Attention-Driven Auditory Stream Segregation using a SOM coupled with an Excitatory-Inhibitory ANN

Michiel Boes, Damiano Oldoni, Bert De Coensel, Dick Botteldooren
Acoustics Group, Department of Information Technology
Ghent University
Belgium
Michiel.Boes@intec.ugent.be

***Abstract* — Auditory attention is an essential property of human hearing. It is responsible for the selection of information to be sent to working memory and as such to be perceived consciously, from the abundance of auditory information that is continuously entering the ears. Thus, auditory attention heavily influences human auditory perception and systems simulating human auditory scene analysis would benefit from an attention model. In this paper, a human-mimicking model of auditory attention is presented, aimed to be used in environmental sound monitoring. It relies on a Self-Organizing Map (SOM) for learning and classifying sounds. Coupled to this SOM, an excitatory-inhibitory artificial neural network (ANN), simulating the auditory cortex, is defined. The activation of these neurons is calculated based on an interplay of various excitatory and inhibitory inputs. The latter simulate auditory attention mechanisms in a human-inspired but simplified way, in order to keep the computational cost within bounds. The behavior of the model incorporating all of these mechanisms is investigated, and plausible results are obtained.**

Keywords – Auditory Attention Model; Computational Auditory Scene Analysis; Auditory Stream Segregation; SOM; Artificial Neural Network; Environmental Sound

I. INTRODUCTION

Recent findings in psychophysics and neurophysiology point out that selective auditory attention is of great importance in human auditory perception [1][2]. Humans easily outperform current computer models in the process of perceiving and analyzing an acoustic environment, called auditory scene analysis (ASA). ASA involves decomposing a complex mixture of incoming sounds, originating from different sources, into individual auditory streams, using different auditory, but also visual and other cues [3]. Auditory attention is not only found to be indispensable in the process of auditory stream segregation itself [4], but, through competitive selection, it also enables listeners to select a single auditory stream for entrance into working memory, where it can be analyzed in detail. Information entering working memory is consciously perceived and can be used to create a mental image of the acoustic environment, and thus can influence the decision making process [5]. These findings make it clear that an auditory attention model is indispensable in computational auditory scene analysis (CASA).

The model described in this paper uses a submodel based on a Self-Organizing Map (SOM) to classify, and differentiate between different sounds, similar to the one described in [6]. On top of this SOM, it places a network of neurons with a combination of simple excitatory and inhibitory inputs. These excitation and inhibition terms implement the concepts of saliency-driven bottom-up attention, top-down attention and inhibition-of-return that are found in other auditory and also visual attention models [5][7], but adapts them for use in the SOM-based neural network. Finally, a submodel for competitive selection and clustering of excited neurons, loosely based on some concepts of Locally Excitatory Globally Inhibitory Oscillator Networks (LEGION, see [8]), is employed in this paper.

As in all models that simulate human brain functions, the proposed model for human auditory attention and stream segregation needs to make certain compromises between computational efficiency and biological accuracy. The present model is designed to be integrated into a large-scale noise monitoring network. One of the goals of this system will be to classify sound events that are potentially noticed by a listener, and to detect and identify conspicuous sound events, within the assessment of potential long-term effects of exposure to environmental sound on quality of life [9][10]. Consequently, the model is aimed to be embedded in low-cost hardware that has to run continuously for weeks to months. This requirement makes it not feasible to use similar but much more detailed models for auditory attention, such as the one presented in [11]. Nevertheless, while some of the submodels presented in the present work behave according to greatly simplified rules, the proposed model's architecture, and the way in which its different submodels interact are strongly based on available knowledge of the human auditory system.

This paper is structured as follows: in the next section, the model's architecture is presented, and its different submodels are discussed. Subsequently, the behavior of the model will be examined and some examples illustrating its use will be given. Finally, our conclusions are presented.

II. MODEL

A. General

The general structure of the model is given in Fig. 1. The central element in the model is the Self-Organizing Map or SOM. A grid of neurons connected to the SOM is used as a greatly simplified model for the auditory cortex, each neuron encoding a prototypical sound. The same groups of neurons on the grid will be excited by similar input sounds, and as such, the SOM classifies different types of sounds. In *a first phase* (see section II.B and the upper right part in Fig. 1), the SOM is trained in an unsupervised way, using features that describe the sound's characteristics (intensity, time contrast, frequency contrast), calculated for a long training sound fragment. *Next* (see section II.C and the upper left part in Fig. 1), the same features, now calculated for the test sounds, are used as input for the grid of neurons coupled to the SOM. The neurons of this grid will be excited to a certain degree, depending on the type of sound they represent, and the type of input sound. *The third part* of the model (see section II.D and the lower left part of Fig. 1) is the core attention model. It implements bottom-up attention as an enhancement of excitation for SOM neurons representing salient sounds, inhibition-of-return as a decrease of excitation of highly and frequently excited nodes and top-down attention as a modulator of inhibition-of-return. *Finally* (see section II.E and central right part in Fig. 1), local excitation and global inhibition mechanisms are used for clustering of excited neurons, and for competitive selection to decide which neurons deliver input to the working memory.

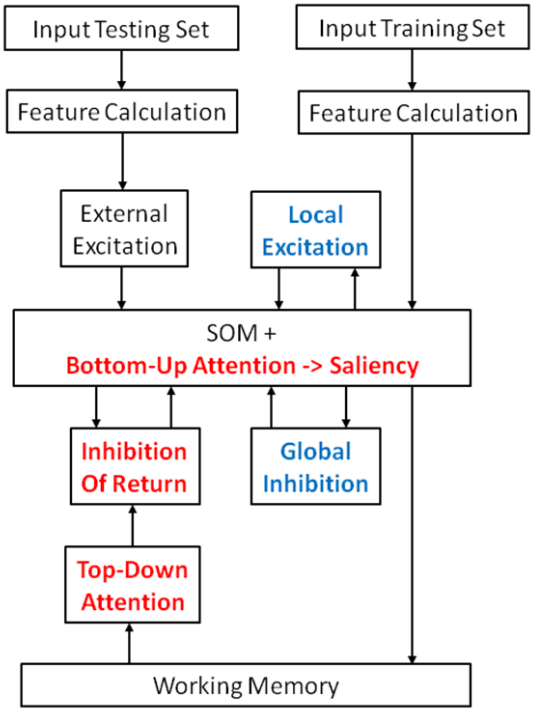

Figure 1. Overview of the model structure, with all of its submodels. The attention submodel in the lower left part is displayed in red, and the clustering and competitive selection submodel in the central right part is displayed in blue.

The latter may trigger further processing of the environmental sound. For example, if the model is incorporated into a measurement network, it may trigger recording of sound, transmission to a central database for automated and more detailed sound source recognition etc.

B. Learning

The learning phase of the model is very similar to the one presented in [6]. For reference, a short overview is given. A first step in the processing of incoming sound is the calculation of features describing it. The model starts from the 1/3-octave band spectrum, calculated with a temporal resolution of 0.125s. Next, a simplified cochleagram is calculated, taking into account energetic masking, using the Zwicker loudness model [12]. This cochleagram covers the complete range of hearable frequencies (0-24 Bark) with a resolution of 0.5 Bark. Thus, 48 spectral values are obtained at a rate of 8 times per second. Inspired by more detailed models for calculating an auditory saliency map [13], features encoding absolute intensity and spectro-temporal variations are calculated, by convoluting resp. Gaussian and difference-of-Gaussian filters with different scales to the cochleagram. Using 4 scales for intensity, 6 for spectral contrast and 6 for temporal contrast, 16 values are obtained for each frequency band in the cochleagram, and thus a 48 x 16 = 768 dimensional feature vector, describing the incoming sound, is calculated at each timestep.

Next, the obtained feature vectors of the training sequence are used as inputs to a Self-Organizing Map (SOM) or Kohonen map [14]. A SOM is a mathematical model often used as an unsupervised technique for nonlinear dimension reduction. It consists of nodes placed in a 2D grid, usually forming a hexagonal lattice, with each node corresponding to a reference vector in a multi-dimensional vector space. After training with the feature vectors of the training sequence, the SOM nodes encode the range of sounds contained in the training sequence by means of their corresponding reference feature vectors.

It should be noted that in order to obtain a well-trained SOM, a large training sequence is required. In light of the application of the model presented in this paper in outdoor sound monitoring, a SOM that is able to recognize most of the daily environmental sounds at a particular fixed outdoor microphone location will require training on several days of consecutive acoustic data. For this purpose standard SOM training techniques are extended with a continuous training algorithm that selects not well-matched and salient new data for further training. For each incoming feature vector a saliency value is calculated, using the same method as in section II.D, and only when this saliency value exceeds a certain threshold the feature vector is taken into account for further learning. Similarly, only incoming feature vectors with a Euclidian distance to the best matching unit exceeding a certain threshold are taken into account.

C. Excitation

With each SOM node, a neuron, sensitive to input sounds represented by that node, can be associated. The SOM thus can be viewed as an abstract model for the auditory cortex. When a specific sound (e.g. caused by a car passing by, people talking …) is used as input, the neurons corresponding to nodes in the SOM with a feature vector most similar to the incoming sound should be excited. In what follows, both the terms 'SOM node' and 'SOM neuron' will be used for the SOM node itself as well as for the neuron of the excitatory-inhibitory neural network associated to it.

A possible way to gather information about the similarity between the incoming feature vector and the reference vectors of the SOM is to calculate the Euclidian distances between both. This method however fails to take into account that certain features tend to have larger values, and to display bigger variations than others. Thus, in order to calculate the excitation of a certain node, first the incoming features are scaled, by dividing them by the maximum value, taken over all nodes' reference vectors, of the corresponding feature. This value gives a good estimation of the maximum value that can be expected for this feature, and thus gives an idea of the range of values that the feature can assume. The reference vectors are then scaled in the same way and the Euclidian distance between the two scaled vectors is calculated.

$$d_i(n) = \sqrt{\sum_{j=1}^{j_{max}} \left(\frac{\varphi_{i,j} - f_j(n)}{\varphi_{max,j}}\right)^2} \qquad (1)$$

Here, $\varphi_{i,j}$ is the value of the j-th element of the feature vector representing the i-th node, $f_j(n)$ is the j-th element of the input feature vector in the n-th timestep and $\varphi_{max,j} = \max_{\forall i} \varphi_{i,j}$. In order to convert this distance to a measure of similarity, i.e. displaying higher values for similar vectors and lower values for dissimilar vectors, a Gaussian-type function is used.

$$E_{0,i}(n) = \exp\left(-\frac{d_i(n)^2}{2\,\sigma_E^2}\right) \qquad (2)$$

The standard deviation σ_E used in this Gaussian-type function is a parameter that can be chosen depending on the desired sensitivity of the neurons.

It should be clear that not all of the 768 features are equally important in the process of recognizing a certain sound source, and some features can even contain confusing information. Take, for example, a neuron sensitive to bird sounds. If the model would give features encoding information at low audio frequencies the same amount of importance as features encoding information at high audio frequencies, in which the bird sound is dominant, a disturbing background sound containing mainly low frequencies would cause the incoming feature vector to have a large (scaled) Euclidian distance from the reference vector representing bird sound, even though the high frequency features are very clear and recognizable. In order to solve this problem, only the 76 features (10% of the total) with the highest values in the scaled reference vector of a node are taken into account when calculating the distance to that node, so (1) becomes:

$$d_i(n) = \sqrt{\sum_{j \in S_i} \left(\frac{\varphi_{i,j} - f_j(n)}{\varphi_{max,j}}\right)^2} \qquad (3)$$

with S_i the set of 76 feature vector indices with the highest scaled value in the reference vector of the i-th SOM node. In the bird sounds example, the features with the highest values in the scaled reference vector will be features encoding information at higher audio frequencies, and thus the presence of lower frequencies won't disturb the recognition of the bird sound anymore.

In order to account for non-instantaneous excitation of neurons, the excitatory mechanism is modeled as a leaky integrator, approaching its goal $E_{0,i}$ in an exponential way, with different time constants for increase and decrease. Thus, excitation is given by:

$$E_i(n) = \sum_{m=0}^{n} \left[E_{0,i}(m)\left(1 - \mathrm{e}^{-\frac{\delta_t}{\tau_E}}\right) e^{-\frac{(\mathrm{m}-\mathrm{n})\delta_t}{\tau_E}}\right] \qquad (4)$$

or, in a recursive form:

$$E_i(n) = E_i(n-1)e^{-\frac{\delta_t}{\tau_E}} + E_{0,i}(n)\left(1 - e^{-\frac{\delta_t}{\tau_E}}\right) \qquad (5)$$

with δ_t the length of the timestep between consecutive input vectors (0.125s in this paper) and τ_E the time constant of the leaky integrator, with different values for increasing and decreasing excitation. The factor $\left(1 - e^{-\frac{\delta_t}{\tau_E}}\right)$ is needed to make the leaky integrator converge towards its goal value $E_{0,i}$.

D. Attention

The next step in the construction of the model is the inclusion of attention mechanisms. Auditory attention can be described as "the cognitive process underlying our ability to focus on specific aspects of the acoustic environment, while ignoring others" [2]. The proposed attention submodel is inspired by the model described by Knudsen [5], involving bottom-up and top-down attention mechanisms, and an inhibition-of-return (IOR) mechanism preventing attention from permanently staying focused on the same sound source. These combined effects interplay and influence the decision about which auditory stream is selected to enter working memory.

Bottom-up attention is a rapidly operating mechanism, independent of any particular tasks the listener might be performing. It facilitates the detection of conspicuous, potentially interesting or dangerous sounds. For example, regardless of the activity in which a listener is engaged, the

sound of a gunshot will almost certainly draw his attention. In the proposed model, this is implemented as a factor that increases the excitation of nodes representing salient sounds, compared to those representing non-salient sounds. This way, nodes representing salient sounds will more easily be excited to higher levels, and thus the detection of salient sounds will be facilitated. In order to calculate a measure of saliency for each node, the method described in [15] is largely followed, with the major adjustment that spectro-temporal orientation and pitch are not considered. Thus, only the features also used by the SOM are needed in order to calculate a saliency map as a function of frequency for each node. In order to obtain a single value for saliency of a node, the values for the different frequencies are simply added together. Next, the saliency values corresponding to the SOM nodes are linearly rescaled, and an offset value can be added in order for the minimum and maximum to reach predefined values. These rescaled and offset saliency values S_i can then be used as an enhancing factor for the excitations of the SOM nodes.

$$E_{0,i}(n) = S_i \exp\left(-\frac{d_i(n)^2}{2\,\sigma_E^2}\right) \tag{6}$$

The predefined minimum and maximum saliency factor values then control for the influence of saliency, as they determine the portion of the node's excitation due to saliency.

Inhibition-of-return is a mechanism that causes attention to attenuate or switch to another sound source after a certain period of time. Because of this, a listener is able to continuously shift attention and scan the acoustic environment. For example, when a listener stands next to a busy road, at first his attention will be drawn to the sound of the cars passing by. After a while, his attention to the car sound will weaken, and other environmental sounds, such as bird sounds, will be paid attention to. As in the case for the activation mechanism, the inhibition-of-return mechanism is modeled as a leaky integrator. As soon as a neuron is activated, inhibition-of-return will rise towards its current excitation. When the neuron is not activated, inhibition-of-return will decrease towards zero:

$$IOR_{0,i}(n) = \begin{cases} \mathrm{E_i(n)} & \text{if } \mathrm{A_i(n-1)} > 0 \\ 0 & \text{if } \mathrm{A_i(n-1)} = 0 \end{cases} \tag{7}$$

$\mathrm{A_i}(n)$ is the activation of the i-th SOM neuron in the n-th timestep. It will be defined in (11) (see below). Inhibition-of-return is thus given by:

$$IOR_i(n) = \sum_{m=0}^{n} \left[IOR_{0,i}(m) \left(1 - \mathrm{e}^{-\frac{\delta_t}{\tau_{IOR}}}\right) e^{-\frac{(\mathrm{m-n})\delta_t}{\tau_{\mathrm{IOR}}}} \right] \tag{8}$$

or, in a recursive form:

$$IOR_i(n) = IOR_i(n-1) e^{-\frac{\delta_t}{\tau_{IOR}}} + IOR_{0,i}(n) \left(1 - e^{-\frac{\delta_t}{\tau_{IOR}}}\right) \tag{9}$$

with parameters defined in analogy with (5). Again, the time constant has different values for increasing and decreasing inhibition-of-return.

Top-down attention is operating more slowly than bottom-up attention. When a noticed sound is considered to be interesting, based on information already held in working memory, the top-down attention mechanism tries to focus sustained attention on this sound source. In the above example where a listener listens to car sounds on a busy road, attention will stay focused on the car sounds if the listener is given the task to try and detect when a particular car is passing by. In the model, this effect can be implemented as a determining factor for the time constant of the inhibition-of-return. When top-down attention needs to be focused on a certain area of the SOM, the inhibition-of-return time constants associated with these SOM nodes can be adapted in such a way that the scanning of the acoustic environment, caused by the inhibition-of-return, will be delayed, or even halted when attention is focused on the area of interest in the SOM.

Finally, competitive selection is needed, in order to decide what information is selected to enter working memory. A simple, but plausible way to achieve this, is to select the most strongly excited neurons, taking into account the external neural excitation and the inhibition-of-return. An issue that arises in this system is the fact that sometimes the most strongly excited neurons are scattered over the SOM, and do not represent one single stream to enter working memory. This issue is addressed by a clustering mechanism explained in the next section.

E. Clustering

In order to achieve clustering and competitive selection, the model uses some concepts of a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION), which has earlier been used for a similar purpose [16]. In contrast to the original LEGION model [8], in the present model, no oscillators are involved, but the concepts of local excitation and global inhibition are implemented to achieve the same goal of clustering and segregation. Global inhibition is used as a simple way to select the information to enter the working memory: only a selection of SOM neurons that are sufficiently excited will be activated when taking into account the global inhibition effect. As only the activated neurons hold meaningful information, these are the only ones sending information to the working memory. Local excitation interplays with this effect, by achieving simultaneous activation of neighboring neurons, which represent similar sounds, and which are thus likely to represent the same auditory stream.

In the implementation, global inhibition is assumed to depend on the total activation of all SOM neurons. When the sum of the activations of all neurons is above a certain predefined value, global inhibition will increase, and vice versa. By calculating the global inhibition as explained above, total activation will always approach the same value. Taking

into account external excitation, inhibition-of-return and global inhibition, activation is calculated as follows.

$$A_i(n) = \max[0, E_i(n) - IOR_i(n) - IG(n)] \quad (10)$$

$IG(n)$ is the global inhibition in the n-th timestep, calculated as explained above. The maximum makes sure that neurons for which total inhibition is higher than excitation have zero activation instead of a non-realistic negative value.

In order to determine the strength of the local excitation, first, in a similar way to the LEGION model, connection weights between the SOM nodes have to be calculated. A first property of the local excitation model is that, like in [16], there are only connections between neighboring SOM nodes, modeling hardwired connections in the network. A second property is that nodes with similar reference vectors have high connection weights, whereas nodes with very dissimilar reference vectors are only weakly connected. To calculate similarity between reference vectors, the distance is taken between the two vectors, scaled as described in section II.C. Also in the same way as in section II.C, a Gaussian-type function is used to convert this distance to a measure for similarity, again with the standard deviation of the Gaussian-type function as a parameter determining sensitivity. This measure for similarity is then used as connection weight between the two SOM nodes. These connection weights are fixed, and have to be calculated only once, as soon as the trained SOM is known. Now, neurons with non-zero activation, according to (10), will provide an extra excitation for their neighboring neurons. This extra excitation term is calculated as the activation of the neighboring neuron, multiplied by the calculated connection weight between the neurons. Thus, each node in the hexagonal SOM lattice, except for the border neurons, will receive 6 additional excitation inputs. As total excitation is altered by these new excitations, global inhibition will increase to obtain the same total activation as before. With the addition of the local excitation and the recalculation of global inhibition, the activation pattern of the SOM neurons has changed, and can now be used again as input for local excitation as explained above, and again, global inhibition will adapt to the added excitation. Repeating this process a predefined number of times for each timestep leads to clustering and only one, or a few, clusters of SOM neurons will finally be activated. Thus, the final formula for the calculation of the node activation becomes:

$$A_i(n) = \max[0, E_i(n) + EL_i(n) - IOR_i(n) - IG(n)] \quad (11)$$

where $EL_i(n)$ is the total local excitation of the i-th node, calculated as explained above, and $IG(n)$ is the total global inhibition adapted to the situation with local excitation, both at the n-th timestep.

III. Results

In this section, an illustration of the above described auditory attention model is given. A SOM consisting of a 75x50 grid of nodes is trained on 768-dimensional feature vectors, calculated with time intervals of 0.125s, based on 10 days of continuously recorded ambient sound. The sound was recorded by a microphone, placed in the city of Ghent, next to a quiet road and a river. The recordings contain mainly urban background noise, as well as some distinct cars driving by, ducks in the river, birds singing occasionally, the humming noise produced by some machinery in a neighboring laboratory, people talking, … To illustrate the model, the SOM neurons are externally excited by an incoming sound fragment with a duration of one minute. For each timestep, an activation pattern is calculated as described in section II. Time constants for external excitation are taken to be 0.01s when increasing, and 0.05s when decreasing. Inhibition-of-return time constants are taken to be 0.2s and 10s for increasing and decreasing values respectively. Experiments with the time constants indicate that these values yield reasonable results, independent of the SOM used, but, depending on the desired behavior of the model, the constants can be adjusted. For instance, in order to simulate a nervous person, the inhibition-of-return time constants can be decreased so the model will switch attention and scan the acoustic environment at a faster rate.

Firstly, to demonstrate the basic mechanisms of the model, excitation is not saliency weighted, eliminating bottom-up attention, and conscious top-down attention is also not taken into account. In Fig. 2, the average activation of the SOM neurons during the one-minute testing fragment is shown. The region on the SOM map around the node indicated by A in Fig. 2 is found to represent silent and non-salient city background noise (as determined by an expert listener). The zone around node B represents similar sounds, but including a humming machinery sound. The region around node C in the map represents the sound of ducks, and finally, the region around node D represents the sound of cars passing by. Fig. 3 shows the evolution, during 15 seconds of the one-minute testing fragment, of the different terms in (11), each behaving according to the model described in section II. Most of the time, the external excitation equals the total excitation, as local excitation only exceeds zero in certain timesteps where the iterative clustering mechanism rises this term to noticeable values. It can be seen that in the first few seconds, neuron D displays much activity, after which inhibition-of-return has grown to such levels that no further activation occurs. After a few seconds neurons A and B become active, indicating non-salient background sounds and machinery noise. Comparatively large zones of the SOM represent very similar non-salient sounds, causing a large amount of neurons to be excited externally, in turn causing the global inhibition to rise, according to the process described in section II.E. Finally, in the last few seconds of the analyzed fragment, the neurons describing duck sound are slightly activated. This activation would be expected to be stronger, as in the sound fragment duck sound is clearly present here. This is found to be due to the training of the SOM, as in the training set duck sound is only sparsely available.

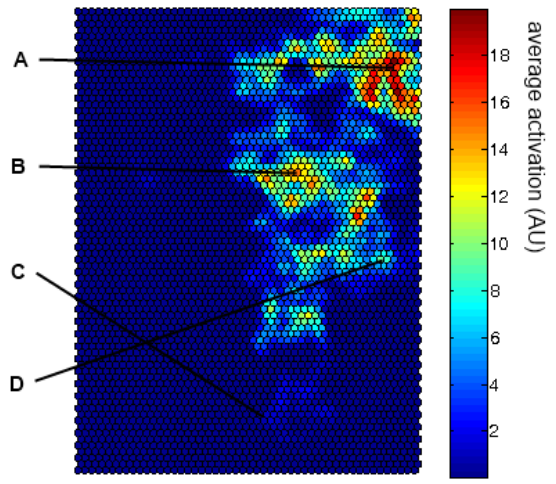

Figure 2. Average SOM-neuron activation during one-minute testing fragment, without bottom-up or top-down attention. Indicated nodes A, B, C and D represent a selection of different prototypical sounds.

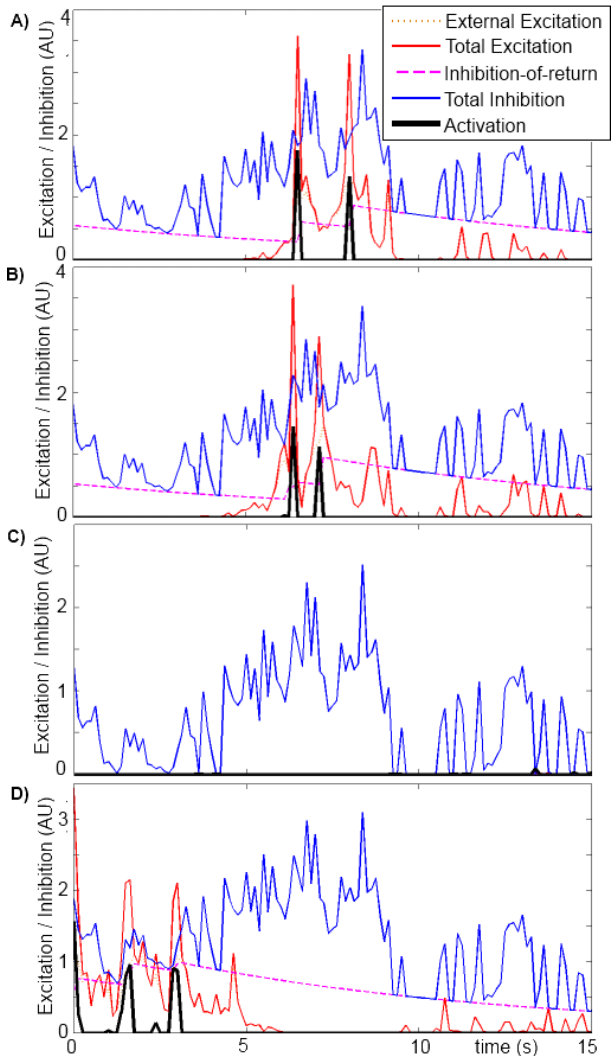

Figure 3. Evolution of different excitation and inhibition terms, as well as total activation of the nodes A, B, C and D as shown in Fig. 2, without bottom-up or top-down attention.

Secondly, the same procedure is repeated, including bottom-up attention by weighing external excitation with node-dependent saliency coefficients as described in section II.D. Fig. 4 shows the average activation and Fig. 5 shows the dynamics of the same four selected nodes as in Fig. 3. It can be seen that some of the activation of the SOM zone around node A moves to the zone around node B. This makes sense, as these two zones are often excited together, and the most salient of them, B, should be the one that is consciously perceived. In the zones around C and D, hardly anything changes. This was also to be expected, as when these sounds occur, they usually dominate the incoming sound, and therefore no other zones of the SOM map are excited simultaneously and saliency weighing cannot influence the node activation to a significant extent.

Finally, top-down attention is included in the simulation. For the neurons in the zone around D, inhibition-of-return time constants are changed to 1s for both increasing and decreasing values. Thus, inhibition-of-return increase is slower and decrease is faster in this zone than in the rest of the ANN, facilitating sustained attention on this zone. The average node activation in this case is given in Fig. 6 and the dynamics of the same nodes A, B, C and D are given in Fig. 7. It can be seen that now, during the whole period that a car is driving by, neuron D is activated, as inhibition-of-return does not switch attention to other zones on the map any more. Also, attention is drawn from the already weakly activated duck sound to a car driving by on a more distant road. Experiments with top-down attention focused on other zones of the map similarly yields plausible results.

IV. Conclusions

In this paper, a human-mimicking computational model of auditory attention is presented. It consists of a series of submodels that are inspired by existing models, but are adapted in order to be combined in one global model, to simulate different features of the human auditory attention focusing process. Common to all submodels is a balance between biological plausibility and computational complexity, as the model is aimed to be run on low-cost hardware for environmental sound monitoring during prolonged periods of time. The aim of the model is to detect and classify sound events of interest in a versatile way. Potential applications of the model are e.g. monitoring traffic noise, habitat monitoring or the computational analysis of urban soundscapes.

Application of the model consists of two phases. In a first phase, the model is trained on the sound at a particular location for a predefined period of time, e.g. a couple of weeks for urban outdoor environments. During this (unsupervised) training process, the model learns to classify the sounds that are present at the location of the microphone on the basis of co-occurrence of features. In the second phase, the model is employed to quickly detect and classify particular sound events of interest. An important feature of the model is that, on top of a bottom-up attention mechanism, it provides the possibility to focus (top-down) attention on those sounds that

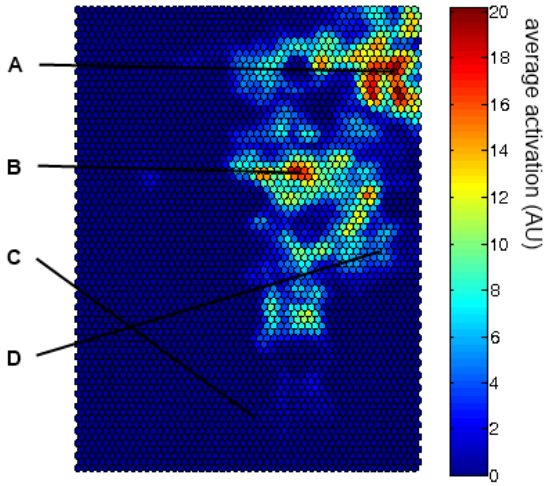

Figure 4. Average SOM-neuron activation during one-minute testing fragment, with only bottom-up attention. Indicated nodes A, B, C and D represent a selection of different prototypical sounds.

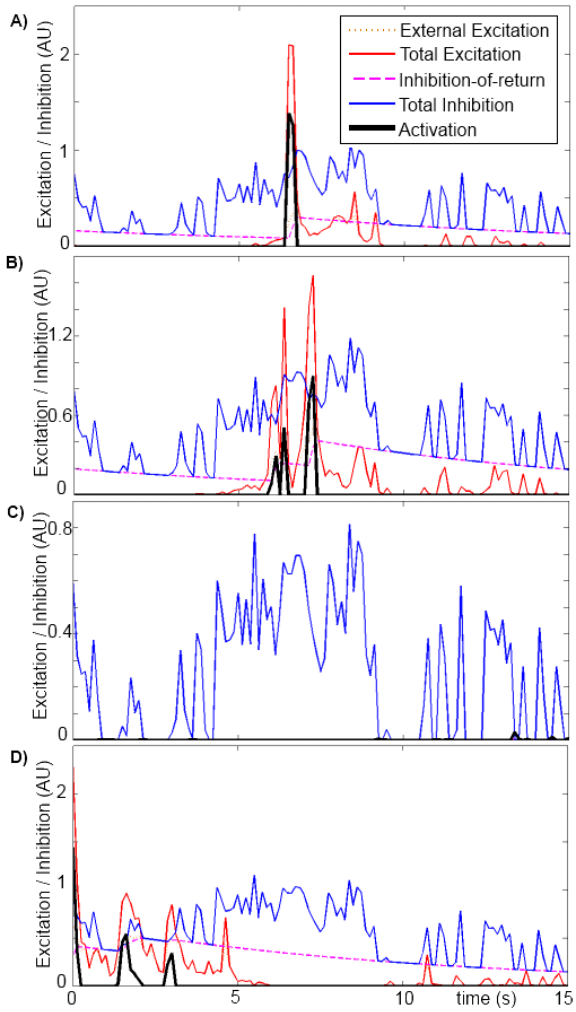

Figure 5. Evolution of different excitation and inhibition terms, as well as total activation of the nodes A, B, C and D as shown in Fig. 4, with only bottom-up attention.

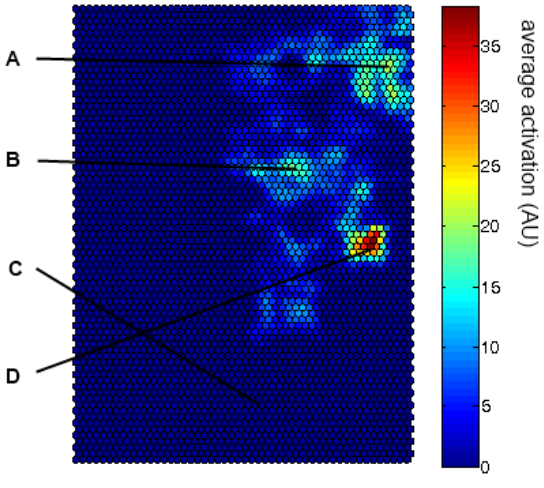

Figure 6. Average SOM-neuron activation during one-minute testing fragment, with bottom-up attention and top-down attention focussed on the zone around node D. Indicated nodes A, B, C and D represent a selection of different prototypical sounds.

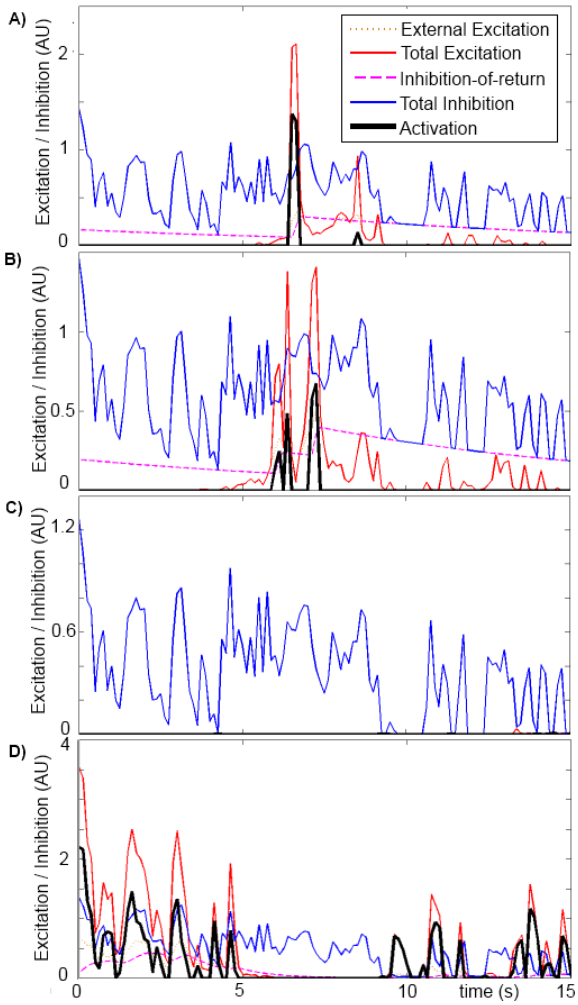

Figure 7. Evolution of different excitation and inhibition terms, as well as total activation of the nodes A, B, C and D as shown in Fig. 2, with bottom-up attention and top-down attention focussed on the zone around node D.

are of interest for the user. An implementation of the integrated model is tested on an actual urban soundscape, and it is shown that the model displays sensible, human-like behavior.

ACKNOWLEDGMENT

Michiel Boes is a doctoral fellow, and Bert De Coensel is a postdoctoral fellow of the Research Foundation–Flanders (FWO–Vlaanderen); the support of this organization is gratefully acknowledged.

REFERENCES

[1] 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.

[2] 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.

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

[4] R. P. Carlyon, R. Cusack, J. M. Foxton, and I. H. Robertson, "Effects of attention and unilateral neglect on auditory stream segregation," J. Exp. Psychol.-Hum. Percept. Perform., vol. 27, no. 1, pp. 115-127, 2001.

[5] E. I. Knudsen, "Fundamental components of attention," Annu. Rev. Neurosci., vol. 30, pp. 57–78, 2007.

[6] 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, Sep. 2011.

[7] B. De Coensel, and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in Proc. of the International Conference on Acoustics (ICA), Sydney, Australia, Aug. 2010.

[8] D. Wang and D. Terman, "Locally Excitatory Globally Inhibitory Oscillator Networks," IEEE Trans. Neural Netw., vol. 6, no. 1, pp. 283-286, 1995.

[9] 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), pp. 2017-2023, Hong Kong, 2008.

[10] 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, Aug. 2009.

[11] 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, Sep. 2004.

[12] E. Zwicker and H. Fastl, "Psychoacoustics. Facts and Models," edited by M. R. Schroeder (Springer, Berlin, 1999).

[13] C. Kayser, C. Petkov, M. Lippert, N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," Curr. Biol., vol. 15, no. 21, pp. 1943-1947, 2005.

[14] T. Kohonen, "Self-Organizing Maps," edited by Teuvo Kohonen (Springer, Heidelberg, 2001).

[15] 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.

[16] 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), pp. 1413-1420, Barcelona, Spain, 2010.