Context-dependent environmental sound monitoring using SOM coupled with LEGION

Damiano Oldoni, *Student Member, IEEE*, Bert De Coensel, *Member, IEEE*,
Michaël Rademaker, *Member, IEEE*, Timothy Van Renterghem, Bernard De Baets and Dick Botteldooren, *Member, IEEE*

***Abstract*— Environmental sound measurement networks are increasingly applied for monitoring noise pollution in an urban context. Intelligent measurement nodes offer the opportunity to perform advanced analysis of environmental sound, but trade-offs between cost and functionality still have to be made. When using a tiered architecture, local nodes with limited computing capabilities can be used to detect sound events of potential interest, which are then further analyzed by more powerful nodes. This paper presents a human-mimicking model for detecting rare and conspicuous sound events. Features encoding spectro-temporal irregularities are extracted from the sound, and a Self-Organizing Map (SOM) is used to identify co-occurring features, which most likely belong to a single sound object. Extensive training allows this map to be tuned to the typical sounds that are heard at the microphone location. A Locally Excitatory Globally Inhibitory Oscillator Network (LEGION) is used to group units of the SOM in order to construct distinct sound objects.**

I. Introduction

Advances in the design of low-cost computing devices and sensors, together with an increase in bandwidth and covering power of low-cost wireless networks, are forming a technological push for the use of wireless sensor networks [1]. Acoustical sensor networks in particular provide a wide range of applications, such as audio surveillance for public security [2], [3], habitat monitoring [4], [5] or environmental noise pollution monitoring [6], [7]. Information retrieved from the latter could be used to assess potential noise annoyance or sleep disturbance, to validate noise maps or even to locally steer activities, e.g. via intelligent traffic systems.

Although the hardware, storage capacity and communication bandwidth needed for building environmental sound measurement networks is increasingly becoming cheaper, trade-offs between cost and functionality still have to be made. For example, it is infeasible to perform advanced sound source recognition using small, cost- and energy-efficient nodes, while it is also infeasible to simply record and transmit the sound at all microphones continuously, due to data storage and transmission bandwidth limitations. A solution for this problem is to use a tiered architecture (see e.g. [5]), in which the spatial resolution of the network is exploited by using cheap local nodes with limited computing capabilities, which select and transmit sound fragments of possible interest to be processed by more powerful nodes (usually centrally located).

One of the most basic techniques for sound event detection is thresholding: when the instantaneous sound pressure level exceeds a predefined threshold, the occurrence of a sound event is assumed, and the node starts recording for a given period of time. In case of adaptive thresholding, the threshold is relative to the background level, which can vary in time slowly [8]. More recently, a number of techniques for selecting salient parts of the auditory scene have been proposed, inspired by the neural mechanisms that guide human attention [9], [10], [11]. However, a major disadvantage of current techniques is that no distinction is made between frequently occurring and thus expected sound events, and rare events. Moreover, the kind of expected sound events depends on the context of the microphone. For example, the sound of birds singing can be expected near a microphone situated inside an urban park, while the sound of cars passing by is expected in a busy street.

The ideal node in an environmental sound measurement network for monitoring noise pollution should, in a computationally efficient way, be able to learn and discern the sounds frequently occuring at the location of the microphone, thus distinguishing between common and rare or conspicuous sound events. In this paper, we show how this goal could be achieved using a simple biologically inspired technique.

Features encoding spectro-temporal irregularities are extracted from standard 1/3-octave band levels, which can be measured with off-the-shelf sound level meters. Subsequently, sound events are discerned using a combination of two types of neural networks: a Self-Organizing Map (SOM) [12] that allows—after extensive training—to identify co-occurring sound features and a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION) [13] for grouping and segregation of corresponding sound fragments. The combination of both neural networks models two essential features of the brain: the SOM mimics the plasticity (during the learning phase) and complex morphology of the network of neurons forming the auditory cortex, while the LEGION approximates the dynamic oscillations between connected neurons.

In Section II we provide a description of the coupled SOM-LEGION network, starting from the sound extraction, and the specific solutions adopted. The model was applied in

Damiano Oldoni, Bert De Coensel, Timothy Van Renterghem and Dick Botteldooren are with the Department of Information Technology, Ghent University, Belgium (phone: +32-9-264-9994; email: damiano.oldoni@intec.ugent.be). Michaël Rademaker and Bernard De Baets are with the Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium.

different real scenarios: the results and some discussions are provided in section III. Finally, a section with conclusions follows in IV.

II. Methodology

A. Sound feature extraction

In a first stage, a *feature vector* is extracted, at regular time intervals, from the sound signal measured by the node microphone. Instead of calculating a detailed time-frequency representation of the raw sound wave, the model starts from the 1/3-octave band spectrum, calculated with a temporal resolution of 1 s. This procedure has the main advantage that off-the-shelf sound measurement equipment can be used as a front-end, which reduces the computational load on the measurement node. The choice of time resolution can be justified by noting that the sounds of main importance for environmental noise pollution monitoring (cars, trains, aircraft, fans etc.) have a relatively slow varying temporal envelope [14], [15]. A simplified cochleagram $s(f,t)$ is then calculated using the Zwicker loudness model [16], which accounts for energetic masking. The complete hearable frequency range is considered (0 to 24 Bark) with a spectral resolution of 0.5 Bark, resulting in 48 spectral values at frequencies $f_j = \frac{1}{2}j$ Bark, for each timestep.

The mechanism for extracting the feature vector, which characterizes the amount of novelty in the sound signal, is inspired by the way the human auditory system biases its attention toward particularly conspicuous events. The auditory system is, next to absolute intensity, also sensitive to spectro-temporal irregularities. Based on existing models for auditory saliency [9], [10], [11], the proposed model calculates measures for intensity, spectral and temporal modulation using a center-surround mechanism, which mimicks the receptive fields in the auditory cortex. In particular, multi-scale features are calculated in parallel by convolving the cochleagram with various 2D gaussian and difference-of-gaussian filters $g_i(f,t)$. The former encode intensity, while the latter subtract between a "center" fine scale and a "surround" coarser scale, and encode the spectral and temporal gradient of the cochleagram at 16 scales (4 for intensity, 6 for spectral contrast and 6 for temporal contrast):

$$r_i(f,t) = (s * g_i)(f,t) \tag{1}$$

with $i = 1, 2, \ldots, 16$. Fig. 1 shows a section of the filters along the time or frequency axis. Finally, a feature vector $\vec{r}(t)$ is constructed at each timestep, consisting of $16 \times 48 = 768$ values:

$$\vec{r}(t) = \sum_{i=1}^{16} \sum_{j=1}^{48} r_i(f_j,t)\vec{e}_{48(i-1)+j} \tag{2}$$

with $\{\vec{e}_k : 1 \leq k \leq 768\}$ the standard basis for the 768-dimensional Euclidean space.

B. Feature co-occurrence analysis: Self-organizing map

The self-organizing map (SOM), an abstract mathematical model of topographic mapping from the (visual) sensors to the cerebral cortex [17], is most often described as an unsupervised technique for the visualization of high-dimensional data [18]. It does so using typically a 2D network of *units* or *nodes*. Their representation in the high-dimensional space is provided through *reference vectors*. After initialization, their coordinates are modified during the training process wherein the following (vastly simplified) steps are repeated until a stopping criterion is met:

1) Feed an input high-dimensional data point to the SOM.
2) Determine the best-matching unit (BMU) i.e. the unit corresponding to the closest reference vector.
3) Move the reference vector corresponding to the BMU and, to a lesser extent, those of the neighbouring units in the 2D grid, closer to the input high-dimensional data point.

In practice, the training process and the resulting SOM are strongly influenced by a number of parameters, such as the size of the SOM, the type of initialization of the units, the strength of learning and the type of neighbourhood considered in the third step, as well as the evolution over time of the learning parameters.

Nevertheless, after training it is clear that the frequency distribution of the input data in the high-dimensional space will be approximated by the reference vectors of SOM units, possibly leading to dense high-dimensional clusters interspaced by regions where the reference vectors of the SOM units are more distant. This emerging order is the basis for the effective visualization in the SOM. Consequently, the SOM can also be considered to perform a kind of abstraction, compressing information while preserving the most important high-dimensional relationships [12]. A trained SOM could then be understood as a nonlinear 2D projection of

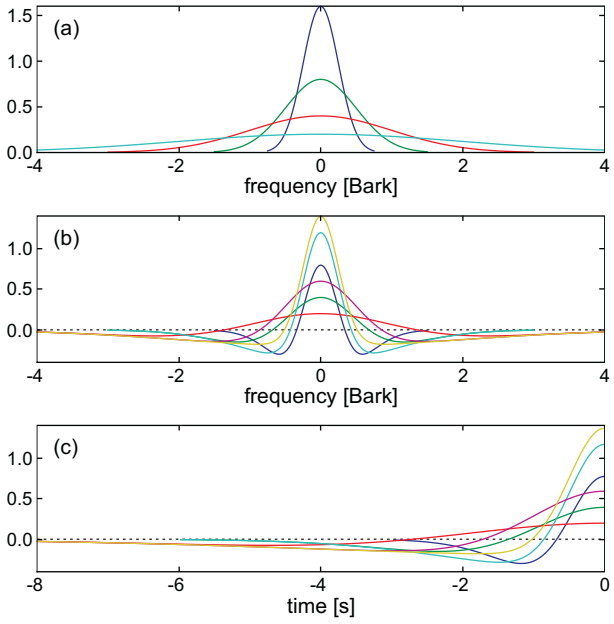

Fig. 1. Cross section of the receptive filters that are used to calculate (a) intensity, (b) spectral contrast and (c) temporal contrast. For the latter, causality is preserved by only convolving with the past.

the probability density function of the high-dimensional input data. An intuitive quantification of the SOM quality is then the average high-dimensional distance of a set of data points to their respective BMUs.

Now, we provide a brief, more formal description of the SOM technique, based on the description in [18]. More formally, we consider an n-dimensional input space $\mathbb{R}^n$, in our application the 768-dimensional space of raw sound features. The SOM units are represented by the reference vectors $\vec{m}_i \in \mathbb{R}^n$, with index i identifying the unit. The M units in the 2D network are aligned on a regular M_x by M_y grid and are represented as $\vec{\mathbf{m}}_i = (\mathbf{m}_x, \mathbf{m}_y) \in \mathbb{R}^2$. As the vectors $\vec{m}_i$ are adapted during training, we will write $\vec{m}_i(t)$ to denote the vector at time-step t during training, and use $\vec{m}_i$ only when training is complete. Input data is represented as $\vec{r} \in \mathbb{R}^n$, and at time-step t, the sample $\vec{r}(t)$ is processed by the SOM. The BMU at time-step t is then found by considering

$$c(t) = \arg\min_i \|\vec{r}(t) - \vec{m}_i(t)\| \ . \tag{3}$$

Thus, at time step t, $\vec{m}_{c(t)}(t)$ denotes the BMU for the input sample $\vec{r}(t)$. Adapting the BMU or, indeed, any unit, is then performed as follows:

$$\vec{m}_i(t+1) = \vec{m}_i(t) + h_{c(t),i}\left(\vec{r}(t) - \vec{m}_i(t)\right) \ , \tag{4}$$

where h, the neighbourhood function, performs a non-linear smoothing selection on the discrete 2D neighbourhood structure. Often used is a Gaussian function of the distance between the BMU at time step t, $c(t)$, and the generic unit i:

$$h_{c(t),i} = \alpha(t) \exp\left(-\frac{\left\|\vec{\mathbf{m}}_i - \vec{\mathbf{m}}_{c(t)}\right\|^2}{2\sigma^2(t)}\right) . \tag{5}$$

The time-step dependent parameters governing the behaviour of this type of neighbourhood function are the learning rate $0 < \alpha(t) < 1$ and the width of the 2D neighbourhood $\sigma(t)$. Both are monotonically decreasing in t:

$$\alpha(t) = \alpha_0 \frac{C}{C+t}, \quad C = \frac{N}{100}, \tag{6}$$

$$\sigma(t) = 1 + (\sigma_0 - 1)\left(\frac{N-t}{N}\right) \tag{7}$$

where N is the number of samples. Observe that $h_{c,i} = \alpha(t)$ only for the BMU, and is strictly decreasing for units farther from it in the 2D grid. Thus, for a constant similarity, the BMU is adapted to a stronger extent than any neighbouring units.

A final point concerns the visualisation of the 2D grid after training. Due to the high dimensionality of the raw feature space, the visualization of the trained map via projection on particular planes is rarely informative—one could argue that if such an approach would lead to satisfactory results, there was less need to apply the SOM algorithm in the first place. Rather, in order to easily identify regions with similar high-dimensional representations, it will be more informative to display how close in the high-dimensional space a unit in the map is to its neighbouring units. In fact, a typical way to visualize the morphology of the map uses the so-called U-matrix [19], which is a matrix of dimensions $[2M_x - 1, 2M_y - 1]$ containing both the distances between the nearest neighbours and their average. Color-coding the units on the map on the basis of their average distance to their nearest neighbours allows distinguishing regions where 2D neighbouring reference vectors are similar, from regions of high variability. We provide an example in Section III.

When training is complete, the SOM quality can be assessed on the basis of two concepts. The first is the average distance between each input vector from a set of test samples and its BMU, the so-called *average quantization error* E. It is computed as follows for a set of test samples $\vec{r}(1), \ldots, \vec{r}(N)$:

$$E = \frac{\sum_{t=1}^{N} \left\|\vec{r}(t) - \vec{m}_{c(t)}\right\|}{N} , \tag{8}$$

with $\vec{m}_{c(t)}$ now denoting the BMU for test sample $\vec{r}(t)$.

The second concept is the *topographic error*, the proportion of test samples for which the BMU and the next-best-matching unit are not neighbours. A low topographic error can be considered to be indicative of a focused SOM, clustering the units around the dense regions in $\mathbb{R}^n$.

These concepts complement each other quite well: if all the units are widely spaced in the $\mathbb{R}^n$ space formed by the eigenvectors, the SOM is very likely to obtain a quite low average quantization error, while the topographic error is likely to be large. If, in contrast, the units are packed too tightly around the densest regions in $\mathbb{R}^n$, the average quantization error can be expected to be high, while the topographical error is expected to be low.

To reduce both the average quantization error and the topographic power of the map, it is usually sufficient to reduce the initial width of the neighbourhood function σ_0 and/or the learning rate α_0, making the map less flexible, while simultaneously increasing the number of training runs to compensate for the slower learning [12].

A hexagonal lattice was used in this paper, allowing a 2D grid of equal-spaced units while maximizing for any value of $\sigma(t)$ the number of neighbours in the grid. The unit reference vectors were initialized by the linear initialization function, resulting in a regular array of vectorial values that lie on the subspace spanned by the eigenvectors corresponding to the two largest principal components of input data used during the training [18]. In our application, the high-dimensional space is composed of the raw sound features, meaning each unit corresponds to an abstract prototype of a sound. The goal is thus to group similar (in the raw feature sense) sound fragments in the SOM. Sound feature values that often arise together, and are thus often part of the same sound fragment, are then expected to have the same BMU, or to even cluster close together in the SOM. In order to allow this behaviour to arise, a proper choice of features is crucial, as well as proper values for the parameters governing the SOM construction, training and resulting performance.

The training phase has to take into account a very large number of input data: in our case 86400 samples (the number of seconds in one day) were used. Afterwards, the trained SOM is ready to receive new data samples and localize the BMU.

As we will now show, a natural link between SOM and LEGION then arises: the similarity of a raw feature vector and a specific SOM unit is, from a neural oscillatory point of view, a measure of the external stimulation that a LEGION oscillator receives. Conceptually the SOM unit and the LEGION oscillator can be considered the same formal neural unit. In fact, the two neural networks are the expression of two different functionalities of ideal neurons: the long term memory formation is modeled by the SOM extensive training while the dynamic oscillatory correlation of sensory cortex neurons excited by an auditory stimulus is schematized by LEGION. The LEGION network model and the details of the SOM-LEGION coupling are developed in the next section.

C. Segregation: LEGION

Increased insight in the oscillatory correlation properties of the neurons in the sensory cortex during the 1980s resulted in an increase in theoretical research on possible computational models of the corresponding biological mechanisms. One of the first models thus constructed, by von der Malsburg and Schneider [20], was later extensively developed in the auditory context by Wang [21], [22] using a so-called *shifting synchronization theory*, based on oscillatory correlation, where neuronal oscillators representing the neuronal counterpart of specific sound features are used.

In that context, each sound object was represented by synchronization of a group of oscillators corresponding to the relative sound features. Contrarily, desynchronization among different groups of oscillators meant that the sound is the sum of different auditory streams.

The Wang model is based on a particular network architecture referred to as LEGION [13]: it is generally composed of a 2D grid of oscillators, in our coupled SOM-LEGION architecture corresponding to the units in the SOM. The dynamics of the i-th oscillator is the combined activity of an excitatory unit x_i and an inhibitory unit y_i:

$$\dot{x_i} = 3x_i - x_i^3 + 2 - y_i + I_i H\left(p_i - \varphi\right) + S_i + \rho\,, \qquad (9)$$

and

$$\dot{y_i} = \epsilon\left(\gamma\left(1 + \tanh\left(x_i/\beta\right)\right) - y_i\right)\,, \qquad (10)$$

where I_i is the external stimulation, H is the Heaviside function, p_i is the so-called lateral potential, φ is a threshold, S_i is the overall coupling contribution due to the near oscillators of the network and $\rho < 0$ is a source of Gaussian noise. There are three regulating parameters: γ, ϵ and β, where the last two are small positive constants.

The external stimulation I_i in (9), together with the permanent connection weights T_{ik} (explained later), entails the core of the SOM-LEGION coupling. I_i depends on the distance between the input raw feature vectors $\vec{r}(t)$ and the i-th unit of the trained SOM, closely related to the quantization error in the SOM. It is computed as

$$I_i\left(t\right) = IH\left[\left\|\vec{r}\left(t\right) - \vec{m}_i\right\|^{-1} - \lambda M_h\left(t\right)\right], \qquad (11)$$

where I is a positive constant, H the Heaviside function, $\left\|\vec{r}\left(t\right) - \vec{m}_i\right\|^{-1}$ is a measure of the similarity of the input vector to the i-th unit, $M_h\left(t\right)$ is the h-order simple moving average of the inverse of the distance of the BMU and $0 < \lambda < 1$ is a relative threshold. Because of the use of H, this formulation of the external stimulation can be referred to as a binarization: oscillators similar enough to the raw feature vector are stimulated, while those too far away are not.

It must be clear by now that all variables in (9)-(10) are dimensionless. It holds true for the variable of integration which is naturally referred as time and that here we call internal time or LEGION time and indicated as $t_{_L}$; at the contrary in (11) the real time is involved. The simplest way to match them is to fix a certain LEGION time interval $\tau_{_L}$ and impose the equality $\tau_{_L} = 1$ s thus avoiding the confusion between two different time scales. Returning to (9)-(10) it means that:

$$\dot{x_i} = \frac{x_i}{dt_{_L}} = \frac{1}{\tau_{_L}}\,\frac{x_i}{dt}, \quad dt = \tau_{_L}\,dt_{_L}, \qquad (12)$$

and the same holds for y_i.

If I_i is positive and $H = 1$, the i-th oscillator produces a near-steady stable orbit between a so-called silent phase (left branch of the $\dot{x}$-nullcline cubic function in (9)) and an active phase (right branch). The passage between them occurs at a faster time scale compared to motion within each phase, thus resulting in a sort of jumping. Finally, the parameter γ in (10) influences the relative time spent in each phase.

The coupling term S_i is typically composed of two terms:

$$S_i = \sum_{k \in N(i)} W_{ik} H\left(x_k - \theta_x\right) - W_z H\left(z - \theta_{xz}\right)\,, \qquad (13)$$

with the first term taking into account the phase of the oscillators in the neighbourhood, $N\left(i\right)$, through the use of *dynamic* connection weights (explained later) and the second term referring to the activity of a global inhibitor z weighted by W_z. If at least one oscillator is in the active phase, $z \to 1$ at a slow time scale whereas $z \to 0$ if all oscillators are in the silent phase, thus allowing the activation of new oscillators (for more details on the form of z and the threshold θ_{xz}, see ([21])).

Terman and Wang [23] formulated a procedure called *dynamic normalization*, significantly speeding up the synchronization within each oscillator block. It involves the dynamic connection weights, which can be assessed from the external stimulation, and the so-called *permanent connection weights*:

$$\dot{u}_i = \eta\left(1 - u_i\right) I_i - \nu u_i\,, \qquad (14)$$

$$\dot{W}_{ik} = W_T\, T_{ik}\, u_i u_k - W_{ik} \sum_{j \in Ni} T_{ij}\, u_i u_j - \omega\nu W_{ik}\,, \qquad (15)$$

where the variable u measures whether the oscillator i is stimulated, the constants $\eta \gg \nu$ are chosen so that u_i tends to 1 quickly if the oscillator i is stimulated, while it relaxes slowly to 0 when it doesn't receive any external stimulation. In (15), W_T is the so-called total dynamic connection weight and the last term, not explicitly dependent on u, is here for the first time introduced as a dissipating term weighted by the parameter ω: this term does not affect appreciably the normalization if $\omega\nu \ll 1$. When using this procedure, all the oscillators belonging to the same externally excited group receive the same amount of coupling from their neighbours, irrespective of whether they are completely surrounded by externally stimulated oscillators or not, being one of oscillators at the border of the group. The T_{ik} are called *permanent* connection weights and, contrarily to the dynamic weights W_{ik}, are fixed between two neighbouring oscillators, being the expression of the hardwired connections in the network. In the SOM-LEGION coupled model, these permanent weights are determined during training, being related to the similarity of two neighbouring units, $\delta_{ik} = \|\vec{m}_i - \vec{m}_k\|^{-1}$:

$$T_{ik} = T_{max}\left[1 + \phi\left(\frac{\delta_{ik} - \delta_{min}}{\delta_{max} - \delta_{min}} - 1\right)\right], \qquad (16)$$

where the constant T_{max} is the maximal permanent connection weight and $\phi < 1$ is a scaling factor in order to have $(1-\phi)\,T_{max} \leq T_{ik} \leq T_{max}$. Thus, the more similar two units of the SOM are, the higher the coupling between the two corresponding oscillators is.

The study of the dynamics of our LEGION network implies solving hundreds of coupled differential equations, rendering impossible any attempt to process in real-time the massive amount of data acquired by a sound measurement network. To speed up the computational process the *singular limit method* developed by Linsay and Wang [24] is extensively used. This method, in the form of an algorithm, allows skipping most of the computation by considering the fact that the oscillatory system feels the effect of oscillator changes only when oscillators jump up or down: only at those moments the lateral potential and global inhibitor values can change. Thus, the only information needed to know the dynamics of the entire system is the branch occupied by each oscillator and the time at which a jump occurs (for more details on the method, see [24]).

The lateral potential, as implemented in [24], is not suited for dynamic external stimulation $I(t)$. In this paper a different and simpler approach was used: at the end of each cycle of the algorithm the active oscillators that do not have at least 1 of 6 neighbours active are forcedly inhibited by moving them to the left branch.

III. Results

In our work we have focused on two different sound scenarios: a typical urban sound environment defined by a mixture of light and heavy traffic noise, labelled as T, and a park, with typical natural sounds and only marginally affected by human presence, labelled P. Two fixed measurement stations, one for each scenario, recorded standard 1/3-octave band levels calculated with a time resolution of 1 s. Different values for some SOM parameters were tested in order to improve the ability of SOM to identify co-occurring sound features. The dimensions of the 2D grid seemed to be not critical above lower limit values. In this paper they were fixed to $M_x = 25$ and $M_y = 100$. The most critical parameters were found to be the length of training runs, the initial value of the learning rate, α_0, and the width of the 2D neighbourhood σ_0. To evaluate the quality of the SOM training, some sound excerpts were recorded at the same scenarios but not used during the training phase. An example is provided in Fig. 2, wherein the distance between the raw feature vector related to a quiet moment at T and the units of two maps trained at T but with different flexibility are plotted. The less flexible map, which is the one trained with smaller α_0 and σ_0, displays a better focusation and is thus preferable.

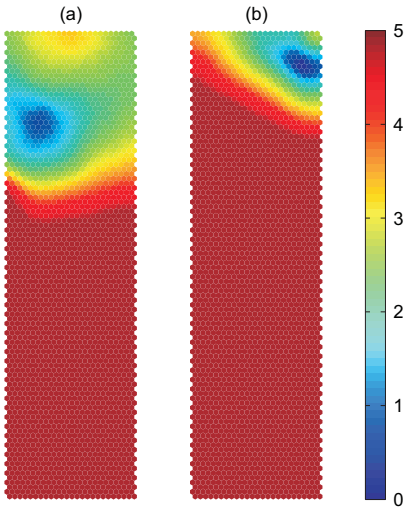

Fig. 2. Distance of the raw feature vector related to a typical sample from T and two maps trained at the same location, but with different initial parameters: (a) $\alpha_0 = 0.6$, $\sigma_0 = 50$, high flexibility; (b) $\alpha_0 = 0.03$, $\sigma_0 = 10$, low flexibility. Training length: 86400 samples.

Training maps in fixed scenarios result in a strong sound-context dependency. Thus, all of the units of a map trained in P are very dissimilar to raw feature vector corresponding to a typical sample from T, as can be seen in Fig. 3(d). Obviously, the units in such a map display a better matching for a quiet natural sound sample, as shown in Fig. 3(b). In contrast, the map trained in T shows good focusation and a low quantization error for both the samples Fig. 3(a) and (c), as even in a road traffic environment, silent periods are present (e.g. during the nocturnal part of the recording used for training). This context dependency allows an intuitive way to distinguish between common and rarely occurring (or even new) sound events, possibly triggering an alert or a more detailed analysis of the sound events: by re-training the map with that specific input, a later occurence of the same sound will no longer trigger an alert. The context dependency

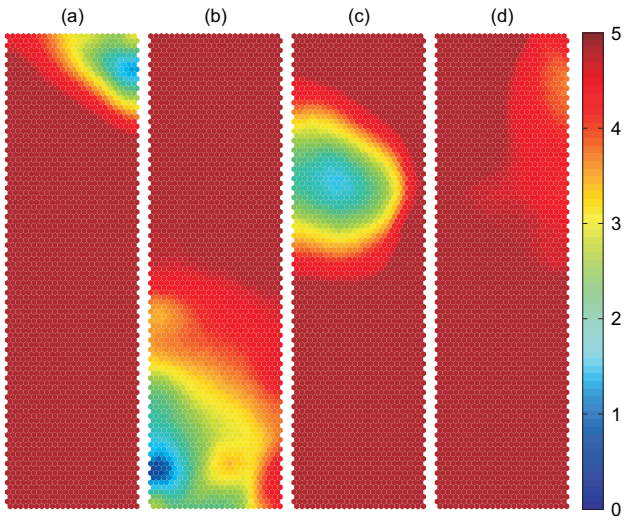

Fig. 3. Distance of the raw feature vector related to a typical sample from P and the units in the SOM trained in (a) T, (b) P. Distance of the raw feature vector related to a typical sample from T and the units in the SOM trained in (c) T, (d) P.

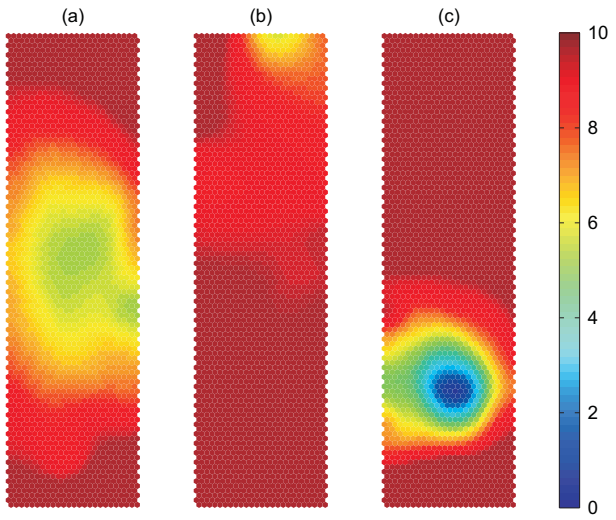

Fig. 4. Distance between the raw feature vector of a sample from a crowded street and the units in the SOM trained in (a) T, (b) P, (c) HT. Training length: 86400 samples, $\alpha_0 = 0.03$, $\sigma_0 = 10$.

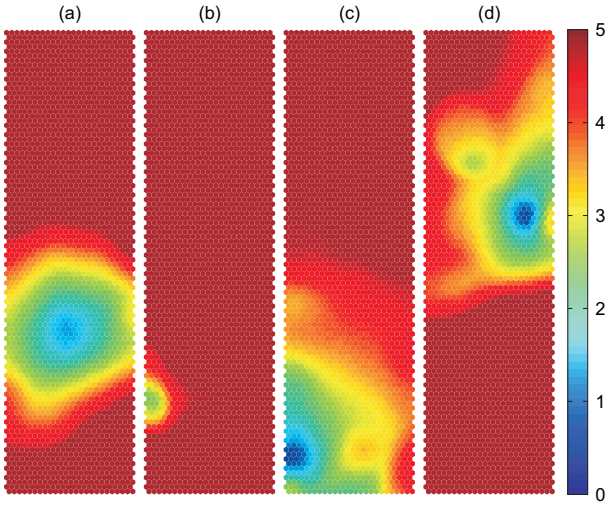

Fig. 5. Distance between the raw feature vector of a typical sample from T and the units of the SOM trained in (a) T, (b) HT. Distance between the raw feature vector of a typical sample from P and the units in the SOM trained in (c) P, (d) HP. Training length: 86400 samples, $\alpha_0 = 0.03$, $\sigma_0 = 10$.

can also be exploited in a different way: feeding a sample to a number of SOMs, each of which was trained on a different context, and comparing the focusations and the quantization errors, can yield information about which context the sample most likely belongs to.

The context dependency can be reduced by training the SOM with excerpts coming from various scenarios. There is an interesting parallel between this situation and the human brain, which is exposed to a lot of different sound contexts during life. To approximate this multi-context learning, a series of 51 sound excerpts of 15 minutes where recorded at various locations in and around the city of Ghent, including traffic-free shopping streets, street canyons with low and high traffic intensity, residential areas, open squares, urban parks and quiet areas at the edge of the city. The new sound samples replaced partly of the night time samples of each scenario, T and P respectively, thus creating two more heterogeneous scenarios called HT and HP. Two new SOMs were trained, one in HT and the other in HP. The units of the old SOMs cover the new sounds only poorly, having been trained exclusively with inputs coming from their specific scenario, T or P. In contrast, the new SOMs are very versatile and can match practically all types of inputs corresponding to the wide range of scenarios they have been trained on. In Fig. 4 this aspect is visualized by taking into account a 1 s sound fragment from a crowded shopping street, where talking passers-by can be heard. Moreover, the new SOMs still show a low quantization error for samples from T or P, as shown in Fig. 5.

The U-matrix of the SOM trained in HT is shown in Fig. 6, revealing how the SOM is composed of regions where neighbouring units are very similar and regions where the opposite holds true. This is common if the SOM has been trained on the basis of a very diverse set of sounds (e.g., coming from very different contexts).

As explained at the end of Section II-B, the LEGION oscillators and the SOM units are two different functional representations of the same neural units. In particular, the units best matching the input can be interpreted as externally excited neuronal oscillators, in accordance with (11). LEGION thus provides:

1) grouping of contiguous excited oscillators representing particular raw feature vectors, by means of coherent oscillations;
2) segmentation of distinct groups of oscillators by introducing a phase among the groups oscillation.

For the simulation shown in Fig. 7 the SOM trained in P was chosen. The parameters for SOM training were set as follows: $\alpha_0 = 0.03$, $\sigma_0 = 10$. The values $h = 3$ and $\lambda = 0.92$ were used for binarization in (11) and the

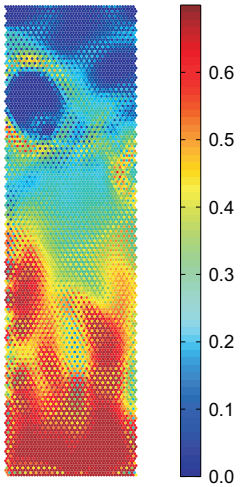

Fig. 6. U-matrix of a SOM trained in HT. Training length: 86400 samples, $\alpha_0 = 0.03$, $\sigma_0 = 10$.

external stimulation I was set respectively to 0.2 and 0 for stimulated and unstimulated oscillators. The neighbourhood was composed of the 6 nearest neighbours. The maximal value of the global inhibitor W_z was set to 1.7. The following values for the parameters regarding the dynamic connection weights W_{ik} in (14)–(15) were used: $\eta = 3.0$, $\nu = 0.1$ and $\omega = 1$. The permanent connection weights, as defined in (16), were calculated using $\phi = 0.5$ and $T = 1.5$. Of the parameters in (9)–(10) governing the dynamics of a single oscillator, only γ is needed if the singular limit method is adopted, and it is set to 6.5 here. Finally, for the LEGION time, the value $\tau_L = 15$ was used.

Fig. 7 shows oscillatory dynamics of LEGION together with the similarity to the SOM units and the external stimulation $I(t)$ for a period of 2 s. It is a clear example of the ability of LEGION to segregate different groups of stimulated oscillators by letting them move to the active phase at different times. In particular, Fig. 7 at t = 4.2 s shows the transient phase wherein the oscillators recombine their dynamic connection weights to adapt themselves to the new external input.

IV. Conclusions

A model for context-dependent environmental sound monitoring, rigidly grounded on neurological mechanisms, was constructed in this paper. The plasticity of the human cortex, in the context of processing spectro-temporal features, was simulated by the use of a Self-Organizing Map (SOM) based on 1 s standard 1/3-octave band levels. Much as human beings do, sounds were learned within the context in which they were usually heard, resulting in a high context dependency and a high tuning of the model on the typical sounds heard in the specific scenario. In other words, how the presence or absence of a sound during training influenced the SOM, depends on the other sounds perceived during training. After training, the map could be used to assess how typical a new sound fragment is by determining its similarity to the units

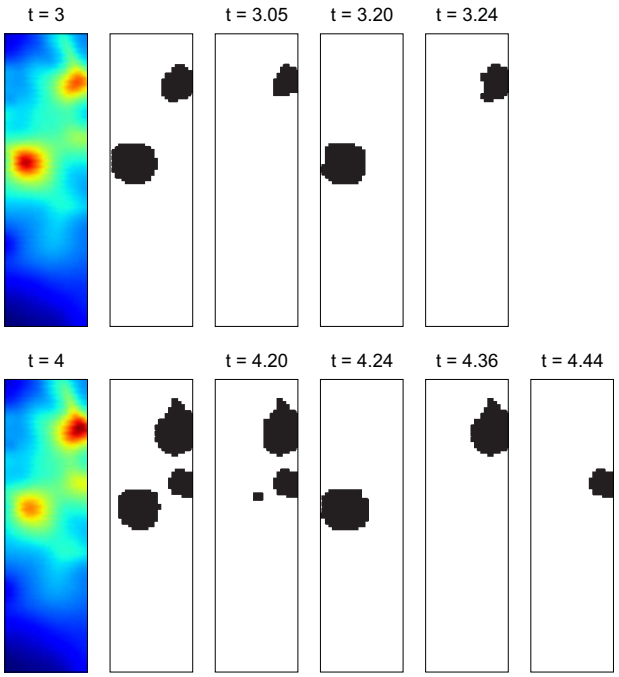

Fig. 7. Left (2 columns): similarity (inverse of the distance) of two input samples at t = 3 s (top) and t = 4 s (bottom), before (1st column) and after (2nd column) binarization ($\lambda = 0.92$, moving average order: $h = 3$). Right (4 columns): some snapshots of LEGION taken at different times. The samples used here are extracted from test input data recorded in scenario P.

of the SOM.

A different manifestation of the context dependency is the number of nodes a SOM devotes to a specific type of sound (e.g. car passages, near-silence, pedestrian chatter). Correspondingly, the more heterogeneous the soundscape on which a SOM is trained, the smaller the number of nodes dedicated to each specific type of sound.

By coupling the SOM to a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION), which simulates the oscillatory correlation activity of the neuronal sensory cortex, we were able to use the coupled model for object formation and segregation tasks, where an object in our context is a group of contiguous units similar to the new sound sample.

The model could be used to distinguish between common and rare sound events in a context-specific manner. Moreover, we feel the model merits further research in order to assess its suitability for specific environmental sound recognition and segregation. In order to do so, future work will have to focus on increasing the time resolution and the sound stream formation ability by reducing the transient time in the LEGION oscillatory dynamics.

A different avenue of interest is increasing the biological plausability of the SOM-LEGION coupling. More to the point, previously unheard sound events can result in little activation of the map, while one would prefer to have a LEGION-segregation between known and unknown components even in such a setting, especially in case of highly salient, though unknown, events. Segregation could also

be performed for unknown sound events by changing the binarization threshold, perhaps by considering local maxima in activation of the SOM. In the current implementation, previously unheard components will likely be ignored due to them having a smaller activation than the known components of the sound event. Another issue regards the training phase. In this paper there is a sharp distinction between training and testing phase, which is not biologically plausible: to a certain extent, connections in the brain remain flexible, and training from external stimuli remains possible. A possible improvement of the model could be to trigger a new SOM learning phase when conspicous but unknown sound events are observed.

ACKNOWLEDGMENT

Bert De Coensel is a postdoctoral fellow of the Research Foundation – Flanders (FWO–Vlaanderen); 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] H. Karl and A. Willig, *Protocols and Architectures for Wireless Sensor Networks*. Chichester, UK: John Wiley & Sons, Ltd., 2005.

[2] M. Mancas, L. Couvreur, B. Gosselin, and B. Macq, "Computational attention for event detection," in *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)*, Bielefeld, Germany, Mar. 2007.

[3] L. Couvreur, F. Bettens, J. Hancq, and M. Mancas, "Normalized auditory attention levels for automatic audio surveillance," in *Proceedings of the 2nd International Conference on Safety and Security Engineering (SAFE 2007)*, Malta, Jun. 2007.

[4] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless sensor networks for habitat monitoring," in *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications (WSNA'02)*, Atlanta, Georgia, USA, Sep. 2002.

[5] H. Wang, D. Estrin, and L. Girod, "Preprocessing in a tiered sensor network for habitat monitoring," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 4, pp. 392–401, 2003.

[6] L. Filipponi, S. Santini, and A. Vitaletti, "Data collection in wireless sensor networks for noise pollution monitoring," in *Proceedings of the 4th IEEE/ACM International Conference on Distributed Computing in Sensor Systems (DCOSS'08)*, Santorini Island, Greece, Jun. 2008.

[7] S. Santini, B. Ostermaier, and R. Adelmann, "On the use of sensor nodes and mobile phones for the assessment of noise pollution levels in urban environments," in *Proceedings of the 6th International Conference on Networked Sensing Systems (INSS'09)*, Pittsburgh, Pennsylvania, USA, Jun. 2009.

[8] 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.*, vol. 126, no. 2, pp. 656–665, 2009.

[9] C. Kayser, C. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Curr. Biol.*, vol. 15, no. 21, pp. 1943–1947, 2005.

[10] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, Aug. 2007, pp. 1941–1944.

[11] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," in *Proceedings of the 15th European Signal Processing Conference (EUSIPCO 2007)*, Poznań, Poland, Sep. 2007, pp. 1206–1210.

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

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

[14] B. De Coensel, D. Botteldooren, and T. De Muer, "$1/f$ noise in rural and urban soundscapes," *Acta Acust. Acust.*, vol. 89, no. 2, pp. 287–295, 2003.

[15] B. De Coensel and D. Botteldooren, "The quiet rural soundscape and how to characterize it," *Acta Acust. Acust.*, vol. 92, no. 6, pp. 887–897, 2006.

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

[17] H. Yin, *The Self-Organizing Maps: Background, Theories, Extensions and Applications*. Springer, 2008, pp. 715–762.

[18] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 1–6, 1998.

[19] A. Ultsch, "Self organized feature maps for monitoring and knowledge acquisition of a chemical process," in *Proc. ICANN*, vol. 93, Amsterdam, the Netherlands, Sept. 1993, pp. 864–867.

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

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

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

[23] D. Terman and D. Wang, "Global competition and local cooperation in a network of neural oscillators," *Physica D: Nonlinear Phenomena*, vol. 81, no. 1-2, pp. 148–176, 1995.

[24] P. Linsay and D. Wang, "Fast numerical integration of relaxation oscillator networks based on singular limit solutions," *IEEE Transactions on Neural Networks*, vol. 9, no. 3, pp. 523–532, 1998.