A computational auditory attention model for urban soundscape design

Damiano Oldoni[a)]
Bert De Coensel[b)]
Michiel Boes[c)]
Timothy Van Renterghem[d)]
Dick Botteldooren[e)]
Acoustic group of the Department of Information Technology (INTEC), Ghent University
Sint-Pietersnieuwstraat 41, Ghent, 9000, Belgium

Urban soundscape design aims to create outdoor spaces with a pleasant sonic environment, and is of special interest where noise level abatement is not feasible or only has a limited effect. A possible approach is to mask unpleasant sounds by adding other sounds that can be considered as desirable by the users of the space. In this context, not only energetic masking but also informational masking should be taken into account. The presented model of auditory attention provides a computational tool to assess the effectiveness of such interventions, thus reducing the need for a listening panel. After an initial training phase for a particular sonic environment, the model provides an acoustic summary containing the sounds that constitute the soundscape. Moreover, the model can be used to simulate how a listener switches attention between the sounds over time. The model, balancing computational efficiency and biological accuracy, provides the urban soundscape designer with a tool for analyzing both real and artificial mixtures. In this way, the perceptual effect of adding pleasant sounds can be assessed.

1 INTRODUCTION

Nowadays the design of urban outdoor spaces cannot prescind from acoustical aspects[1-3], especially if the purposes of such spaces are psychological restoration and general well-being. Composing pleasant acoustic environments is one the final goals of the soundscape designer:

[a)] email: damiano.oldoni@intec.ugent.be
[b)] email: bert.decoensel@intec.ugent.be
[c)] email: michiel.boes@intec.ugent.be
[d)] email: timothy.van.renterghem@intec.ugent.be
[e)] email: dick.botteldooren@intec.ugent.be

"desired" sounds should be often heard whereas "undesired" sounds should not be noticed by the listener. Knowledge of human auditory perception and new techniques for soundscape analysis are thus needed.

In this paper, a human-mimicking computational model for urban soundscape design is presented. It includes a model of auditory attention based on a self-organized mapping of acoustical features. The model, through extensive training, is tuned to a given sound environment and involves the construction of an acoustic summary, useful to provide an overview of the typical sounds composing the specific soundscape. By means of modeling auditory attention, the model can also simulate how listeners would switch their attention over time between different sounds. As such, this model can be a very useful tool for the soundscape designer in order to forecast the effects of soundscape interventions in a specific location.

There is a substantial difference between the present model and most of the existing computational auditory scene analysis models (see Wang and Brown[4] for an overview). Such models aim to extract sound samples as clean as possible for each auditory object of which the auditory scene is composed. Furthermore, the considered objects are defined a priori. A typical application is separating speech or other target sounds from the background, which is typically defined as all other sounds. Contrarily, the present model aims at analyzing the scene as accurately as a human listener would be capable of. A second important difference is the compromise between biological accuracy and computational efficiency due to the integration of the model in long-term sound measurement equipment.

The next section presents the computational framework, followed by a case study showing the applicability of the model in the soundscape design process. The paper ends with a section presenting conclusions and perspectives.

2 THEORETICAL FRAMEWORK

In this section, existing knowledge on human auditory perception of environmental sounds is worked out into a computational framework. The input of the model is the recorded sound signal at a particular location, the output is a measure of the potential attention given by a synthetic listener to different sound sources. As mentioned before, the model has been developed for long-term deployment: monaural signals are used and simplified models for loudness calculation, masking, auditory saliency and auditory attention are proposed.

The model can be decomposed into three submodels, illustrated in Fig. 1. First, the peripheral auditory processing is modeled and the calculation of a measure of the auditory saliency is performed (a), then the acoustical features are mapped based on co-occurrence (b) and lastly the auditory attention is modeled (c). In the next paragraphs a short description of each submodel is provided.

2.1 Peripheral Auditory Processing

The first stage aims to extract acoustical features in a way that mimicks human peripheral auditory processing. In particular, a feature vector is extracted, at regular time intervals, from the sound signal measured by the microphone. Instead of a detailed and computationally demanding time-frequency representation, the model starts from the 1/3-octave band spectrum of the sound pressure level with temporal resolution of 1 s. The advantage of such limited data rate (31 numbers per second) is that it can be implemented in a large-scale measurement network and data can be stored for very long periods of time.

Next, energetic masking is simulated by a cochleagram, calculated using the Zwicker loudness model[5]. The complete range of the audible frequencies (0-24 Bark) with a resolution of 0.5 Bark is considered, thus resulting in 48 spectral values.

To simulate the human auditory system, the absolute intensity and the spectro-temporal variations form the basis of the calculated acoustical features. Based on existing models for auditory saliency[6,7], they are calculated by means of a center-surround mechanism mimicking the receptive fields in the auditory cortex. Multiscale features are calculated by convolving 2D Gaussian (for intensity) and difference-of-Gaussian filters (for spectral and time gradients) with the cochleagram. For intensity 4 different scales are used, while spectral and time contrast are both encoded by 6 different scales. Figure 2 shows a section of the filters along the time or frequency axis. Thus, a feature vector is constructed at each timestep, consisting of 768 values (4+6+6=16 different filters times 48 spectral values).

2.2 Feature Co-occurrence Mapping

The second stage of the presented model consists of mapping the acoustical features of the incoming sounds based on co-occurrence. In order to obtain this mapping, a self-organizing approach has been chosen. The human auditory cortex shows, in fact, a clear tonotopic organization[8,9]: neurons next to each other are typically excited by similar stimuli. A Self-Organizing Map (SOM) or Kohonen Map[10] is thus used.

A SOM is formed by several units or nodes placed in a two dimensional array, forming a hexagonal lattice. Each node has a corresponding reference vector representing the unit position in the 768-dimensional sound feature space. After initialization, the coordinates of the reference vectors are modified during an unsupervised learning phase based on The Original Incremental SOM Algorithm[10]. In the latter algorithm, each iteration consists of two steps. First, for an input sound feature vector the closest reference vector is found, generally called the best-matching unit (BMU); second, the reference vector corresponding to the BMU and, to a lesser extent, those of the neighboring nodes, are moved closer to the input feature vector. After this unsupervised training phase, the distribution of the input data is nonlinearly mapped by the reference vectors of the SOM units. This implies that some regions of the high dimensional sound feature space will be densely and accurately mapped by the reference vectors of the SOM units whereas other regions will be sparsely and poorly represented.

However, the algorithm as described above does not yet take into account auditory attention, which influences human learning. In fact, a human listener would never describe a soundscape exclusively based on the rate of occurrence of particular sounds. Therefore, a second specific training algorithm is needed which accounts especially for highly salient sounds, likely able to attract attention. This training algorithm, called continuous selective learning[11], can be seen as a series of short learning periods: a new learning phase is triggered only if the distance between the input feature vector and the BMU is higher than a threshold T_1 (activation threshold) and it ends if the distance is less than a second threshold T_2 (deactivation threshold). This strategy allows the SOM to learn less frequently occurring sounds. Moreover, in order to promote the learning of salient sounds, a measure of the overall saliency, calculated from the sound feature vector (see De Coensel and Botteldooren[14] for details), is used to modulate the learning strength. Continuous selective learning significantly improves the capability of the SOM to identify, in terms of distance to the BMU, most of the sounds from the selected location. Each SOM unit encodes, by means of its reference vector in the sound feature space, an abstract sound prototype from the given location. In order to decode such information into hearable sound excerpts, a sound recording session is performed. A series of 5-second sound samples are extracted based on the

distance between their sound feature vectors and their BMUs of the SOM. The compilation of these representative sounds can be called the "acoustic summary" of the given soundscape[11]. The sounds are not automatically labeled. For this purpose, an expert listener can identify regions in the map for which the corresponding sound samples belong to specific sound sources.

2.3 Computational Model of Auditory Attention

To take into account auditory attention, an artificial neural network is introduced, coupled to the trained SOM. To each unit of the SOM, an artificial neuron is linked, and its excitation reflects the similarity of the incoming sound to the sound represented by its corresponding unit. In order to achieve this, feature vectors of the incoming sound are calculated, with a temporal resolution of 1s as indicated in Section 2.1. Then, the Euclidian distance between the incoming sound feature vector and the SOM units' reference vectors is calculated, and a Gaussian function is used convert this distance into a measure of similarity between the two vectors (approaching the value of one for very similar vectors and zero for very dissimilar vectors). Using this result as an excitation factor for the corresponding artificial neurons, zones on the map representing sounds similar to the incoming sound will be more strongly excited than others.

Next, bottom-up saliency-driven attention is introduced. This type of attention is a fast mechanism, promoting conspicuous, salient sounds, and operating in an unconscious way. It is implemented by calculating a saliency factor for each node, and weighing external excitations of the artificial neurons with this factor.

Concepts of a Locally Excitatory Globally Inhibitory Oscillator Network[12] (LEGION) are used in order to achieve the attention mechanism of competitive selection[13]. As in a LEGION, local excitation and global inhibition of neurons are introduced, but in order to keep the computational cost within bounds, no oscillators are used. By means of an iterative mechanism, alternating between adding excitation terms to neighboring neurons in the network proportional to the excitation of the neuron itself and adding a global inhibition term to all neurons in the network, only one or a few clusters of neurons are finally excited more strongly than they are inhibited, and are thus activated.

As in De Coensel and Botteldooren[14], inhibition-of-return (IOR) is introduced in order for the model to be able to scan the acoustic environment. By adding an inhibition term to neurons that were activated in the previous time step, continuous activation of the same nodes is made impossible, and attention is automatically shifted to another sound source after a certain time. Finally, conscious, top-down attention can be implemented in the model by modulating the time constants of the inhibition-of-return mechanism. A region of the SOM representing a sound of specific interest can be given higher IOR time constants causing attention to stay in this zone for a longer period of time before shifting to another sound source.

3 CASE STUDY

3.1 Description

In this section the model is tested and its application in the design of soundscaping measures is illustrated. In particular, the perceptual effects of attractive songbirds are assessed; the introduction of small green quiet areas, ideal for bird population, in the urban environments is in fact a feasible solution for increasing the pleasantness of the soundscape.

The sound in an urban street of Ghent was monitored by a fixed microphone. The sonic environment at this location is composed of road traffic noise (mainly car and tram noise) and

different kinds of noises produced by human activities due to the proximity of several shops and one school. The measurement station continuously recorded 1/3-octave band spectra at 1 s time intervals. Data measured during 3 weeks has been used for training the computational attention model as explained in the previous section.

A one-hour sound recording (L_{Aeq} = 68.2 dB(A)) was performed at the above described location during a working day within the training period. In order to mimic the effect of the capacity of songbirds to attract attention, 30 artificial one-hour sonic environments were subsequently created by mixing the original recording with an increasing number of randomly occurring bird vocalizations. Background-free bird sounds were used, for which the peak level was adjusted in order to agree with the sound level of the few bird vocalizations present in the original sonic environment. The sound level of the artificial sonic environments encompasses all possible situations, from a few sporadic vocalizations (46.3 dB(A), SNR = -21.9) to continuous bird chorus (75.8 dB(A), SNR = +7.6).

3.2 Results

The sound feature vectors and saliency values related to the 3 week period are calculated as explained in Section 2.1. Afterwards, a SOM composed of 3750 nodes (75 x 50) placed in a hexagonal grid was trained in three stages using the training algorithms explained in Section 2.2. First, the standard incremental SOM training algorithm was applied to the sound features related to 14 consecutive hours of the first day of measurements. The second phase consists of the continuous selective learning based on the 3 weeks measurement period, while in the third phase the same learning strategy was applied to the 30 artificial sound mixtures randomly ordered. The effect of the three successive learning phases is shown in Figure 3 by means of the U-matrix[15], which allows to visually distinguish regions of the map composed by units whose reference vectors are similar (small distance) from regions presenting high variability. After the continuous learning training phase, the map structure is richer and reveals cluster formation, thus showing a better adaption to the given sonic environment. From now on we refer exclusively to the last, fully trained SOM.

Starting from the 30 artificial soundscapes and several hours of recordings at the considered location, an acoustic summary was created and an expert listener marked the SOM units related to bird sound samples as in Fig. 4. It is found that most of the units representing bird vocalizations can be grouped in two distinct regions, schematically related to single bird chirps, region 1, and a chorus of bird vocalizations, region 2. Although the sound source is the same, the sound features related to single chirps and bird chorus are different, and are thus being mapped to different regions of the SOM.

The distribution of the occurrence of the BMU among the SOM units for the original sonic environment and the artificial sound mixtures is plotted in Fig. 5. It can be seen that the introduction of bird vocalizations progressively modifies the natural sonic environment. Such effect can be easily quantified by calculating the percentage of time the BMU belongs to region 1 or region 2, as a function of the SNR, as shown in Fig. 6. In particular, the percentage of time the bird chirp sound features are dominant (BMU belongs to region 1) increases until a peak is reached at a SNR equal to -2 dB. For sound mixtures with higher SNR, isolated single bird chirps are less likely, i.e. the percentage of time the BMU belongs to region 1 decreases monotonically. At the contrary, episodes of quasi-continuous bird chorus occur more and more often and the percentage of time BMU belongs to region 2 increases.

Subsequently, the attention mechanisms explained in Sect. 2.3 are taken into account (except for top-down attention as this would require a model of working memory) and the same

procedure is repeated. The fraction of time that attention is focused on bird sound is shown in Fig. 7. For sound mixtures with lower SNR, the fraction of time that bird vocalizations attract attention is slightly higher than in Fig. 6, while the opposite is true for higher SNR. On the one hand, at lower SNR the bird chirps trigger attention due to their high saliency without sensibly activating inhibition-of-return because they do not occur very often. On the other hand, for higher SNR, bird sound is always detectable, thus triggering the inhibition-of-return, i.e. shifting attention away from it. The results obtained from this computational model are in accordance with the results obtained by De Coensel et al.[17]: adding bird vocalizations to an urban sonic environment characterized by road traffic noise would increase the pleasantness of the soundscape already at an SNR of -10 dB.

4 DISCUSSION AND CONCLUSIONS

Human auditory perception forms the basis of the soundscape approach to the acoustic design of the living environment. However, the soundscape designer does not yet have a lot of methods and techniques at his/her disposal, that are based on mechanisms underlying human auditory perception and attention. In this paper, a computational auditory attention model for analyzing the urban soundscape was presented. It consists of simplified implementations of existing submodels for auditory saliency, topographic mapping, learning and auditory attention. In particular, it implements processes as bottom-up selective attention and learning and it is able to simulate how listeners would switch attention among the several sounds present in an urban sonic environment. The strong point of this model lies in its capability to analyze very long sound periods, an essential feature of potential computational models for soundscape analysis.

The model can be applied to extract an acoustic summary of a specific soundscape, a compilation of the typical sounds audible at a given location. Bird sounds, the sound from fountains etc., are typically considered to have a restorative effect[18, 19]. For this reason, the model can be even more useful for the soundscape designer due to its capability to assess the potential positive effect of adding sounds aimed to distract attention away from undesired sounds. By means of a case study, the effect of adding bird vocalizations to an urban sonic environment was studied and the results are in accordance with empirical results.

Modeling the mechanisms leading to top-down auditory selective attention is beyond the scope of this paper, as was modelling the process of meaning attachment to sounds. The latter would involve solving several linguistic issues[20] and taking into account the effects of inter-individual differences. These problems, together with automated labeling of the acoustic summary, are main issues for future work.

5 ACKNOWLEDGEMENTS

Bert De Coensel is a postdoctoral fellow, and Michiel Boes 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).

6 REFERENCES

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)

2. L. Brown and A. Muhar, “An approach to the acoustic design of outdoor space”, *J. Environ. Plann. Manage.*, **47**, (2004)

3. M. Zhang and J. Kang, “Towards the evaluation, description and creation of soundscapes in urban open spaces”, *Environ. Plann. B*, **34**, (2007)

4. D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, John Wiley & Sons, Inc., (2006)

5. E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models 2nd Edition*, Springer-Verlag (1999)

6. C. Kayser, C. Petkov, M. Lippert and N. K. Logothetis, “Mechanisms for allocating auditory attention: An auditory saliency map”, *Curr. Biol.*, **15**, (2005)

7. O. Kalinli and S. Narayanan, “Prominence detection using auditory attention cues and task-dependent high level information”, *IEEE Trans. Audio Speech Lang. Process.*, **17**, (2009)

8. 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**, (2004)

9. C. Humphries, E. Liebenthal and J. R. Binder, “Tonotopic organization of human auditory cortex”, *NeuroImage*, **50**, (2010)

10. T. Kohonen, *Self-Organizing Maps, 3rd Edition*, Springer-Verlag (2001)

11. 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”, *Proc. Internoise*, (2011)

12. D. Wang and D. Terman, “Locally excitatory globally inhibitory oscillator networks”, *IEEE Trans. Neural Netw.*, **6**, (1995)

13. A. Baddeley, “Working memory: looking back and looking forward”, *Nat. Rev. Neurosci.*, **4**, (2003)

14. B. De Coensel and D. Botteldooren, “A model of saliency-based auditory attention to environmental sound”, *Proc. ICA*, (2010)

15. A. Ultsch, “Self organized feature maps for monitoring and knowledge acquisition of a chemical process”, *Proc. ICANN*, **93**, (1993)

16. 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**, (2010)

17. B. De Coensel, S. Vanwetswinkel, and D. Botteldooren, “Effects of natural sounds on the perception of road traffic noise”, *J. Acoust. Soc. Am.*, **129**, (2011)

18. B. De Coensel and D. Botteldooren, “The quiet rural soundscape and how to characterize it”, *Acta Acust. Acust.*, **92**, (2006)

19. J. Kang and M. Zhang, “Semantic differential analysis of the soundscape in urban open public spaces”, *Build. Environ.*, **45**, (2010)

20. 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**, (2006)

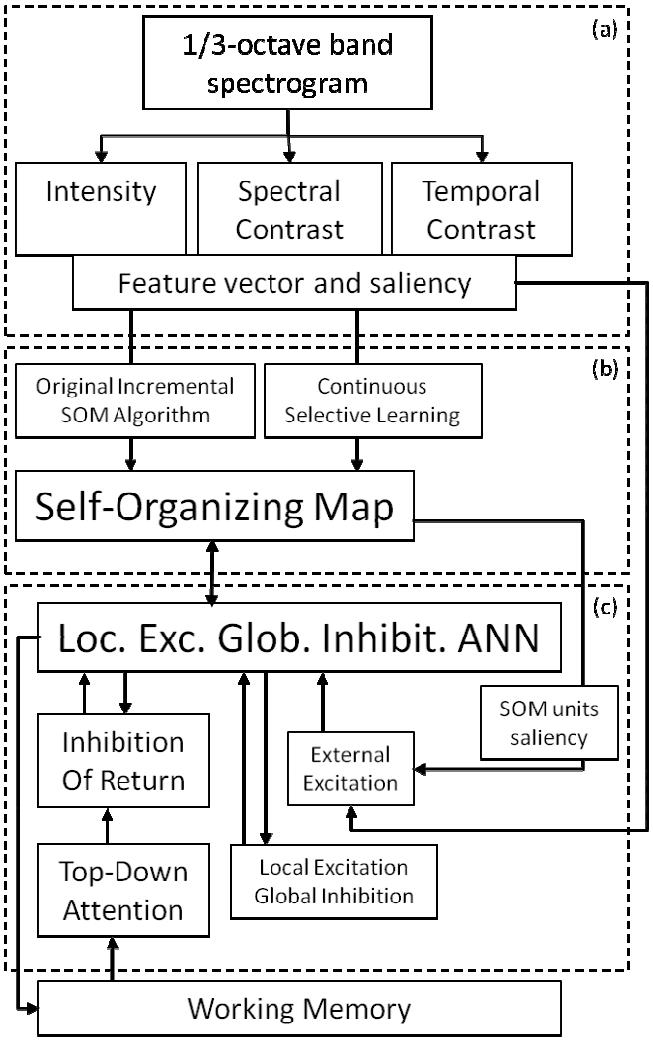

Fig. 1 – Overview of the proposed computational model: (a) peripheral auditory processing, (b) self-organized mapping of acoustical features based on co-occurrence, and (c) auditory attention.

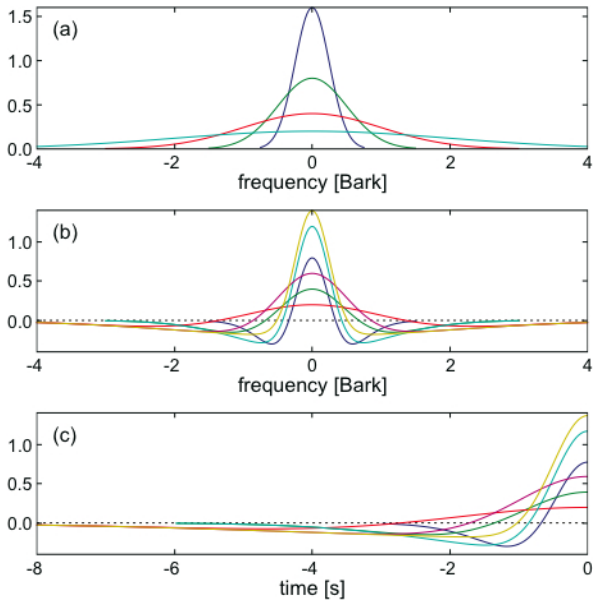

Fig. 2 – Cross section of the modelled receptive filters used to calculate (a) intensity, (b) spectral contrast and (c) temporal contrast. In (c) the convolution is performed only with the past in order to preserve temporal causality.

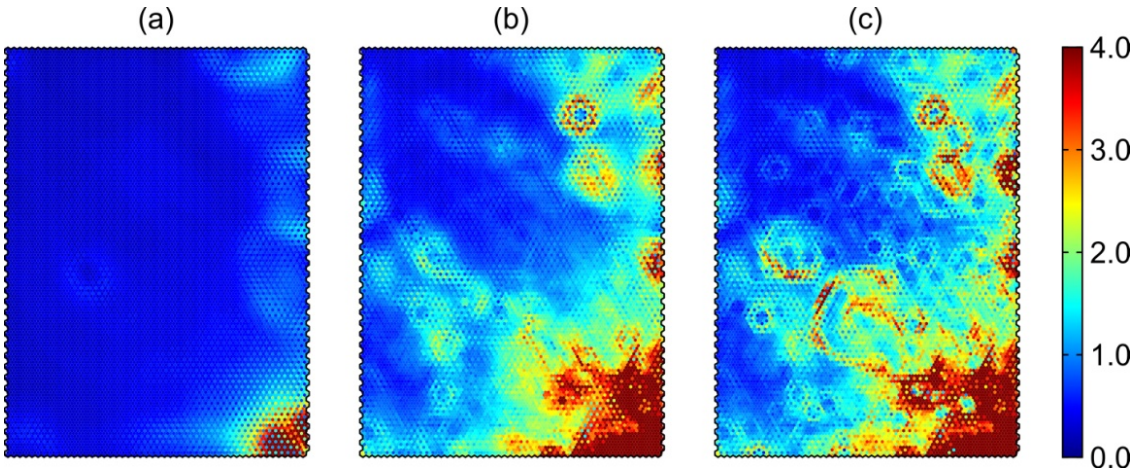

Fig. 3 – U-matrices showing the distance between the nearest units composing the SOM after each of the three stages of the training: the initial phase using the standard incremental SOM training algorithm applied to the sound features related to 14 consecutive hours of the first day of measurements, continuous selective learning based on (b) three weeks of data and (c) 30 one-hour artificial sound mixtures presenting a different number of randomly occurring bird vocalizations added to the original recording.

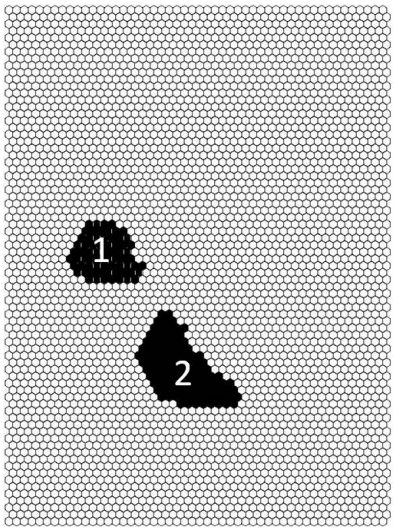

Fig. 4 – Regions of the SOM, related to individual bird chirps (region 1) and a chorus of bird song (region 2) as marked by an expert listener.

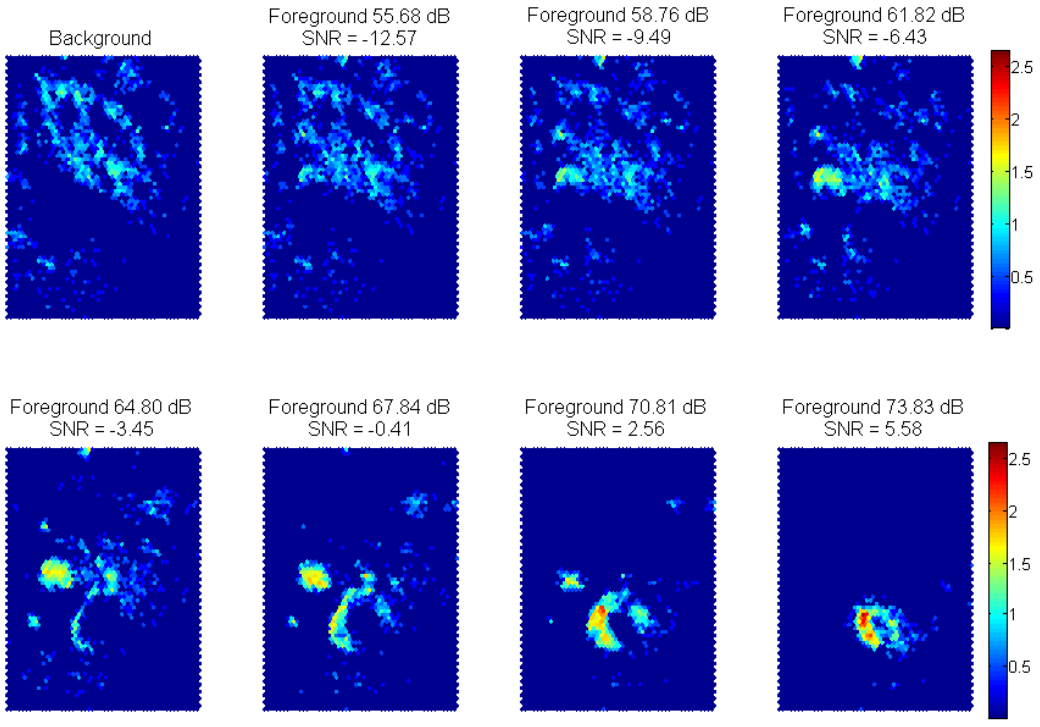

Fig. 5 – Distribution of the occurrence of the BMU among the SOM units for different sound scenarios: background, i.e. natural soundscape (upper left), artificial soundscapes, in which bird vocalizations (foreground) are progressively added to the background. For each sound scenario, 3600 testing samples (one hour) were used.

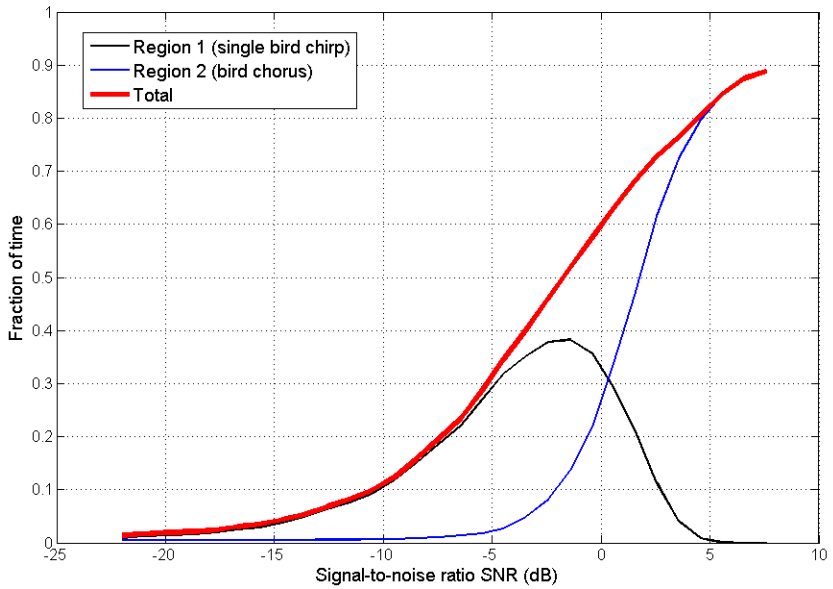

Fig. 6 – Fraction of time the BMU is found in region 1 (bird chirp, black line), region 2 (bird chorus, blue line) and their sum (red line) as a function of the SNR between the foreground (introducing bird vocalizations) and the background (without introducing bird vocalizations). 3600 sound samples (one hour) were used.

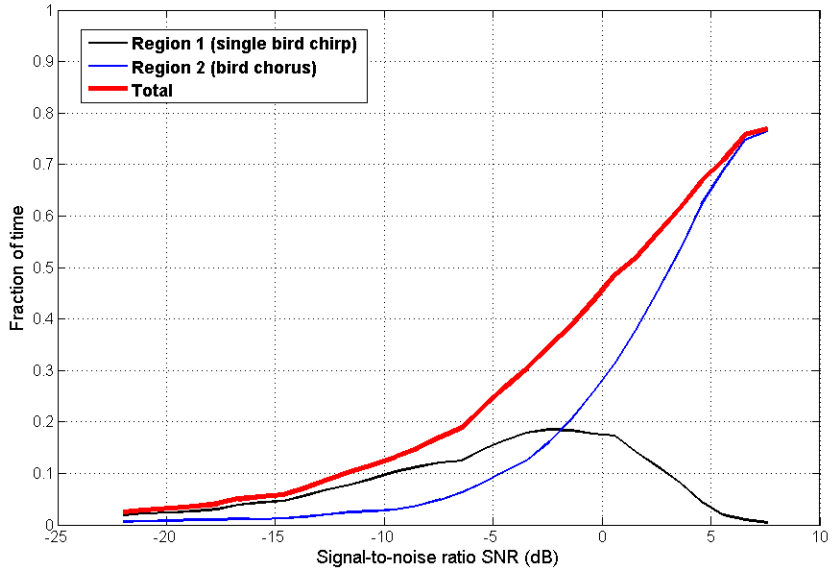

Fig. 7 – Fraction of time the simulated auditory attention is given to bird chirp sounds in region 1 (black line), bird chorus in region 2 (blue line) and their sum (red line) as a function of the SNR between the foreground (introducing bird vocalizations) and the background (without introducing bird vocalizations). 3600 sound samples (one hour) were used.