2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Long-Term Learning Behavior in a Recurrent Neural Network for Sound Recognition

Michiel Boes, Damiano Oldoni, Bert De Coensel and Dick Botteldooren

***Abstract*— In this paper, the long-term learning properties of an artificial neural network model, designed for sound recognition and computational auditory scene analysis in general, are investigated. The model is designed to run for long periods of time (weeks to months) on low-cost hardware, used in a noise monitoring network, and builds upon previous work by the same authors. It consists of three neural layers, connected to each other by feedforward and feedback excitatory connections. It is shown that the different mechanisms that drive auditory attention emerge naturally from the way in which neural activation and intra-layer inhibitory connections are implemented in the model. Training of the artificial neural network is done following the Hebb principle, dictating that *"Cells that fire together, wire together"*, with some important modifications, compared to standard Hebbian learning. As the model is designed to be on-line for extended periods of time, also learning mechanisms need to be adapted to this. The learning needs to be strongly attention- and saliency-driven, in order not to waste available memory space for sounds that are of no interest to the human listener. The model also implements plasticity, in order to deal with new or changing input over time, without catastrophically forgetting what it already learned. On top of that, it is shown that also the implementation of short-term memory plays an important role in the long-term learning properties of the model. The above properties are investigated and demonstrated by training on real urban sound recordings.**

I. Introduction

THE perception and analysis of an acoustic environment, also known as auditory scene analysis (ASA), is an important process in which the human brain easily outperforms current computer models. The process involves the decomposition of complex sound mixtures into individual auditory streams, and the attribution of meaning to these sound streams. In order to accomplish this, the human brain makes use of not only a number of different auditory cues, but also visual and other sensory input [1]. The process of auditory stream segregation is believed to be largely dependent on auditory attention [2][3][4]. By means of competitive selection, attention determines which sensory input, and thus which sound stream, is selected for further analysis by the brain. A single auditory stream is thus selected for entry into working memory, where it is consciously perceived and used in the formation of a mental image of the listener's acoustic environment [5]. Attributing a meaning to sound streams requires recognition of patterns in the auditory input. As these patterns for known input sound streams can and do evolve over the course of time, and new, previously unheard, sound streams can occur, there is a need for flexibility to adapt to changing patterns and to learn new ones, without completely forgetting already known patterns. Thus, both auditory attention and flexibility in learning are indispensable for any computational ASA model.

In previous work by the same authors, a biologically inspired neural network model for auditory scene analysis, incorporating both auditory attention and learning flexibility, was developed [6][7], upon which the model presented in this paper builds. The model consists of 3 neural layers, connected to each other by feedforward excitatory connections as well as feedback excitatory connections between the last two layers. Excitatory and inhibitory connections within each of the layers are implemented as a K-winner-takes-all process, modelling the mechanism of competitive selection, which is essential to auditory attention. Learning and adaptation of the synaptic connection weights between the neurons is done following the Hebbian principle of *"Cells that fire together, wire together"*, with significant modifications, compared to standard Hebbian learning. On top of this, also long-term and short-term synaptic plasticity are implemented in a biologically inspired way. Whereas in previous work focus has been on the structure of the model, and the way in which all important auditory attention mechanisms emerge naturally from the biologically inspired implementation of neuron behavior in the network, this paper will focus on the learning mechanisms in the model. Both long- and short-term memory are implemented in the model and their behavior and mutual interaction will be investigated. Also, the way the model handles the so-called 'stability-plasticity' problem [8][9] - how to design a system that is sensitive to new or changing patterns, without catastrophically forgetting previously learned ones - will be examined.

The present model is designed to be integrated into a large-scale noise monitoring network, in which its goal would be to detect and interpret conspicuous sound events that would have been noticed by a human listener, in order to assess the environmental sound quality [10][11]. Consequently, the model should be able to run continuously for weeks to months on low-cost hardware, with limited computational power. Thus, as in all models that simulate human brain functions, compromises between biological accuracy and computational efficiency have to be made. Nevertheless, even though strong simplifications of biological mechanisms have been made, the structure of the model, and the way the different sub-models interact are strongly based on available knowledge of the human brain.

The authors are with the Department of Information Technology (INTEC), Ghent University, Belgium (email: Michiel.Boes@intec.ugent.be).

Michiel Boes is a doctoral fellow, and Bert De Coensel is a postdoctoral fellow of the Research FoundationFlanders (FWOVlaanderen); the support of this organization is gratefully acknowledged.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

The remainder of this paper is structured as follows. The next section contains a description of the model: its architecture, the learning mechanisms it uses and the way it includes auditory attention. In Section III, the model's learning systems are investigated in more detail, with a special focus on the 'stability-plasticity' problem. Finally, in Section IV, our conclusions are given.

II. Model Description

The model used in this paper is the same as the one presented in [7], with some subtle but nonetheless important modifications. For reference, it is described below, with special emphasis on the changes compared to the original version.

A. Model Architecture and Input

In Fig. 1 a schematic overview of the model's architecture is given. Essentially, it is an artificial neural network consisting of three layers: a layer encoding sensory input ϕ, a middle layer χ, and a layer encoding sound concepts ψ. Excitation of the ϕ-layer is given by a highly simplified simulation of sound processing in the human ear, as described further on in this section. Feedforward excitatory connections exist between the ϕ- and χ-layers, as well as between the χ- and ψ-layers. These connections serve as an instantaneous sound classification, directly based on the input from the ear. As the temporal aspect is of great importance in sound stream identification, this is included in the model by means of feedback connections from the ψ- to the χ-layer, with a delay of one timestep (for the current paper $\Delta t = 0.125$s). Thus, the identified sound in the previous timesteps will influence the way the sound in the current timestep is perceived.

The input to the model is calculated following the method described in [12]. A $1/3$-octave band spectrum of the input sound is calculated with a temporal resolution of Δt, after which a cochleagram over the complete frequency range of human hearing is calculated (0-24 Bark with a resolution of 0.5 Bark, thus resulting in 48 values), where the Zwicker loudness model is used to take energetic masking into account [13]. Subsequently, Gaussian and difference-of-Gaussian filters with different scales are applied to the obtained cochleagram, resulting in features encoding intensity (4 different filters), spectral contrast (6 different filters) and temporal contrast (6 different filters). Thus, per timestep and per bark, $4+6+6=16$ features describing the input sound are obtained, resulting in a total of 768 (48 spectral values $\times$ 16 features per spectral value) values for each timestep, that are used as excitatory values for the artificial neurons in the ϕ-layer. Thus, the number of neurons in the ϕ-layer is fixed at 768, while the amount of neurons in the other layers can be chosen freely (in the current paper χ contains 1000 neurons and ψ 100). It is important to note that the above described sound features, that are used as input for the model, are closely linked to measures for auditory saliency as calculated in [12] and [14]. As described in Section II-C, this plays an important role in the way auditory attention is incorporated in the model. Excitation of neurons in the χ-

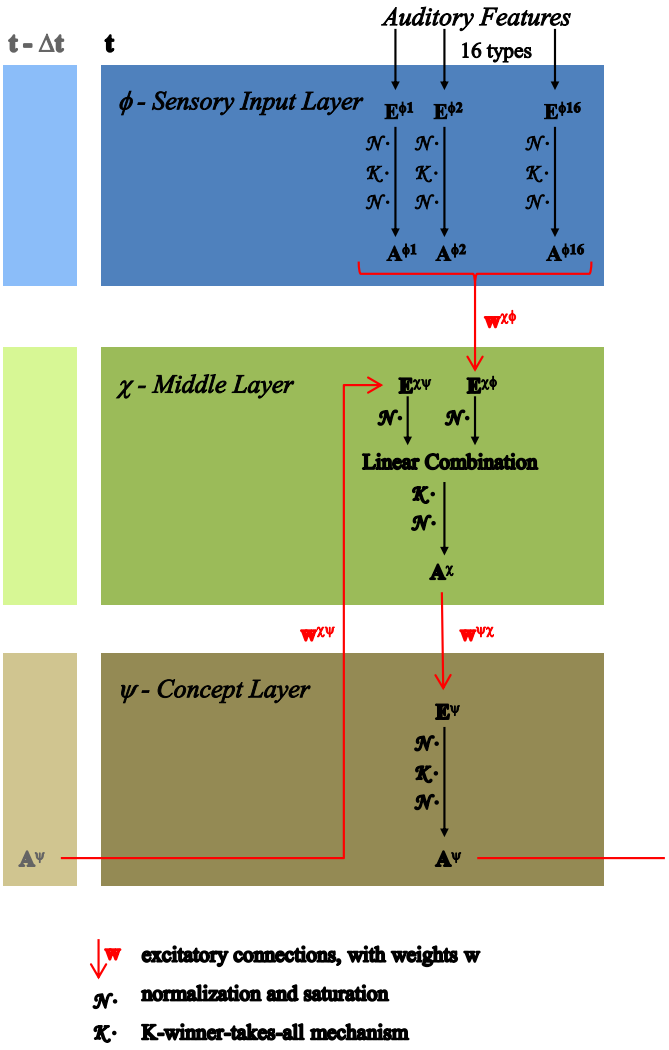

Fig. 1. A schematic overview of the model, displaying its layered structure (three layers: ϕ, χ and ψ), its temporal structure (the link with neural activations on the previous timestep), and both the inter-layer (excitatory connections) and intra-layer dynamics (normalization and saturation, and K-winners-take-all mechanisms)

and ψ-layers is calculated as the sum of activations of the neurons the excitatory connections originate from, weighed by their corresponding connection weights:

$$E_j^Y(t) = \sum_i w_{ji}^{YX} A_i^X(t), \tag{1}$$

where $A_i^X(t)$ is the final activation of the ith neuron in layer X, w_{ji}^{YX} is the connection weight from neuron i in layer X to neuron j in layer Y, and $E_j^Y(t)$ is the excitatory input to neuron j in layer Y. As was mentioned before, in the case of the feedback connections from ψ to χ, an extra time delay of one timestep Δt is taken into account. In this formula, X and Y represent two generic layers which are linked by connections originating from X and going to Y (X and Y can thus be interpreted as ϕ and χ, as χ and ψ or as ψ on the

previous timestep and χ), and this naming convention will also be used further on in this paper.

B. Neural Dynamics

In each of the neural layers, first, the excitatory inputs are normalized. As the system is designed to run continuously it is impossible to calculate a normalization constant on the complete input set, and thus normalization is done by dividing the excitatory input to each layer by a factor that is calculated as a leaky integral, slowly (for this work, $\tau = 1000$s is used) following the maximal excitation value within this layer. This process attempts to keep excitation values in the interval $[0, 1]$, without eliminating the natural temporal variation in excitation strength. Still, it does not guarantee values between 0 and 1, and thus, in order to obtain this, a saturation function is applied:

$$E_i'^X(t) = \frac{E_i^X(t)/\nu^X(t)}{1 + E_i^X(t)/\nu^X(t)}, \tag{2}$$

in which $E_i^X(t)$ is the excitation of neuron i in layer X before, and $E_i'^X(t)$ after normalization and saturation, and $\nu^X(t)$ the normalization factor for layer X at time t.

As the feature values used as input to the ϕ-layer have different dynamical ranges for the different Gaussian and difference-of-Gaussian filters used in their calculation, the above normalization process is not applied to the ϕ-layer as a whole, but to the different feature types separately. This in contrast to the original model described in [7], in which, because of the imbalance between the different filters, certain feature types were always dominant. Also, in the χ-layer, bottom-up, feedforward excitation, coming from the ϕ-layer is normalized independently on the normalization of the top-down, feedback excitation, coming from the ψ-layer at the previous timestep, before the two are linearly combined (in the current paper, the combination weights are 0.6 for bottom-up excitation, and 0.4 for top-down excitation).

Subsequently, a mechanism similar to K-winner-takes-all is applied, simulating excitatory and inhibitory connections within the layer in order to obtain competitive selection between the neurons inside this layer. As will be explained in Section II-C, this is an essential mechanism in auditory attention. The threshold e_K, above which neural excitation survives the competitive selection, is calculated as a leaky integral, following the value of the K'th most strongly excited neuron after normalization. The remaining excitation is the calculated as follows:

$$E_i''^X(t) = \begin{cases} E_i'^X(t) - e_K & E_i'^X(t) > e_K \\ 0 & E_i'^X(t) < e_K \end{cases}, \tag{3}$$

where, $E_i'^X(t)$ is the normalized excitation of node i in neural layer X at timestep t, $E_i''^X(t)$ is the same excitation after competitive selection, and e_K is the threshold value. In contrast to the original model, maximal excitation is not kept constant under competitive selection. Thus, in case of high competition between the neurons, the value of the K'th most strongly excited neuron will be relatively close to the strongest excitation, and competitive selection will attempt to leave final neural excitation relatively low. This reflects the uncertainty in the model in case of highly competitive input. On the other hand, in order for the model's attention mechanisms to work, excitation strengths, or more specifically temporal changes in excitation strengths, need to be conserved throughout the different layers, as will be explained in Section II-C. Thus, the time constant in the leaky integrator determining the threshold e_K needs to be considerably larger than one timestep.

In order to calculate the final activation of the neurons, $A_i^X(t)$, another normalization procedure, exactly alike the first one, is executed on the resulting excitation values after competitive selection, in order to obtain activation values, making use of the full $[0, 1]$ interval.

$$A_i^X(t) = \frac{E_i''^X(t)/\mu^X(t)}{1 + E_i''^X(t)/\mu^X(t)}, \tag{4}$$

in which $A_i^X(t)$ is the final activation of neuron i in layer X, $E_i''^X(t)$ the excitation value of the same neuron after normalization and competitive selection, and $\mu^X(t)$ the corresponding normalization factor for layer X at time t.

C. Auditory Attention

Human attention is believed to consist of an interplay between bottom-up, saliency-based attention and top-down, voluntary attention, and a competitive selection mechanism on top of that [1][4]. The concept of inhibition-of-return is usually introduced in human attention models in order to prevent attention from permanently staying fixed on one single item [12]. Whereas these attention mechanisms could be included in existing human perception models by artificially adding extra parameters and submodels [6], in the current model, they automatically arise from the way in which biological neural behavior has been implemented.

The sound features that are used as input to the model, as explained in Section II-A, are very similar to the ones used in [12] and [14] in order to generate auditory saliency maps. In addition, the normalization procedure and K-winner-takes-all mechanism used in the current model show important similarities to further processing of the features in order to obtain saliency values in the before mentioned works. Thus, it can be concluded that neural activation in the ϕ-layer is closely linked with auditory saliency. Furthermore, linearity in the calculation of neural excitation in χ- and ψ-layers, in addition to sufficiently large time constants in the calculation of both normalization and competitive selection, ensure that changes in saliency of the input sound are reflected in changes in activation in all neural layers. This process represents saliency-driven bottom-up attention.

Top-down auditory attention, on the other hand, is a conscious mechanism, based on known information about the listener's environment. In the current model, this is implemented in the form of feedback connections from ψ to χ: the perception of the input sound is thus biassed by the recognised sound streams in the previous timestep. In addition, a biasing top-down excitation to neurons in the ψ-layer, representing sounds in which the listener is currently

interested, can be implemented. Finally, competitive selection will determine which neurons will be selected to be activated, and thus which auditory streams are being paid attention to.

Finally, synaptic fatigue serves to implement the concept of inhibition-of-return. In case of persistent stimulation of a synaptic connection, the neurotransmitter vesicle regeneration rate will not be able to keep up with its release rate, and the synapse will be temporarily inhibited [15]. In the current model, the evolution of the concentration of the vesicles is modelled as follows:

$$v_i^X(t+\Delta t) = \min[1, v_i^X(t) + \Delta t(R - v_i^X(t)A_i^X(t))], \quad (5)$$

where $v_i^X(t)$ is the concentration of vesicles in the synapses leaving from neuron i in layer X on time t, R is the regeneration rate of vesicles, and $A_i^X(t)$ the activation of neuron i, which, multiplied with the concentration of available vesicles, can be interpreted as the rate at which vesicles are released. Note that the vesicle concentration is limited to a maximum value of 1. Finally, by multiplying each synaptic connection weight w_{ji}^{YX} by its corresponding vesicle concentration v_i^X, synaptic fatigue is taken into account in the model. And this effectively prevents neurons from staying permanently strongly activated, thus simulating the effect of inhibition-of-return.

D. Learning and Memory

Learning in the current artificial neural network model is done following the Hebb principle. In the original version of the model, all connection weights were initialized at values close to their maximum (1), and learning consisted of reducing the weights of connections between neurons that were not firing together, and thus did not obey Hebb's rule. Forgetting, on the other hand, was embedded in the model by slow, random synaptogenesis or synapse growth, which, by virtue of its randomness, decreased the contrast in the learned connection patterns at a low rate. Although this system has some very interesting and desirable properties, it also has a particular flaw, as described in section III, and in the current model it is generalized, in order to address this problem.

In the model presented in the present paper, connection weights are initialized at random values, around a base level, close to, but still lower than the maximum connection strength (for the current work, a base level of 0.8 is used). Learning is still mainly done by pruning away the undesired connections, but in addition to this, also a slight growth in connection strength for connections that are firing together is introduced.

$$w_{ji}'^{YX} = w_{ji}^{YX} + \eta[A_j^Y(A_i^X - \theta)], \quad (6)$$

where w_{ji}^{YX} and $w_{ji}'^{YX}$ are the synaptic connection weights from neuron i in layer X to neuron j in layer Y, respectively before and after learning, A_i^X and A_j^Y the activations of the neurons connected with this weight, η a constant, defining the learning rate, and θ a threshold, determining whether learning will cause the weights to increase or decrease: when activation of the neuron in X exceeds θ, the corresponding connection weights will be strengthened by learning, otherwise they will be weakened.

Random synapse growth in the original model is now replaced by random connection strength convergence towards the base level, thus still decreasing the contrast in learned patterns, and still effectively simulating forgetting in the model. This is done as follows:

$$w_{ji}''^{YX} = w_{ji}'^{YX} + \zeta R(B - w_{ji}'^{YX}), \quad (7)$$

where $w_{ji}'^{YX}$ and $w_{ji}''^{YX}$ are the synaptic connection weights from neuron i in layer X to neuron j in layer Y, respectively before and after the forgetting step, B is the model's base level, R is a (pseudo-)random number between 0 and 1, and ζ is a parameter, determining the speed of forgetting. Thus, the current Hebbian learning and forgetting model can be seen as a generalization of the original in [7], which can be found by using a base level of 1. Although synaptic pruning and synaptogenesis are no longer strictly correct names for these two effects, we will still use these terms further on in this work.

Short-term memory is included in the model as temporary modifications of synaptic connection weights, whereas synaptic pruning and synaptogenesis permanently change these weights, thus representing long-term memory. Short-term memory is implemented through a leaky integrator for each synaptic connection, converging towards a preset maximal value M_{STM} (in the current work $M_{STM} = 1.5$), when the two nodes connected by this synapse are activated simultaneously, and converging to its default value of 1 when these nodes do not fire together, at a faster rate than pruning and synaptogenesis. By multiplying this factor for each connection with its corresponding weight w_{ji}^{YX}, the resulting short-term memory adjusted connection weights will learn new patterns at a significantly faster rate, but will also forget these newly learned patterns quickly if they are not repeated, thus effectively simulating short-term memory in the human brain.

III. Learning, Forgetting and Flexibility

In case of a learning base level of 1 (the original model, as described in [7]), learning will decrease connection strengths of synapses between neurons that do not fire together, but will not increase any connection strengths. Thus, for neurons j in layer Y, this process indeed forms patterns in the connection strengths originating from layer X. However, all connections will still be weaker than their base level, which is also their maximum strength. Thus, as can be seen in Fig. 2, a pattern similar to the just learned one will excite an untrained Y-neuron more strongly than the one trained to the corresponding pattern, as all connections leading to the untrained neuron will be stronger than those to the trained one. From Eq. 6, it can be seen that only connections leading to an activated neuron in layer Y ($A_j^Y > 0$) will be updated by learning, thus, neurons which have only comparatively weak connections leading to them will never be updated. As a consequence of this, in the original model's early learning

phase, each timestep, new Y neurons would be activated and have their connections updated, no matter how similar the input was to any previous ones. After a number of timesteps, this process had taken up all 'memory space' in Y for all relatively similar patterns. In order to solve this problem, the trained Y neurons need to be kept competitive against untrained base level Y neurons, and consequently, it is necessary that learning not only weakens undesirable connections, but also strengthens desirable connections. Thus, the model generalization with arbitrary base levels, as explained in section II-D, was introduced. The base level B and growth threshold θ determine the growth strength, and this, in turn, determines how similar the input to a layer has to be to an already memorized pattern in order to activate the corresponding Y neuron and thus to contribute to the learning process of this pattern, as can be seen in Fig. 3. This is rather similar to the vigilance parameter introduced in Adaptive Resonance Theory (ART) networks [16]. Low growth rates for learning will lead to narrow generalization, fine categories, and detailed memories, as incoming patterns will have to match the learned ones very good in order to excite the corresponding neurons more strongly than base level neurons. Exactly the same effect is seen for high vigilance values in ART. High growth rates, on the other hand, will lead to broad generalization, coarse categories, and abstract memories, similarly to low vigilance values in ART. In case of a base level of 1, or the corresponding case of maximum vigilance in ART, category learning is reduced to exemplar learning, as long as there is unused memory (Y neurons with only base level connections leading to them) left.

For testing purposes, the model was trained on several hours of consecutive recordings in a typical urban park, thus containing mainly chirping sounds of birds, human speech from passing pedestrians, yelling children and some traffic noise. Training was done with a base level $B = 1$, and a second time with $B = 0.8$ and $\theta = 0.1$. For both cases, a measure for similarity in the trained network was calculated as follows: for each neuron in the middle layer χ, first, the correlation between its own trained pattern originating from ϕ and the patterns of all other neurons in χ from ϕ is calculated, and subsequently these correlation values are averaged. This averaged correlation value is a measure that describes how unique this particular trained pattern is. Thus, in the case of $B = 1$, we expect all trained patterns to be relatively similar, thus resulting in high correlations, for reasons given in the previous paragraph. In the case of $B = 0.8$, on the other hand, lower correlations and more unique trained patterns would be expected, and each neuron will represent a broader category of patterns. Indeed, in Fig. 4, a distribution of neurons as a function of their average correlation with the other neurons is plotted for both cases, and the obtained results turn out to be as expected: the most common average correlation value decreases from 0.6 to around 0.5, indicating a more varied set of learned patterns, for the same set of input patterns.

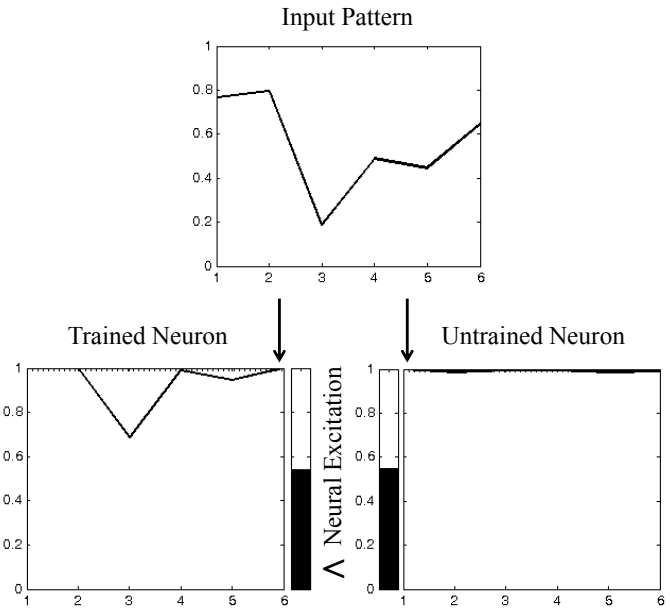

Fig. 2. An example input pattern (activation values of neurons in the X layer) and two stored patterns (connection weights from neurons in X layer to the two neurons in the Y layer), in case of a base level of 1.0. The left neuron is already trained to the input pattern, and its total excitation is slightly smaller than the total excitation of the right neuron, which is untrained (connection weights close to base level), in contrast to the case shown in Fig. 3.

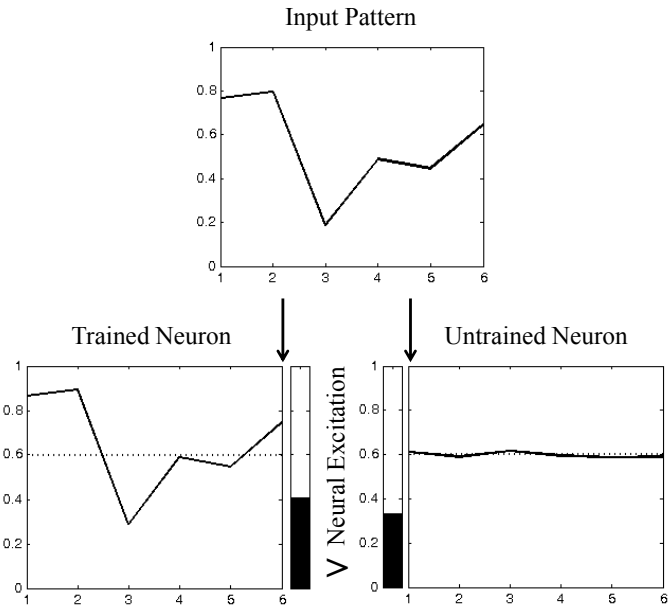

Fig. 3. An example input pattern (activation values of neurons in the X layer) and two stored patterns (connection weights from neurons in X layer to the two neurons in the Y layer), in case of a base level of 0.6. The left neuron is already trained to the input pattern, and its total excitation is larger than the total excitation of the right neuron, which is untrained (connection weights close to base level), in contrast to the case shown in Fig. 2.

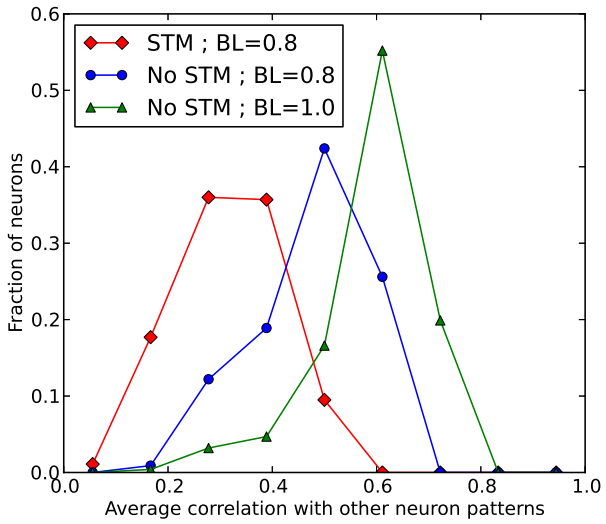

Fig. 4. Distribution of neurons in ϕ layer according to the average correlation between its trained pattern and those of the other neurons in the layer. This is given for the case with short-term memory enabled and a connection strength base level of 0.8 (red graph with diamond markers), for the case with short-term memory disabled and a base level of 0.8 (blue graph with circular markers) and for the case with short-term memory disabled and a base level of 1.0 (green graph with triangular markers).

For the above training, short-term memory was not enabled, as it also has an effect on the vigilance of the network, as it is called in the context of ART. Short-term memory will temporarily strengthen the desirable connections, but not weaken the undesirable ones, and thus make the corresponding neuron easier to activate, and effectively decrease vigilance for this particular neuron. As can be seen in Fig. 4, enabling short-term memory reduces the average correlation value even further, to about 0.3, as was to be expected. It is interesting to note that the short-term memory mechanism can easily be adapted to also, or even only, temporarily weaken undesirable connections, with potentially complex influence on the model's vigilance. However, further research is needed in order to fully explore the possibilities this would create.

As the model is designed to run in a large-scale noise monitoring network, it will be learning continuously on days, weeks, and even months of data. As these data are real recordings of environmental sounds in various urban locations, for large periods of time, low-saliency urban background sound will be processed. However, less frequent and more salient sound events are often of much greater interest to a listener, and should thus certainly be learned by the model, even if the amount of available input data for these events is considerably lower. Thus, learning should be very attention- and saliency-driven, in order to reach a balance between the very frequent but less salient sounds and the uncommon but salient sounds in the trained network. The amount of influence saliency has on learning is mainly determined by the time constants of e_K in competitive selection. If these are relatively short, e_K quickly moves to the value of the Kth most strongly excited neuron, and the resulting final activation of the neuron will be mainly determined by the difference between the most strongly excited neuron and Kth most strongly excited one, and not by saliency of the input. The longer these time constants are, the longer it takes for e_K to reach the value of the Kth most strongly excited neuron, and the longer saliency will play an important role. However, if the values of the time constants are too high, there will be long periods of time in which all neurons will be activated (because e_K is increasing very slowly, and thus staying too low for a long period of time) or no neurons at all will be activated (because e_K is decreasing very slowly, and thus staying too high for a long period of time), and thus, nothing sensible can be recognised by the model in the incoming sound. The effect of a change in competitive selection time constants for the same training fragment as above, can be seen in figure 5. As can be seen, the peak around the average correlation value of 0.3 remains in the case with small time constants, but is significantly smaller. This peak is caused by well-trained neurons, representing clearly distinct categories. In addition to that, in the low time constants case, a high peak at an average correlation value of around 0.8 can be seen, caused by barely trained neurons, with connections strengths all still very close to their base values, and thus all representing very similar patterns. Thus, as was to be expected, the frequently occurring background sounds are still well-trained in the small time constants case, causing the peak at a correlation value of 0.3, whereas the less frequently occurring, but more salient sounds, are not well-trained anymore, causing an extra peak at a correlation value of 0.8.

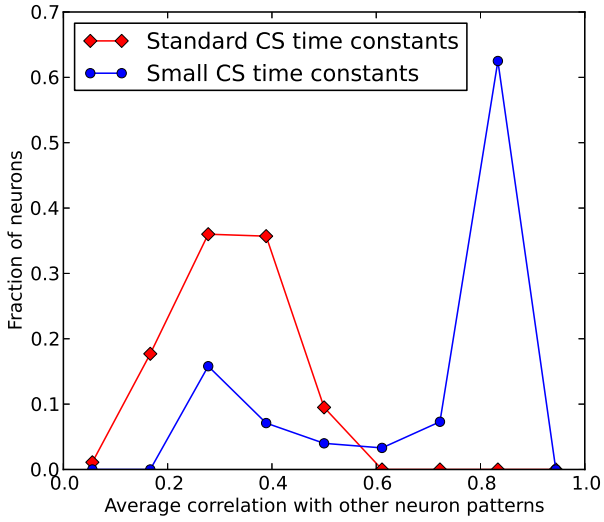

Fig. 5. Distribution of neurons in ϕ layer according to the average correlation between its trained pattern and those of the other neurons in the layer. This is given for the case of standard competitive selection time constants: $\tau_\phi = 2$s, $\tau_\chi = 0.2$s and $\tau_\psi = 0.02$s (red graph with diamond markers), and for the case of ten times smaller, and thus almost immediate, time constants (blue graph with circular markers).

Another property of the model which plays an important role in long-term learning is the process of forgetting. As can be seen in Eq. 7 this is implemented as a slow convergence of synapse strengths toward their base level, at a rate proportional to a random factor and the difference between current synapse strength and the base level. Thus, as long as a neuron is not activated, its trained pattern will slowly become less pronounced, but never disappear completely, as convergence will become slower as the connection strengths approach their base level. While the pattern will never disappear, the neuron will still become progressively more sensitive to input not completely matching the learned pattern, as synapses that are of no importance to the learned pattern become stronger and important ones become weaker while converging to their base level. Thus, once a neuron has forgotten sufficiently, it can be activated by a different input pattern and consequently be trained on this pattern, and the more the neuron has forgotten, the more different the new pattern can be to the original. The forgetting behavior of a single neuron is shown in Fig. 6. The first graph shows a measure for the level of training of the neuron: the normalized difference between synapse strength and its base level ($= (w_{ji}^{YX} - B)/(1-B)$ if $w_{ji}^{YX} > B$, and $= (B - w_{ji}^{YX})/(B-1)$ otherwise), averaged over all synapses leading to the neuron. A well-trained neuron which has mostly connection strengths close to their maximum value and to their minimum value (thus forming a very pronounced pattern) will have a 'level of training' close to 1, whereas a badly trained neuron with connection strengths around base level will have a 'level of training' close to 0. The second graph shows the average strength of all synapses leading to the neuron, which is a good measure for the level of generalization in the recognition of the pattern: a higher average synapse strength causes the neuron to be more easily excited, even by input not completely matching its learned pattern. The third graph, finally, shows the average neuron activation over a one minute timespan. All three graphs are plotted as a function of time, expressed in minutes. As can be seen, activation of the neuron causes the level of learning to rise, and the level of generalization to decrease, while in periods without activation, the level of learning slowly decreases and the level of generalization slowly increases, completely in line with the properties of forgetting as explained above.

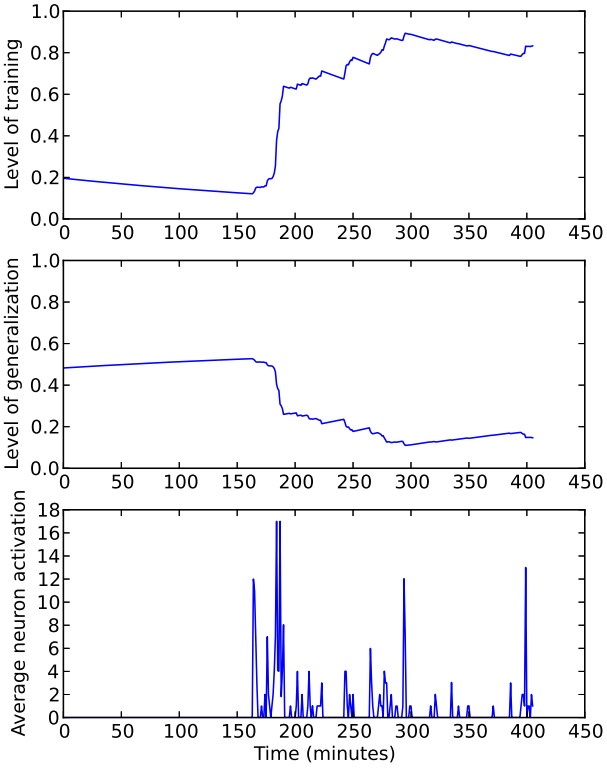

Fig. 6. The level of training (above), the level of generalization (middle) and the average activation over a one minute timespan (below) of a neuron as a function of time, expressed in minutes.

IV. Conclusions

In this paper, an artificial neural network model for sound recognition, aimed at long-term noise monitoring, is presented. It consists of three neural layers, connected by feedforward and feedback excitatory connections. It is shown to include auditory attention mechanisms, emerging naturally from the biologically inspired implementation of neural dynamics and intra-layer connections. In this work special focus is given to the long-term learning properties of the model. It is found that strong attention-based learning plays an important role in guiding the system to use its memory space optimally. Also short-term memory, as it is implemented in the model, is seen to play an important role in long-term learning, as it has profound influence on the coarseness of the learned categories. Finally, also forgetting is shown to be indispensible in long-term learning, as it enables neurons representing sounds that do not occur any more to be used to learn new patterns. The above effects are studied and demonstrated by training the model on real recordings in urban park environments.

References

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts, USA: The MIT Press, 1994.

[2] R. P. Carlyon, R. Cusack, J. M. Foxton, and I. H. Robertson, "Effects of attention and unilateral neglect on auditory stream segregation," *J. Exp. Psychol.-Hum. Percept. Perform.*, vol. 27, no. 1, pp. 115–127, 2001.

[3] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLoS Biol.*, vol. 7, no. 6, p. e1000129, 2009.

[4] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention focusing the searchlight on sound," *Curr. Opin. Neurobiol.*, vol. 17, no. 4, pp. 437–455, 2007.

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

[6] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "Attention-driven auditory stream segregation using a SOM coupled with an excitatory-inhibitory ANN," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, Brisbane, Australia, 2012.

[7] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, Dallas, TX, USA, 2013.

[8] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends Cogn Sci*, vol. 3, no. 4, pp. 128–135, 1999.

[9] S. Grossberg, Ed., *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition and Motor Control*. Boston, MA, USA: Reidel, 1982.

[10] D. Botteldooren and B. De Coensel, "A model for long-term environmental sound detection," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, Hong Kong, 2008, pp. 2017–2023.

[11] B. De Coensel, D. Botteldooren, T. De Muer, B. Berglund, M. 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.

[12] D. Oldoni, B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem, and D. Botteldooren, "A computational model of auditory attention for use in soundscape research," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 852–861, 2013.

[13] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*, M. R. Schroeder, Ed. Berlin: Springer, 1999.

[14] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1941–1944.

[15] N. S. Simons-Weidenmaier, M. Weber, C. F. Plappert, P. K. D. Pilz, and S. Schmid, "Synaptic depression and short-term habituation are located in the sensory part of the mammalian startle pathway," *BMC Neuroscience*, vol. 7, pp. 38–38, 2006.

[16] M. A. Arbib, Ed., *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 1998.