A Biologically Inspired Recurrent Neural Network for Sound Source Recognition incorporating Auditory Attention

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

Abstract—In this paper, a human-mimicking model for sound source recognition is presented. It consists of an artificial neural network with three neuron layers (input, middle and output) that are connected by feedback connections between the output and middle layer, on top of feedforward connections from the input to middle and middle to output layers. Learning is accomplished by the model following the Hebb principle, dictating that "cells that fire together, wire together", with some important alterations, compared to standard Hebbian learning, in order to prevent the model from forgetting previously learned patterns, when learning new ones. In addition, short-term memory is introduced into the model in order to facilitate and guide learning of neuronal synapses (long-term memory). As auditory attention is an essential part of human auditory scene analysis (ASA), it is also indispensable in any computational model mimicking it, and it is shown that different auditory attention mechanism naturally emerge from the neuronal behaviour as implemented in the model described in this paper. The learning behavior of the model is further investigated in the context of an urban sonic environment, and the importance of short-term memory in this process is demonstrated. Finally, the effectiveness of the model is evaluated by comparing model output on presented sound recordings to a human expert listeners evaluation of the same fragments.

I. INTRODUCTION

HUMAN brains excell at the process of auditory scene analysis (ASA). ASA involves perceiving an acoustic environment, analyzing it, decomposing complex sound mixtures into auditory streams originating from a single sound source and assigning a meaning to these sound streams. The human brain uses different auditory cues to accomplish this, and also visual and other sensory cues play a role in this process [1]. Auditory attention mechanisms are found to be of great importance in human auditory perception, and even essential in the process of auditory stream segregation [2][3][4]. By means of competitive selection it selects a single auditory stream for entrance into working memory, where it is further analyzed, a meaning is assigned to it, and it is used to form a mental image of the (acoustic) environment [5]. Thus, auditory attention plays a vital role in auditory stream segregation, sound recognition and human ASA in general and therefore it is also indispensable in any model for computational ASA.

The model described in this paper can be seen as a generalization of and expansion on previous models by the same authors, and as such, it incorporates similar concepts and submodels[6][7]. It has a structure consisting of 3 neural layers, connected to each other by feedforward excitatory connections as well as feedback excitatory connections. In each of the layers, K-winner-takes-all processes, representing competitive selection as a consequence of both excitatory and inhibitory connections within the layer, and normalization processes, ensuring stability of the model, take place. Training of the synaptic connection weights between neurons in the different layers is done following Hebbian theory, with some significant differences to normal Hebbian learning, inspired by neurological findings on human brain development. Also, long-term and short-term synaptic plasticity are implemented, modelling respectively short-term memory and inhibition-of-return. As will be shown, the model includes important (auditory) attention mechanisms such as saliency-driven bottom-up attention, volitional top-down attention, inhibition-of-return and competitive selection in a very natural way.

As the human brain is a highly complicated structure, continuously processing enormous amounts of data, it is inevitable that compromises have to be made between biological accuracy and computational efficiency of any model trying to describe it. The model presented in the current paper is designed to be used in a large-scale sound monitoring network, consisting of a high number of measurement nodes. The final goal of the model within this network would be to detect and classify sound events that would be noticed by a human listener, within the assessment of potential long-term effects of exposure to environmental sound on quality of life [8]. Thus, the model is designed to be used on low-end hardware, running continuously for long periods of time. As a consequence of this hardware requirement, strong simplifications of the biological reality have been made. Nevertheless, the structure of the model, and the way its different components interact have all been based on available knowledge of the human auditory brain.

The remainder of this paper is structured as follows. In the next section, the model is described in detail: the neural architecture is discussed, the learning mechanisms are presented and an overview of how the different auditory attention mechanisms are incorporated into the model is given. In section III the behavior of the model is illustrated and results describing its accuracy are presented. Finally, in section IV, conclusions are presented.

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 Foundation Flanders (FWO Vlaanderen); the support of this organization is gratefully acknowledged.
II. MODEL DESCRIPTION

A. Network Architecture and Neural Activation

In Fig. 1 an overview of the structure of the model as described in the current paper is shown. In its essence, it is an artificial neural network, consisting of 3 neural layers: an input layer (I) that encodes input sound features, a hidden layer in the middle (M), and an output layer (O) in which each neuron represents a concept, or label, that can be attributed to the input sound. Feedforward excitatory connections are present from the input layer to the middle layer and from there on to the output layer. Feedback excitatory connections are present from the output layer to the middle layer, with a delay of one timestep ($\Delta t = 0.1s$). The middle layer has important similarities with Self-Organizing Maps (SOM) employed in previous work by the authors, as it also serves to categorize different types of sounds, based on prototypical sounds encoded in the synaptic weights from the input to middle layer [7]. As will be shown further on in this section, also neuronal behaviour in the middle layer of the current model is not unlike mechanisms employed in classical SOM training. Because of these similarities, a hexagonal topology is chosen for the middle layer, as is customary in SOMs [9].

Initially, features are extracted from the input sound signal following the method described in [10]: based on 1/3-octave band spectra of the input sound, calculated with a temporal resolution of 0.1s, a simplified cochleagram, covering the complete frequency range of human hearing (0-24 Bark with a resolution of 0.5 Bark, resulting in 48 spectral values), is obtained, taking into account energetic masking by use of the Zwicker loudness model [11]. For each timestep and each spectral value, Gaussian and difference-of-Gaussian filters with different scales are convolved with the cochleagram, resulting in features encoding absolute intensity (4 differently scaled Gaussian filters) and spectral and temporal contrast (both 6 differently scaled difference-of-Gaussian filters). Thus, per timestep, a feature vector is obtained containing $48 \times (4 + 6 + 6) = 768$ elements. These elements then serve as excitatory input to the first neural layer in the model, each element representing the excitation of a single neuron in this layer. Note that these features are strongly related to measures for auditory saliency as calculated in [10] and [12], and, as shown in section II-C this plays a major role in the auditory attention system in the model described in this paper.

Firstly, these excitation inputs are normalized and a saturation function is applied in order to limit their values to the interval $[0,1]$. The normalization factor is calculated as a leaky integral, slowly ($\tau_v = 100s$) following the maximum excitation value:

$$\nu^I(t) = \alpha \nu^I(t - \Delta t) + (1 - \alpha) \max_i(E^I_i(t)),$$

where $\nu^I(t)$ is the normalization factor of the input neural layer at time $t$, $E^I_i(t)$ is the $i$th element of the 768-dimensional vector containing the excitatory inputs to the neurons at time $t$ and $\alpha = e^{-\Delta t/\tau_v}$. Dividing $E^I_i(t)$ by $\nu^I(t)$ for all $i$ attempts to keep its values within bounds, without completely eliminating the variation in time of the overall excitation strength, which is a highly important property for attention mechanisms in the model, as explained in section II-C. However, it does not guarantee values between 0 and 1. In order to achieve this, a saturation function is applied:

$$E'^I_i(t) = \frac{E^I_i(t)/\nu_I(t)}{1 + E^I_i(t)/\nu_I(t)}$$

(2)

for all $i$.

Next, a K-winner-takes-all mechanism is applied, leaving only the most strongly excited neurons to be activated [13]. This mechanism simulates competition between the different neurons in the layer by means of internal excitation and inhibition effects in a highly simplified manner, compared to more detailed approaches in [6][7]. Neuronal activation is then calculated as follows:

$$A^I_i(t) = \begin{cases} \max(E'^I_i(t)) & \text{if } E'^I_i(t) > e_K \\ 0 & \text{if } E'^I_i(t) < e_K \\ \frac{(E'^I_i(t) - e_K)}{\max(E'^I_i(t)) - e_K} & \text{otherwise} \end{cases}$$

(3)
where $E_{j}^{M}(t)$ is defined as before, and $e_{K}$ is the $K$'th largest value of $E_{j}^{M}(t)$ for all $i$. As will be further clarified in section II-C, this $K$-winner-takes-all mechanism plays an important role in the implementation of auditory attention into the current model.

As can be seen in Fig. 1, neural connections are present from the input layer to the middle layer. Excitatory input to the middle layer originating from the input layer is calculated as follows:

$$E_{j}^{MI}(t) = \sum_{i} w_{ji}^{MI} A_{i}(t),$$

(4)

where $w_{ji}^{MI}$ is the connection weight of the neural synapse between the $i$th neuron of the input layer and the $j$th neuron of the middle layer. The middle layer, however, does not only receive excitatory input from the input layer, but also from the output layer (with a time delay of $\Delta t$), which is calculated similarly:

$$E_{j}^{MO}(t) = \sum_{k} w_{jk}^{MO} A_{k}(t - \Delta t),$$

(5)

with naming conventions similar to those above. In order to combine these, and calculate one activation value per middle layer neuron first both excitation patterns are normalized and saturated following the exact same method as used for the excitatory pattern of the input layer. These normalized patterns are then linearly combined into a total normalized excitation pattern as follows:

$$E_{j}^{M}(t) = \gamma E_{j}^{MI}(t) + (1 - \gamma) E_{j}^{MO}(t),$$

(6)

where $\gamma$ is a fixed combination parameter (for the results in this work, $\gamma = 0.7$ is used). Finally, in order to calculate the activation of the middle layer neurons, a global inhibition term is applied, calculated in the same way as explained above, in order to select the most strongly excited neurons. In addition, as the well-defined topology of the middle layer allows for it, extra activation is added to each neuron, proportional to the activation of its neighbors in the hexagonal grid. This process is similar to the local excitation in [7] and serves to achieve clustering in the training of neural connections to the middle layer, resembling the way this is done in classical SOM training [9].

Finally, the process to calculate the excitation and activation values of the neurons in the output layer is very similar to the processes described above. Excitation is calculated as a sum over all middle layer neural activations, weighed with the corresponding synapse strengths $w_{ji}^{OM}$, these values are normalized and saturated, and finally, by means of a global inhibition term, the most strongly excited neurons are activated.

B. Synaptic Strengths, Learning and Memory

Learning of the synaptic connection weights in the model is done following the Hebb principle: "Cells that fire together, wire together.". However, there are some significant differences to traditional Hebbian learning, attempting to simulate human learning more accurately. Initially, all connection weights are set to their maximum value of 1, reflecting the overabundance of synaptic connections as a result of progressive growth in the early phases of human brain development [14]. During development, useless connections in the human brain are gradually eliminated in a process called synaptic pruning, leaving only well-used connections, resulting in a more precise and mature circuitry [15]. This is modelled by decreasing the synaptic strengths of connections from activated neurons to non-activated neurons (thus, useless connections) after each timestep as follows:

$$w_{ji}^{YX} = (1 - \eta A_{i}^{X}) w_{ji}^{YX},$$

(7)

in which $\eta$ is a constant, defining the pruning rate (in this work, $\eta = 0.02$ is used), $X$ and $Y$ are two different neural layers, $w_{ji}^{YX}$ and $w_{ji}^{YX}$ are the connection strengths from the $i$th neuron of layer $X$ to the $j$th neuron of layer $Y$ respectively before and after pruning, and $A_{i}^{X}$ is the activation of the $i$th neuron of layer $X$. Thus, the more strongly the neuron in layer $X$ is activated, the faster its connections to inactive neurons will be reduced.

While the growth of new synapses (synaptogenesis) slows down after early brain development phases in humans, it does not stop. This is reflected in the model by slowly and randomly increasing the strength of all connections at each timestep as follows:

$$w_{ji}^{YX} = w_{ji}^{YX} + \zeta R(1 - w_{ji}^{YX}),$$

(8)

in which $\zeta$ is a constant describing the synaptogenesis rate (here $\zeta = 10^{-5}$ is used), $R$ is a (pseudo-)random number between 0 and 1, and $w_{ji}^{YX}$ and $w_{ji}^{YX}$ are the connection strengths from the $i$th neuron of layer $X$ to the $j$th neuron of layer $Y$ respectively before and after synaptogenesis. The factor $1 - w_{ji}^{YX}$ limits connection strengthening at its maximum value of 1, thus modelling the neurological process of synaptic scaling, which limits the total synaptic input to a neuron in the human brain, in a highly simplified way.

Because the synaptic pruning rate is higher than the synaptogenesis rate, the total strength of neural connections in the network will decrease during learning, just like the total number of synapses in a human brain decreases during development [14]. The advantage of this method of training the neural network, as compared to traditional Hebbian learning, is its relatively high resistance to forgetting. Once a pattern has been learned, it is difficult to change this into a different pattern, as this would involve synaptogenesis, which is relatively slow. Neurons that are not yet trained to recognise a specific pattern, on the other hand, still have maximum strength connections, and only need relatively fast synaptic pruning to learn. This process can best be compared to carving a statue out of stone: when starting from a raw stone block, any shape can be reached relatively easily, but once a particular shape been carved from the stone, it is very difficult to change this into another.

To guide learning in the model, a teacher can be implemented as additional excitatory input for the output layer. Neurons that are chosen to represent concepts which are
considered to be related to the input sound at the corresponding timestep can be given extra excitatory input, and thus a higher chance to be activated (e.g. when an input sound in which birds can be heard is presented to the model, the teacher will add an extra excitatory input to the concept neurons representing 'birds' and 'nature sound'). This excitatory input can be seen as coming from working memory, where the human brain temporarily stores information about its environment, originating from different senses (visual information, smell, speech of a human teacher, ...). Thus, if a concept is frequently present in working memory when a certain sound is heard, it is highly likely that this concept is linked to the sound, and an excitatory input is sent from working memory to the corresponding concept layer neurons. This, in turn, will cause the learning mechanism to enhance connections between these neurons and middle layer neurons associated with the input sounds, following Hebb’s principle. Real acoustic environments are mixtures consisting of different sounds originating from more than one source, and it is almost impossible for a human teacher to select the exact timesteps in which each source is dominant. Thus, it is inevitable that at certain timesteps, the teacher will give additional excitatory input to output layer neurons that do not describe the input sound during the same timestep. For instance, when considering a sound fragment in which footsteps can be heard, the output layer neuron associated to footsteps will receive extra excitatory input from the teacher, but during a number of timesteps in between two footsteps, the sound that is actually present is not footstep sound, but background sound. To make sure this output neuron is not trained on both footstep and background sound, but only on footstep sound, the excitatory input from the teacher needs to be kept relatively small compared to the excitatory input originating from the middle layer. This way, only when doubt arises between different concepts, the teacher will determine which neurons are finally activated. In the initial phase of learning, this will always be the case, as the model did not learn anything yet, but as learning continues, more and more sounds will be clearly recognised, without external help, and the relative influence of the teacher will decrease.

Long-term depression (LTD) and long-term potentiation (LTP) are important forms of synaptic plasticity, lasting for minutes of more, that are thought to be related to short-term memory [16][17]. These effects affect synaptic connection weights temporarily, whereas synaptic pruning and synaptogenesis are permanent modifications to these weights. In the current model LTD and LTP are implemented in a highly simplified manner as follows: LTD occurs in a synapse when the two neurons it connects repeatedly fire together, while LTD occurs in a connection when its destination neuron is repeatedly activated, while its origin neuron is not. This translates into the following mathematical rule:

\[
L^Y_{ji}(t) = \begin{cases} 
\xi_{LTP}LT^Y_{ji} + (1 - \xi_{LTP})M_{LTP} & A^X(t) > 0 \land A^Y(t) > 0 \\
\xi_{LTD}LT^Y_{ji} & A^X(t) = 0 \land A^Y(t) > 0 \\
\xi_{LTD reset}LT^Y_{ji} + (1 - \xi_{LTD reset}) & \text{else}
\end{cases}
\]

where \(LT^Y_{ji}\) and \(LT^Y_{ji}\) are synapse strength modification factors for the synapse between the \(i\)th neuron in layer \(X\) and the \(j\)th neuron in layer \(Y\), describing long-term synaptic plasticity, respectively before and after updating its values, \(\xi_{LTP} = \exp(-\Delta t/\tau_{LTP}), \xi_{LTD} = \exp(-\Delta t/\tau_{LTD}), \xi_{LTD reset} = \exp(-\Delta t/\tau_{LTD reset})\) and \(M_{LTP}\) a parameter describing the maximal value for \(LT^Y_{ji}\) (in this work \(M_{LTP} = 1.5\) is used). Time constants \(\tau_{LTP}\) and \(\tau_{LTD}\) are given smaller values than \(\tau_{LTD reset}\), as this enables the long-term synaptic plasticity effect to build up in a limited amount of time compared to the time during which the effect lasts (in this work \(\tau_{LTP} = 10s\) and \(\tau_{LTD reset} = 100s\) are used). In order for the mechanism to take effect in the model, \(w^Y_{ji}LT^Y_{ji}\) are used as connection weights when calculating neural excitations, instead of just \(w^Y_{ji}\). Thus, this effect can indeed be seen as short-term memory: it temporarily strengthens "good" connections, and weakens "bad" connections, and when these connections are used during this period (if the same concept is repeated) it enhances the learning effect on these synapses, and the information is thus transferred into long-term memory (permanently altered connection weights).

A last form of synaptic plasticity that is taken into account in the model is short-term synaptic depression, or synaptic fatigue. This involves temporary (seconds to tens of seconds) inhibition of synapses, due to persistent stimulation [18]. It is implemented into the model as follows:

\[
F^Y_{ji} = \begin{cases} 
\phi_{SF}F^{YX}_{ji} & A^X(t) > 0 \\
\phi_{F reset}F^{YX}_{ji} + (1 - \phi_{F reset}) & \text{else}
\end{cases}
\]

where \(F^Y_{ji}\) and \(F^Y_{ji}\) are connection weight modification factors for synaptic fatigue, \(\phi_{SF} = \exp(-\Delta t/\tau_{SF})\) and \(\phi_{F reset} = \exp(-\Delta t/\tau_{F reset})\), similarly to naming conventions for long-term synaptic plasticity. Also similarly to the case of long-term synaptic plasticity, \(\tau_{F reset}\) is chosen to be larger than \(\tau_{SF}\) (in this work \(\tau_{SF} = 1s\) and \(\tau_{F reset} = 10s\), and \(w^Y_{ji}LT^Y_{ji}F^Y_{ji}\) are used as connection weights when calculating neural excitations). As will be shown in the next section, this effect plays an important role in the models attention mechanism.

C. Attention Mechanisms

Most theories of human attention include the interplay of bottom-up, saliency-based attention and top-down, volitional
attention, combined with the mechanism of competitive selection [4] [1]. In addition, often, the concept of inhibition-of-return is introduced, in order to prevent attention from permanently staying focussed on one single item [10]. In contrast to introducing these mechanisms by artificially adding extra parameters or submodels to existing models [7], in the current model, they emerge naturally from the way in which biological neural behavior has been implemented.

As noted in section II-A, the features used as input to the model are the same as the ones used in [10], and very similar to the ones in [12] to calculate auditory saliency maps. The calculation of these saliency maps further involves scaling of the feature values to the $[0, 1]$ interval, and a mechanism of internal excitation and inhibition to simulate competition between different salient features. The similarity between these two mechanisms and respectively the normalization procedure and K-winner-takes-all mechanism as employed in this work, ensures that total activation of the input layer can be interpreted as a measure for auditory saliency of the incoming sound. Furthermore, the linearity in the calculation of excitation of both the middle and output layers, combined with the way in which the normalization and K-winner-takes-all mechanisms are implemented, ensure that changes in saliency values of the input sound will be reflected in changes in activation intensity in all neural layers of the system, thus reflecting the effect of bottom-up, saliency-driven attention.

Top-down, voluntary attention can be modelled as a biasing, extra excitatory input to the output layer, very similar to the way a teacher is implemented in the learning phase of the model. This will increase excitation of neurons in the output layer, connected to certain concepts deemed interesting by higher level brain functions, and thanks to feedback connections this will also be reflected in the middle layer. Thus, activation of neurons connected to the concept of interest will be facilitated and increased in strength. Next, the K-winner-takes-all mechanism will select the most strongly excited neurons to be activated, thus modelling competitive selection. The attention mechanism of inhibition-of-return, finally, is incorporated in the model by means of synaptic fatigue, effectively preventing neurons from permanently staying activated, and thus also preventing attention from staying focussed on one single concept.

### III. Simulation Results

In order to test the model, several hours of recordings are made in a typical urban environment, containing mainly traffic noise (cars, busses and trucks passing by, both at close distance and further away), the chirping sound of some birds and human speech from passing pedestrians and yelling children. For learning and evaluating purposes, 302 fragments from these recordings, with an average duration of 6s are randomly selected and labeled by a human expert listener. This labeling, however, is not done in a structured, regulated way, but in an as natural as possible way: the expert listener is allowed to use any labels he judges to be fitting to the sound fragment, and is not limited to a certain list of possible labels. Because of this, the total number of labels is relatively high (99 labels for 302 sound fragments) and for each label, a number of synonyms, or labels with a highly similar meaning, will be included (for instance: car, truck, accelerating car, engine sound, traffic sound, ...).

From these 302 labeled urban environment sound fragments, 202 are randomly selected to be used for training, and the other 100 are used for evaluating the performance of the model. Training is done as follows: the 202 training sounds are processed and used as input for the model, while, as described in section II-B, extra excitatory input is given to the concepts neurons corresponding to the labels given to the sound fragment by the expert listener. In order to ensure
Fig. 4. Excitation pattern of the concept layer (without excitation from the teacher) as a function of time in a late learning phase. A limited number of neurons is activated significantly more strongly than average.

Fig. 5. Activation pattern of the concept layer as a function of time in a late learning phase, corresponding to the excitation pattern shown in Fig. 4. The teacher’s excitatory input is on the same neurons as in Fig. 3, but the trained model’s neural excitations are decisive for the activation pattern.

At each timestep, most neurons have comparable excitation values, and thus, the choice of which neurons to activate will be decided by the excitation of the concept neurons for ‘birds’ and ‘nature sounds’ given by the teacher. Indeed, in the final neural activation pattern for this fragment, as seen in Fig. 3, the concept neuron representing ‘nature sounds’ is dominant from 0.1s to 0.3s, and the concept neuron representing ‘birds’ is dominant from 0.4s until the end of the fragment. The second time this fragment is used, though, network learning is already in an advanced phase, as can be seen in the neural excitation pattern shown in Fig. 4. Two concepts are clearly excited more strongly than the others: ‘birds’ and ‘low frequency background noise’. In this case, the network recognizes the input sound well, no doubt arises, and the influence of the teachers extra excitation is limited, as can be concluded from the fragment’s final activation pattern, as shown in Fig. 5.

In a developing human brain, a decrease in the total number of synapses is observed, and, as the learning mechanism described above aims to mimic its biological example, similar behavior is expected during training of the model. In Fig. 6, the evolution of the sum of all connection weights from the middle to the output layer and vice versa is shown, in two training cases: one in which short-term memory was included in the model, and one without short-term memory. Indeed, in both cases, an initially strong decrease is seen, followed by a saturation phase in which the decrease rate stagnates. In the case without short-term memory, however, this decrease is significantly less than otherwise, indicating that the model’s neural circuitry develops more slowly and less thoroughly. In Fig. 7 the evolution of sum of all connection weight changes per timestep is shown on a logarithmic scale, for the same two cases as before, and a clear learning convergence can be observed for both cases.

Finally, in order to evaluate the model’s sound source recognition capability, the 100 labeled sounds not used for training are given as input to the model, this time without the teachers excitatory input to the output layer neurons. This way, for each timestep, output layer neuron activations are obtained and for each fragment, these activations are summed over all timesteps and then normalized in such a way that their sum becomes 1. Next, a labeling threshold is set, and if the summed and normalized activation value of an output layer neuron exceeds this threshold, its corresponding label is attributed to the sound fragment. Similarly, a threshold is set to decide on the absence of a certain sound source or the absence of attention for it. In order to eliminate disturbing influence of synonyms or labels with similar meanings, first all labels are manually grouped into four categories: ‘traffic’, grouping all traffic sounds, ‘nature’, mainly consisting of bird chirping sounds, ‘human’, containing human speech and shouting children and ‘other’, in which all labels not belonging to one of the other categories are grouped. By comparing the category of the labels given to each of the sound fragments by the model with those given by the expert listener, a measure for the models sound source recognition
quality can be calculated. More specifically, the probability for the expert listener to judge a sound to belong to a certain category in case the model attributed labels from this category to the sound is calculated (true positives), and similarly, also the probability for the human listener not to mention a certain category in case the model did not, is calculated (true negatives). The results of these calculations as a function of threshold value are shown in Fig. 8 both with and without short-term memory included in the model, for all categories except for ‘other’, as this category does not contain a consistent set of similar labels, but just a collection of all labels that do not belong to the other categories. It can be seen that the true positive rate is between 0.6 and 0.7 for all categories and with a decision threshold between 0.35 and 0.45, in case short-term memory is included, while these values become significantly lower for the ‘nature’ and ‘human’ categories in case short term memory is disabled. This demonstrates the positive effect of short-term memory on learning, certainly for input stimuli that occur less frequently, considering that for traffic sound, that is nearly always present in the recordings, no significant difference is seen. On the other hand, no significant difference between the two cases is observed in the true negative rates. In both cases, values between 0.8 and 0.9 are obtained for nature sounds, between 0.6 and 0.8 for human sounds and between 0.4 and 0.6 for traffic noise. A possible reason for these relatively low scores for traffic noise true negatives is that the model tends to focus more on salient sounds such as bird chirps or human speech, compared to less salient traffic background, whereas the human listener did mention traffic sounds in almost all fragments.

Fig. 7. The evolution of the total change in neural connection weight strength per timestep as a function of the number of training timesteps. The red graph is in the case without short-term memory, the blue graph is with the inclusion of short-term memory.

IV. CONCLUSIONS

In this paper, an artificial neural network model for auditory sound source recognition, inspired by the human brain, is presented. It consists of three neural layers (input layer, middle layer and output layer), connected to each other by feedforward connections, and in addition, also feedback connections between output and middle layer are present. Neural behavior is modelled in an as biologically accurate as possible, but still highly simplified way, as the model is designed to run on low-cost hardware. It is shown that auditory attention effects emerge naturally from the neural behavior as implemented, without the need to artificially introduce an extra attention system outside the source recognition model. A Hebbian-like learning algorithm is used to establish the connection weights, with some significant differences as compared to standard Hebbian learning, in order to prevent the model from forgetting known patterns while learning new ones. In addition, long-term synaptic plasticity effects are included into the model, and it is shown that these can be seen as an implementation of short-term memory.

In order to test the model, it is first trained on a number of human-labeled sounds. The influence of the teacher in learning is shown to decrease as the model’s neural circuits mature, and the model learns to recognise the presented sounds without the help of the teacher. Also, a decrease in total connection weight strength during learning is observed, in accordance to a decrease in number of synapses seen in a developing human brain. Finally, a measure for the quality
of the sound recognition done by the model is calculated, and the positive influence of the short-term memory effects included in the model on learning is demonstrated.

REFERENCES


