Neurocomputing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation

Jindong Liu [a,*], David Perez-Gonzalez [b], Adrian Rees [b], Harry Erwin [a], Stefan Wermter [c]

[a] School of Computing and Technology, University of Sunderland, Sunderland SR6 0DD, UK
[b] Institute of Neuroscience, The Medical School, Newcastle University, Newcastle upon Tyne NE2 4HH, UK
[c] Department of Informatics, University of Hamburg, 22527 Hamburg, Germany

ARTICLE INFO

Keywords:
Spiking neural network
Sound localisation
Inferior colliculus
Interaural time difference
Interaural level difference
Intelligent robotics

ABSTRACT

This paper proposes a spiking neural network (SNN) of the mammalian subcortical auditory pathway to achieve binaural 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 a sound source over a wide frequency range. 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 (θ), respectively. The neurons in each group are tonotopically arranged to take into account the frequency organisation of the auditory pathway. To reflect the biological organisation, only ITD information extracted by the MSO is used for localisation of low frequency (<1 kHz) sounds; for sound frequencies between 1 and 4 kHz the model also uses ILD information extracted by the LSO. This information is combined in the IC model where we assume that the strengths of the inputs from the MSO and LSO are proportional to the conditional probability of $P(\theta|ITD)$ or $P(\theta|ILD)$ calculated based on the Bayes theorem. The experimental results show that the addition of ILD information significantly increases sound localisation performance at frequencies above 1 kHz. Our model can be used to test different paradigms for sound localisation in the mammalian brain, and demonstrates a potential practical application of sound localisation for robots.

1. Introduction

Humans and other animals show a remarkable ability to localise sound sources using the disparities in the sound waves received by the ears. This has inspired researchers to develop new computational auditory models to help understand the biological mechanisms that underlie sound localisation in the brain. The project described in this paper discusses sound processing in the mammalian brain and aims to build a computational model that can be tested on biomimetic mobile robots to validate and refine models for focused hearing.

During the last decades, the structure and function of pathways in the auditory brainstem for sound localisation have been extensively studied [9]. Binaural sound localisation systems take advantage of two important cues [6] derived from the sound signals arriving at the ears: (i) interaural time difference (ITD), and (ii) interaural level difference (ILD). Assuming a pure tone sound source is positioned on the left side, the sound signal at the left ear is represented by $a\sin 2\pi ft$ (where a is the sound amplitude and f the sound frequency) while the sound at the right ear is represented by $(a-\Delta a)\sin 2\pi f(t+\Delta t)$, where Δa and Δt represent, respectively, the level difference (ILD) caused by the shadowing effect of the head, and the additional time (ITD) required for the sound wave to travel the further distance to the right ear. Using these two cues sound source direction can be estimated in the horizontal or azimuthal plane.

The ranges over which these cues operate depend on head size. In humans the ITD cue is effective for localising low frequency sounds (20 Hz $\sim$1 kHz) [7], however, the information it provides becomes ambiguous for frequencies above $\sim$1 kHz. In contrast, the ILD cue has limited utility for localising sounds below 1 kHz, but is more efficient than the ITD cue for mid- and high-frequency (>1 kHz) sound localisation [7]. 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 [15].

Several hypotheses for ITD and ILD processing have been proposed [6,12,2], with one of the most influential being a model

* Corresponding author.
E-mail addresses: jindong.liu@gmail.com (J. Liu), davidpg@usal.es (D. Perez-Gonzalez), adrian.rees@ncl.ac.uk (A. Rees), harry.erwin@sunderland.ac.uk (H. Erwin), wermter@informatik.uni-hamburg.de (S. Wermter).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.10.030

advanced by Jeffress [6]. In his model, ITDs are extracted by a mechanism in which neural activity elicited by sound from each ear travels through a number of parallel delay lines, each one of which introduces a different delay into the signal and connects with a particular MSO cell. One of these delays compensates for the interaural delay of the sound waves, thus causing the signal from both ears to arrive coincidentally at a neuron that fires maximally when it receives simultaneous inputs. Smith et al. [12] provided partial evidence for Jeffress's model in the cat with the description of axons that resemble delay lines for the signal arriving at the MSO from the contralateral ear, but they found no evidence for delay lines for the MSO input from the ipsilateral side. For ILDs, physiological evidence suggests this cue is encoded in the neuronal firing that results from the interaction of an excitatory input from the side ipsilateral to the LSO, and an inhibitory input driven by the sound reaching the contralateral side. Thus, as the sound moves from one side to the other, the firing rate of the neurons decreases in one LSO and increases in the other.

Modellers have taken different approaches to represent this system. In an engineering study, Bhadkamkar [1] proposed a system to process ITDs using a CMOS circuit, while Willert [14] built a probabilistic model which separately measures ITDs and ILDs at a number of frequencies for binaural sound localisation. Recently, Voutsas and Adamy [13] realised a multi delay-line model using spiking neural networks (SNN) which incorporate realistic neuronal models. This model only takes into account ITDs and while it gives good results for low frequency sounds, it is not effective for frequencies greater than 1 kHz. Some models seek to incorporate the contribution of the inferior colliculus. In an engineering model, Rodemann [10] applied three cues for sound localisation, however, it did not take advantage of the biological connection between the superior olivary complex (SOC) and the IC. Willert [14] and Nix [8] implemented a probabilistic model to estimate the position of the sound sources, which includes models of the MSO, LSO and IC and uses the Bayesian theorem to calculate the connections between them. However, the model did not use SNN to simulate realistic neuronal processing.

This paper presents a model designed to identify sound source direction by means of a SNN. It is the first to employ an SNN that combines both ITD and ILD cues derived from the SOC in a model of the IC to cover a wide frequency range. To simulate the biological connection between the MSO/LSO and the IC, we propose a model which applies Bayes probability theorem to calculate the synaptic strength of the connection between cells in these nuclei. 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 a sound source.

The rest of this paper is organised as follows. Section 2 presents the neurophysiological organisation of the mammalian auditory pathway as derived mainly from cat and guinea pig experimental data. It also presents an IC model which takes account of the projection from MSO and LSO. Section 3 proposes a system model which simulates the mammalian auditory pathway from the cochlea up to the IC. In Section 4, experimental results are presented to show the feasibility and performance of the sound localisation system. Finally, conclusions and future work are considered in Section 5.

2. Biological fundamentals and assumptions

When sound waves arrive at the external ear, they enter the auditory meatus and vibrate the tympanic membrane, or ear drum, to then be propagated through the auditory ossicles in the middle ear to the cochlea of the inner ear. There the vibrations generate a travelling wave of displacement that propagates along the basilar membrane inside the cochlea, such that the point of maximum displacement is dependent on the frequency of the sound, thus leading to a spatial separation of the frequencies in the stimulus. The motion of the basilar membrane activates the inner hair cells arrayed along its length, which in turn trigger action potentials in auditory nerve (AN) fibres that transmit the spike encoded information to the central nervous system. Each auditory nerve fibre is maximally sensitive at a characteristic frequency (CF) which is determined by the location of the inner hair from which it receives its input [15]. This tonotopic representation of frequency is maintained in subsequent nuclei of the ascending auditory pathway.

In addition to this tonotopic representation, the AN fibres also encode temporal information about the sound waveform. The hair cells act as halfwave rectifiers so that the probability of AN fibre excitation is maximal during the peak phase of the sound waveform. This phase locking occurs at frequencies of 20 Hz $\sim$ 5 kHz, and is an essential step in the extraction of ITDs, because it represents the basis for comparing the relative timing of the waveforms at the ears. Fig. 1 shows an example of spikes phase-locked to the peaks of the sound waveform (t_1^l, t_1^r, t_2^l and t_2^r).

As the sound pressure level (SPL) increases, the discharge rate of most AN fibres increases sigmoidally over a relative range of $\sim$ 30 dB. In order to cover the wide range of SPL to which we are sensitive, e.g. 120 dB, the relative operating range changes adaptively according to the background sound levels. There is also a smaller population of AN fibres that have a higher threshold and a wider dynamic range. For simplicity, in this paper we do not model the biological details of the encoding of sound amplitude, but rather we use the measured SPL (e.g. p_i^l and p_i^r in Fig. 1) in the first stages of ILD processing.

After encoding the temporal and amplitude information, the spike-encoded information from each ear is transmitted via the cochlear nuclei to the SOC to extract ITDs and ILDs in the MSO and LSO, respectively [15] (Fig. 2). The MSO on one side receives excitatory inputs from the anteroventral cochlear nucleus (AVCN) from both the ipsilateral and contralateral sides. An ITD-sensitive cell in the MSO fires when the contralateral excitatory input leads the ipsilateral by a specific time difference. According to hypotheses based on Jeffress's original model, activation of these coincidence detectors is thought to occur when the contralateral delay line network compensates for the time delay of the sound in the ipsilateral ear, i.e. ITD. These ITD-sensitive cells in the MSO can be idealised as a coincidence cell array where each cell receives a delay-line input, and they are assumed to be

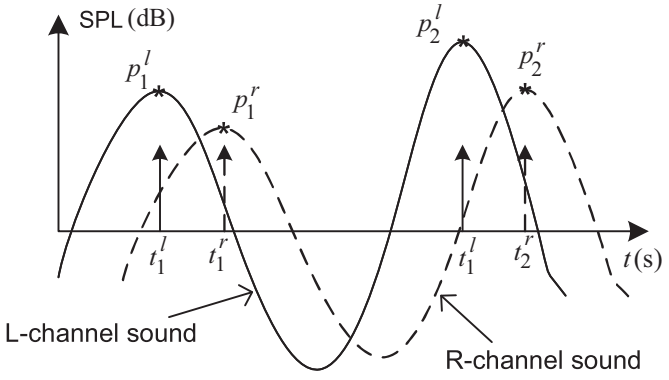


Fig. 1. An example of sound signals arriving at both ears (left, continuous line; right, dashed line), and the phase-locked spikes (t_1^l, t_1^r, t_2^l and t_2^r) triggered by them. The signal corresponding to the right ear is delayed and has a smaller amplitude than the left one, indicating that the origin of the sound was on the left side of the head. $p_i^{l/r}$ is the sound pressure level at which the spikes are generated.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

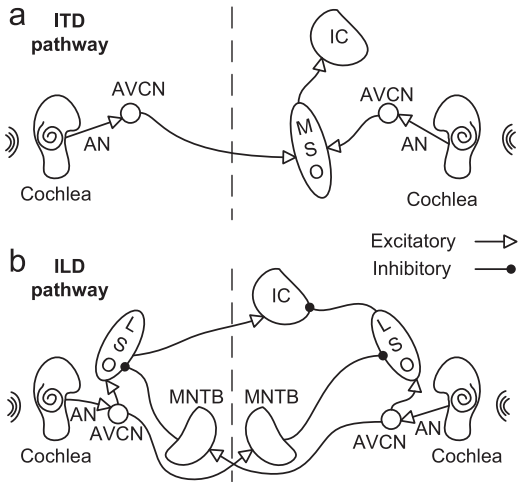


Fig. 2. Schematic diagram of the mammalian auditory pathway from the cochlea to the inferior colliculus (IC): (a) pathway for the processing of interaural time differences (ITD) and (b) pathway for the processing of interaural level differences (ILD). Note how in the case of the ILD pathway the IC receives information from both LSOs: excitatory (white arrowheads) from the contralateral and inhibitory (black circles) from the ipsilateral. AN, auditory nerve; AVCN, anteroventral cochlear nucleus, LSO, lateral superior olive; MNTB, medial nucleus of the trapezoid body; MSO, medial superior olive.

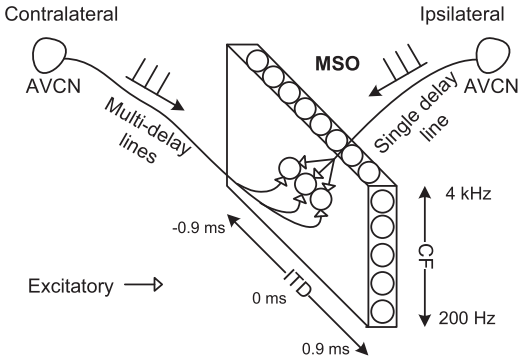


Fig. 3. Schematic diagram of the MSO model used in our system. While all the spike trains from the ipsilateral AVCN share the same delay, the ones originated in the contralateral side are subjected to variable delays. The difference between the ipsilateral and contralateral delays makes each cell in the MSO model to be most sensitive to a given ITD. This array of ITD-sensitive cells is repeated across frequency channels. Our system could detect ITDs from −0.9 to 0.9 ms. The MSO model contained neurons sensitive to frequencies between 200 Hz and 4 kHz.

distributed along two dimensions: CF and ITD [5] (see Fig. 3). The output of the MSO cells is transmitted to the ipsilateral IC.

Recently, since this work was begun, an alternative to the delay-line hypothesis has come to the fore that explains ITD sensitivity on the basis of inhibitory mechanisms [2], however, the precise mechanism underlying ITD sensitivity is beyond the scope of this paper, and in any case does not impact on the processing that we model in the IC, and does not impact on the processing that we model in the IC.

For ILD processing, cells in the LSO are excited by sounds in a level dependent manner at the ipsilateral ear and inhibited at the contralateral ear [15]. For instance, in response to a sound on the left, the left LSO receives excitation from the ipsilateral AVCN, but inhibition from the contralateral side, meditated by the medial nucleus of the trapezoid body (MNTB) which transforms excitation from the contralateral AVCN to inhibition (Fig. 2). In contrast to the MSO, there is no evidence for delay lines projecting to the LSO. Although the mechanisms of ILD processing are not fully understood we know the spike rate of LSO neurons depends on

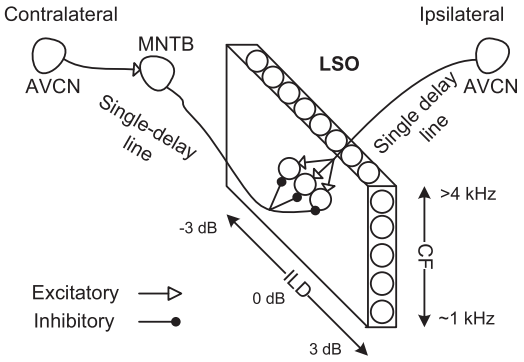


Fig. 4. Schematic diagram of the LSO model used in our system. Similarly to Fig. 3, we assume that there are cells most sensitive to a given ILD and frequency. The ILD sensitivity is caused by the interaction of excitatory (ipsilateral) and inhibitory (contralateral) inputs. Our system could detect ILDs from −3 to 3 dB. (The LSO model contained neurons sensitive to frequencies between ~1 kHz and 4 kHz.)

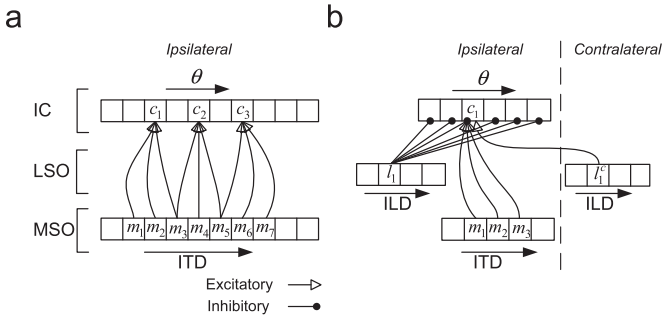


Fig. 5. Schematic diagram of the distribution of inputs to our IC model, assuming that there are no connections across frequencies. (a) In the range of 20 Hz to 1 kHz, the IC model only receives inputs from the MSO model. (b) From 1 to 4 kHz, the IC model receives inputs from the MSO model and both LSO. The distributions of the projections were calculated using Bayesian statistics (see Section 3 for details).

the sound level difference between both ears. In this paper, we represent the cells in the LSO distributed across two dimensions, CF and ILD, in an analogous manner to the MSO (Fig. 4). The LSO sends an excitatory output to the contralateral IC and an inhibitory output to the ipsilateral IC.

The cells in the MSO and LSO operate over different frequency ranges. For example, in cat the MSO is a low-frequency structure with most of its neurons in the range from 20 Hz to about 5 kHz [5], while the LSO is a high-frequency structure with little representation below 1 kHz [4]. The inferior colliculus (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 connections between cells with the same CF. Consequently, in our model the laminae of the IC with low CF only receive projections from the MSO, while the laminae 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.

Taking into account this biological evidence, we propose an IC model for sound source localisation as outlined in Fig. 5. The IC model consists of different components according to the frequency domain: in the low frequency domain, as shown in Fig. 5(a), only the ipsilateral MSO is involved in sound localisation; while in the middle frequency domain, shown in Fig. 5(b), the ipsilateral MSO and both LSOs contribute inputs to the IC.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

The cells in the IC receive excitatory inputs from the ipsilateral MSO and contralateral LSO, and inhibitory inputs from the ipsilateral LSO. The connection type between the MSO and the IC is many-to-one and one-to-many, while the inhibitory input from the LSO is a one-to-all projection. The input from the contralateral LSO is composed by excitatory connections to the IC assumed to be mainly few-to-one.

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) 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 with a single compartment (the soma) and several synapses which connect the elements of the network.

The postsynaptic current $I(t)$ of a neuron, triggered by a synaptic input (spike) at a time $t=t_s$, can be modelled as a constant square current with an amplitude (or weight) of w_s, starting at a delay or latency l_s relative to the timing of the incoming input, and lasting a time τ_s. The excitatory or inhibitory effect of each input is modelled using a positive or negative $I(t)$, respectively. The response of the soma to the synaptic inputs is modelled using a leaky integrate-and-fire model [3]:

$$u(t) = u_r \exp\left(-\frac{t-t_s}{\tau_m}\right) + \frac{1}{C}\int_0^{t-t_s} \exp\left(-\frac{s}{\tau_m}\right) I(t-s)\, ds \qquad (1)$$

where $u(t)$ is the membrane potential of the neuron relative to the resting potential, u_r is the initial membrane potential, and τ_m is a time constant, which will affect the temporal integration of the inputs. In this paper, the value of τ_m is 1.6 ms. C is the capacitance which is charged by $I(t)$, in order to simulate the procedure of the postsynaptic current charging the soma. The soma model has one more parameter, the action potential threshold φ. When $u(t)=\varphi$, the soma will fire a spike; then $u(t)$ is reset to 0, the relative resting potential of the neuron.

A schematic structure for the sound localisation procedure is shown in Fig. 6. 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 200 Hz and 4 kHz.

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 Fig. 6. This simulates the halfwave rectified receptor potential of the inner hair cells in the cochlea that lead to phase-locked spikes in AN fibres. Every positive peak in the waveform triggers a phase locked spike to feed into the MSO model. 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 Fig. 3). We denote the delayed spike sequence as $S_C(\Delta t_i, f_j)$, where C stands for the contralateral ear, Δt_i for the delay time and f_j for the frequency channel j. Similarly, $S_I(\Delta T, f_j)$ represents the spike train of the ipsilateral ear, with a single fixed delay time ΔT. $S_C(\Delta t_i, f_j)$ and $S_I(\Delta T, f_j)$ are the inputs to the MSO model. The cells in the MSO are modelled as coincidence cells (see Fig. 7(a) and Table 1 for details).

Owing to the delay line input, each cell fires maximally at a certain ITD. The output the MSO model is represented as $S_{ITD}((\Delta t_i - \Delta_T), f_j)$. If $S_{ITD}((\Delta t_i - \Delta_T), f_j)$ results in spikes for a given cell, it indicates that the sound is arriving at the ear contralateral to the MSO earlier than at the ipsilateral by $\Delta t_i - \Delta_T$ seconds. Once the ITD calculation is implemented for all frequency channels, a map can be plotted to reflect the spike rate, combining the information from both MSOs. See Section 4 and Fig. 10 for an example of the ITD spike map. In the ITD map, the x-axis

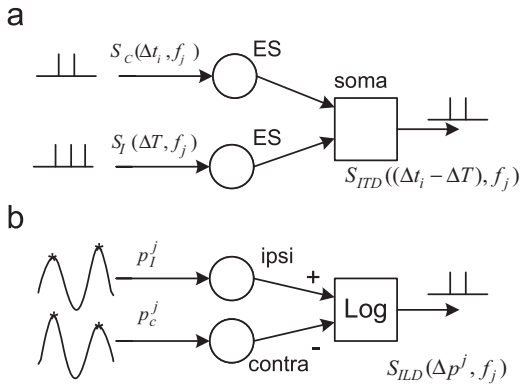


Fig. 7. (a) Model of the ITD processing and (b) model of the ILD processing. ES stands for excitatory synapse. Ipsi for ipsilateral and contra for the contralateral gammatone frequency channel.

Table 1
Parameters for the MSO model.

	Synapse			Soma			Δ_{t_i} Step	Δ_{t_i} Range
	l_s	τ_s	w_s	φ	τ_m	C		
MSO	2.1	0.08	0.1	8e-4	1.6	10	2.26e-2	[0 0.9]

Note: $\Delta_T = 0$. The unit of l_s, τ_s, τ_m and Δ_{t_i} step/range, is ms. The units of C, w_s and φ are mF, A, and V, respectively. There is one MSO model for each side.

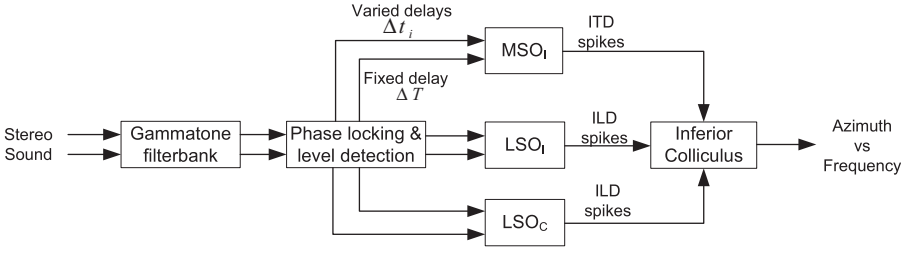


Fig. 6. Flowchart of the biologically inspired sound localisation system. This example only shows one IC; note that there is a symmetric model for the contralateral IC. MSO_I: ipsilateral MSO model; LSO_I: ipsilateral LSO model; and LSO_C: contralateral LSO model.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

represents ITD, the y-axis the frequency channel, and the z-axis the spike rate. When combining information from both MSOs in one map, we represent the ITD as $ITD = t_r - t_l$, where t_r and t_l are the sound arrival time at the right and left ear, respectively. Thus, the left side of the ITD map is calculated from the left MSO, and vice versa.

The ILD pathway is not modelled using a leaky integrate and fire 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 (Fig. 4). When a certain ILD is detected at a given frequency, the corresponding LSO cell fires a spike (see Fig. 7(b)). The level difference is calculated as $\Delta p^j = \log(p_I^j / p_C^j)$, where p_I^j and p_C^j are the ipsilateral and contralateral sound pressure level for the frequency channel j. Similar to the MSO model, the output of LSO model can be represented as a map of the spike rate, $S_{ILD}(\Delta p^j, f_j)$, see Section 4 and Fig. 11. Analogous to the case of the ITD map in the MSO, the ILD map is built using the information from both LSOs. Here, we define the ILD as $ILD = \log(p_r / p_l)$, where p_r and p_l are the sound level at the right and left ear, respectively. Therefore, the left side of the ILD map is calculated from the left LSO, and vice versa.

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 Fig. 6. 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 an 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 Eq. (1)) of each connection. Inspired by Willert's work [14], we used an approach based on conditional probability to calculate these gains, as shown in the following functions:

$$e_{m_i\theta_j} = \begin{cases} p(\theta_j|m_i,f) & \text{if } p > 0.5 \max_j(p(\theta_j|m_i,f)) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$e_{l_i\theta_j} = \begin{cases} p(\theta_j|l_i,f) & \text{if } p > 0.8 \max_j(p(\theta_j|l_i,f)), \quad f \geq f_b \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$c_{l_i\theta_j} = \begin{cases} 1-p(\theta_j|l_i,f) & \text{if } p < 0.6 \max_j(p(\theta_j|l_i,f)), \quad f \geq f_b \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

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 implies 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 (see below), f_b should be around 850 Hz.

$c_{l_i\theta_j}$ 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. For example, $p(\theta_j|m_i,f)$ can be calculated by

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

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 coming from 7 discrete azimuthal angles (from -90° to 90° in 30° 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 Eq. (5) to Eq. (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. Fig. 8 shows the gain calculated for the MSO projection at two different frequencies: (a) 449 Hz and (b) 2888 Hz.

A similar procedure is used to calculate the gains of the LSO projection to the IC. Fig. 9 shows the gains calculated for this projection, at 2888 Hz. Fig. 9a shows the excitatory contralateral projection, while Fig. 9b shows the inhibitory ipsilateral projection.

Eqs. (2)–(4) which are used to calculated the gains for the projections between the MSO or LSO and IC have two features: (i) Eqs. (2) and (3) map the excitatory connections of each MSO and LSO cell to the IC cells representing the most likely azimuths, while Eq. (4) maps 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. (ii) The equations also reflect the distribution of projections from the MSO and LSO to the IC. For example, Eq. (2) implies that there can be multiple m_i that have an active connection to a single IC cell θ_j. For example, in Fig. 8 a sound coming from a single azimuth (e.g. 30°) causes multiple MSO cells to respond to different extents (e.g. cells tuned at -0.54 to -0.09 ms ITD). Furthermore, Eq. (3) defines a few-to-one projection from the contralateral LSO to the IC (Fig. 9a), while Eq. (4) shows a one-to-all projection from the ipsilateral LSO to the IC (Fig. 9b).

The output of the IC model represents the azimuth within each frequency band and this information would be directed to the thalamocortical part of the auditory system which is beyond the scope of this study.

4. Experimental results

The model was tested in conjunction with the robot head using real sounds. Two types of broadband sound sources were employed: noise and speech presented at different azimuths to the head. The speech sounds include five words in the English language: "hello", "look", "fish", "coffee" and "tea".

Fig. 10 shows the results of the ITD calculation for a presentation of the word "fish" presented at 60° to the left and 1.28 m away from the head (see the spectrogram of the recording

[1] Sounds were recorded in a low noise environment (~5 dB SPL background noise). The distance of the sound source to the centre 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 44,100 Hz. Sound duration was 1.5 s, with 10 ms of silence at the beginning.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

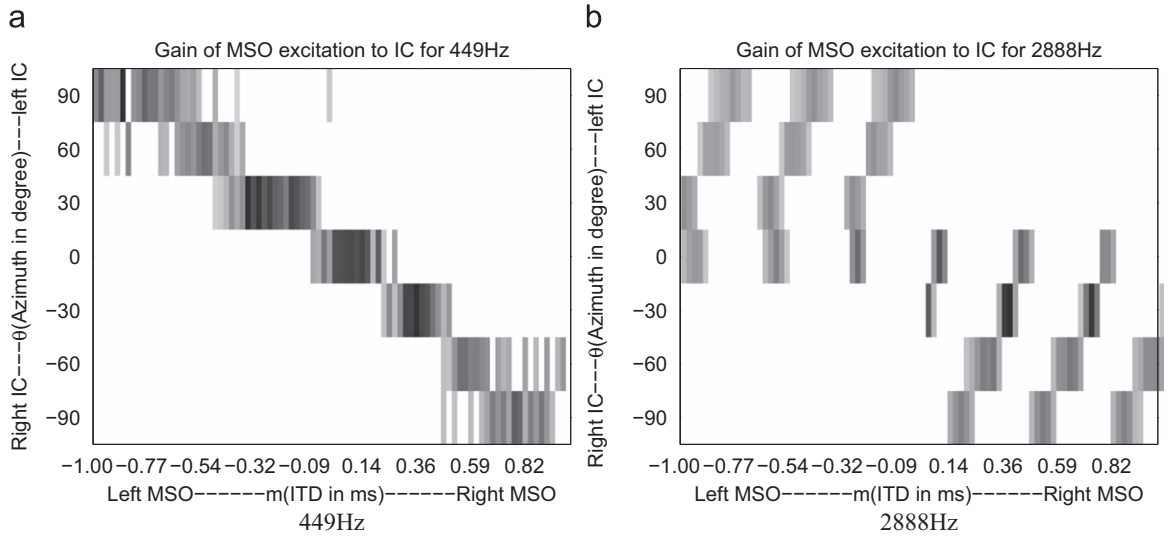


Fig. 8. Gain of the projection from the ipsilateral MSO to the IC, at 449 Hz (a) and 2888 Hz (b). Each coordinate represents the gain of the connection from each of the 89 MSO cells (characterised by their best ITD, abscissa) to a given IC cell (characterised by its best azimuth, ordinate). Dark areas indicate high gain values.

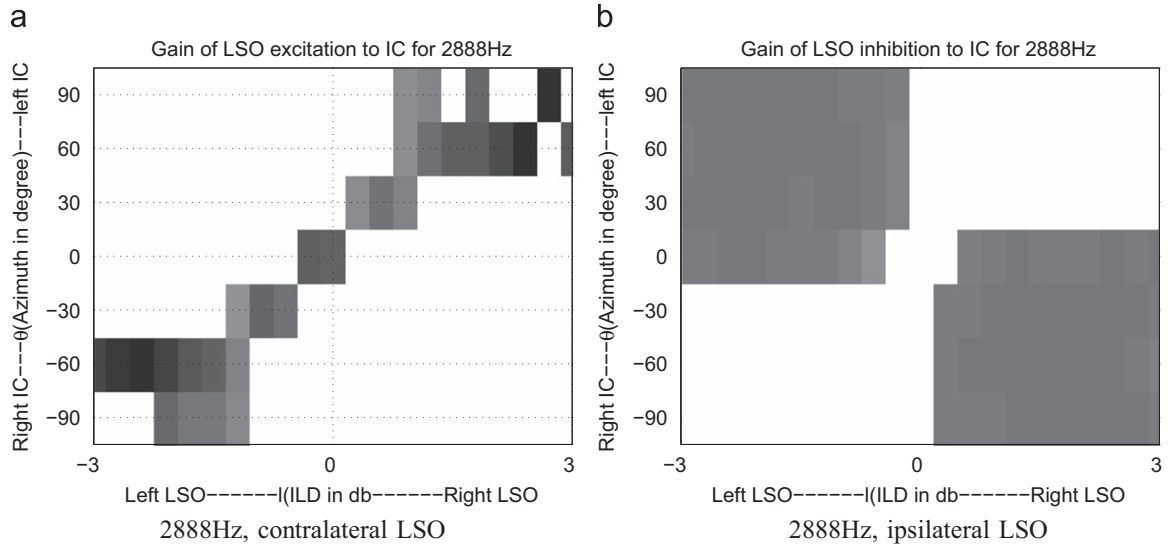


Fig. 9. Gain of the projection from the contralateral ((a) excitatory) and ipsilateral ((b) inhibitory) LSO to the IC, at 2888 Hz. Each coordinate represents the gain of the connection from each of the 22 LSO cells (characterised by their best ILD, abscissa) to a given IC cell (characterised by its best azimuth, ordinate). Dark areas indicate high gain values.

in Fig. A1). Recall that the ITD was defined as $ITD = t_r - t_l$ for the ITD map; therefore the real ITD of the sample sound is around 0.6 ms, which is successfully reflected by peaks in Fig. 10 over most frequency channels. However, it is also noticeable that there are multiple peaks above 850 Hz. These false detections also occur in biological systems and are due to the cyclic nature of sound, which generates ambiguities when the wavelength of the sound is more than twice the distance between the ears [7]. In our model, with the microphones separated by 15 cm, the limit frequency is about 850 Hz.

When the same stimulus is processed by the LSO model, the peak activity in the ILD map (Fig. 11) appears on the left, consistent with the fact that the sound originated from the left side and activated cells in the left LSO. However, below 1 kHz the activity is not lateralised. This is explained by the diffraction of the sound waves when their wavelengths exceed the dimensions of the head so that the sounds suffer minimal attenuation when they arrive at the contralateral ear [7].

The IC combines the inputs from the MSO and LSO models, applying the previously calculated gains to generate a frequency-azimuth map. Fig. 12 shows the output of the IC using our system model under the same stimulus conditions described above. The behaviour of the model when only the MSO input to the IC is used is depicted in Fig. 12a, while Fig. 12b includes both MSO and LSO inputs. Including the ILD information removes the false detections at 0°, 30° and 60°, generated by the MSO model at frequencies above 850 Hz.

The data in Fig. 13 shows the final step in the modelling process, which is to combine the azimuth data across frequencies to obtain a single estimate of the sound source location. In these plots the *x*-axis represents the azimuth of the stimulus sound

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

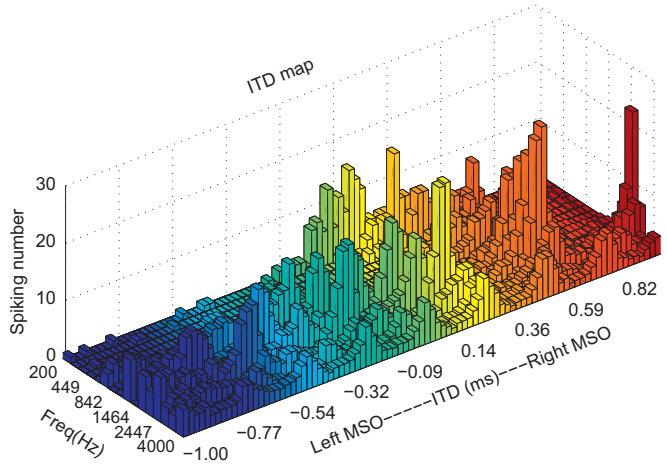


Fig. 10. Output of the MSO model when presenting a sound (the word "fish") from the left side at -60° and a distance of 1.28 m. In our system, this azimuth corresponds to an ITD of ~ 0.60 ms.

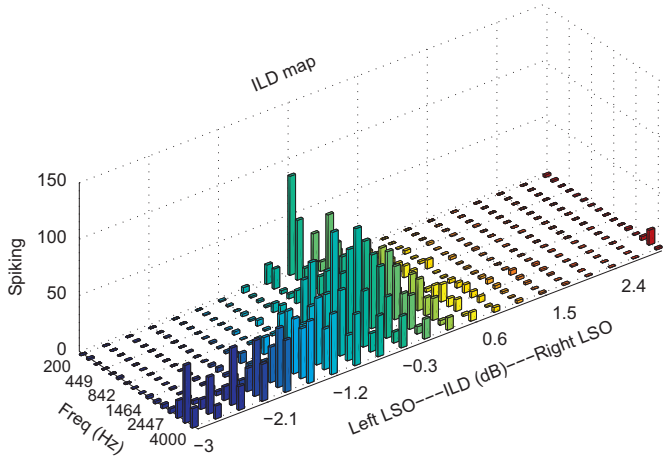


Fig. 11. Output of the LSO model when presenting the same sound as in Fig. 10.

source while the *y*-axis is the estimated azimuth calculated by the model. The size of the square symbols is proportional to the sum of the spikes over all frequencies at the corresponding estimated angle. For example, in Fig. 13b, the square at $(-30\ -30)$ is larger than the others for the same stimulus position, indicating that for a sound placed at -30° the IC neurons coding for -30° in our model responded with a higher spike rate than the neurons coding for other azimuths.

The localisation efficiency of the system is defined as the percentage of the total number of spikes that occur at the correctly estimated azimuth. In this experiment we compare the performance of the IC model under three conditions. (i) MSO input only, (ii) LSO input only and (iii) both MSO and LSO inputs. In Fig. 13a and b, the estimation is only based on ITD information from the MSO model. While the response of the model is best at the correct angles for sounds in the range -90° to 90°, there is also activity at several false estimations, especially at the most lateral azimuths. This is because the test sounds contain high frequency components that introduce ambiguities into the MSO model. Under these conditions the localisation efficiency across all sound types and angles is only about 25%. A similar problem is apparent in the LSO only condition (Fig. 13c and d) where although the correct azimuths are well estimated in most the cases, there is a significant spread of estimates.

When using both the MSO and the LSO inputs so that the model operates on both ILD and ITD information (Fig. 13e and f), the distribution of the responses gives almost precise localisation. The overall localisation efficiency increases to 80%, and 90% for sound signals between -45° and 45°. The highest localisation efficiency occurred at 0°. The efficiency decreases when the sound moves to the sides. At the most extreme positions (-90° and 90°) the azimuth was not always correctly estimated.

We also tested the ability of our system to localise multiple simultaneous sound sources using two speakers. Fig. 14 shows the model response with two kinds of sound source combinations: a pure tone combined with a speech sound, and two speech sounds together (see their spectrogram in Figs. A2 and A3). The results are encouraging and the model correctly localises two sound sources. Comparing the two columns in Fig. 14, the results

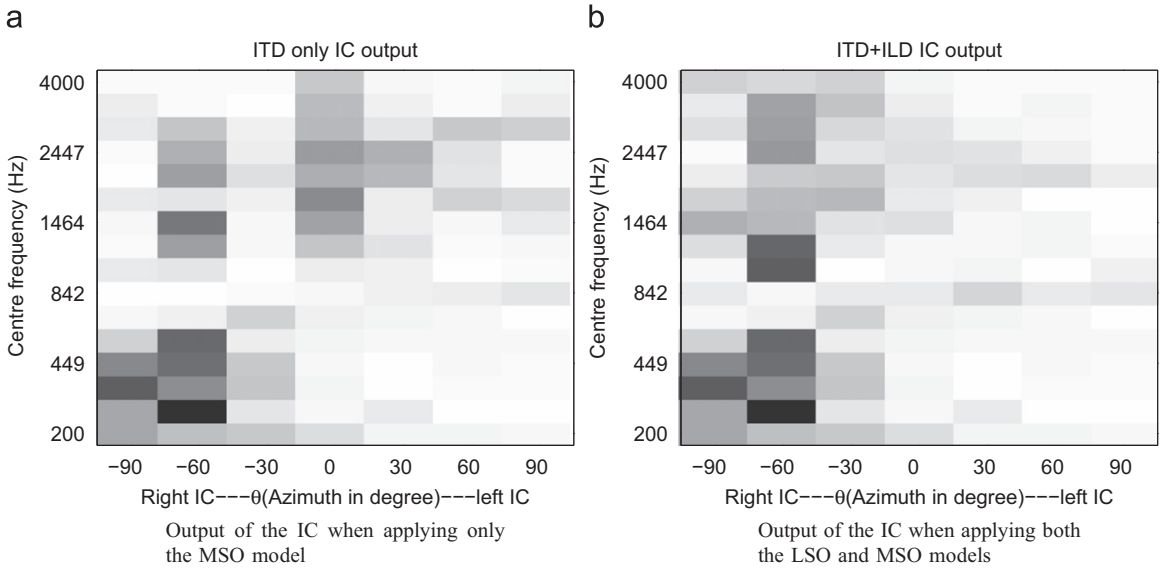


Fig. 12. Frequency-azimuth map calculated from the IC output. Dark areas represent a high spike rate, indicating the detection of a possible sound source at that azimuth and frequency. This example used the same sound sample as in Fig. 10, i.e. the word fish played at -60°. Note how the inclusion of the LSO model improves the localisation eliminating false detections at high frequencies.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

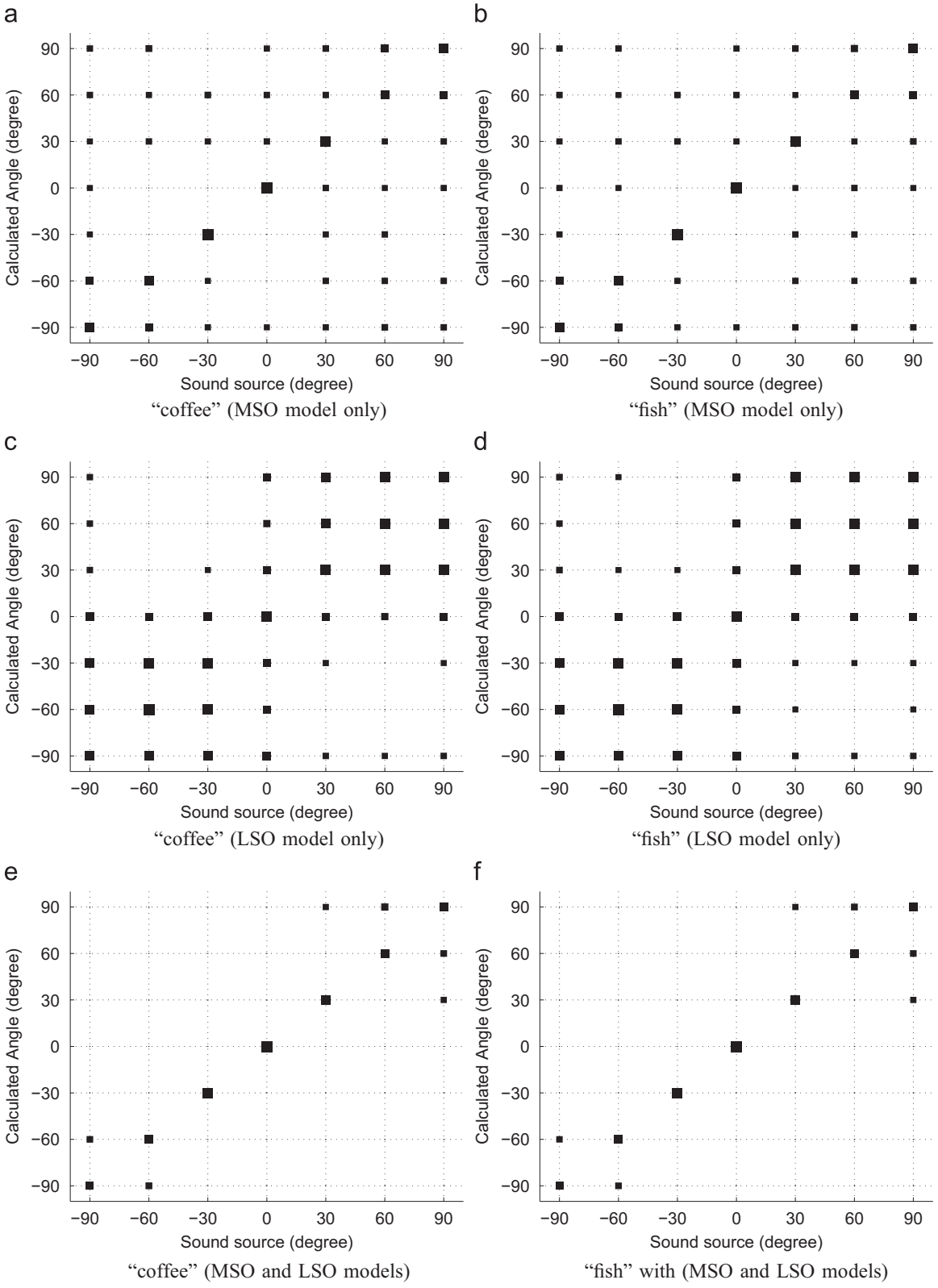


Fig. 13. The results of the sound localisation system for words: “coffee” (a, c, e) and “fish” (b, d, f). The size of the squares represents the proportion of spikes in the corresponding angle estimation.

with the LSO input are more precise and clearer than those only with MSO input.

Our experimental results match the biological evidence [15] that (i) the ITD and ILD cues have the highest efficiency for sound localisation when the sound source is in front of the observer, (ii) the utility of the ITD cue is limited at frequencies above 1 kHz, and (iii) the ILD is the main cue for the high frequency sound localisation.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

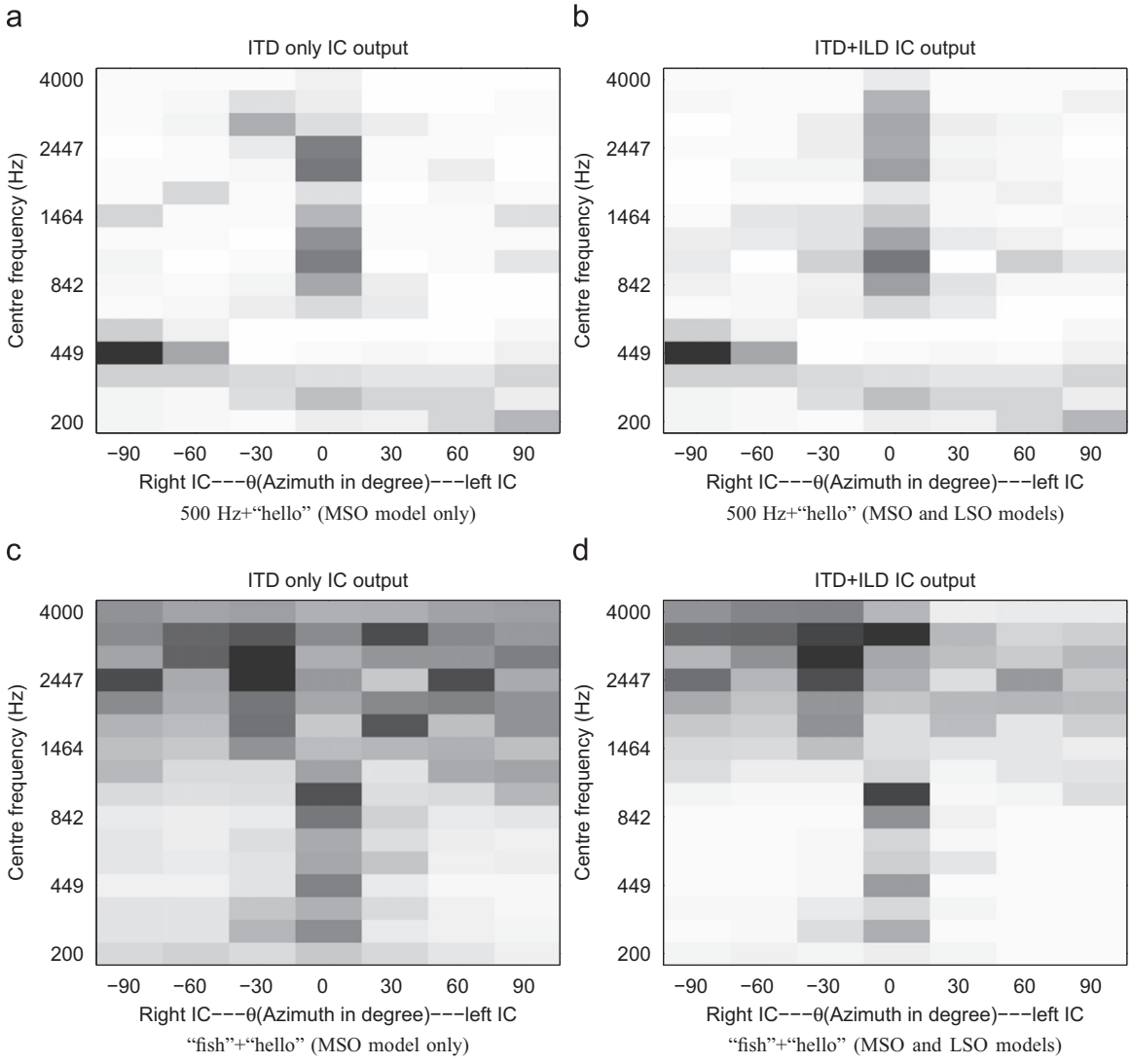


Fig. 14. The results of the sound localisation system for two sound sources played simultaneously. (a and b) 500 Hz pure tone at -90° and the word "hello" at 0° and (c and d) words "fish" at -30° and "hello" at 0°.

The whole system is built on a PC with Intel Dual-core 1.66 GHz CPU and 2.0 GB memory. The programming tools are based on Matlab 2007b. The average computation time for one second stereo sound data (sample rate 44,100 Hz) is 3.1 s. Considering Matlab's low efficiency in code compiling, our system could run at least twice as fast if using C/C ++ language, i.e. 1.5 s computation time for one second sound data. So that we can say, our system have potential practical applications for robot sound localisation. For further details, see [16].

5. Conclusion and future work

This paper describes the design and implementation of a sound localisation model that uses an SNN inspired by the mammalian auditory system. In this system, both ITD and ILD pathways were modelled based on neurophysiological theories and data. ITD and ILD spikes were computed in the MSO and LSO models, and they were directed to the IC in a similar manner to the biological system where they were merged together to achieve broadband sound localisation. The experimental results showed that our system can localise a broadband sound source from the -90° to 90° azimuth for frequencies between 200 to 4000 Hz. The effect of frequency and sound source position on localisation efficiency showed a high correspondence with neurophysiological data.

One of the main limitations of our system is its reduced efficiency when the sounds come from very lateral positions. The accuracy of the system decreases when the sound azimuth is $>60^\circ$. This problem can be traced back to the gain distributions calculated for the MSO and LSO projections (Figs. 8a and 9a) which indicate that something in the sound calibration or the environment influenced the initial calculation of the ITD and ILD gains.

In the future, active sound localisation, which can specify the feature frequencies of an interesting object, will be the next area of our research. For the application of our system to a mobile robot, we plan to implement a self-calibrating sound localisation system which can adaptively adjust the synapse and soma parameters according to the environment. Such a system would allow a robot to focus actively on a speaker, and improve its speech recognition performance in a noisy environment.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030

Acknowledgement

This work was supported by EPSRC (EP/D055466 to SW and HE and EP/D060648 to AR). We would also like to thank Chris Rowan for building the robot head.

Appendix A. Spectrograms of the recorded speech sounds used in this paper

Figs. A1–A3 show the spectrograms of the sound recordings used in this paper. In each figure, the spectrogram of the recording from each ear is shown.

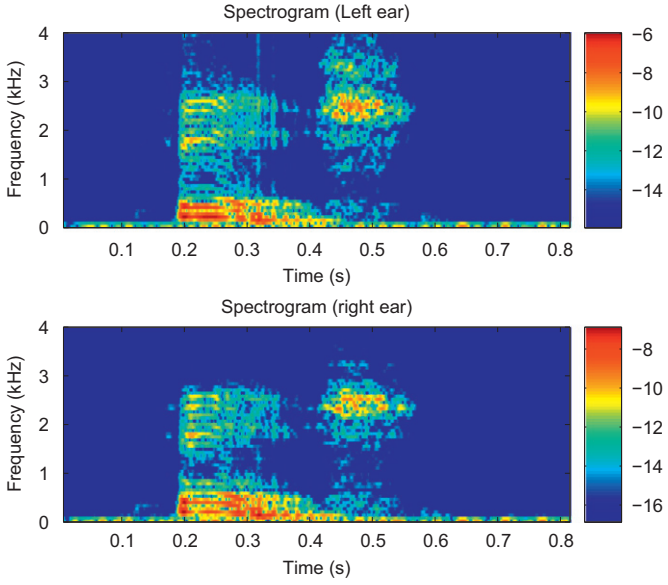


Fig. A1. The spectrogram of the recording when "fish" was played at -60°.

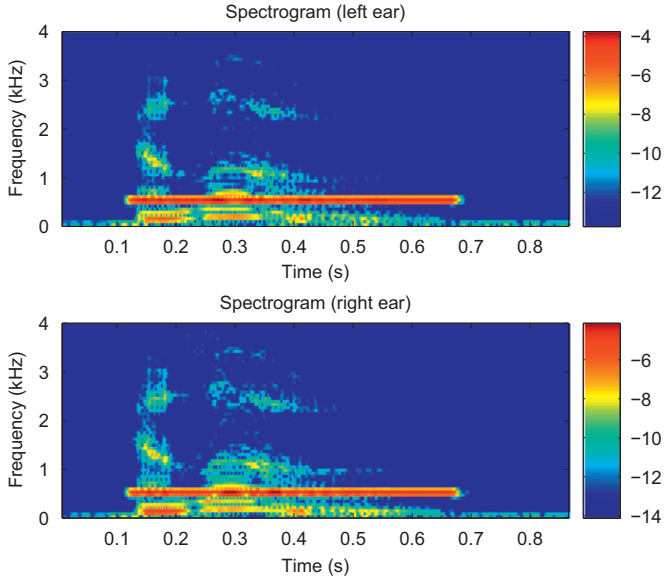


Fig. A2. The spectrogram of the recording when "hello" was played at 0° and pure tone 500 Hz at -90°.

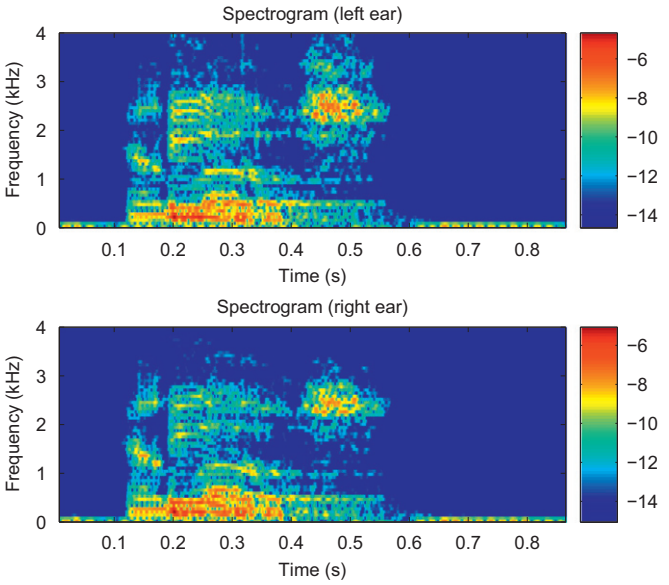
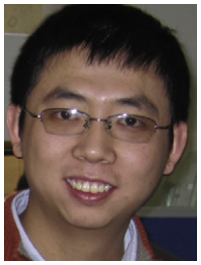


Fig. A3. The spectrogram of the recording when "fish" was played at -30° and "hello" played at 0°.

References

[1] N.A. Bhadkamkar, Binaural source localizer chip using subthreshold analog cmos, Proceedings of the 1994 IEEE International Conference on Neural Networks Part 1 (of 7), vol. 3, IEEE, Orlando, FL, USA1994, pp. 1866–1870.
[2] A. Brand, O. Behrend, T. Marquardt, D. McAlpine, B. Grothe, Precise inhibition is essential for microsecond interaural time difference coding, Nature 417 (6888) (2002) 543–547.
[3] W. Gerstner, W.M. Kistler, Spiking Neuron Models, Single Neurons, Populations, Plasticity, Cambridge University Press, 2002.
[4] K. Glendenning, R. Masterton, Acoustic chiasm: efferent projections of the lateral superior olive, Journal of Neuroscience 3 (8) (1983) 1521–1537.
[5] J.J. Guinan, S.S. Guinan, B.E. Norris, Single auditory units in the superior olivary complex: II: locations of unit categories and tonotopic organization, International Journal of Neuroscience 4 (3) (1972) 147–166.
[6] L. Jeffress, A place theory of sound localization, Journal of Comparative & Physiological Psychology 41 (1948) 35–39.
[7] B. Moore (Ed.), An Introduction to the Psychology of Hearing, Academic Press, San Diego, 2003.
[8] J. Nix, V. Hohmann, Sound source localization in real sound fields based on empirical statistics of interaural parameters, The Journal of the Acoustical Society of America 119 (2006) 463–479.
[9] D. Oertel, R. Fay, A. Popper (Eds.), Integrative Functions in the Mammalian Auditory Pathway, Springer, New York, 2002.
[10] T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, B. Scholling, Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple Mapping, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006, pp. 860–865.
[11] M. Slaney, An efficient implementation of the Patterson–Holdsworth auditory filter bank, Apple Computer Technical Report 35 (1993).
[12] P. Smith, P. Joris, T. Yin, Projections of physiologically characterized spherical bushy cell axons from the cochlear nucleus of the cat: evidence for delay lines to the medial superior olive, The Journal of Comparative Neurology 331 (1993) 245–260.
[13] K. Voutsas, J. Adamy, A biologically inspired spiking neural network for sound source lateralization, IEEE Transactions on Neural Networks 18 (6) (2007) 1785–1799.
[14] V. Willert, J. Eggert, J. Adamy, R. Stahl, E. Körner, A probabilistic model for binaural sound localization, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics 36 (5) (2006) 982–994.
[15] T. Yin, Neural mechanisms of encoding binaural localization cues in the auditory brainstem, Integrative Functions in the Mammalian Auditory Pathway (2002) 99–159.
[16] M. Queiroz, R. Berrêdo, A. Pádua Braga, Reinforcement learning of a simple control task using the spike response model, Neurocomputing 70 (2006) 14–20.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030



Jindong Liu received the MSc degree in Control Theory and Control Engineering from the Institute of Automation, Chinese Academy of Sciences in 2002 and the PhD degree in Robotics from University of Essex, UK in 2007.

Dr. Liu's research interests are in the area of biomimetic robotics, artificial intelligence, computational neuroscience of the auditory system and image processing. He was a Research Scientist in the Faculty of Applied Science, University of Sunderland, UK from 2007 to 2010. He is now a research associate in Department of Computing, Imperial College London, UK.

David Perez-Gonzalez received the BSc degree in Biology and the PhD degree in Neuroscience from the University of Salamanca, Salamanca, Spain in 2002 and 2007, respectively.

His research interests include the organisation and function of the inferior colliculus and the auditory system. He is now a Research Associate at the Instituto de Neurociencias Castilla y León, University of Salamanca, Spain.

Adrian Rees received his MA and DPhil degrees in Physiological Sciences from Oxford University. His research interests are in auditory neuroscience with a particular focus on the structure and functional organisation of the midbrain auditory nuclei, and the encoding of complex sounds. He pursues these questions using a multidisciplinary approach that ranges from single neuron recording to human psychophysics. He is currently Reader in Auditory Neuroscience in the Institute of Neuroscience at Newcastle University, UK.

Harry Erwin received his BSc degree in Mathematics from UC Davis in 1968 and an MSc in Mathematics from UC San Diego in 1971. His PhD was awarded in Computational Biology by George Mason University in 2000. Dr. Erwin's research interests are in the computational neuroscience of the auditory system, biomimetic robotics and sensorimotor integration in bats. He is a Senior Lecturer in the Faculty of Applied Science, University of Sunderland.

Stefan Wermter is Full Professor of Knowledge Technology at the University of Hamburg and Director of the Centre for Knowledge Technology. He studied Computer Science with a minor in Medicine at the Universities of Dortmund and Bochum. He also holds an MSc from the University of Massachusetts in Computer Science, and PhD and Habilitation in Computer Science from the University of Hamburg. He has been a Research Scientist at the International Computer Science Institute in Berkeley before accepting the Chair in Intelligent Systems at the University of Sunderland. His main research interests are in neural networks, hybrid systems, cognitive neuroscience, cognitive robotics and natural language processing.

Please cite this article as: J. Liu, et al., A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing (2010), doi:10.1016/j.neucom.2009.10.030