大阪
inter-noise 2011
Osaka Japan September 4-7

Soundscape analysis by means of a neural network-based acoustic summary

Damiano Oldoni[1], Bert De Coensel[2], Michiel Boes[3], Timothy Van Renterghem[4], Samuel Dauwe[5], Bernard De Baets[6], Dick Botteldooren[7],

[1,2,3,4,7] Ghent University, Department of Information Technology,

Sint-Pietersnieuwstraat 41, Gent, B-9000 BELGIUM

[5] Ghent University - IBBT, Department of Information Technology,

Gaston Crommenlaan 8 (Bus 201), Gent, B-9000, BELGIUM

[6] Ghent University – KERMIT, Research Unit Knowledge-based Systems,

Coupure links 653, Gent, B-9000, BELGIUM

ABSTRACT

The importance of detecting and characterizing sound events in the context of soundscape analysis is more and more understood among acousticians, architects and urban planners. A compilation of typical sounds for a given location, supplemented with a selection of rare but salient sound events, can quickly and efficiently provide an overview of the specific soundscape. The main subject of this paper is to construct a computational model to automatically create such an acoustic summary, using a biologically inspired auditory processing model. In particular, a neural network approach is taken, using a self-organizing map (also called a Kohonen map). A series of psychoacoustic features extracted from the sound forms the input data of the model at each time step. Typically, the learning algorithm of self-organizing maps is strictly dependent on the frequency of occurrence of the input data. However, human perception and retrospective assessment of soundscapes depend on more than only the rate of occurrence of particular sounds. For example, salient events will be remembered more vividly, even if they occur only rarely. To take into account this effect, a specific unsupervised continuous learning algorithm is used, which promotes memorizing less often occurring sounds too. At the end of training, the self-organized map represents an exhaustive acoustic summary of the soundscape at a particular location, and can be used to retrieve and listen to particular sound events.

Keywords: Soundscape, Self-Organizing Map

1. INTRODUCTION

Recently, the role played by sound in the urban planning process has grown and can now be considered as relevant as factors like visual aesthetics, safety and mobility [1][2][3]. In particular, it becomes more and more clear how the human perception of environmental sound plays a primary role, thus leading to a crucial shift in the research on community noise [4][5]. Inspired by the work of Shafer [6], a holistic approach, often referred as soundscape planning, has shown great potential in several urban spaces [7][8].

The current challenge for acousticians is thus to develop specific tools in order to efficiently assess the soundscape at a given location. An effective tool is to describe the soundscape in a

[1] damiano.oldoni@intec.ugent.be
[2] bert.decoensel@intec.ugent.be
[3] michiel.boes@intec.ugent.be
[4] timothy.vanrenterghem@intec.ugent.be
[5] samuel.dauwe@intec.ugent.be
[6] bernard.debaets@ugent.be
[7] dick.botteldooren@intec.ugent.be

qualitative but direct way by means of a compilation of sounds called acoustic summary; this work focuses on a computational model to automatically create such an acoustic summary composed of common sounds and a selection of rare but conspicuous sound events. Such model, based on a biologically inspired auditory processing model and a neural network called Self-Organizing Map (SOM), can quickly and efficiently provide an overview of the specific soundscape.

An important aspect to be pointed out immediately is the innate context dependency of the system. The SOM is trained based on temporal correlation of the occurrence of sound features. The clustering thus emerging will be more detailed and sensitive for sounds that occur more often in the area but remain ignorant on other sounds.

The next section is dedicated to describing the different building blocks of the model, while Section 3 contains results obtained applying such model to a case study. The article ends with a brief discussion of the results and important short term advances.

2. METHODOLOGY

2.1 Overview

The model starts by extracting several sound features encoding loudness and spectro-temporal irregularities from standard 1/3-octave band levels. Such sound features, designed to extract the information particularly useful for potentially triggering the so-called bottom-up attention [9], are calculated at each time step (1s) thus forming a time dependent sound feature vector. A SOM is sequentially trained on such vectors, but, contrary to classical implementations [10], the initial static learning phase is followed by an unsupervised selective continuous learning phase in order to imitate human continuous learning. After such training the units composing the SOM are encoding prototypical sounds by means of their sound features. The final step of the model focuses on retrieving sounds whose features are as similar as possible to the sound features encoded in the SOM units. The final set of retrieved sounds is what we call an acoustic summary.

2.2 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 or shorter. Next, energetic masking is simulated by means of a cochleagram, calculated using the Zwicker loudness model [11]. The cochleagram covers the complete range of the hearable frequencies (0-24 Bark) with a resolution of 0.5 Bark, thus resulting in 48 spectral values. To simulate the human auditory system the absolute intensity and the spectro-temporal variations are the most important features that are to be detected. Based on existing auditory saliency models [12][13][14], the proposed model uses the centre-periphery mechanism, thus simulating the receptive fields in the auditory cortex. Gaussian and difference-of-Gaussian filters are applied through convolution to the cochleagram. The former encode intensity, while the latter subtract between a “center” fine scale and a “surround” coarser scale. The spectral and temporal gradient of the cochleagram is thus encoded at 16 different scales (4 for intensity, 6 for both spectral contrast and temporal contrast) as shown in **Error! Reference source not found.**. At each time step t, the sound feature extraction results in a sound feature vector or simply feature vector consisting of 16×48 = 768 values.

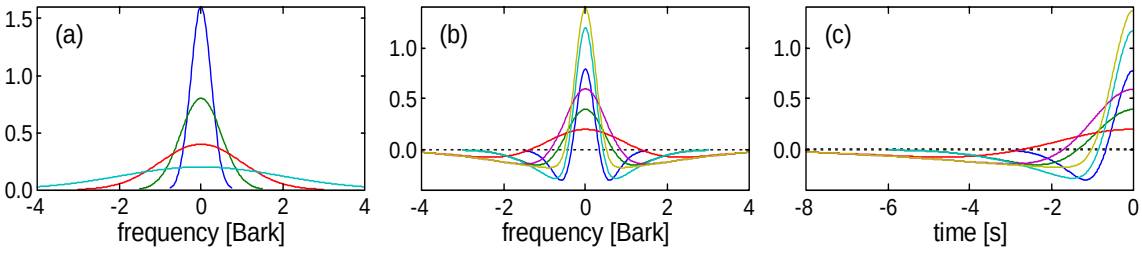

Figure 1– Intensity filters (a), spectral contrast (b) and temporal contrast filters (c). In (c) temporal causality is preserved by only convolving with the past.

2.3 Learning typical feature combinations

The Self-Organizing Map, sometimes called Kohonen map [15], is a neural network often used as a nonlinear dimension reduction tool. The units composing the network are typically placed in a 2D grid, usually forming a hexagonal lattice. Each unit has a corresponding reference vector which represents the node position in the high-dimensional sound feature space. After initialization based on principal component analysis [10][15], the training phase can modify the position of the reference vectors to better represent the sound feature vectors calculated as explained in 2.2 at a given instance in time. The training algorithm consists of two steps and it is applied to each sound feature vector of the training set. At first an input sound feature vector is fed to the SOM and the nearest reference vector, whose unit is called the Best Matching Unit (BMU), is found. Next, the reference vector of the BMU and, to a lesser extent the reference vectors of its neighbours, are moved towards the position of the input sound feature vector. After training, the SOM units have learned through the position of their reference vectors what features have often co-occurred. In particular, sounds co-occurring very often during training cannot be segregated by SOM. For this reason SOM can be seen as a context-dependent sound segregation technique. For more details about the mathematical model of SOM in this context see Oldoni [10].

After this training phase, the SOM could be seen as a nonlinear 2D projection of the probability density function of the high-dimensional input data. In other words, if a new sound feature vector is provided, the distance to its BMU is an indirect measure (assessed through sound features) of how often the corresponding sound occurred: the smaller the distance to the BMU, the more often that sound occurred in the training phase and it is said to be well recognized by the SOM.

However, a typical soundscape description provided by attentive listeners is not a mere projection of the rate of occurrence of particular sounds. For this reason, a second, longer training has been performed. In this second phase a continuous selective learning strategy is used: a new learning phase starts only if the distance of the input sound feature vector is larger than an activation threshold, T1, and it continues until the distance is less than a deactivation threshold T2, with $T2 \leq T1$. Continuous selective learning can greatly improve the SOM matching power to new or not often occurred sounds without losing too much information about what has been previously learned. After the training is finished, the SOM units encode, by means of their reference vectors, all the information related to the training input sounds. Although continuous learning can be conceived, at least theoretically, as a livelong training phase, practically it is still not yet feasible (see 4 for technical advances). However it has been experimentally found that after some days (more than 500000 samples) SOM has learned a sufficiently wide gamma of not often occurring sounds or outliers and thus being able to well characterize the sound environment.

2.4 Acoustic summary formation: sound samples retrieval

The position of the feature reference vectors in the high-dimensional sound feature space can be interpreted as the description of a prototypical sound. It is clearly impossible to convert the point in feature space back into sound samples. Therefore, after training the SOM, sound samples are recorded in the same environment. Several sound sample retrieval methods could be implemented; however the common point is to find sounds whose sound feature vectors are as similar as possible to the reference vectors of the SOM units.

The simplest method is exclusively based on best matching criterion: the closest sound feature vector for each SOM unit reference vector is found and the corresponding 1s sound sample is selected. For listening purposes a fixed temporal window, centered on the selected sound sample is extracted. In this work 5s long sound excerpts have been generated.

More complex retrieval methods, attempting to find onset and offset of sound events have been tested. They are more selective and thus need long sound recording periods to find satisfactory sound excerpts for most of the SOM units. A solution could be a hybrid retrieval method.

3. RESULTS

This model has been tested on different locations. We present a comparison of two sites, S and D, in a typical urban environment characterized by a mixture of quiet and noise from light and heavy traffic and human activities. The sound at each location has been constantly monitored by two measurement stations which recorded continuously standard 1/3-octave band levels with a time resolution of 1s. Based on this data, sound feature vectors have been calculated as in 2.2. Two SOMs have been used, one for each location. Before starting the SOM training, the position of the reference vectors of the SOM units have been initialized using PCA-based technique using twelve

hour data (43200 1-s samples, from 8am to 8pm); the first training has been carried out using samples corresponding to an entire day, that is 86400 1-s samples. As explained in 2.3 this training has been followed by a longer continuous selective learning phase using sound feature vectors collected in two weeks.

Given the high-dimensionality of the sound feature space, it is impossible to visualize the position of the SOM unit reference vectors. However, the Unified Distance Matrix, briefly called U-matrix [15], can give an idea of the reference vector positions by means of the Euclidean distance between the reference vectors of neighboring units. In Figure 2 the U-matrixes of the two SOMs before and after continuous selective learning are shown.

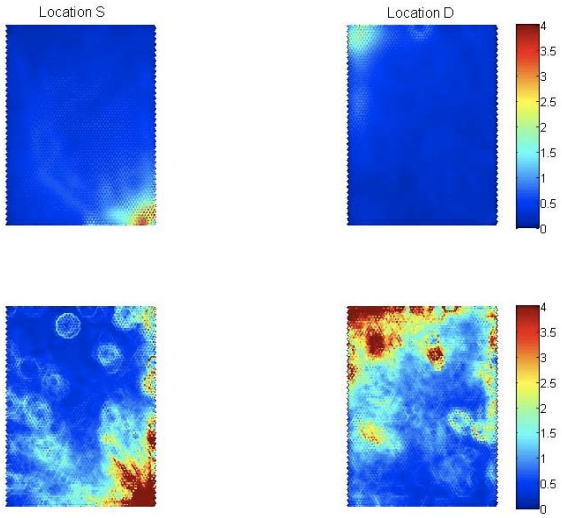

Figure 2 - Visualization of the SOMs before (top) and after (bottom) selective learning phase in S (left) and D (right) by means of their U-matrix. The Euclidean distance between each unit reference vector and its neighbors is shown as a color scale.

The effects of continuous selective learning are noticeable: several clusters (blue areas separated from the rest by yellow or red, indicating larger distance) grouping combinations of features that co-occur are created. The initial large non-selective (blue) area of the map has now become more specialized. It turns out that more units are used to encode loud events and so-called outliers. The importance of such sounds for human soundscape perception is great and can be computationally quantified by means of auditory saliency [13]. **Error! Reference source not found.** shows the saliency for the feature vectors corresponding to the units of the final SOM. In the upper left corner of map S and the lower right corner of map D saliency is low. Low saliency can perceptually be associated to quietness. Thus the initial maps focused too much on silent periods which are indeed occurring more often at the selected locations.

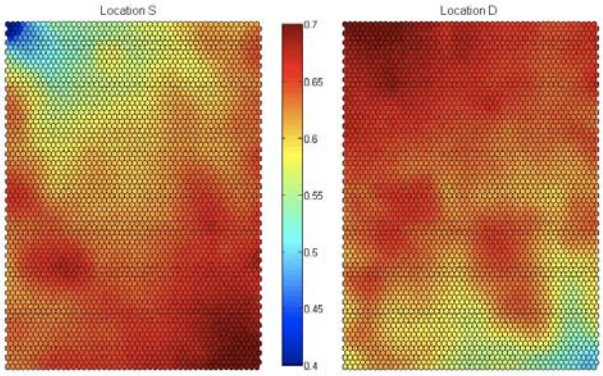

Figure 3 – Saliency of the reference vectors of the units of the SOM trained in S (left) and in D (right). The nodes encoding quietness (top-left in S and bottom-down in D) are less salient than the nodes in the opposite angle of the SOM encoding loud specific

The SOM categorizes the sounds that are observed at the selected locations, but, particularly when continuous selective learning is used, information on the frequency of occurrence is lost. It is however a straight-forward task to detect which combinations of features, which sounds, are present at different times of the day and how often they occur. Such temporal information can be visualized by means of the temporal distribution of the occurrence of the BMU among the SOM units, as shown in Figure 4 (top) for location S on one particular day.

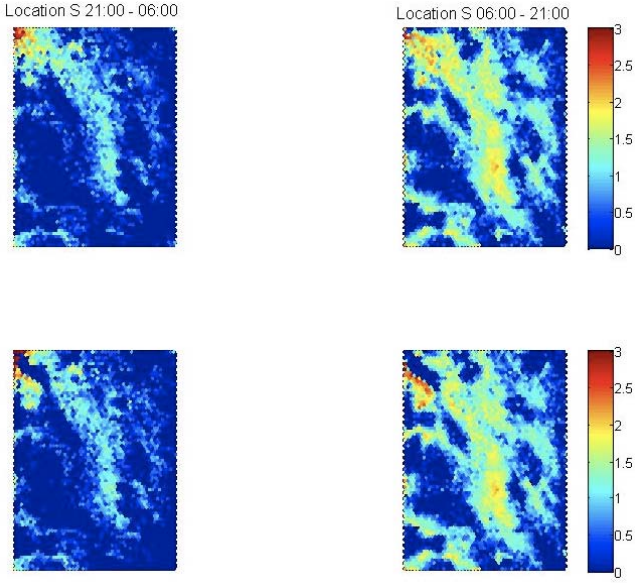

Figure 4 – Night (left) and day (right) distribution of the occurrence of the Best Matching Unit (top) and of the Best Saliency-Weighted Matching Unit (bottom) of the SOM trained in S. The distribution is displayed in a logarithmic scale and has been obtained using sound feature vectors from location S related to one working day (86400 samples) and divided in a night (9pm-6am) and a day period (6am-9pm).

However, as explained in 2.3, the sound event occurrence doesn't reflect appropriately how humans perceive soundscape: the saliency of a sound determines how strongly it attracts the attention of the listener. For this reason the criterion to assess the BMU can be modified as follows: given an input sound feature vector, the proximity to each SOM unit is calculated and multiplied by the saliency of that SOM unit. The unit related to the maximum is taken. A Gaussian proximity function ($\mu=0$, $\sigma=3$) has been used to translate distance in feature space to proximity. The unit found with this approach will be called Best Saliency-Weighted Matching Unit (BSMU). The distribution of the occurrence of the BSMU for location S is shown in Figure 4 (bottom). General commonalities are shared with the BMU occurrence distribution, as the differences between night-day. However, a very noticeable difference regards the distribution of the occurrence for sound feature vectors related to "daily" quietness, whose BMUs are typically next to the very top-left corner units encoding "nocturnal" quietness.

These last maps should guide the user of the acoustic summary to those locations in the map that are of importance and allow him or her to listen to the typical sound hidden behind this unit.

4. CONCLUSIONS

A tool for computational soundscape analysis has been presented in this paper. Based on sound feature extraction technique mimicking bottom-up attention mechanisms and continuous selective learning paradigm for outliers learning promotion, an acoustic summary can been created. A specific type of neural network called Self-Organizing Map (SOM) clusters the sounds that are picked up by the microphone. Once training is ended, a sound fragment can be associated as a prototype to each unit of the SOM based on a minimum distance criterion in the high dimensional sound feature space. The user of the summary can listen to these sounds to aurally explore the sonic environment and attach some meaning. Frequency of occurrence at different times of the day, weighted by saliency give the user some indication on the importance of each sound fragment in terms of common sound

or specific outlier occurring under certain circumstances only.

The proposed system has been tested at several locations. Two locations in Gent characterized by typical urban sounds from private and public transport and human activities are shown in this written paper. This shows the applicability and computational feasibility of the presented approach.

Three main advances will be carried out in the short term, covering both theoretical and technical aspects. The time resolution of the 1/3-octave band spectrum calculation of 1s is not suitable for characterizing fast time fluctuating sounds as speech. In fact, the amount of SOM units encoding human activities as speaking is smaller than it should, especially considering how big is the influence of the human presence on the sonic environment at both the locations. Although the selective continuous learning boosted the creation of many clusters for specific sound events, using saliency to guide the training could focus the map more on sounds that are important for the soundscape perception. The major technical short term advance will be the implementation of the model in a wide intelligent distributed sensor network. Controlling the sensors and collecting and processing data in real time will increase the quality of the final acoustic summary. Two aspects of the model would particularly benefit from it: the SOM training and the sound sample retrieval. Life-long like training phases will be possible and sound sample retrieval will rely on a trigger-based system, thus reducing the high storage load due to continuous sound recording sessions in the interests of the acoustic summary accuracy and completeness.

ACKNOWLEDGEMENTS

This research is part of the IDEA (Intelligent, Distributed Environmental Assessment) project, a 4-year strategic basic research project, financially supported by the IWT-Vlaanderen (Flemish Agency for Innovation by Science and Technology).

Bert De Coensel is a postdoctoral fellow of the Research Foundation – Flanders (FWO–Vlaanderen); the support of this organisation is gratefully acknowledged.

REFERENCES

[1] M. D. Adams, W. J. Davies and N. S. Bruce, “Soundscapes: an urban planning process map,” Proc. INTER-NOISE 09, (2009).

[2] T. B. Hellström, “Noise design: Architectural modelling and the aesthetics of urban acoustic space,” PhD Dissertation, Royal Institute of Technology, (Stockholm, 2003).

[3] S. Olafsen, “Using planning guidelines as a tool to achieve good soundscapes for residents,” Proc INTER-NOISE 09, (2009).

[4] P. Lercher and B. Schulte-Fortkamp, “The relevance of soundscape research to the assessment of noise annoyance at the community level,” Proc. ICBEN, 225-231 (2003).

[5] B. Schulte-Fortkamp and D. Dubois, “Recent advances in soundscape research,” ActaAcust.Acust. 92(6) 5-8 (2006).

[6] Schafer, R. M. “The soundscape: our sonic environment and the tuning of the world,” (Destiny Books, Rochester, Vermont, 1994).

[7] J. Kang, “A systematic approach towards intentionally planning and designing soundscape in urban open public spaces,” Proc. INTER-NOISE 07, 28-31 (2007).

[8] A. L. Brown, A. Muhar, “An approach to the acoustic design of outdoor space,” Journal of Environmental planning and Management, 47(6), 827-842 (2004).

[9] E. I. Knudsen, “Fundamental components of attention,” Annu. Rev. Neurosci., 30, 57–78 (2007).

[10] 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 Proc. of the IEEE International Joint Conference on Neural Networks, 1413-1420 (2010).

[11] Eberhard Zwicker and Hugo Fastl, “Psychoacoustics. Facts and Models,” edited by M. R. Schroeder (Springer, Berlin, 1999).

[12] 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).

[13] O. Kalinli and S. Narayanan, “A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech,” in Proc. 8th Annual Conference of the International Speech Communication Association, 1941–1944 (2007).

[14] V. Duangudom and D. V. Anderson, “Using auditory saliency to understand complex auditory scenes,” Proc. 15th European Signal Processing Conference, 1206–1210 (2007).

[15] Teuvo Kohonen, “Self-Organizing Maps,” edited by Teuvo Kohonen (Springer, Heidelberg, 2001).