Proceedings of the Institute of Acoustics & Belgium Acoustical Society
Noise in the Built Environment, Ghent, 29-30 April 2010

IMITATING HUMAN AUDITORY PROCESSING FOR URBAN SOUNDSCAPE MEASUREMENT

D Oldoni	INTEC Ghent University, Ghent, Belgium
B De Coensel	INTEC Ghent University, Ghent, Belgium
M Rademaker	KERMIT Ghent University, Ghent, Belgium
T Van Renterghem	INTEC Ghent University, Ghent, Belgium
D Botteldooren	INTEC Ghent University, Ghent, Belgium
B De Baets	KERMIT Ghent University, Ghent, Belgium

Abstract

Intelligent measurement nodes offer the opportunity to perform advanced soundscape analysis, instead of just logging the sound pressure level. A model which mimics human auditory system is proposed and applied to analyze urban soundscapes. It is constructed as a combination of two types of neural networks: a Self Organizing Map that allows –after extensive training– to identify co-occurring sound features and a Locally Excitatory Globally Inhibitory Oscillator Network that is able to segregate them. The model takes into account the context of the listener and can be tuned to classify typical sounds of the soundscape at the location of the microphone.

1 INTRODUCTION

There is a growing awareness of the important role played by the context in human perception of sound. During the last decade, scholarly interest on soundscapes has grown sensibly. One of the open research problems is how soundscapes can be computationally analyzed in a way that mimics how humans perform this task, taking into account the context of the listener. Solutions to this problem could be readily applied in acoustical sensor networks. Intelligent measurement devices offer the opportunity to perform advanced analysis of the soundscape, instead of just logging the sound pressure level. However, although the functional knowledge on general auditory perception is rather advanced, up to date there exist no satisfying computational models for auditory scene analysis of general sound environments. Most effort has been spent at implementing computationally efficient models for performing highly specific tasks (such as removing background noise from speech signals), rather than on building biologically plausible, human-mimicking models that can be applied to a wide range on environmental sounds.

In this paper a biological plausible model for human auditory processing is proposed. It is constructed as a combination of two types of neural networks: the first is a Self-Organizing Map (SOM) that allows –after extensive training– to identify co-occurring sound features and the second is a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION) that is able to perform stream segregation of sound. The training of a Self-Organizing map mimics the human learning phase while its final structure is the computational counterpart of the complex morphology of the auditory cortex. This novel model, even if not yet sufficiently accurate to provide detailed sound recognition, reveals to be future proof, especially considering the inclusion of the major role played by the context of the listener. The model provides the possibility to be implemented in a sound measurement network for the detection of rare events and for general environmental sound monitoring.

In Section 2 we provide a description of the different stages which compose the model. Its application to different soundscapes, results and discussions are presented in Section 3. Finally, Section 4 offers some general conclusions about the model and the provided results, and an overview on the short and long term future developments of this model and its possible applications.

2 METHODOLOGY

2.1 Sound feature extraction

The model starts from the sound signal measured by the microphone, and calculates the 1/3-octave band spectrum with a resolution of 1s. Such quite low time resolution is chosen considering the fact that typical urban sounds (cars, trains, tram etc.) show relative slow temporal variations[1,2]. To simulate energetic masking a cochleagram is calculated using the Zwicker loudness model[3]. The cochleagram is intended to cover the complete range of the hearable frequencies, that means from 0 to 24 Bark, and with a resolution of 0.5 Bark, thus resulting in 48 spectral values at specific frequencies f_j = 0.5 j Bark. To simulate the human auditory system the absolute intensity and the spectro-temporal variations are the most important features that are to be detected. Models for measuring the auditory saliency already exist[4,5,6]. The proposed model uses centre-periphery mechanism, thus simulating the receptive fields in the auditory cortex. Two dimensional gaussian and difference-of-gaussian filters are applied through convolution to the cochleagram. Figure 1 shows a section of the receptive filters in the time and frequency domain separately. The sound features extraction results in a feature vector consisting of 16×48 = 768 values.

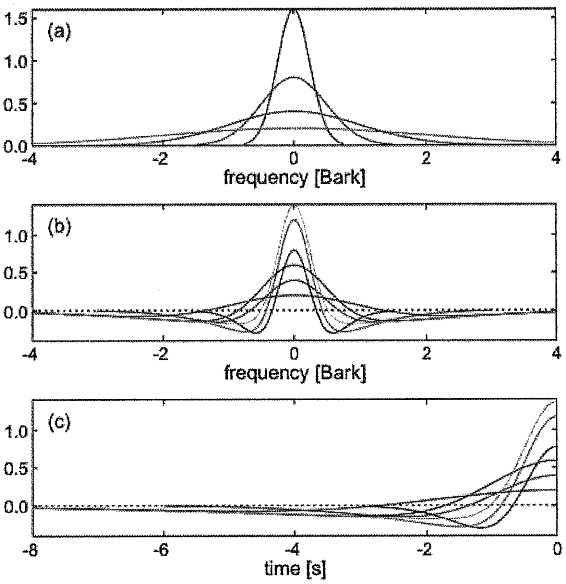

Figure 1 Filters used to calculate (a) intensity, (b) spectral contrast and (c) temporal contrast. In (c) temporal causality is preserved by only convolving with the past.

2.2 Learning and feature co-occurence analysis: Self-Organizing Map

Some regions of the cerebral cortex are specialized to decode specific sensory information as visual or auditory signals while other regions are organized to process specific tasks as speech control or eye movements. Within several of these regions, like the somatosensory or the visual area, a detailed *topological preservation* between the sensory signals features and a specific cortex region exists. These fine structures are generally called *brain maps*. The map of the auditory stimuli, called *tonotopic map* is found to be almost exactly logarithmically ordered with respect to the frequency, representing a statistical mapping of the occurence of the different tones[7].

The Self-Organizing Map (SOM) was originally thought to mimic such neurological mapping of the sensory stimuli into the cerebral cortex. The most important feature of such artificial neural network is the topological preservation which combined with the usage of a 2D network results in a dimensionality reduction of the mapped space[7]. The typical network is formed by several *units* or *nodes* placed in a 2D array, often forming an hexagonal lattice. Each node has a corresponding *reference vector* which represents the node position in the high-dimensional space: in our case it is the 768-dimensional sound features space. Each node can be thought of as an abstract sound feature prototype. After initialization, the coordinates of the reference vectors are modified during the training phase which consists of the iterative application of the following algorithm:

1. for a high-dimensional input vector the closest reference vector is found. The corresponding node in the 2D grid is usually called the best-matching unit (BMU).
2. Move the reference vector corresponding to the BMU and, to a lesser extent, those of the neighboring nodes, closer to the input high-dimensional vector.

These steps are repeated for each input vector. In our case at least 86400 samples (the seconds in a day) are used without repetitions. Re-feeding the SOM with the same samples, to achieve a likely better training, would contradict the intent to build a model grounded on biological and psychological knowledge: in fact, it would imply an artificial time discontinuity and the unrealistic situation to listen multiple times to a single sound. The final position of the reference vectors are strongly influenced by several parameters, for instance the size of the neighborhood of the BMU and the learning factor which rules how much the reference vectors near to the input need to move. To reduce the time of training, linear initialization of the reference vectors is preferred to the random initialization[7].

After training the set of the reference vectors positions can be seen as a nonlinear projection of the probability density function of the high-dimensional input space. In our case, a SOM, extensively trained with saliency features vectors, has learned through the position of its reference vectors what features have been often co-occurred together. Thus, if a new saliency feature vector is provided, the distance to its BMU is an indirect measure (assessed through saliency features) of how often the corresponding sound occurred. By the means of a threshold, it is possible to select what nodes of the map are near to the new saliency feature vector. It is like a comparison between the features of an input sound and a set of "abstract sound samples" with certain features.

The nodes of the map that are sufficiently near to the new saliency feature vector form clusters, usually one, two, rarely three. If more than one zone is near to the input sample, it means that the observation contains a rarely occurring superimposition of different saliency features values. These areas of the map can be interpreted as different zones of the auditory cortex having been excited by external stimulation. To segregate such areas in a biologically plausible way, one has to take the neuronal oscillatory correlation theory into account.

2.3 Segregation: LEGION

Oscillatory correlation properties of the neurons in the visual sensory cortex were discovered and intensively investigated since the end of 1980s[8,9]. Evidences of synchronous oscillations were discovered in auditory cortical cortex[10,11] too. The first computational model was created by von der Malsburg[12] and extensively developed in the so-called *shift-synchronization theory* by Wang[13,14]. The main point of the Wang's model is to represent each sound object as a synchronized group of neuronal oscillators corresponding to specific sound features. If sounds with different features are contemporarily present, the two groups of neural oscillators are internally synchronized but desynchronization between the two groups occurs. The general model developed by Wang is called LEGION[15] and it was used for static image segregation and simple speech retrieval from noisy contexts. It is normally a two-dimensional grid of neuronal oscillators, whose dynamics are, for each oscillator, ruled by the combination of an excitatory unit x_i and an inhibitory unit y_i:

$$\dot{x}_i = 3x_i - x_i^3 + 2 - I_i + S_i + \rho$$
$$\dot{y}_i = \varepsilon(\gamma(1+\tanh(x_i/\beta)) - y_i)$$

where I_i is the external stimulation, H is the Heavyside function, S_i is the total coupling due to the near oscillators' state, ρ is a source of Gaussian noise, γ, ε and β are three regulating parameters.

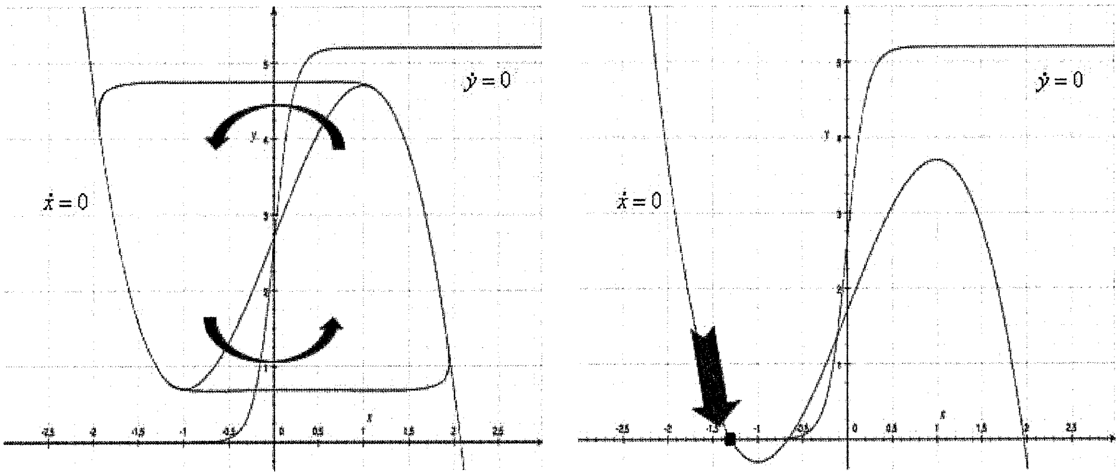

Figure 2 Left: Stable limit cycle of an enabled oscillator. It oscillates between a silent phase or left branch (x<0) and an active phase or right branch (x>0). The two phases are linked by jumps, colored in red. Right: If no external stimulation occurs, the oscillator relaxes to a stable fixed point, in red.

The curve $\dot{x}=0$ is a cubic curve and $\dot{y}=0$ a sigmoid. If there is no external excitation, a fixed stable point occurs (see Figure 2, right). Contrarily, if an external excitation is present $I_i > 0$, the two curves intersect only in a point and the oscillator is moving in a stable limit cycle (Figure 2, left). Such periodic solution can be described as an oscillation between a silent phase or left branch and an active phase or right branch. The synchronization within each group of oscillators and the desynchronization among different groups is entailed by the coupling term S_i, composed of two terms: a *local coupling* with any active nearby oscillators and a *global inhibition* when at least one oscillator is active (for more details see the detailed description of the LEGION model[13,16]).

To conclude, the most important insights on the SOM - LEGION coupling are:

- there is a one-to-one correspondence between the oscillators of the LEGION and the nodes of the SOM.
- The oscillators corresponding to nodes, whose distance to the input saliency feature vector is less than the fixed threshold, receive a positive external stimulation (I_i>0).
- Each second a new saliency feature vector is provided. It means that the external stimulation is not static.

3 RESULTS

This model was tested on two different soundscapes: an urban environment with a mixture of quiet, light and heavy traffic noise, labelled as T and a botanic garden, with a limited human presence, labelled as P. The soundscape for each scenario was constantly monitored by two measurements stations which recorded continuously standard 1/3-octave band levels at time interval of 1s. Two SOMs, one for each soundscape, were trained with samples corresponding to an entire day, that is 86400 1s samples. During the learning phase the two maps could learn, based on saliency, the typical features of the corresponding soundscape, providing a sort of picture of the sound scenario at T or P. It is clear that such training produces strong sound-context dependent maps. This can be seen in Figure 3.

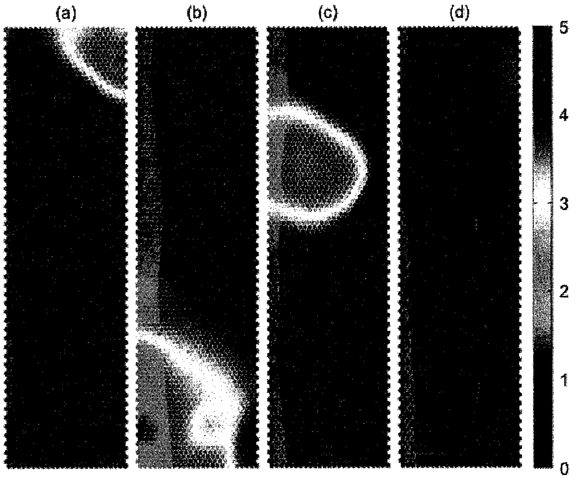

Figure 3 Distance between a quiet sound sample from P and the nodes of the SOM trained in (a) T, (b) P; distance between a noisy sample excerpt of passage of a car in T and the nodes of the SOM trained in (c) T, (d) P.

An excerpt taken during the passage of a car in T cannot be recognized by the SOM trained in P, Figure 3(d). At the contrary, the SOM trained in T is sufficiently adapted and a well localized group of nodes is found to have very similar saliency features values, Figure 3(c). The same does not hold if the input is a typical quiet sound sample from P: although not at the same extent as the SOM in P, the SOM trained in T can recognize such input, Figure 3(a-b).

The context-focused SOMs of the previous example are not well corresponding to the everyday human experience. We are usually used to various soundscapes: we can in general recognize easily cars or trucks or the bark of a dog. We can thus roughly model the human auditory system as a SOM trained in different sonic environments. A multi-context learning was provided by the means of 51 sound excerpts of 15 minutes recorded at various locations in and around Ghent, like shopping streets, streets with low and high traffic intensity, street canyons, industrial sites, quiet residential areas and urban parks. Two new SOMs were trained separately using the new data and partly the training samples from T and P respectively. As showed in Figure 4, the old SOMs, trained in T or P, are not sufficiently able to recognize input samples coming from new sound scenarios. The example provided in the figure is an excerpt from a crowded shopping street in the city centre.

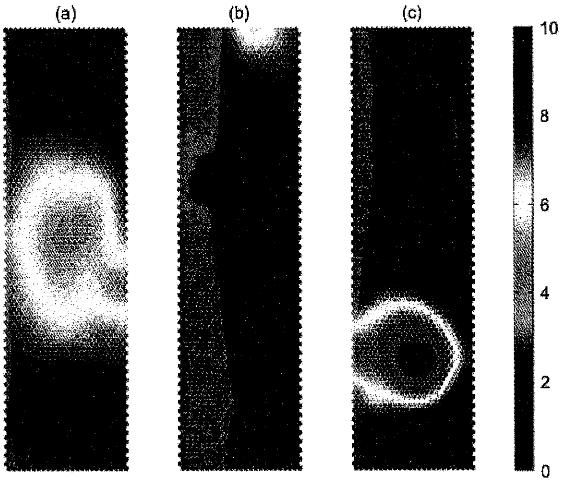

Figure 4 Distance between the saliency features vector of a sample taken from a crowded street and the units of the SOM trained in (a) T, (b) P and (c) the multi-context scenario composed of 51 different locations and part of the samples from T.

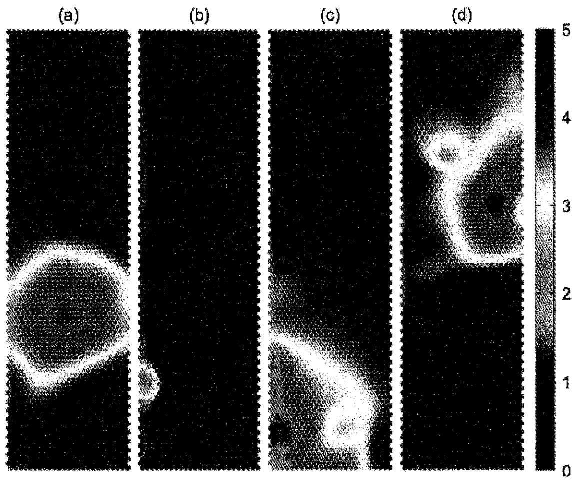

Figure 5 Distance between the saliency features vector of a sample taken from the passage of a car in T and the SOM trained in (a) T, (b) the multi-context scenario composed of 51 different locations and part of the samples from T; distance between a quite sound sample in P and the SOM trained in (c) P, (d) the multi-context scenario composed of 51 different locations and part of the samples from P.

At the contrary the new maps reveal to be very versatile, being able to recognize samples both from the new scenarios (Figure 4) and from T or P (Figure 5).

Like explained in the previous section, the function of the SOM is to simulate the formation and the activity of the auditory cortical map. After the learning phase, every saliency features vector stimulates one or more regions of the SOM: the distance to the nodes is used to univocally select such zones by means of a distance threshold. Once the stimulated zones of the map are localized, we simulate the neuronal activity by LEGION, as explained in 2.3. In Figure 6 the binarization and the oscillatory properties of LEGION are shown for an excerpt of 2 seconds. The segregation of the different groups of oscillators is an emergent property of the network, usually preceded by a transient phase.

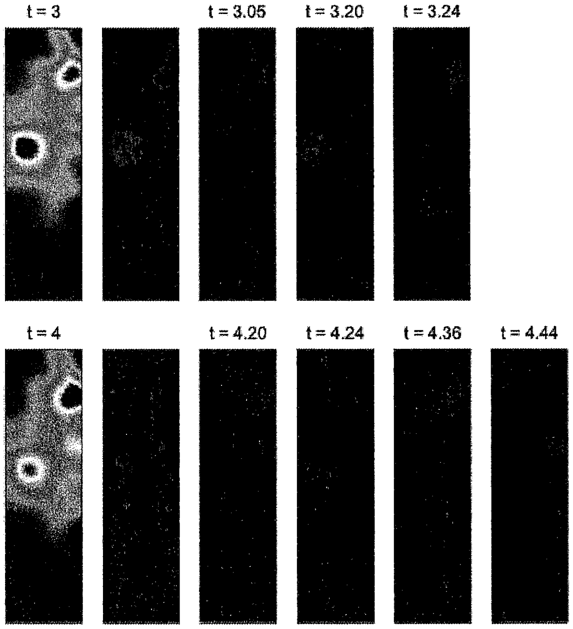

Figure 6 Left (2 columns): the inverse of the distance between two consecutive input samples (t=2s at the top, t=3s at the bottom) from P and the SOM trained in the same location before (first column) and after binarization (second column). Right: snapshots of the oscillatory activity of LEGION taken at different times.

4 DISCUSSION AND CONCLUSIONS

The aim of this paper was to model the human auditory processing in soundscape analysis. The first important stage of the model is the extraction of the features which the human auditory system is sensitive to. The absolute intensity and the spectro-temporal variations are calculated based on 1s standard 1/3-octave band levels. The learning phase of the human auditory cortex was simulated by the training of a Self-Organizing Map (SOM), which mimics the construction of a tonotopic map. Finally, the oscillatory activity of the auditory cortex during listening is modelled by a specific network of oscillators called LEGION grounded in the oscillatory correlation theory. In this section we discuss the obtained results and the challenging theoretical and practical problems to be solved.

From Figure 3 and Figure 5, it is evident that the more heterogeneous the soundscape on which a SOM is trained, the smaller is the area dedicated to each specific type of sound. This is clear comparing Figure 3(a) and Figure 3(b): a quiet sample activates a bigger area in the SOM trained on the park scenario (P) than the SOM trained on the road traffic scenario (T). The latter soundscape is more heterogeneous having moments of quiet alternated with typical road noise. The same behaviour can be seen in Figure 5 where a comparison between the SOMs trained on the multi-context scenarios and the previous SOMs is provided. Although the multi-context map is versatile, a side effect is that the distance of the input samples to their BMU is greater. The problem is that every SOM was trained using the same number of samples, that is 86400: thus, the multi-context SOMs could use only 40500 samples from T or P, less than half the number of samples used by the SOMs trained exclusively on T or P. A solution is to assess the number of training samples roughly depending on the number of contexts.

Although the extensive training on heterogeneous sound scenarios mimics the typical human experience, the importance of context-dependent maps has not to be underestimated: such specific maps can be useful to simulate the activity of specific auditory cortex areas of some individuals involved in activities which need a detailed knowledge of a particular soundscape, e.g. bird hunters, instruments makers, mechanics, etc. Moreover, involving context dependency can be useful for soundscape recognition. Let consider a set of SOMs, each one trained on a different specific soundscape. Given a sample from one of these soundscapes, the comparison of the focusation and the distance to the BMU of each SOM will yield information about which soundscape the sample is coming from.

In our model all the training samples are equally weighted. However, the importance of the bottom-up attention triggered by auditory saliency has been extensively studied and modelled[17]. Low salient sound events are considered less important, not triggering the attentional process that is a prerequisite for active learning. The attention should play a role also in the LEGION activity, being the attention very important to select what group of oscillators is dominant. However it supposes the equivalence between a group of neuronal oscillators and a sound object which is still not yet completely demonstrated in our model. This problem can be likely solved by a much longer training: 86400 samples are not so much if compared with the learning phase and the plasticity of the human auditory cortex.

Our model presents however a conceptual inconsistency with the real human auditory system, which is triggered when a new sound, not recognizable, is heard. In our model it would often mean very high distance to every node of the SOM as seen in Figure 3 and Figure 4, thus corresponding to a complete absence of neuronal oscillatory activity. Considering the LEGION activity an instance of higher mental functions does not solve the problem: even a new sound event is subject to the same processes of the auditory scene analysis, being segmentation, integration and segregation[18]. The paradox can be solved by only considering the rule that the attention should play in the binarization threshold assessment. In our model the threshold was proportional to the 3-order simple moving average of the inverse of the distance of the BMU. The attention should be able to modulate the coefficient of proportionality, thus letting to the new sound to elicit a group of oscillators. Another typical feature of the human auditory cortex not yet implemented in our model is the *dynamic learning*: if a new sound occur and the attention level is high, the learning should be triggered.

s

5 REFERENCES

1. B. De Coensel and D. Botteldooren, 1/f noise in rural and urban sound soundscapes, ActaAcust.Acust. vol 89(2) 287-295. (2003).
2. B. De Coensel and D. Botteldooren, The quiet rural soundscape and how to characterize it, ActaAcust.Acust., vol 92(6) 887-897. (2006).
3. E. Zwicker and H. Fastl, Psychoacoustics. Facts and Models, 2nd ed Springer Series in Information Sciences Springer-Verlag. Berlin (1999).
4. C. Kayser, C.Petkov, M. Lippert and N.K. Logothetis, Mechanisms for allocating auditory attention: an auditory saliency map, Curr.Biol. 15(21) 1943-1947. (2005).
5. O. Kalinli and S. Narayanan, A saliency-based auditory attention model with applications to unsupervised prominentsyllable detection in speech, Proc. 8th Interspeech, 1941-1944. Antwerp (August 2007).
6. V. Duangudom and D.V. Anderson, Using auditory saliency to understand complex auditory scenes, Proc. 15th EUSIPCO, 1206-1210. Poznań (September 2007).
7. T. Kohonen, Self-Organizing Maps, 3rd ed Springer. (2001).
8. R. Eckohorn et al., 'Coherent oscillations: A mechanism of feature linking in the visual cortex', Biol.Cybern., 30 169-180. (1988).
9. C.M. Gray, P. König, A.K. Engel and W. Singer, 'Oscillatory responses in cat visual cortex exhibit intercolumnar synchronization which reflects global stimulus properties', Nature 338 334-337. (1989).
10. P.E. Maldonado and G.L. Gerstein, 'Neuronal assembly dynamics in the rat auditory cortex during reorganization induced by intracortical microstimulation', Exp.BrainRes. 112 431-441. (1996).
11. M. Brosch, E. Budinger and H. Scheich, 'Stimulus-related gamma oscillations in primate auditory cortex', J.Neurophysiol. 87 2715-2725. (2002).
12. C. von der Malsburg, 'The correlation theory of the brain function', Max-Planck-Institute for Biophysical Chemistry, Internal Report 81(2). München (1981).
13. D.L. Wang, 'Auditory stream segregation based on oscillatory correlation', Proc. 4th IEEE Workshop NNSP, 624-632. Ermioni (September 1994).
14. D.L. Wang, 'Primitive auditory segregation based on oscillatory correlation', Cogn.Sci. 20(3) 409-456. (1996).
15. D.L. Wang and D. Terman, 'Locally excitatory globally inhibitory oscillator networks', IEEE Trans.Neural Net. 6(1) 283-286. (1995).

16 D. Terman and D.L. Wang, 'Global competition and local cooperation in a network of neural oscillators', Physica D 81(1-2) 148-176. (1995).

17. 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(2) 656-665. (2009).
18. A.S. Bregman, Auditory scene analysis: The perceptual organization of sound, The MIT Press. (1994).