is found.
(2) The reference vector corresponding to the BMU and those of units near to the BMU are moved closer to the input feature vector.

The second step underlies the topological preservation. After training, the reference vectors of the SOM units tend to a nonlinear discrete mapping of the distribution of the input data. Some regions of the feature space will be densely mapped by the reference vectors of the SOM units, while other regions will only be sparsely represented. This way, the high-dimensional relationships underlying the input feature data are projected on a 2D map.[76] Once the projection is sufficiently accurate, as quantified by the stopping criterion, training stops.

Machine learning purely based on co-occurrence does not account for the influence of several factors influencing human learning, such as attention, that were mentioned in Sec. II. Therefore, the basic SOM training is extended with a second training phase that accounts for saliency and novelty of the sound, thus attributing more weight to sounds that are likely to attract attention. The implemented strategy, called continuous selective learning,[14] can be seen as a series of much shorter learning periods, triggered whenever the distance between the new feature vector and the BMU is higher than an activation threshold, T_1, and halted when less than a deactivation threshold, T_2, with $T_2 \leq T_1$. Moreover, in order to give more importance to salient sound events, the overall saliency as calculated in Sec. III B is used as a modulator of the learning strength. It is observed that after a couple of weeks of continuous selective learning, the SOM is capable of identifying—in terms of distance to the BMU—most of the sounds occurring in a specific acoustic environment. In other words, after such training, the reference vector of each SOM unit corresponds to a representative sound prototype. In order to translate the information encoded in the SOM into hearable sound samples, a sound recording session can be used, during which representative 5-s sound samples with feature vectors closest to each SOM unit are stored. We call this compilation of sounds the "acoustic summary" of the given soundscape.[14] Note that the sound samples of the acoustic summary are not labeled automatically in this work. Instead, this can be performed by an expert listener (e.g., an acoustician acquainted with the soundscape of the given location), who explores the acoustic summary and identifies regions in the map corresponding to specific classes of sounds used to present the results in Sec. IV B.

D. Modeling auditory attention

In order to identify sounds that will be heard on the basis of a trained SOM, an excitatory-inhibitory artificial neural network (ANN), simulating the auditory cortex, is introduced. With each unit of the SOM, a neuron is associated, to be excited by input sounds with feature vectors that are similar to the reference feature vector of the corresponding SOM unit. In order to achieve this, first, a measure for similarity between the input feature vector and the SOM reference vectors needs to be calculated. This is done by calculating the Euclidean distance between the two vectors. Low values of this distance indicate high similarity and vice versa, but, as high excitation is desired in case of high similarity, a Gaussian-like function, centered around zero, is used to convert the Euclidean distances to excitation values, resulting in excitation values approaching 1 for highly similar, and 0 for dissimilar vectors. To take into account the fact that the excitation process is not instantaneous, a leaky integrator is used, with different time constants for increasing and decreasing excitation values.

Bottom-up attention, as explained in Sec. II C, is a rapidly operating process and is independent of the activity the listener is involved in, facilitating the detection of conspicuous and salient sounds. This is implemented in the model by weighing neuron excitation with a saliency factor, calculated based on the reference vector of the corresponding SOM unit.

As in De Coensel and Botteldooren,[12] IOR is introduced to prevent auditory attention from staying focused on one particular source, thus enabling a listener to scan his/her auditory environment. At each timestep, only a certain number of neurons will finally be activated, indicating that attention is focused on these neurons. In the current model, IOR is implemented as an increasing inhibition term for these neurons, causing activation to decrease and eventually to fall back to zero. For neurons that are not activated, and thus are not a candidate to get attention, IOR decreases to zero, such that activation is made possible again. This way, IOR causes attention to be continuously shifted from one zone to another. As with excitation, a leaky integrator is used for the implementation, again with different time constants for increasing and decreasing values.

The effect of top-down or outward oriented attention is implemented as a factor modulating the IOR mechanism. By changing the IOR time constants for neurons related to certain zones of the SOM, the shifting of attention can be delayed or even halted when it is focused on neurons corresponding to one of these zones. This way, sustained attention on the sounds represented by these zones is facilitated. Modeling the cause of top-down attention itself is far beyond the scope of current computational models.

Finally, concepts of a Locally Excitatory Globally Inhibitory Oscillator Network[77] (LEGION) are used to implement clustering and competitive selection, to indicate which sound receives attention and thus is entered in working memory. In order to minimize the computational load of the model, there are no oscillators as in a LEGION, but local excitation and global inhibition terms are still used for clustering and competitive selection, respectively. Local excitation is added to the input of each neuron, based on the excitation of its neighboring neurons, weighted with precalculated connection weights that depend on the similarity of the reference vectors of the two corresponding SOM units. Neighboring neurons which represent very similar sounds are strongly connected, while connection is weak when the neurons represent dissimilar sounds. A preliminary unit activation can be calculated as the sum of excitation terms minus the IOR, with negative values being set to zero. Global inhibition now adds a new inhibition term to each neuron in the network, calculated based on the sum of these preliminary activations of all neurons. When this summed activation exceeds a certain preset value, global inhibition will rise, and vice versa. By subtracting this inhibition term from the preliminary activation and setting negative values to zero, the final activation is calculated. Thanks to the clustering effect of local excitation, at each timestep, only one or a few clusters will have positive values for their final activation. These clusters represent the sounds that receive attention, and for which information is sent to working memory.

IV. CASE STUDY

A. Overview

In this section, a proof of concept of the computational framework presented in Sec. III is provided. A fixed sound measurement station was installed in the city of Ghent, next to an urban road, carrying about 3000 vehicles/day during a typical work day. The sonic environment at the chosen location mainly consists of a mixture of road traffic noise due to private and public transport, and noise from pedestrians due to the proximity of several shops and one educational institution. A standard $\frac{1}{3}$-octave band spectrum at 1s time intervals was measured during 3 weeks and is used to train the computational attention model (see Sec. IV B).

The aim of the case study was to assess the perceptual effects of attracting songbirds at the microphone location, a measure that is often proposed to increase the pleasantness of a soundscape.[78] For this, a 1-h sound recording was performed during a work day (but not during the latter 3-week period used for training). The L_{Aeq} during this 1-h period was 68.2 dB(A). Subsequently, a series of 30 artificial 1-h sonic environments were created by mixing the original recording with an increasing number of bird sounds at random instances in time. For this, a series of bird vocalizations without background noise, with a duration of up to a few seconds, were used, for which the peak level was adjusted to match the peak level of the few bird sounds present in the original recording. The 1-h L_{Aeq} of the added bird sound ranged from 46.3 dB(A), representing a few sporadic vocalizations, to 75.8 dB(A), representing a quasi-continuous bird chorus, resulting in a signal-to-noise ratio (SNR) for bird sound versus background ranging from −21.9 dB to 7.6 dB.

B. Results

A first assessment of the effect of adding bird sound would be to check the audibility of the bird sound above the background noise. Figure 3 shows the average short-term partial loudness (STPL) of the bird sound above the background, for the series of artificial sound mixtures, as a function of signal-to-noise ratio, as calculated with the model of Glasberg and Moore.[31] The average partial loudness rises monotonically with signal-to-noise ratio, and starts to increase with a higher rate between −5 and 0 dB, marking the range in which the individual bird vocalizations, which can be partially energetically masked if considered

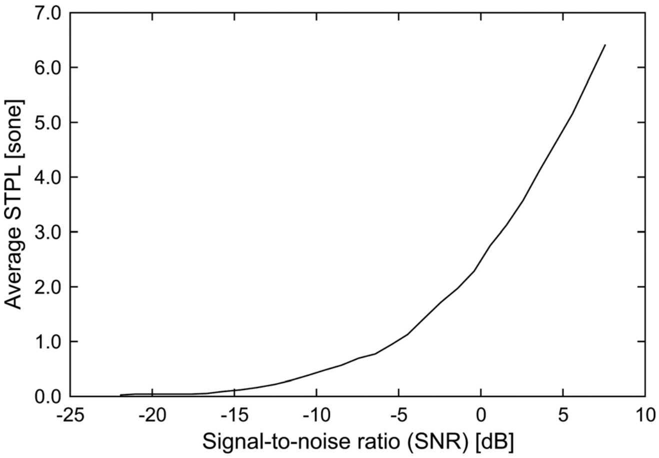


FIG. 3. The average short-term partial loudness (STPL) of the bird sound above the background noise, as a function of signal-to-noise ratio.

separately, start to form a chorus that is audible continuously. Note that the energetic masking model by Glasberg and Moore has only limited applicability in evaluating the effect of acoustical design measures *in situ*, because it requires that separate recordings for foreground and background sound are available (thus only artificial sound mixtures can be used), and that, due to its computational complexity, fragments are short—for the results of Fig. 3, only the first minute of sound was used.

To demonstrate the performance of the auditory attention model presented in this work, first, acoustical feature vectors and instantaneous saliency values were calculated for the 3-week measurement period, using the algorithm presented in Sec. III B. Subsequently, based on this data, a SOM, composed of 50 × 75 = 3750 hexagonally placed units, was trained in three phases. During the first phase, the incremental SOM training algorithm, as presented in Sec. III C, was applied to the features calculated during 14 h of the first day of the measurement period. During the second phase, the selective learning algorithm was applied to the remaining 3 weeks of measurement data. During the third phase, the artificial sound mixtures containing bird vocalizations were used in random order. Training a SOM on the sounds at a specific location results in a strong sound context dependency.[13] In particular, sounds not present in the training set cannot be easily classified (they will have a large distance to the BMU). Therefore, the third training phase is needed to get the SOM acquainted with the new bird sounds added to the background. From now on, we will refer exclusively to this fully trained SOM.

An acoustic summary has been created as mentioned in Sec. III C, based on several hours of recording at the given location and the 30 artificial soundscapes. Next, SOM units related to bird sounds are marked by an expert listener, and these are shown in Fig. 4(h). They are mainly grouped into two different regions, related to individual bird chirps (region 1) and a chorus of bird song (region 2). In light of Sec. III C, the presence of multiple SOM regions devoted to bird sounds should not be surprising: the sound of a single chirp and the sound produced by many birds in chorus result in different sound features, and thus in different regions of the map. Figures 4(a)–4(g) shows how often each of the units of the SOM become the BMU when the original sound and each of the artificial sound mixtures is presented to the model.

As expected, units inside both regions corresponding to bird sounds are more frequently the BMU as the SNR of bird sound increases. This behavior can be quantitatively evaluated by calculating the percentage of time the BMU belongs to either region 1 or region 2 as a function of the SNR. Figure 5 shows that the percentage of the time that individual bird chirp features are dominant (BMU belonging to region 1) increases monotonically, until a peak is reached at a SNR equal to −2 dB. At that point, the percentage of the time that bird chorus features (BMU belonging to region 2) are dominant starts to increase, while the time that individual bird chirp features are dominant falls back to zero with increasing SNR, marking a quasi-continuous bird chorus present throughout the corresponding artificial sound mixtures.

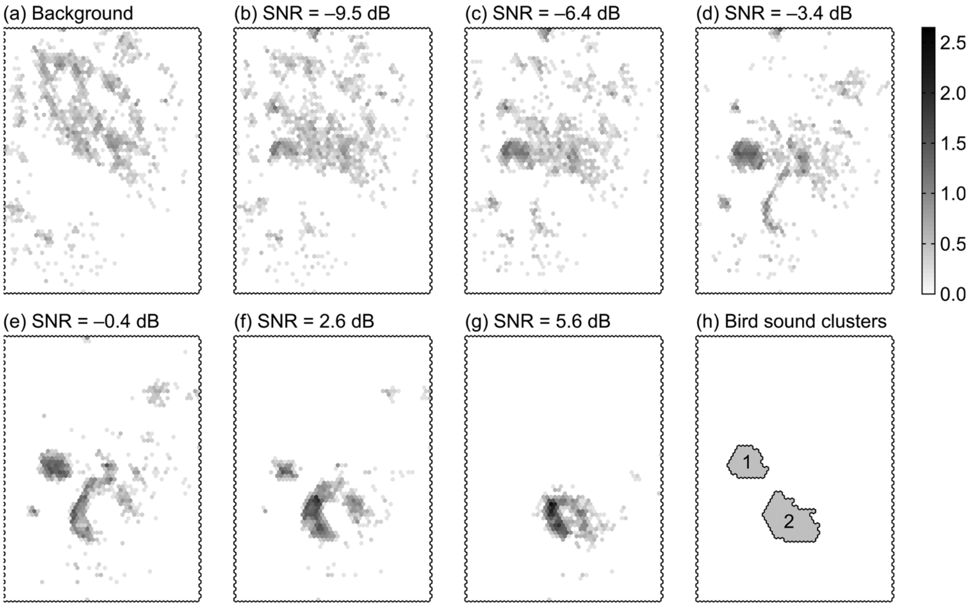


FIG. 4. Logarithmic distribution of the occurrence of the BMU among the SOM units for different scenarios: (a) background, (b)–(g) artificial soundscapes, in which bird vocalizations are progressively added to the background. For each sound scenario, 1 h (3600 testing samples) has been used. (h) The two regions of the SOM related to individual bird chirps (region 1) and bird chorus (region 2).

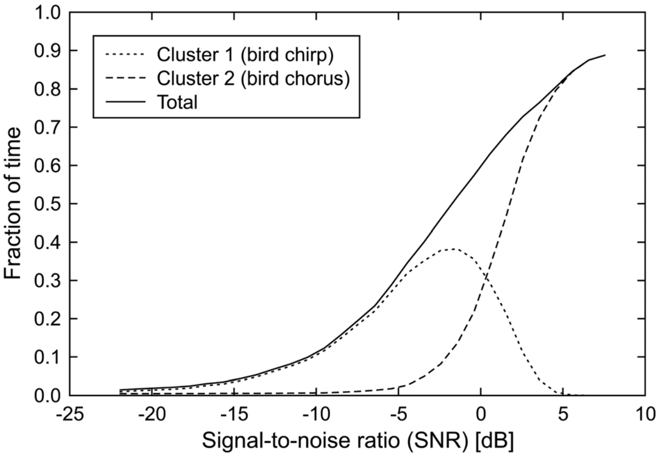


FIG. 5. Evolution of the fraction of time the BMU is located in region 1 (bird chirp, dotted line), region 2 (bird chorus, dashed line) and their sum (total, continuous line) as a function of SNR between background and foreground. For each sound scenario, 1 h (3600 testing samples) has been used.

Now, the same procedure is repeated, taking into account attention mechanisms. Although implemented in the general computational model (see Sec. III D), the effect of top-down attention is not taken into account, as this would require a model for working memory, which is outside the scope of this paper. Consequently, IOR time constants are the same for all neurons. The neuron with the strongest activation is now taken at each timestep to represent the sound (i.e., the combination of sound features) that receives attention, and in the same way as before, a distribution of occurrence is calculated. From this distribution, the same two clusters are used to calculate the percentage of time the most strongly activated neuron is located in each of the regions, thus approximating the fraction of time that attention is focused on bird sound. The results are displayed in Fig. 6.

It can be seen that for lower SNR, the percentage of time that attention is paid to birds is slightly higher than in Fig. 5, while for higher SNR, this percentage is lower. This indeed is the expected behavior, as for lower SNR, each time bird sound is detectable, it will get attention because its saliency is higher than the background, and because inhibition-of-return will be very low. For higher SNR, bird sound will be continuously detectable, and inhibition-of-return will cause attention to shift away from it. Considering that sounds need to be audible and be paid attention to, in order to contribute to the appraisal of a soundscape, these results are also in accordance with empirical results reported by De Coensel *et al.*[29] There, it was found that, already at an SNR of -10 dB, adding (salient and intermittent) bird sound to a sonic environment dominated by road traffic noise would increase the pleasantness of the soundscape significantly, more than adding the sound of a continuously flowing fountain at various SNR, although the latter may be more suited to energetically mask road traffic noise.[23,29] The sounds produced by bird vocalizations and fountains are generally considered to be positive in urban and rural environments.[79,80] Consequently, in the context of soundscape planning, the presented model can be helpful to quantify the potential positive effect of introducing additional sounds in the sonic environment, e.g., through the use of audio islands.[81]

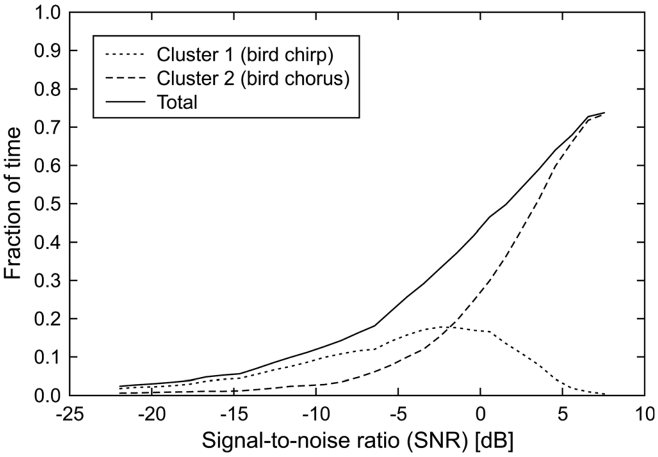


FIG. 6. Evolution of the fraction of time the auditory attention is located in region 1 (bird chirp, dotted line), region 2 (bird chorus, dashed line) and their sum (total, continuous line) as a function of SNR between background and foreground. For each sound scenario, 1 h (3600 testing samples) has been used.

V. CONCLUSIONS

Taking into account the mechanisms underlying human auditory perception of environmental sound is a fundamental principle in soundscape design. However, models and techniques that would assist the soundscape designer in achieving this goal are still lacking. In this work, a computational model for soundscape analysis was presented, which implements processes such as bottom-up selective attention and learning, with the goal of simulating how listeners would switch their attention over time between different sounds. The model consists of simplified implementations of several already existing submodels for auditory saliency, topographic mapping, learning, and auditory attention. It complements already existing models of attention-based auditory scene analysis[7] although it does not provide the same level of detail. However, the novelty of this model lies in its capability to process long stretches of sound in order to accommodate for the huge variation in environmental sounds that characterize the typical urban outdoor environment. The model can be applied to construct an acoustic summary of a soundscape, i.e., a collection of the typical sounds that can be heard at a particular location, and allows assessment of the influence of soundscaping measures such as adding additional sounds to distract attention away from undesired sounds. The latter use was illustrated through a case study, in which the effect of adding bird sound to an urban sonic environment was investigated, and in general, accordance with empirical results was found for this particular case. An unexpected model outcome was the emergence of two regions in the map as more and more bird sounds were entered. It was confirmed by listening to the samples that these highly activated regions corresponded to what could be labeled “bird chirps” at the one hand, and a “bird chorus” at the other.

The presented model does not take into account cross-sensory or high-level cognitive effects that lead to top-down auditory selective attention, or meaning attachment to sounds. The latter would involve accounting for the influence of inter-individual differences, and solving linguistic issues.[28] Indeed, a sound can be described at two different levels, either by its source or by the action generating it, although such levels are not always clearly separated. A description of the physical properties of either the sound source or the sound itself is provided only when the listener is not able to identify the source or the activity generating the sound. In order to explain the complexity of labeling and categorizing the sounds, observe the following two examples: the sounds of a tram passing by a stopping place and the sound produced by birds. In the first example, a listener would typically label each sound based on the action that generates the sound (braking, opening the doors, warning sounds before closing the doors, accelerating), while it would be unlikely that the specific sources (brakes, engine, or loudspeaker) are mentioned. In this case, activity categorization will thus be dominant. Obviously, listeners would also very likely mention the tram as a whole, referring to the sound source. In the second example, the sound produced by birds, the label "bird chirping" is generally used, thus showing again a mixture of the two levels: sound source (birds) and the activity producing such sounds (chirping). Moreover, a listener may refer to the number of birds: while one bird chirping or several birds chirping together denote the same activity, the (number of) sound sources changes. Automated labeling of the acoustic summary as compiled with the present model thus provides a challenge for future research.

ACKNOWLEDGMENTS

B.D.C. is a postdoctoral fellow, and M.B. is a doctoral fellow of the Research Foundation–Flanders (FWO–Vlaanderen); the support of this organization is gratefully acknowledged. This work was carried out within the framework of the IDEA project, supported by IWT Vlaanderen (grant IWT-080054). The authors would like to thank Samuel Dauwe for his technical contribution in collecting measurement data.

[1] B. Hellström, "Noise design: Architectural modelling and the aesthetics of urban acoustic space," Ph.D. thesis, School of Architecture, Royal Institute of Technology, Stockholm, Sweden (2003), 272 pp.

[2] A. L. Brown and A. Muhar, "An approach to the acoustic design of outdoor space," J. Environ. Plann. Manage. **47**, 827–842 (2004).

[3] M. Zhang and J. Kang, "Towards the evaluation, description, and creation of soundscapes in urban open spaces," Environ. Plann. B **34**, 68–86 (2007).

[4] P. Grahn and U. K. Stigsdotter, "The relation between perceived sensory dimensions of urban green space and stress restoration," Landscape Urban Plan. **94**, 264–275 (2010).

[5] E. Öhrström, A. Skånberg, H. Svensson, and A. Gidlöf-Gunnarsson, "Effects of road traffic noise and the benefit of access to quietness," J. Sound Vib. **295**, 40–59 (2006).

[6] A. Gidlöf-Gunnarsson and E. Öhrström, "Noise and well-being in urban residential environments: The potential role of perceived availability to nearby green areas," Landscape Urban Plan. **83**, 115–126 (2007).

[7] *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, edited by D. Wang and G. J. Brown (John Wiley and Sons, Hoboken, NJ, 2006), 395 pp.

[8] L. Yu and J. Kang, "Factors influencing the sound preference in urban open spaces," Appl. Acoust. **71**, 622–633 (2010).

[9] A. L. Brown, J. Kang, and T. Gjestland, "Towards standardization in soundscape preference assessment," Appl. Acoust. **72**, 387–392 (2011).

[10] D. Botteldooren and B. De Coensel, "A model for long-term environmental sound detection," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN′08)* (Hong Kong, 2008), pp. 2017–2023.

[11] 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. **126**, 656–665 (2009).

[12] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *Proc. ICA* (Sydney, Australia, 2010), pp. 3480–3487.

[13] D. Oldoni, B. De Coensel, M. Rademaker, B. De Baets, and D. Botteldooren, "Context-dependent environmental sound monitoring using SOM coupled with LEGION," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN′10)* [CD-Rom] (Barcelona, Spain, 2010).

[14] D. Oldoni, B. De Coensel, M. Boes, T. Van Renterghem, S. Dauwe, B. De Baets, and D. Botteldooren, "Soundscape analysis by means of a neural network-based acoustic summary," in *Proc. Internoise* (Osaka, Japan, 2011), pp. 3796–3801.

[15] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA, 1994), 792 pp.

[16] E. S. Sussman, "Integration and segregation in auditory scene analysis," J. Acoust. Soc. Am. **117**, 1285–1298 (2005).

[17] E. S. Sussman, J. Horváth, I. Winkler, and M. Orr, "The role of attention in the formation of auditory streams," Percept. Psychophys. **69**, 136–152 (2007).

[18] 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. **30**, 643–656 (2004).

[19] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**, 114–123 (2011).

[20] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—Focusing the searchlight on sound," Curr. Opin. Neurobiol. **17**, 437–455 (2007).

[21] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*, Springer Series in Information Sciences, 2nd ed. (Springer-Verlag, Berlin, Germany) (1999), Vol. 22, 428 pp.

[22] D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538 (2001).

[23] M. E. Nilsson, J. Alvarsson, M. Rådsten-Ekman, and K. Bolin, "Auditory masking of wanted and unwanted sounds in a city park," Noise Control Eng. J. **58**, 524–531 (2010).

[24] C. S. Watson, "Some comments on informational masking," Acta Acust. Acust. **91**, 502–512 (2005).

[25] N. Durlach, "Auditory masking: Need for improved conceptual structure," J. Acoust. Soc. Am. **120**, 1787–1790 (2006).

[26] T. Y. Lee and V. M. Richards, "Evaluation of similarity effects in informational masking," J. Acoust. Soc. Am. **129**, EL280–EL285 (2011).

[27] C. Lavandier and B. Defréville, "The contribution of sound source characteristics in the assessment of urban soundscapes," Acta Acust. Acust. **92**, 912–921 (2006).

[28] D. Dubois, C. Guastavino, and M. Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," Acta Acust. Acust. **92**, 865–874 (2006).

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

[30] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," J. Audio Eng. Soc. **50**, 331–342 (2002).

[31] B. R. Glasberg and B. C. J. Moore, "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds," J. Audio Eng. Soc. **53**, 906–918 (2005).

[32] R. Leech, B. Gygi, J. Aydelott, and F. Dick, "Informational factors in identifying environmental sounds in natural auditory scenes," J. Acoust. Soc. Am. **126**, 3147–3155 (2009).

[33] B. Gygi and V. Shafiro, "The incongruency advantage for environmental sounds presented in natural auditory scenes," J. Exp. Psychol. **37**, 551–565 (2011).

[34] J. W. Lewis, W. J. Talkington, A. Puce, L. R. Engel, and C. Frum, "Cortical networks representing object categories and high-level attributes of familiar real-world action sounds," J. Cogn. Neurosci. **23**, 2079–2101 (2011).

[35] L. R. Engel, C. Frum, A. Puce, N. A. Walker, and J. W. Lewis, "Different categories of living and non-living sound-sources activate distinct cortical networks," NeuroImage **47**, 1778–1791 (2009).

[36] N. M. Weinberger, "Reconceptualizing the primary auditory cortex: Learning, memory and specific plasticity," in *The Auditory Cortex*, edited by J. A. Winer and C. Schreiner (Springer, New York, 2011), Chap. 22, pp. 465–491.

[37] R. G. Rutkowski and N. M. Weinberger, "Encoding of learned importance of sound by magnitude of representational area in primary auditory cortex," Proc. Natl. Acad. Sci. USA **102**, 13664–13669 (2005).

[38] K. M. Bieszczad and N. M. Weinberger, "Representational gain in cortical area underlies increase of memory strength," Proc. Natl. Acad. Sci. USA **107**, 3793–3798 (2010).

[39] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Neurosci. **6**, 1216–1223 (2003).

[40] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks," J. Neurosci. **25**, 7623–7635 (2005).

[41] T. Raij, L. McEvoy, J. P. Mäkelä, and R. Hari, "Human auditory cortex is activated by omissions of auditory stimuli," Brain Res. **745**, 134–143 (1997).

[42] F. W. Ohl, H. Scheich, and W. J. Freeman, "Change in pattern of ongoing cortical activity with auditory category learning," Nature **412**, 733–736 (2001).

[43] J. Pekkola, V. Ojanen, T. Autti, I. P. Jääskeläinen, R. Möttönen, A. Tarkiainen, and M. Sams, "Primary auditory cortex activation by visual speech: an fMRI study at 3T," NeuroReport **16**, 125–128 (2005).

[44] P. Heil, R. Rajan, and D. R. F. Irvine, "Topographic representation of tone intensity along the isofrequency axis of cat primary auditory cortex," Hearing Res. **76**, 188–202 (1994).

[45] A. Morel and J. H. Kaas, "Subdivisions and connections of auditory cortex in owl monkeys," J. Comp. Neurol. **318**, 27–63 (1992).

[46] C. I. Petkov, C. Kayser, M. Augath, and N. K. Logothetis, "Functional imaging reveals numerous fields in the monkey auditory cortex," PLoS Biol. **4**, 1213–1226 (2006).

[47] T. M. Talavage, P. J. Ledden, R. R. Benson, B. R. Rosen, and J. R. Melcher, "Frequency-dependent responses exhibited by multiple regions in human auditory cortex," Hearing Res. **150**, 225–244 (2000).

[48] T. M. Talavage, M. I. Sereno, J. R. Melcher, P. J. Ledden, B. R. Rosen, and A. M. Dale, "Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity," J. Neurophysiol. **91**, 1282–1296 (2004).

[49] C. Humphries, E. Liebenthal, and J. R. Binder, "Tonotopic organization of human auditory cortex," NeuroImage **50**, 1202–1211 (2010).

[50] R. M. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World* (Destiny Books, Rochester, VT, 1994), 320 pp.

[51] 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. **7**, e1000129 (2009).

[52] E. I. Knudsen, "Fundamental components of attention," Annu. Rev. Neurosci. **30**, 57–78 (2007).

[53] S. Shamma, "On the role of space and time in auditory processing," Trends Cogn. Sci. **5**, 340–348 (2001).

[54] C. Kayser, C. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," Curr. Biol. **15**, 1943–1947 (2005).

[55] A. Baddeley, "Working memory: looking back and looking forward," Nat. Rev. Neurosci. **4**, 829–839 (2003).

[56] G. Sperling and E. Weichselgartner, "Episodic theory of the dynamics of spatial attention," Psychol. Rev. **102**, 503–532 (1995).

[57] C. Spence and J. Driver, "Auditory and audiovisual inhibition of return," Percept. Psychophys. **60**, 125–139 (1998).

[58] D. J. Prime, M. S. Tata, and L. M. Ward, "Event-related potential evidence for attentional inhibition of return in audition," NeuroReport **14**, 393–397 (2003).

[59] E. Sussman, I. Winkler, and E. Schröger, "Top-down control over involuntary attention switching in the auditory modality," Psychon. Bull. Rev. **10**, 630–637 (2003).

[60] C. Escera, K. Alho, I. Winkler, and R. Näätänen, "Neural mechanisms of involuntary attention to acoustic novelty and change," J. Cogn. Neurosci. **10**, 590–604 (1998).

[61] C. Escera, E. Yago, M.-J. Corral, S. Corbera, and I. Nunez, "Attention capture by auditory significant stimuli: Semantic analysis follows attention switching," Eur. J. Neurosci. **18**, 2408–2412 (2003).

[62] J. Domínguez-Borras, M. Garcia-Garcia, and C. Escera, "Negative emotional context enhances auditory novelty processing," NeuroReport **19**, 503–507 (2008).

[63] J. Domínguez-Borras, S.-A. Trautmann, P. Erhard, T. Fehr, M. Herrmann, and C. Escera, "Emotional context enhances auditory novelty processing in superior temporal gyrus," Cereb. Cortex **19**, 1521–1529 (2009).

[64] M. Garcia-Garcia, C. Escera, I. SanMiguel, and I. Clemente, "COMT and ANKK-1 gene-gene interaction accounts for distraction effect and resetting of the gamma neural oscillations to novel sounds," Int. J. Psychophysiol. **77**, 231 (2010).

[65] M. Garcia-Garcia, F. Barceló, I. C. Clemente, and C. Escera, "The role of DAT1 gene on the rapid detection of task novelty," Neuropsychologia **48**, 4136–4141 (2010).

[66] M. M. Bradley and P. J. Lang, "Affective reactions to acoustic stimuli," Psychophysiology **37**, 204–215 (2000).

[67] G. Thierry and M. V. Roberts, "Event-related potential study of attention capture by affective sounds," NeuroReport **18**, 245–248 (2007).

[68] S. N. Wrigley and G. J. Brown, "A computational model of auditory selective attention," IEEE Trans. Neural Netw. **15**, 1151–1163 (2004).

[69] B. De Coensel, D. Botteldooren, and T. De Muer, "1/*f* noise in rural and urban soundscapes," Acta Acust. Acust. **89**, 287–295 (2003).

[70] D. Botteldooren, B. De Coensel, and T. De Muer, "The temporal structure of urban soundscapes," J. Sound Vib. **292**, 105–123 (2006).

[71] B. De Coensel, D. Botteldooren, K. Debacq, M. E. Nilsson, and B. Berglund, "Clustering outdoor soundscapes using fuzzy ants," in *Proc. IEEE Congress on Evolutionary Computation (CEC'08)* (Hong Kong, 2008), pp. 1556–1562.

[72] E. Zwicker, H. Fastl, and C. Dallmayr, "BASIC-Program for calculating the loudness of sounds from their 1/3-oct band spectra according to ISO 532 B," Acustica **55**, 63–67 (1984).

[73] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," IEEE Trans. Audio Speech Lang. Process. **17**, 1009–1024 (2009).

[74] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," J. Electron. Imaging **10**, 161–169 (2001).

[75] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)* (Antwerp, Belgium, 2007), pp. 1941–1944.

[76] T. Kohonen, *Self-Organizing Maps*, 3rd ed. (Springer-Verlag, Heidelberg, Germany, 2001), 521 pp.

[77] D. Wang and D. Terman, "Locally excitatory globally inhibitory oscillator networks," IEEE Trans. Neural Netw. **6**, 283–286 (1995).

[78] M. Raimbault and D. Dubois, "Urban soundscapes: Experiences and knowledge," Cities **22**, 339–350 (2005).

[79] B. De Coensel and D. Botteldooren, "The quiet rural soundscape and how to characterize it," Acta Acust. Acust. **92**, 887–897 (2006).

[80] J. Kang and M. Zhang, "Semantic differential analysis of the soundscape in urban open public spaces," Build. Environ. **45**, 150–157 (2010).

[81] B. Schulte-Fortkamp, "The tuning of noise pollution with respect to the expertise of people's mind," in *Proceedings of Internoise* (Lisbon, Portugal, 2010), pp. 5744–5752.