A neural network approach to auditory streaming

Ulrike Dicke*
Department of Computer Science,
Trinity College Dublin, Ireland.

Abstract

This paper suggests a neural network architecture for the streaming of an acoustic input using fundamental grouping cues of the mammalian auditory system. The auditory periphery is simulated using biologically relevant models, and the primary auditory cortex is represented by a neural network layer, in which three different features of the input are mapped and streamed. Finally, an attentional searchlight that is driven by the cortical network neurons, focuses on a selected cortical area representing a listener who focuses their attention on a distinct acoustic input.

1 Introduction

The biological auditory system provides the best example for developing a model that performs auditory scene analysis since no man–made system offers a comparable solution for this complex task. The one-dimensional input received by the auditory system via the tympanic membrane vibration contains all available information about the acoustic environment. In order to extract a distinct auditory stream, for example the voice of a soprano while she is accompanied by an orchestra, the auditory system performs a number of complex processing steps, in which the input is decomposed and transmitted into the primary auditory cortex (AI-cortex) where it is represented via a distributed pattern of neural activity. Inside the AI-cortex the components of the decomposed input have to be grouped again. This grouping is called *auditory scene analysis* [1]. The model that is presented in this paper uses biologically motivated models, describing the peripheral preprocessing

*email: Ulrike.Dicke@cs.tcd.ie

of an acoustic input and a neural network layer to represent the AI-cortex in which the input components are mapped. An attentional searchlight, that is driven by the cortical network neurons, simulates auditory scene analysis by following the input stream.

2 The Mammalian Auditory System

A sound entering the auditory system via the outer ear is transmitted into the inner ear, where it is transformed into a wave propagating through the fluid filled cochlea. The cochlea of the inner ear is bound by the basilar membrane on which the hair cells of the inner ear are situated. An individual input frequency leads to resonance motions at a certain cochlea location and causes the hair cells situated at this location to bend. This mechanism results in a mechanical frequency decomposition of the acoustic input. The hair cells of the inner ear are innervated by a number of auditory nerve fibers, which are activated by the neurotransmitters a vibrating hair cell releases. At the synaptic gap between hair cells and auditory nerve fibers the input signal is transduced into neural signals which are transmitted via a chain of processing centers into the AI–cortex. The tonotopical representation of frequencies along the basilar membrane is preserved along the auditory pathway into the AI–cortex, where a number of frequency representations have been found [5]. The cortical frequency representations consist of tonotopically ordered neurons that show a maximum response to their individual best frequency (BF). The cortical best frequency neurons form isofrequency panels; however it is not clear so far which different features of an acoustic input are encoded by different neurons of one isofrequency panel [5]. For further details on the biology see for example [2].

3 Auditory Scene Analysis

Auditory scene analysis describes the grouping of the components of an acoustic input in a way that a meaningful representation of the acoustic environment can be achieved and has been extensively studied by Al Bregman [1]. Following a number of psychophysical experiments Bregman suggests a number of competing grouping cues to decide whether individual components of an auditory input are originating from the same sound source or not. The auditory system uses at least 12 competing grouping cues, including proximity of frequency and intensity, common amplitude modulation, harmonicity, common onset and offsets, binaural cues etc. Although the proposed model implements only three of these grouping cues, namely frequency proximity, onsets and offsets, it is already enabling an attentional searchlight, to follow a selected input stream.

4 The Model

The biological system and the problem of auditory scene analysis have been introduced in the previous sections to demonstrates the motivation for the architecture and mechanisms of the presented model. The input shown in figure 1 is preprocessed in a first simulation step using a Gammatone filter bank and the Meddis hair cell model. The output of the hair cell model then serves as input to a cortical network layer, which is concerned with auditory streaming.

4.1 The Auditory Periphery

The Gammatone Filter Bank. The cochlea response to an acoustic input is simulated using a fourth order Gammatone filter bank, which was suggested as a biologically relevant cochlea model by Patterson and Holdsworth [4]. The filters are spaced according to [6] and the bandwidth of each filter is determined following the equivalent rectangular bandwidth (ERB), with respect to the so–called critical bandwidth of the cochlea filters measured in psychophysical experiments.

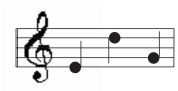

Figure 1: The acoustic input played by a recorder. It was recorded using the sampling frequency $f_s = 8$kHz and had the duration $T = 0.77s$.

The Hair Cell Model. The output of the Gammatone filters serves as input to the Meddis hair cell model [3], which describes the transduction of the mechanical signal into neural signals taking place at the synaptic gap between hair cell and auditory nerve fiber. Each model hair cell contains a pool of neurotransmitters that leak through its membrane into the synaptic gap with a leakage rate proportional to the hair cell bending. As the firing probability of the postsynaptic auditory fiber depends on the quantity of neurotransmitters inside the synaptic gap, it reaches a maximum at locations of maximum hair cell bending (see [3] for further details).

4.2 The Cortical Network Layer

The cortical neural network layer consists of 3 panels of binary neurons with each panel encoding one of the three input features: frequency, onset and offset. The model architecture can be seen in figure 2.

The frequency panel consists of 100 neurons with each neuron receiving its sensory input from the hair cell with the corresponding center frequency and a feedback input from the searchlight

$$h_i^{freq}(t) = \sum_{t_f=t-\tau}^{t} n_i^{hc}(t_f) + w_{max,i} n_{max}^{search}(t). \tag{1}$$

The first term of eq. 1 describes the sensory input to the frequency neuron i. The activity of the corresponding hair cell $n_i^{hc}(t_f)$ is summed up over the time period τ, which was chosen to be 2ms, the time of an excitatory postsynaptic potential (EPSP). The second term is a feedback input coming from the attentional searchlight. The activity of the searchlight $n_{max}^{search}(t)$ focusing on the isofrequency panel max is weighted by a distance dependent Gaussian function $w_{max,i}$. A neuron of the frequency panel fires, $n^{freq}(t+\delta t) = 1$, if its excitation $h^{freq}(t)$ exceeds the firing threshold Θ^{freq}.

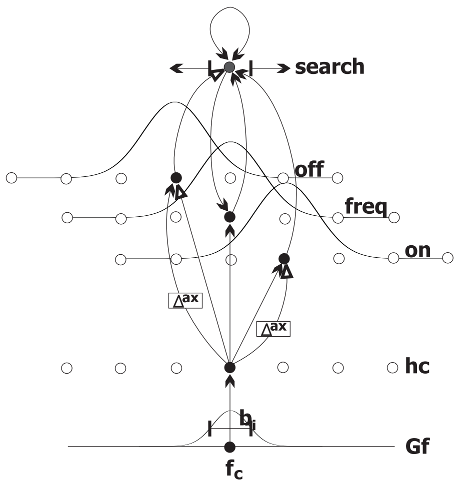

Figure 2: The model architecture. The Gammatone filter with the center frequency f_c provides input to the corresponding hair cell, which gives input to the neurons of the frequency, onset and offset panel (inhibitory connection: open arrows; excitatory connections: black arrows). The attentional searchlight receives inputs from the frequency, the onset and the offset neurons as well as an excitatory feedback from itself and focuses on the isofrequency panel providing maximum excitation.

The onset panel consists of 100 neurons which receive two synaptic inputs from their corresponding hair cells: an excitatory synaptic input arriving immediately from the hair cell and an inhibitory input arriving after an axonal delay Δ^{ax}

$$h_i^{on}(t) = \frac{\sum_{t_f=t-2\text{ms}}^{t} n_i^{hc}(t_f) - \sum_{t_f=t-2\text{ms}}^{t} n_i^{hc}(t_f - \Delta^{ax})}{\sum_{t_f=t-2\text{ms}}^{t} n_i^{hc}(t_f) + \sum_{t_f=t-2\text{ms}}^{t} n_i^{hc}(t_f - \Delta^{ax})}.$$

In order to detect onsets independently from the intensity present in that frequency channel, the synaptic inputs arriving from the hair cell are weighted by the sum of the arriving inputs. The excitation of the onset neurons thus depends only on the relative change in the hair cell input. An onset neuron creates an action potential $n^{on}(t + \delta t) = 1$, if $h^{on}(t) > \Theta^{on}$.

The offset panel consists of 100 neurons which receive a similar input to the onset neurons, with the input, arriving after an axonal delay, being excitatory, while the immediately arriving input is inhibitory

$$h^{off}(t) = \frac{\sum_{t_f=t-2\text{ms}}^{t} n^{hc}(t_f - \Delta^{ax}) - \sum_{t_f=t-2\text{ms}}^{t} n^{hc}(t_f)}{\sum_{t_f=t-2\text{ms}}^{t} n^{hc}(t_f) + \sum_{t_f=t-2\text{ms}}^{t} n^{hc}(t_f - \Delta^{ax})}.$$

The firing condition for an offset neurons is $n^{off}(t + \delta t) = 1$, if $h^{off}(t) > \Theta^{off}$.

The attentional searchlight receives excitatory synaptic inputs from the frequency neurons and the onset neurons. It also an excitatory feedback from itself and an inhibitory synaptic input from the offset neurons

$$h_j^{search}(t) = \sum_i w_{ji} n_i^{freq}(t) + \sum_i w_{ji} n_i^{on}(t) - \sum_i w_{ji} n_i^{off}(t) + \delta(i,j).$$

The searchlight compares the inputs coming from each isofrequency panel i and focuses on the channel providing it with the strongest input max, $n_{max}^{search}(t) = 1$.

5 Results

The recorder input shown in Figure 1 is preprocessed using a Gammatone filter bank consisting of 100 filters with center frequencies between 100 Hz and 4 kHz. The output of the Gammatone filter bank is used as input to 100 model hair cells, which give input to the 3×100 cortical neurons.

Figure 3 shows the output of the network neurons, sampled every 2ms (an approximation of the human temporal auditory resolution), with every active neuron indicated by a dot. The figure on the top left shows the output

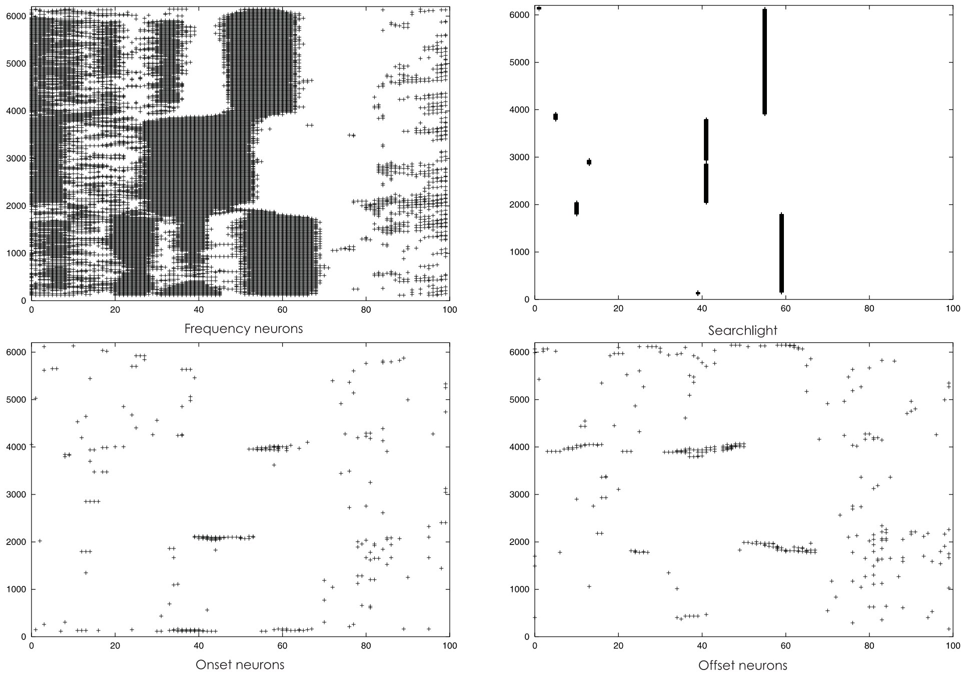

Figure 3: The output of the neuron panels over time.

of the frequency neurons over time, when they are exclusively receiving the hair cell input. In order to maintain the ability to detect low intensity inputs, the firing threshold of the frequency neurons was chosen to be relatively low, which results in a broad activity inside the frequency panels. The recorder input can still easily be distinguished. The figure on the top right shows the output of the searchlight neuron over time which easily follows the three notes, played by the recorder, using the competing inputs coming from the frequency, onset and offset panels. During the second tone, the searchlight switches, which is due to onsets and the following offsets detected in the other channels. The figure on the bottom left shows the onset map and the figure on the bottom right shows the offset map. Despite the noise in these maps it is easy to ascertain both the onset and the offset of each note.

6 Outlook and Conclusion

The objective of this work is to present a first step towards the development of a biologically oriented neural network model which enables the imple-

mentation of auditory grouping cues used for auditory scene analysis. The preprocessing of an acoustic input is performed using a Gammatone filter bank [4] and the Meddis hair cell model [3], which are both considered to be biologically relevant models. The frequency representations found inside the AI-cortex [2, 5] provide the basis for the representation of an acoustic inputs in the presented model. The attentional searchlight is driven by inhibitory and excitatory inputs coming from the cortical neurons. The couplings between the cortical neurons and the searchlight are determined using a Gaussian function, which is biologically plausible as such a form of distance dependent couplings have been verified in various biological experiments. Receiving its input from the cortical neurons the attentional searchlight follows the input stream in a perfect manner using only three out of the 12 suggested grouping cues, used by the mammalian auditory system. The introduced model will be further developed by adding more cortical panels, in which other important features of the auditory input will be represented and grouped. A fourth panel of cortical neurons will be used to map the amplitude modulation. An additional development will be the implementation of a second attentional searchlight, in order to stream two simulataneous auditory inputs.
In concluding it can be said that the proposed relatively simple model has shown a very good performance in following the recorder input.

U. D. is funded by the SMARTII project of the TMR Network, project number FMRX-CT96-0052.

References

[1] A.S. Bregman. *Auditory scene analysis: the perceptual organization of sound.* MIT Press, Cambridge, 1990.

[2] J.P. Kelly. Hearing. In E.R. Kandel, J.H. Schwartz, and T.M. Jessell, editors, *Principles of Neural science.* Prentice-Hall, London, 1991.

[3] R. Meddis, M.J. Hewitt, and T.M. Shackleton. Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse. *J. Acoust. Soc. Am.*, 87:1813–1816, 1990.

[4] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. An efficient auditory filterbank based on the gammatone function. APU report 2341, Cambridge: Applid Psychology Unit, 1988.

[5] S.A. Shamma. Auditory Cortex. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, 1995.

[6] M. Slaney. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Apple Computer Technical Report #35. Apple Computer, Inc., 1993.