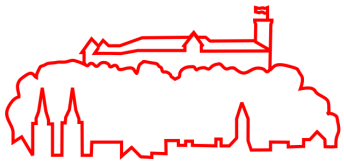

1st EAA – EuroRegio 2010
Congress on Sound and Vibration
15 - 18 September 2010, Ljubljana, Slovenia
With Summer School for Young Researchers from 13 - 15 September 2010

Computational soundscape analysis based on a human-like auditory processing model

Damiano OLDONI[a], Bert DE COENSEL[a], Michael RADEMAKER[b],
Timothy VAN RENTERGHEM[a], Dick BOTTELDOOREN[a] and Bernard DE BAETS[b]

[a]Department of Information Technology, Ghent University,
St.-Pietersnieuwstraat 41, B-9000 Ghent, BELGIUM, e-mail: damiano.oldoni@intec.ugent.be
[b]Department of Applied Mathematics, Biometrics and Process Control, Ghent University,
Coupure Links 653, B-9000 Ghent, BELGIUM

ABSTRACT

In the context of computational soundscape analysis, few attempts have been made to imitate the human approach to soundscape perception. This paper introduces a biologically plausible and human-like auditory processing model for general soundscape analysis. The model is based on two types of neural networks. The first is a Self-Organizing Map (SOM) that models the plasticity during learning and the complex morphology of the human auditory cortex and allows to recognize co-occurring sound features. The second is a Locally-Excitatory-Globally-Inhibitory-Oscillator-Network (LEGION) simulating the oscillatory correlation property of neurons of the auditory cortex. The model is shown to be highly context dependent. The potential of SOM in soundscape analysis and its use as a sound classifier are extensively studied. The proposed methodology can easily distinguish between common and rare sounds of a given soundscape, thus drawing up its acoustic summary.

1. INTRODUCTION

Soundscape analysis is a quite new resesarch field whose computational counterpart has so far remained largely unexplored. The reasons are manifold, but most of them stem from the difficulty to reproduce the human psychological mechanisms that allow labelling a sound as typical or atypical for the given context. Some models that do not draw on humans mimicking techniques for sound recognition tasks, aim to provide generic environmental sound recognition, but they do not attempt to provide soundscape description. The psychoacoustical processes underlying the detection or the recognition of a sound are governed by numerous psychological variables, i.e. sound context perception, attention level and habituation.
In this paper a biologically and psychologically plausible model for computational soundscape analysis is proposed. It is based on a combination of two types of neural networks that imitate different stages of human auditory processing. Factors like context dependency and learning from experience are taken into account.

2. METHODOLOGY

Overview

The model starts by extracting sound features encoding loudness and spectro-temporal irregularities from standard 1/3-octave band levels. Contrary to other methods focused merely on sound recognition and deploying complicated tone detectors by means of harmonic analysis, such sound features are designed to extract only the information useful for potentially triggering the so-called *bottom-up attention* [1]. Subsequently, the sounds forming the soundscape are discerned using a combination of two types of neural networks: a Self-Organizing Map (SOM) [2] that allows—after extensive training—to identify co-occurring sound features and a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION) [3] for grouping and segregation of corresponding sound feature clusters. Contrary to [4], here we dynamically train a SOM in order to imitate human continuous learning. After an initial static learning phase, a new learning phase is triggered whenever a bad matching between the SOM and the new sound feature vectors occurs.
The structure of the SOM resembles for certain aspects the complex morphology of the human auditory cortex. The combination of SOM and LEGION models two essential features of the brain: the SOM mimics the plasticity and the learning abilities of the network of neurons forming the auditory cortex, while the LEGION approximates the dynamic oscillations between connected neurons. This combines part of the auditory experience into “a sound”.

Sound feature extraction

The model starts from the 1/3-octave band spectrum of the sound pressure level with temporal resolutions of 1 s. The energetic masking is the first psychoacoustical phenomenon to be taken into account: it is simulated calculating a simplified cochleagram $s(f,t)$ via the Zwicker loudness model [5] with a frequency range from 0 to 24 Bark and a frequency resolution of 0.5 Bark for a total of 48 frequency values.
The second step highlights the loudness changes and spectro-temporal irregularities embedded in the sound signal. This is done by convolving the cochleagram with various 2D gaussian and difference-of-gaussian filters: the gaussian filters are suitable for intensity encoding, the latter for spectral and temporal contrast. In particular 16 different scales are used: 4 for intensity, 6 for spectral and temporal contrast each. By means of convolution, $16 \times 48 = 768$ values are extracted at each timestep; such values can be seen as a vector called *sound feature vector* [4]. From now on we will work only with such vectors and not with the sound spectrogram.

SOM: sound context and learning

The Self-Organizing Map (SOM), called also *Kohonen map* [2], is a neural network often used as a nonlinear dimension reduction technique. The network is composed of several nodes placed in a 2D grid, usually forming an hexagonal lattice. Each node has a corresponding *reference vector* which represents the node position in the high-dimensional sound feature space. After initialization, the training can modify the position of the reference vectors by means of the sound feature vectors calculated from the initial 1/3-octave band spectrogram as previously explained. The training is a vast number of iterations of a 2-step algorithm. The first step of the algorithm is to feed the SOM with an input sound feature vector and determine its nearest reference vector, whose corresponding node is called best matching unit (BMU). In the second step the reference vector of the BMU and, to a lesser extent the reference vectors of its neighbours, are moved closer to the position of the input sound feature vector. For more details about the SOM mathematical model in the current context see Oldoni [4].

After this training, the reference vectors of the SOM units provide a nonlinear projection of the probability density function of the input data. Furthermore feature vectors occurring very often during the training phase will be very close to the corresponding BMUs and are said to be well recognized. However, sound feature vectors that occur rarely are usually not well represented by the trained SOM, although they are also important for a comprehensive and more realistic soundscape representation. To address this problem, dynamic learning can be adopted: during SOM operation, a new learning phase is initiated when the distance of the input sound feature vector is greater than an activation threshold, T_1, and it continues until the distance is less than a disactivation threshold T_2, with $T_2 < T_1$. After such training the updated SOM is used until a next dynamic training session is triggered.

The learning parameters and the initial number of degrees of freedom (number of nodes) in the SOM have to be chosen carefully in order to not lose information from previous training sessions. This kind of dynamic learning aims to reproduce the continuous human learning based on experience. An alternative method, not yet tested, could be the insertion of new nodes to enhance learning and avoid any loss of previously acquired information.

Usually a SOM trained on a certain soundscape cannot match typical sounds of other soundscapes, due to the innate context dependency of the training phase. However, applying the dynamic learning strategy, a SOM can improve its matching power to new sounds without losing information about what was previously learned.

LEGION: synchronization and segregation

During the last thirty years oscillatory correlation properties of sensory cortical neurons were intensively studied [6] and in particular evidence of synchronous oscillation was discovered in the auditory cortical cortex [7, 8]. Based on the work of von der Malsburg and Schneider [9], Wang developed a computational model that is called *shifting synchronization theory* [10]. The main point of the oscillatory correlation findings and Wang's model is to represent specific sound features as a synchronized group of neurons. If sounds with different features are contemporarily present, the two groups of neurons are internally synchronized but desynchronization between the two groups occurs. This idea is at the base of LEGION [3], the network of oscillators used by Wang.

LEGION is usually composed of a 2D grid of oscillators: in our coupled SOM-LEGION model there is a one-to-one correspondence between the nodes of SOM and the oscillators of LEGION. The dynamics of the single oscillator are simple: if there is no external excitation, the oscillator falls in a fixed stable point, whereas, if the external excitation is greater than zero, the oscillator moves in a stable limit cycle, alternating between a so-called *active phase* and a *silent phase*. There are also 2 different coupling terms: the local coupling between each oscillator and its neighbours and the coupling with an external global inhibitory source that is active when at

least one oscillator is in the active phase. These two couplings are responsible for the oscillatory synchronization within each group of stimulated oscillators and of the desynchronization among different groups (for more details see the exhaustive description of the LEGION model [11]). Some features of the SOM-LEGION coupling should be kept in mind: there is a one-to-one correspondence between the nodes of SOM and the oscillators of LEGION; a LEGION oscillator receives a positive external stimulation only if the distance of the corresponding SOM node to the input sound feature vector is less than a fixed threshold; each time step a new sound feature vector is provided resulting in a time-dependent external stimulation.

3. RESULTS

This model was tested on two different soundscapes: an urban environment with a mixture of quiet, light and occasionally heavy traffic noise, labelled as T, and a botanic garden, with an only limited human presence, labelled as P. The soundscape for each location was monitored by two measurements stations, at 55 m from each other, which recorded standard 1/3-octave band levels at time interval of 1 s. Two SOMs, one for each soundscape, were trained with sets of sound feature vectors corresponding to an entire day, that is 86400 samples. It is clear that such training produces sound context-dependent maps. This can be seen in Fig. 1 in which the SOM trained in the botanic garden does not recognize a 1 s sample extracted from the passage of a car in T. Like humans learn from experience, dynamic learning based on data from T can

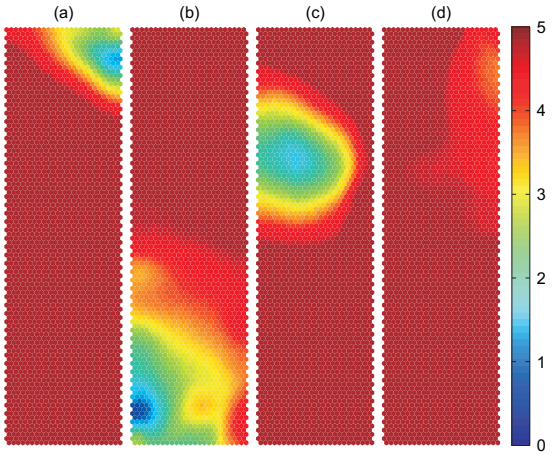

Figure 1. Distance of the sound feature vector related to a typical sample from P (quiteness) and the units in the SOM trained in (a) T, (b) P. Distance of the sound feature vector related to a typical sample from T (during the passage of a car) and the units in the SOM trained in (c) T, (d) P.

improve the ability of the SOM previously trained in P to match urban sounds. A comparison between the distance to the BMU over one hour of input vectors (3600 samples) from T without or with dynamic learning shows such improvement (see Fig. 2 and Tab. 1). Learning new sound features could introduce a sort of forgetting term reducing the performance on samples from the previous soundscape. If the parameters controlling the dynamic learning are chosen correctly such phenomenon does not produce significant collateral effects, as normally occurs in humans. The LEGION oscillators corresponding to the SOM nodes, whose distance to an input feature vector is less than a certain threshold, receive an external excitation. The (de)synchronization properties of LEGION are showed in Fig. 3. The change of the external stimulation in time can result in a transient in which the oscillators recombine their dynamic connection weights to adapt themselves to the new external input (see Fig. 3 at $t = 4.2$ s). However, such phase disruption is not present if the external stimulation is a quasistatic function. In such cases stream formation can be observed.

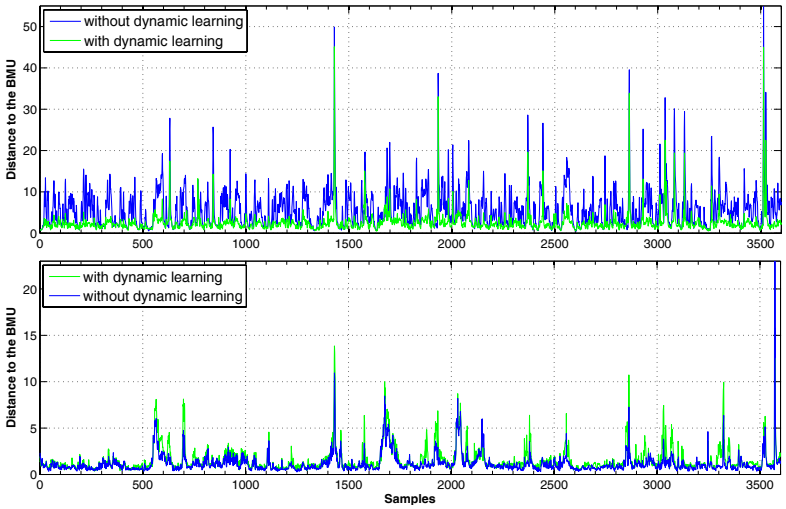

Figure 2. Top: distance before (blue) and after (green) dynamic learning of the BMU of SOM initially trained on P to 3600 samples (1 hour) input sound feature vectors taken from T. Bottom: distance before (blue) and after (green) dynamic learning of the BMU of SOM initially trained on P to 3600 samples (1 hour) input sound feature vectors taken from P. In both cases dynamic learning is carried out using 86400 samples (1 day) from T.

Table 1. Effects of dynamic learning on matching. Dynamic learning is carried out using 86400 samples (1 day) from T: it can improve significantly the matching of the SOM previously trained in P without strongly affecting the ability to match sounds from the same location.

Initial training location	Dynamic learning	Average distance to the BMU	
		input samples from T	input samples from P
P	no	5.76	1.19
	yes	2.60	1.53
T	no	1.94	1.87

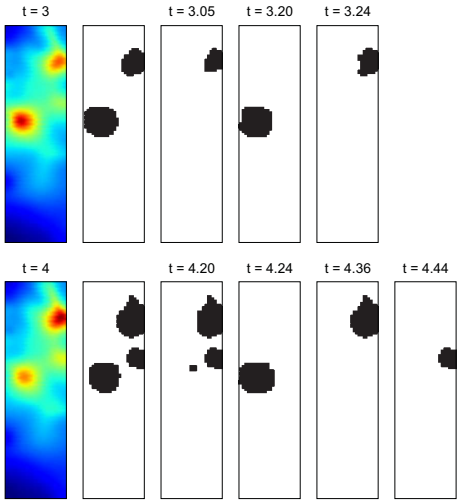

Figure 3. Left (2 columns): similarity (inverse of the distance) of two input samples at $t = 3$ s (top) and $t = 4$ s (bottom), before (1st column) and after (2nd column) binarization ($\lambda = 0.92$, moving average order: $h = 3$, see [4]). Right (4 columns): some snapshots of LEGION taken at different times. The samples used here are extracted from test input data recorded in scenario P previously used for SOM testing.

4. CONCLUSIONS AND PERSPECTIVES

A computational model for soundscape analysis is constructed based on human auditory processing and pyschological aspects as context dependency and learning from experience. Sound features related to the degree of novelty of the sound are calcuated at each time step. Based on such sound features, a self-organizing map, SOM, can learn which combinations of features are typical for the soundscape of a particular location by means of extensive training. In this paper a new dynamic learning algorhithm is described and applied to enhance the ability of the SOM to match more sounds from the same location as time goes on or to adapt its operation to other locations. Next, LEGION imitates the oscillatory (de)synchronization properties of excited neurons. Conceptually the SOM unit and the LEGION oscillator can be considered the same formal neural unit. The coupling model represents two different functionalities of ideal neurons: the memory formation, modeled by SOM training and the dynamic oscillatory correlation of sensory cortex neurons excited by an auditory stimulus, schematized by LEGION.
The presence of a transient in LEGION prohibits continuous stream formation. However, the transient is sensibly reduced or even removed if the region of oscillators that are externally stimulated changes only slowly in time. This could be achieved using a shorter time step in the sound feature extraction and a lifelong dynamic learning of the SOM. Furthermore, attentional mechanisms can be simulated at the LEGION level: attention to a particular type of sound could be simulated by a mask enabling only the oscillators in a precise region of the network to oscillate.

ACKNOWLEDGMENTS

Bert De Coensel is a postdoctoral fellow of the Research Foundation – Flanders; the support of this organisation is gratefully acknowledged. This work was supported in part by the IWT Vlaanderen Project IDEA (IWT-080054) and FWO.

REFERENCES

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

[2] T. Kohonen, *Self-Organizing Maps*. Heidelberg, Germany: Springer-Verlag, 3rd ed., 2001.

[3] D. Wang and D. Terman, "Locally excitatory globally inhibitory oscillator networks," *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 283–286, 1995.

[4] D. Oldoni, B. De Coensel, M. Rademaker, T. Van Renterghem, B. De Baets and D. Botteldooren, "Context-dependent environmental sound monitoring using som coupled with legion," in *Proceedings of the IEEE World Congress on Computational Intelligence*, (Barcelona, Spain), July 2010.

[5] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*. No. 22 in Springer Series in Information Sciences, Berlin, Germany: Springer-Verlag, 2nd ed., 1999.

[6] C. M. Gray and W. Singer, "Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex," *Proc. Natl. Acad. Sci. USA*, vol. 86, no. 5, pp. 1698–702, 1989.

[7] S. M. R. Galambos and P. J. Talmachoff, "A 40-hz auditory potential recorded from human scalp," in *Proc. Natl. Acad. Sci. USA*, vol. 78, pp. 2643–2647, 1981.

[8] E. B. M. Brosch and H. Scheich, "Stimulus-related gamma oscillations in primate auditory cortex," *Journal of neurophysiology*, vol. 87, no. 6, p. 2715, 2002.

[9] C. von der Malsburg, "The correlation theory of the brain function," Internal Report 81-2, Max-Planck-Institute for Biophysical Chemistry, 1981.

[10] D. Wang, "Primitive auditory segregation based on oscillatory correlation," *Cognit. Sci.*, vol. 20, no. 3, pp. 409–456, 1996.

[11] D. Wang, "Auditory stream segregation based on oscillatory correlation," in *Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop*, (Ermioni, Greece), pp. 624–632, Sept. 1994.


1st EAA – EuroRegio 2010
Congress on Sound and Vibration
15 - 18 September 2010, Ljubljana, Slovenia
With Summer School for Young Researchers from 13 - 15 September 2010
European Acoustics Association
EAA
Alps
Adria
Acoustics
Association
Website: http://lab.fs.uni-lj.si/sda/euroregio/
ISBN 978-961-269-283-4
SDA
Slovenian Acoustical Society