# Human-mimicking environmental sound measurement

Dick Botteldooren, Damiano Oldoni, Bert De Coensel

*Acoustics Group, Department of Information Technology, Ghent University, Belgium*

## Introduction

Classical environmental noise measurements determine the acoustic intensity at the point of observation. Attempts to measure more accurately what persons perceive have mainly focussed on using indicators such as loudness, sharpness, roughness, etc. that were proven to correlate well to perception of sound that test persons attentively listened to in lab environments. The listening context is however very different in everyday life. Therefore a different approach is proposed namely to mimic human processing of environmental sound in computational models and eventually in measurement equipment. The amount of scientific knowledge on the different stages of (environmental) sound perception is huge and continuously growing. However, while trying to implement this knowledge in computational models one has to face the limitations of computers and measurement equipment. The use of supercomputers in environmental noise measurement is not practical since it would scale poorly with the number of measurement microphones. This paper reports on continuous efforts to construct just-accurate-enough human mimicking processing, rigidly grounded in psychological and biological knowledge. In selecting models for every step of this complex process, biological plausible solutions are opted for, which may or may not be the most efficient on current computers, but are surely most future proof.

## Computational modelling framework

In previous work[1-3], modelled environmental noise was used as a starting point to analyse how the combined exposure to unwanted sounds such as traffic noise and more pleasant natural sounds (birds, water, etc.) affected soundscape perception. A very simplified notice-event model already shows that including temporal fluctuation at a time scale of seconds can explain some of the observations related to annoyance perception of different sound sources[4]. Thus, time will play an important role. A second crucial factor in environmental noise perception is attention. Both inward and outward oriented attention have to be included in the model[1-2]. Outward oriented attention depends on activity and intentions of the modelled observer. The only possibility to include some of the effects of this form of attention consists in modelling a large number of virtual observers at any given location and look for the average effects. Inward oriented attention focussing depends on features of the environmental sound that attract attention, so called saliency[3]. In modelling the effect of attention on environmental sound perception, the interplay between activation and inhibition-of-return proved essential for obtaining a stable and robust model.

When using modelled or artificially mixed sounds, the problem of auditory stream segregation is avoided. Here we will mainly focus on the additions to the model that are proposed for object formation and stream segregation. The proposed model starts from basic features, extracted from a time-frequency representation of the sound. Currently, specific (Zwicker) loudness *vs.* time is considered as a time-frequency representation, because it can be calculated relatively fast from standard 1/3-octave

band levels. The extracted features mimic the information processing stages in the central auditory system. In particular, the human auditory system is, next to absolute intensity, also sensitive to spectro-temporal irregularities (i.e. contrast on the frequency scale, and changes in time). Intensity, spectral and temporal features are calculated by convolving the specific loudness *vs.* time with (difference-of-)gaussian filters with varying width, and thus encode the intensity, spectral and temporal gradient. These feature extraction mechanisms are largely inspired by those used in the calculation of more complex auditory saliency maps[5,6].

The current model extracts 16 intensity, spectral and temporal feature vectors per second, with each vector consisting of 48 values (2 per critical band). This multitude of features has to be organized and lowered in dimension, mainly based on co-occurrence. Features that always occur together most likely belong to a single sound object. Groups of features that are present at the same time, but not usually co-occur should be separated into several streams. Once these streams are formed, one can attend more to one of them. This additional attention may eventually lead to further segregation.

The determination of co-occurrence of features is modeled by constructing a self-organizing map[7] (SOM) based on long periods of observations at a location under study. The resulting two-dimensional maps might differ depending on the region where the training data was gathered. Once training has been achieved, every new observation projects onto a particular area of the map. If the observation contains one set of features that regularly occur together, the projection will be well focussed on a particular region of the map. This corresponds to one *sound* being heard. If however the observation contains various sets of features that do not regularly co-occur several areas of the map get activated. An example is shown in Figure 1, left.

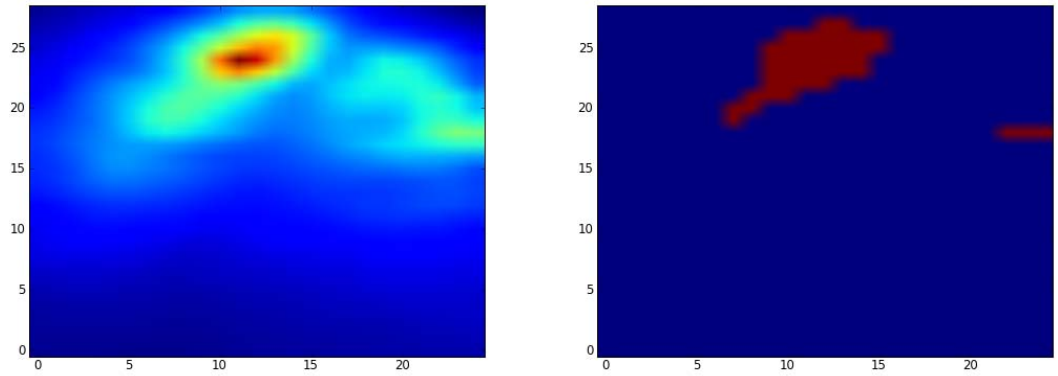


**Figure 1. Left: Projection of a new feature set on the map after training. Blue zones correspond to great distance between the new event feature content and the particular nodes of the map, whereas red zones indicate a small distance (high similarity). One big area is visible in the upper center and one smaller is visible on the upper right hand. Right: Binarization of the similarity map on the left. This map corresponds to the external stimulation provided to the LEGION.**

To identify whether a sound fragment contains one, two or more streams of sound, another processing step is needed. A Locally Excitatory Globally Inhibitory Oscillator Network[8] (LEGION) is used for segregating areas in a two-dimensional map. The previous map, after being binarized through the use of a threshold, provides the external stimulation to a LEGION network (See Figure 1, right). A threshold permits to mark the interested regions, possibly modulated by attention mechanisms. , The oscillatory correlation at the base of LEGION mimics the biological activity of the

brain: coherently perceived objects are linked together by various feature detecting neurons via their specific temporal correlation[9]. Because this type of network is mostly used with static images, it had to be adapted to work with the temporally changing exited regions in the SOM. An example of the model output is shown in Figure 2.

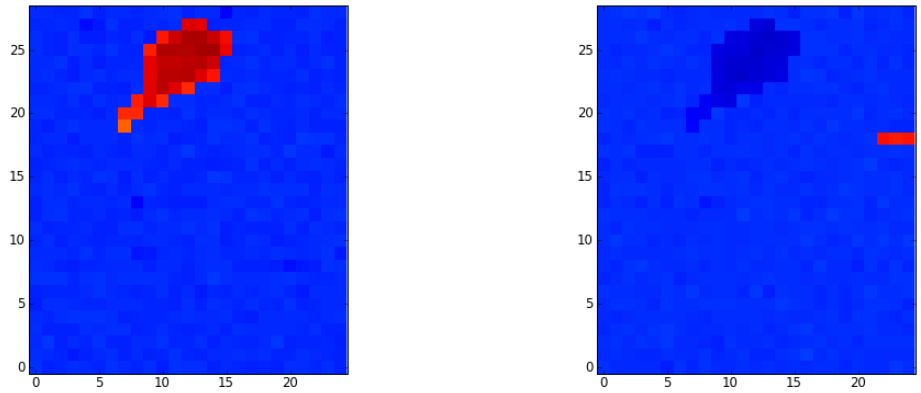


**Figure 2. Two snapshots of network oscillatory activity taken shortly after the start. The two areas are well separated, exploiting the oscillatory correlation properties of LEGION, resulting in a segragation-grouping activity.**

The LEGION network identifies one or more streams. At this point, an attention mechanism can be deployed, leading to focussing on one of the streams. In refs. 1-3, a possible model is described, which implements a winner-take-all mechanism between streams using a balance between activation and inhibition-of-return for each stream.

## Conclusions

In this paper, a framework for modelling human processing of environmental sound was proposed, which could ultimately be used to enhance noise measurement equipment to mimic human perception. In particular, submodels for extracting features and segregating streams using self-organizing maps and LEGION networks were focused on. These submodels could be used as a basis for implementing models for attention focussing and eventually source recognition.

## References

[1] D. Botteldooren, B. De Coensel, B. Berglund, M. E. Nilsson, P. Lercher, "Modeling the role of attention in the assessment of environmental noise annoyance," In *Proceedings of the 9th International Congress on Noise as a Public Health Problem (ICBEN)*, Foxwoods, Connecticut, USA (2008).

[2] D. Botteldooren, B. De Coensel, "A model for long-term environmental sound detection," In *Proceedings of the 5th IEEE World Congress on Computational Intelligence (WCCI)*, Hong Kong (2008).

[3] B. De Coensel, D. Botteldooren, B. Berglund, M. E. Nilsson, "A computational model for auditory saliency of environmental sound," *J. Acoust. Soc. Am.* **125**(4):2528 (2009).

[4] B. De Coensel, D. Botteldooren, T. De Muer, B. Berglund, M.E. Nilsson, P. Lercher, "A model for the perception of environmental sound based on notice-events," *J. Acoust. Soc. Am.* (in press).

[5] C. Kayser, C. I. Petkov, M. Lippert, N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology* **15**:1943-1947 (2005).

[6] O. Kalinli, S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," *Proceedings of Interspeech*, Antwerp, Belgium (2007).

[7] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, Berlin (2001).

[8] D. L. Wang, D. Terman, "Locally excitatory globally inhibitory oscillator networks," *IEEE Trans. Neural Net.* **6**(1):283-286 (1995).

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

ISBN 978-0-946754-56-4

NPL
National Physical Laboratory

# Conference Digest

## MINET Conference:
## Measurement, sensation and cognition

10 - 12 November 2009
National Physical Laboratory, London, UK

*Measuring the Impossible*