Multiple Sound Source Localisation in Reverberant Environments Inspired by the Auditory Midbrain

Jindong Liu[1]*, David Perez-Gonzalez[2], Adrian Rees[2],
Harry Erwin[1], and Stefan Wermter[1]

[1]Dept. of Computing and Technology,
University of Sunderland, Sunderland, SR6 0DD, United Kingdom
[2]Institute of Neuroscience, The Medical School,
Newcastle University, NE2 4HH, United Kingdom
{jindong.liu,harry.erwin,stefan.wermter}@sunderland.ac.uk
{david.perez-gonzalez,adrian.rees}@newcastle.ac.uk
http://www.his.sunderland.ac.uk

Abstract. This paper proposes a spiking neural network (SNN) of the mammalian auditory midbrain to achieve binaural multiple sound source localisation. The network is inspired by neurophysiological studies on the organisation of binaural processing in the medial superior olive (MSO), lateral superior olive (LSO) and the inferior colliculus (IC) to achieve a sharp azimuthal localisation of sound sources over a wide frequency range in a reverberant environment. Three groups of artificial neurons are constructed to represent the neurons in the MSO, LSO and IC that are sensitive to interaural time difference (ITD), interaural level difference (ILD) and azimuth angle respectively. The ITD and ILD cues are combined in the IC to estimate the azimuth direction of a sound source. To deal with echo, we propose an inter-inhibited onset network in the IC, which can extract the azimuth information from the direct path sound and avoid the effects of reverberation. Experiments show that the proposed onset cell network can localise two sound sources efficiently taking into account the room reverberation.

Key words: Spiking neural network, sound localisation, inferior colliculus, reverberation

1 Introduction

Humans and other animals show a remarkable ability to localise multiple sound sources using the disparities in the sound waves received by the ears. For example, humans can localise as many as six concurrent sources [1] and cancel out echoes using two ears [2]. This has inspired researchers to develop new computational auditory models to help understand the biological mechanisms that

* This work is supported by EPSRC (EP/D055466 and EP/D060648)

underlie sound localisation in the brain. Binaural sound localisation systems take advantage of two important cues [3] derived from the sound signals arriving at the ears: (i) interaural time differences (ITD), and (ii) interaural level differences (ILD). Using these two cues sound source direction can be estimated in the horizontal plane.

In humans the ITD cue is effective for localising low frequency sounds (20 Hz ∼1.5 kHz) [4], however, the information it provides becomes ambiguous for frequencies above ∼1 kHz. In contrast, the ILD cue has limited utility for localising sounds below 1.5 kHz, but is more efficient than the ITD cue for frequencies higher than this [4]. The ITD and ILD cues are extracted in the medial and lateral nuclei of the superior olivary complex (MSO and LSO), which project to the inferior colliculus (IC) in the midbrain. In the IC these cues are combined to produce an estimation of the azimuth of the sound [5]. The cells in the IC are classified into 6 main types among which onset and sustained-regular cells play the main role for sound azimuth detection even in an echo environment.

Several hypotheses and models for ITD and ILD processing have been proposed [3][6][7], with one of the most influential being a model advanced by Jeffress [3]. However, all above models only work in an anechoic environment. To deal with reverberation, Litovsky [8] proposed a model of the precedence effect which applies an onset detector to inhibit the localisation cues from the indirect sound path. Palomäki [9] simplified Litovsky's model by using envelope extraction. However, these models did not exploit the biological pathways from the MSO/LSO to the IC, such as the inhibition from the ipsilateral LSO to the IC. These pathways are believed the crucial key for a sharp localisation over broadband frequency.

This paper presents a model designed to identify multiple sound source directions by means of a spiking neural network (SNN). It is the first to employ a SNN that combines both ITD and ILD cues derived from the SOC in a model of the IC to cover a wide frequency range and to target a reverberant environment. This model incorporates biological evidence on the inputs from the MSO and LSO to the IC, and is able to build a sharp spatial representation of sound source. To cope with an reverberant environment, onset cells in the IC are modelled and interconnected to each other by inhibition projection.

2 Biological Fundamentals and Assumptions

When sound waves arrive at ears, the temporal and amplitude information is encoded and transmitted to the MSO and LSO in order to extract ITDs and ILDs respectively [5]. According to Jeffress's original model, the ITD-sensitive cells in the MSO can be idealised as a coincidence cell array where each cell receives a delay-line input, and the cells are assumed to be distributed along two dimensions: CF and ITD (see Figure 1). A cell in the MSO fires when the contralateral excitatory input leads the ipsilateral by a specific time difference.

For the LSO model, we represent the cells in the LSO distributed across two dimensions, CF and ILD, in an analogous manner to the MSO (Figure 2), but

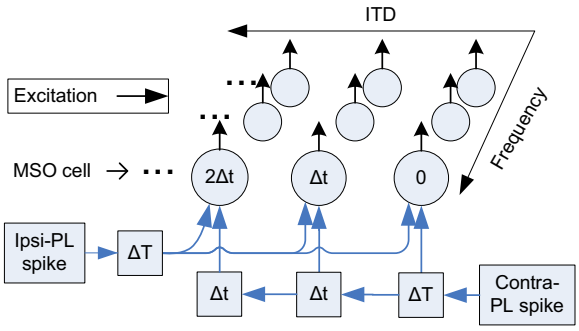


Fig. 1: Schematic diagram of the MSO model. While all the spike trains from the ipsilateral side share the same delay, the ones originating from the contralateral side are subjected to variable delays. The difference between the ipsilateral and contralateral delays makes each cell in the MSO model most sensitive to a specific ITD.

without any interaural delay. Each LSO cell compares the input levels from two ears and generates a spike if the level difference is equal to the characteristic ILD of the cell.

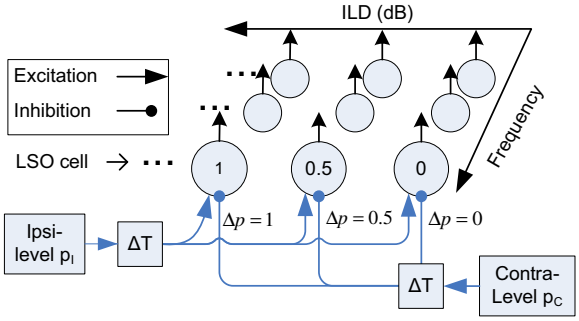


Fig. 2: Schematic diagram of the LSO model used in our system. Analogous to Figure 1, we assume that there are cells most sensitive to a given ILD and frequency. When the exact level difference $\Delta p = \log(\frac{p_I}{p_C})$ is detected, the corresponding LSO cell fires a spike.

All the outputs of the MSO and LSO are projected to the inferior colliculus (IC). The IC is also tonotopically organised, and contains a series of iso-frequency laminae, which span the whole range of frequencies perceived by the animal. In this model, we assume for simplicity that there are only projections from cells with the same CF. Consequently in our model the laminae of the IC with low CF (200 Hz to 1 kHz) only receive projections from the MSO, while the laminae

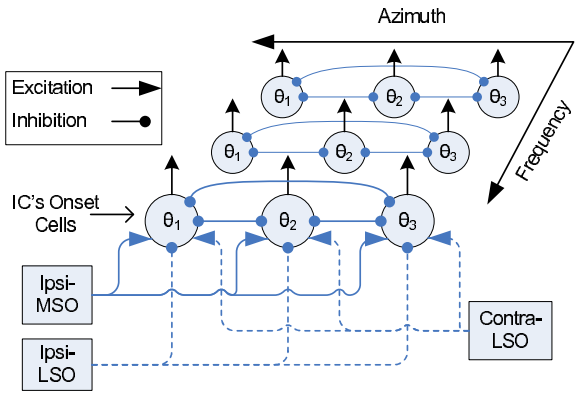


Fig. 3: Schematic diagram of the IC model. In the range of 20 Hz to 1 kHz, the IC model only receives inputs from the MSO model (solid line). From 1 kHz to 4 kHz, the IC model receives inputs from the MSO model (solid line) and both LSOs (dash line).

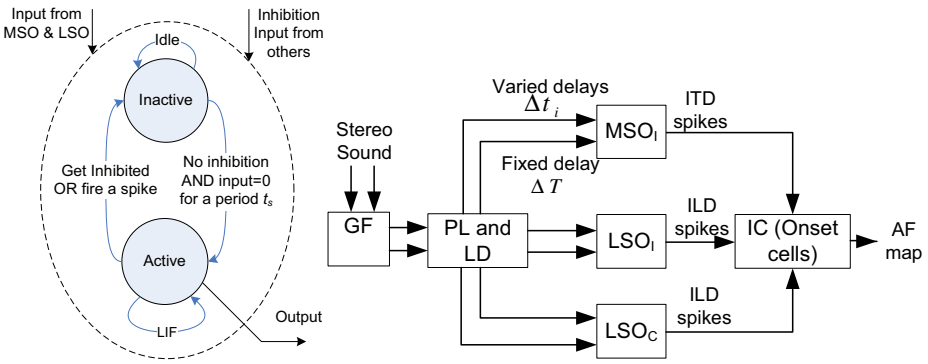
with higher frequencies (up to 4 kHz) receive projections from both the MSO and LSO. The laminae with a CF above 4 kHz would only receive inputs from the LSO, but our model does not include this range of frequencies.

The cells in the IC can be classified into 6 physiological types [10]: sustained-regular, rebound-regular, onset, rebound-adapting, pause/build and rebound-transient. The sustained-regular cells generate regular spikes when their input is kept positive and is can detect ongoing sounds. We hypothesise that these cells could encode sound source locations in the free field in the absence of echoes. However, in a reverberant environment, an echo is added to the sound taking the direct path and this causes a detection error in the output of sustained-regular cells. In contrast, onset cells only generate one spike when the input current changes from 0 to positive and then cease firing as long as the input is kept positive. This property would useful when the IC model is locating a sound in an echoic environment, because its output spike is only related to the sound taking the direct path.

Taking into account this biological evidence, we propose an IC model for sound source localisation as outlined in Figure 3. Analogous to the biology evidence, the IC model consists of different components according to the frequency domain. In addition, we suppose that onset cells with the same CF suppress one another, i.e. an early spike from one onset cell will inhibit other onset cells which have the same characteristic frequency. We will describe this inhibitory network in detail in the next section.

3 System Model of Sound Localisation

Inspired by the neurophysiological findings and the proposed models presented in Section 2, we designed our model to employ spiking neural networks (SNNs)



(a) Onset Cell (b) System Flowchart

Fig. 4: 4a: The IC's onset cell. 4b: Flowchart of the biologically inspired sound localisation system. This example only shows one IC; note that there is a symmetric model for the contralateral IC. $\mathrm{MSO_I}$ ipsilateral MSO model; $\mathrm{LSO_I}$ ipsilateral LSO model; $\mathrm{LSO_C}$ contralateral LSO model; GF Gammatone filterbank; PL phase locking; LD level detection; SR sustained-regular and AF azimuth-frequency.

that explicitly take into account the timing of inputs and mimic real neurons. The cues used for sound localisation, such as time and sound level, are encoded into spike-firing patterns that propagate through the network to extract ITD and ILD and calculate azimuth. Every neuron in the SNN is modelled using a leaky integrate-and-fire (LIF) model. The response of a neuron to spike inputs is modelled by:

$$\begin{aligned} C\tfrac{du}{dt} &= \sum_k I_k(t) - \tfrac{C}{\tau_m}u \\ t_f : u(t_f) &= \phi \end{aligned} \tag{1}$$

where $u(t)$is the membrane potential of the neuron relative to the resting potential which is initialised to 0, and τ_m is a time constant. C is the capacitance which is charged by $\sum_k I_k(t)$ from multiple inputs, where $I_k(t)$ is a current input which is a constant square with amplitude w_s and duration τ_s in response to a spike input. k is the number of input connections to the neuron. The action potential threshold φ controls the firing time t_f. When $u(t) = \varphi$, the soma will fire a spike; then $u(t)$ is reset to 0. Afterwards, the soma will be refractory for a period $t_r = 1$ ms during which it will not respond to any synaptic inputs. After the refractory period, the soma returns to the resting state.

The LIF model can be used to represent the cells in the MSO and the LSO and the sustained-regular cell in the IC. However, the LIF model cannot directly model the onset cell because it constantly responds to continuous inputs, rather than just the initial onset input. Instead, for the onset cell, we propose a hybrid model of LIF and a state machine. Each onset cell has two states: active and inactive. When the cell is active, the cell is implemented as a LIF neuron until a

spike is fired, or the cell receives an inhibitory input, after which the cell's state becomes inactive. The cell goes back to active state only if there is no inhibition and the input is 0 (no spike) for a period t_s (see Figure 4a).

A schematic structure for the sound localisation procedure is shown in Figure 4b. The frequency separation occurring in the cochlea is simulated by a bandpass filterbank consisting of 16 discrete second-order Gammatone filters [11], that produce 16 frequency bands between 200Hz and 4kHz. After the Gammatone filterbank, the temporal information in the waveform in each frequency channel is encoded into a spike train by the phase locking module in Figure 4b. Every positive peak in the waveform triggers a phase-locked spike to feed into the MSO model. The sound level is detected in the same module but directed to the LSO model.

To calculate the ITD, the phase-locked spike trains are then fed into the MSO model. A series of delays are added to the spike trains of the contralateral ear to simulate the delay lines Δt_i (see Figure 1). The spike train of the ipsilateral ear reaches the MSO with a single fixed delay time ΔT. The parameters of cells in the MSO are set: l_s = 2.1ms, τ_s=0.08ms, τ_m=1.6ms, φ=8e-4, w_s=0.1A and C = 10mF

The ILD pathway is not modelled using a LIF model; rather the sound levels previously detected for each side are compared and the level difference is calculated. The LSO model contains an array of cells distributed along the dimensions of frequency and ILD (Figure 2). When a specific ILD is detected at a given frequency, the corresponding LSO cell fires a spike. The level difference is calculated as $\Delta p^j = \log(p_I^j / p_C^j)$, where p_I^j and p_C^j stand for the ipsilateral and contralateral sound pressure level for the frequency channel j.

After the basic cues for sound localisation have been extracted by the MSO and LSO models, the ITD and ILD spikes are fed into the IC model, as shown in Figure 4b. The IC model merges the information to obtain a spatial representation of the azimuth of the sound source. According to the model proposed in Section 2, we need to define the connection strength between the ITD-sensitive cells (m_i) in the MSO and the azimuth-sensitive cells (θ_j) in the IC, and the connection between the ILD-sensitive cells (l_i) in the LSO and θ_j. In a SNN, each of the inputs to a neuron (in this case in the IC) produces a post-synaptic current $I(t)$ in the modelled cell. The post-synaptic currents of all the inputs are integrated to calculate the response of the neuron. To modify the weight of each input we assign a different gain to the amplitude w_s of the post-synaptic current $I(t)$ (in Equation 1) of each connection. Inspired by Willert's work [6], we used an approach based on conditional probability to calculate these gains, as shown in the following functions:

$$e_{m_i\theta_j} = \begin{cases} p(\theta_j \,|m_i\,, f) \text{ if} p > 0.5 \max\limits_j(p(\theta_j \,|m_i\,, f)) \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{2}$$

$$e_{l_i\theta_j} = \begin{cases} p(\theta_j \,|l_i\,, f) \text{ if} p > 0.8 \max\limits_j(p(\theta_j \,|l_i\,, f)), f >= f_b \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{3}$$

$$c_{l_i\theta_j} = \begin{cases} 1 - p(\theta_j \,|l_i\,, f) \text{ if} p < 0.6 \max\limits_j(p(\theta_j \,|l_i\,, f)), f >= f_b \\ 0 \qquad\qquad \text{otherwise} \end{cases}, \tag{4}$$

where $e_{m_i\theta_j}$ and $e_{l_i\theta_j}$ represent the gain of the excitatory synapse between the MSO and LSO respectively and the IC. If $e_{m_i\theta_j}$ is 0, it is equivalent to no connection between m_i and θ_j. Similarly, $e_{l_i\theta_j} = 0$ indicates no connection between l_i and θ_j. The term f_b is the frequency limit between the low and middle frequency regions and is governed by the separation of the ears and the dimensions of the head of the "listener". Based on the dimensions of the robot head used in this study, f_b should be around 850Hz.

$c_{l_i\theta_j}$ represents the gain of the inhibitory synapse between the LSO and the IC. f stands for the centre frequency of each frequency band. $p(*)$ stands for a conditional probability, which can be calculated by Bayesian probability and $p(\theta_j \,|m_i\,, f)$ can be calculated by:

$$p(\theta_j \,|m_i\,, f) = \frac{p(m_i \,|\theta_j\,, f)p(\theta_j \,|f\,)}{p(m_i \,|f\,)} \tag{5}$$

In a physical model, the conditional probability $p(m_i \,|\theta_j\,, f)$ is obtained from the statistics of sounds with known azimuths. To obtain such data, we recorded a 1s-sample of white noise sounds coming from 37 discrete azimuth angles (from -90 to 90 degrees in 5 degree steps) using a robot head. The head had dimensions similar to an adult human head and included a pair of cardioid microphones (Core Sound) placed at the position of the ears, 15 cm apart from one another.[1]

These recordings were processed through our MSO model to obtain an ITD distribution for each azimuth, which was then used to calculate $p(m_i \,|\theta_j\,, f)$. Finally, we applied Equation 5 to Equation 2 to calculate the gain, $e_{m_i\theta_j}$, of the connection between the MSO cells and the IC cells. These gains are further adjusted to leave only components consistent with the known anatomy of the pathway, i.e. there is no significant projection from the contralateral MSO to the IC. A similar procedure is used to calculate the gains of the LSO projection to the IC.

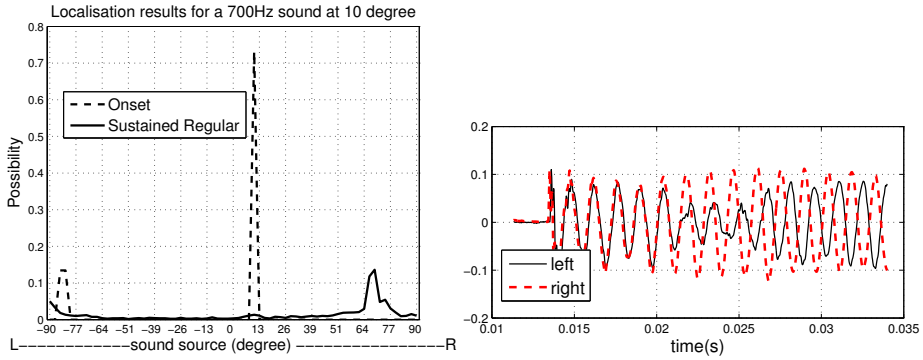
Equations 2 and 3 map the excitatory connections of each MSO and LSO cell to the IC cells representing the most likely azimuths, while Equation 4 maps

[1] Sounds were recorded in a low noise environment with 5 dB SPL background noise. The distance of the sound source to the center of the robot head was 128 cm and the speakers adjusted to produce 90±5 dB SPL at 1 kHz. Recordings were digitalised at a sample rate of 44100 Hz. Sound duration was 1.5s, with 10 ms of silence at the beginning.

the inhibitory LSO projection to cells representing azimuths in the hemifield opposite to the sound source. This inhibition counteracts the effects of false ITD detection at high frequencies.

4 Experimental Results

In this section, we first verify our model by locating a pure tone in a reverberant environment. We then implement our model to locate three groups of two concurrent sound sources in the same environment, and compare the results using sustained regular cells with these using onset cells.

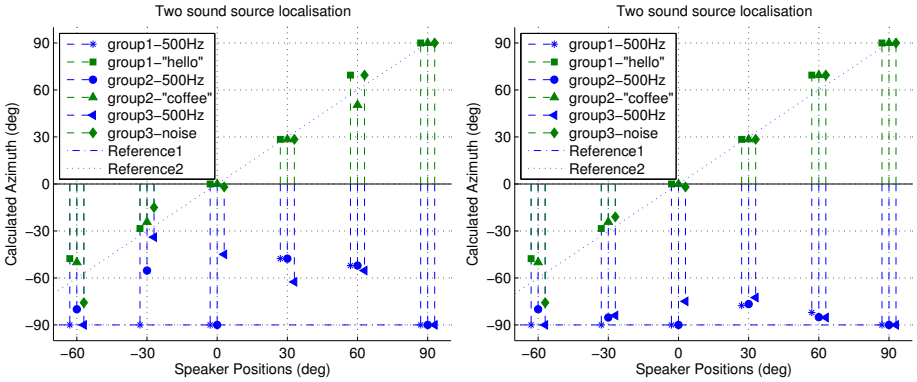


(a) Azimuth distribution (b) Effect of echo

Fig. 5: 5a: Azimuth distribution of single sound localisation in a reverberant environment. 5b: The effect of room reverberation to the recording. The sound sample is a 700Hz pure tone played at 10 degree from the midline.

Figure 5a compares the localisation results obtained using onset and sustained-regular cells. The sound sample was a 700Hz pure tone played at 10 degrees in a reverberant environment (echo delay 6 ms). The possibility of sound source azimuth is calculated as the division of the number of spikes of the IC cells in one azimuth angle of all frequency channels by the total number of spikes. Two methods were tried for the same sound: (i) only using sustained-regular cells in the IC with no inhibitory connection between the cells, and (ii) use the onset cells in the IC and the inhibitory network proposed earlier in this paper. Note that the first method is equivalent to the most conventional methods of sound localisation which are based on ongoing sound detection. Figure shows that the result from onset cells has a peak around 12 degrees which is very close to the real sound azimuth, while the peak of the results from the sustained-regular cells is around 70 degrees which is far from the true location. The main reason is that the reverberation interfered with the sound wave reaching the microphone and changed the ITD and ILD cues.

Figure 5b shows the effect of room reverberation on the recordings of a 700Hz pure tone presented at the front. It shows that the sound reaches the microphone at 0.013s and the signals from both microphones match each other peak by peak. However, from about 0.02s, the reverberant sound arrived and started to interfere with the recorded sound. As a result, the signals shifted by about 0.007ms. This peak shifting caused the localisation error that occurred when using the modelled sustained-regular cells.



(a) Sustained-Regular Cells (b) Onset Cells

Fig. 6: Sound localisation results for three groups of concurrent sound sources. Reference1 stands for the ideal azimuth of first sound source 500Hz and reference2 for the second sound source's ideal azimuth.

To test our model for a mixture of two concurrent sound sources, we designed three test groups and used a 500Hz pure tone from -90 degree as the first sound source for all three groups. The second source sources in three groups were designed as speech "hello", "coffee" or white noise. The second sound source is presented from 7 positions from -90 to 90 degree for every 30 degrees. All the sound sources are 1.28m far from the robot head. During recording, two sound sources are played at the same time. Each test point includes 10 sound sets and the final results are the average value. Figure 6a shows the localisation results using sustained-regular cells. In the figure, the second sound source azimuths were calculated accurately, however the results for the first sound source show a big offset from the ideal detection. In contrast, Figure 6b shows more accurate localisation results in both sound sources.

5 Conclusion and Future Work

This paper describes the design and implementation of a sound localisation model that uses a SNN inspired by the mammalian auditory system for a rever-

berant environment. In this system, both ITD and ILD pathways were modelled and computed in the MSO and LSO models, and the ITD spike and ILD spikes were projected to the IC in a way similar to the biological system where they were merged together to achieve broadband sound localisation. Onset IC cells are modelled and an inhibitory onset network is proposed to eliminate the echo. The experimental results showed that our system can localise two concurrent sound sources in a reverberant environment especially for pure tones with azimuths between -90 and 90 degrees. Our model's success casts light on the mobile robot application in real world application where reverberation is unavoidable. In the future, other IC cell types will be tested in the model For the application of our system to a mobile robot, we plan to implement a sound separation system based on sound source direction in order to improve speech recognition in a noisy environment

References

1. Bronkhorst, A., Plomp, R.: Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. The Journal of the Acoustical Society of America **92** (1992) 3132–3139
2. Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization. Mit Press (1997)
3. Jeffress, L.: A place theory of sound localization. J. Comp. Physiol. Psychol. **41** (1948) 35–39
4. Moore, B.: An Introduction to the Psychology of Hearing (ed.). San Diego: Academic Press (2003)
5. Yin, T.: Neural mechanisms of encoding binaural localization cues in the auditory brainstem. Integrative Functions in the Mammalian Auditory Pathway (2002) 99–159
6. Willert, V., Eggert, J., Adamy, J., Stahl, R., E., K.: A probabilistic model for binaural sound localization. IEEE Trans Syst Man Cybern Part B Cybern **36**(5) (2006) 982–994
7. Voutsas, K., Adamy, J.: A biologically inspired spiking neural network for sound source lateralization. IEEE Trans Neural Networks **18**(6) (2007) 1785–1799
8. Litovsky, R., Colburn, H., Yost, W., Guzman, S.: The precedence effect. Journal of the Acoustical Society of America **106**(4 I) (1999) 1633–1654
9. Palomäki, K., Brown, G., Wang, D.: A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Communication **43**(4) (2004) 361–378
10. Sivaramakrishnan, S., Oliver, D.: Distinct K Currents Result in Physiologically Distinct Cell Types in the Inferior Colliculus of the Rat. Journal of Neuroscience **21**(8) (2001) 2861
11. Slaney, M.: An efficient implementation of the patterson-holdsworth auditory filter bank. Apple Computer Technical Report **35** (1993)