A Biomimetic Spiking Neural Network of the Auditory Midbrain for Mobile Robot Sound Localisation in Reverberant Environments

Jindong Liu, David Perez-Gonzalez, Adrian Rees, Harry Erwin and Stefan Wermter

Abstract— **This paper proposes a spiking neural network (SNN) of the mammalian auditory midbrain to achieve binaural sound source localisation with a mobile robot. 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 source over a wide frequency range in situations where there is auditory clutter and reverberation. 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 using Bayes's theorem to estimate the azimuthal direction of a sound source. Two of known IC cells, onset and sustained-regular are modelled. The azimuth estimations at different robot positions are then used to calculate the sound source position by a triangulation method using an environment map constructed by a laser scanner. The experimental results show that the addition of ILD information significantly increases sound localisation performance at frequencies above 1 kHz. The mobile robot is able to localise a sound source in an acoustically cluttered and reverberant environment.**

Index Terms*— *Spiking neural network, sound localisation, inferior colliculus, interaural time difference, interaural level difference, mobile robotics

I. 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 discussed in this paper aims to explore sound processing in the mammalian brain and 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 and better elucidated [1]. Binaural sound localisation systems take advantage of two important cues derived from the sound signals arriving at the ears: (i) interaural time difference (ITD), and (ii) interaural level difference (ILD) [2] . 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$1.5 kHz) [3], however, the information it provides becomes ambiguous for frequencies above $\sim$1 kHz. In contrast, the ILD cue has limited utility for localising sounds below 1 kHz, but is more efficient than the ITD cue for mid- and high- frequency ($>$1.5 kHz) sound localisation [3]. 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 [1].

Several hypotheses for ITD and ILD processing have been proposed [2][4][5], with the most influential being a model advanced by Jeffress [2]. In his model, ITDs are extracted by a mechanism in which neural activity triggered 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 [4] 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. More recently an alternative to the delay-line hypothesis has been proposed to explain ITD sensitivity based on evidence that it is generated by inhibitory mechanisms [5], however the precise mechanism underlying ITD sensitivity is beyond the scope of this paper.

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 the 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 [6] proposed a system to process ITDs using a CMOS circuit, while Willert [7] built a probabilistic model which separately measures ITDs and ILDs at a number of frequencies for binaural sound localisation. Recently, Voutsas and Adamy [8] built a multi

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

J. Liu, H. Erwin and S. Wermter are in the Faculty of School of Computing and Technology, University of Sunderland, Sunderland, SR6 0DD, United Kingdom jindong.liu, harry.erwin, stefan.wermter@sunderland.ac.uk, www.his.sunderland.ac.uk

D. Perez-Gonzalez, A. Rees are in the Institute of Neuroscience, The Medical School, Newcastle University, NE2 4HH, United Kingdom david.perez-gonzalez, adrian.rees@newcastle.ac.uk

978-1-4244-3553-1/09/$25.00 ©2009 IEEE

delay-line model using spiking neural networks (SNN) incorporating realistic neuronal models. Their 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 multiple cues: for example, Rodemann [9] applied three cues for sound localisation, however this model did not take advantage of the biological connections between the superior olivary complex (SOC) and the IC. Willert [7] and Nix [10] 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, their model did not use spiking neural network 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's 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 model was tested in a reverberant environment, using IC cells with two different firing patterns: onset and sustained-regular. To verify our model, it was used to direct a mobile robot to search for a sound source in an acoustically cluttered environment.

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

II. Biological Fundamentals and Assumptions

When sound waves arrive at the ears they are transduced by the cochlea into spikes in auditory nerve (AN) fibres which transmit the encoded information to the central nervous system. Each auditory nerve fibre is maximally sensitive to a characteristic frequency (CF) [1]. 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 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 later extraction of ITDs because it represents the basis for comparing the relative timing of the waveforms at the ears. Figure 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).

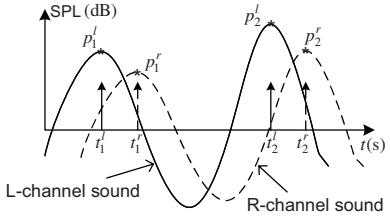


Fig. 1: An example of sound signals arriving at both ears (left, continuous line; right, dashed line), and the phase-locked spikes (t_1^l, t_1^r, t_2^l and t_2^r) triggered by them. The signal corresponding to the right ear is delayed and has smaller amplitude than the left one, indicating that the origin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

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 Figure 1) as the input to the ILD processing.

After the temporal and amplitude information is encoded and extracted, the spikes from each ear are transmitted to the MSO and LSO in order to extract ITDs and ILDs respectively [1]. The MSO on one side receives excitatory inputs 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 Jeffress's original model, activation of these coincidence detectors occurs 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 distributed along two dimensions: CF and ITD [11] (see Figure 2). The output of the MSO cells is transmitted to the ipsilateral 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 [1]. 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 (Figure 3). 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 yet, we know the spike rate of LSO neurons depends on 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 (Figure 3). 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 [11], while the LSO is a high-frequency structure with little representation below 1 kHz

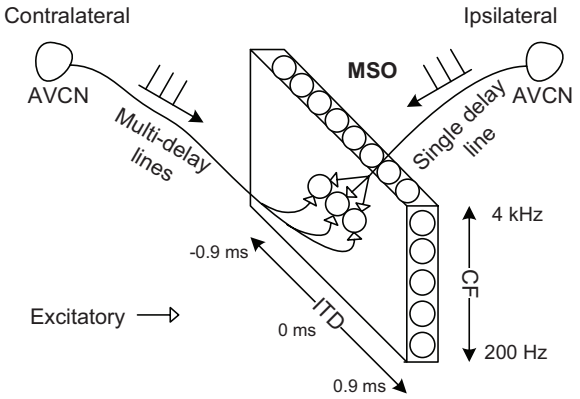


Fig. 2: Schematic diagram of the MSO model used in our system. While all the spike trains from the ipsilateral AVCN share the same delay, the ones originated in the contralateral side are subjected to variable delays. The difference between the ipsilateral and contralateral delays makes each cell in 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.

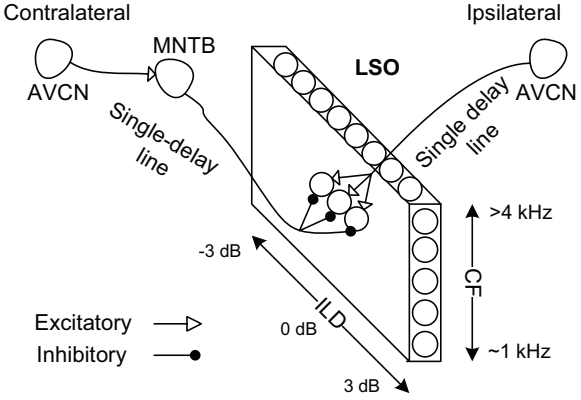


Fig. 3: Schematic diagram of the LSO model used in our system. Similarly to Figure 2, we assume that there are cells most sensitive to a given ILD and frequency. The ILD sensitivity is caused by the interaction of excitatory (ipsilateral) and inhibitory (contralateral) inputs. Our system could detect ILDs from -3 to 3 dB. (The LSO model contained neurons sensitive to frequencies between $\sim$1 kHz and 4 kHz.

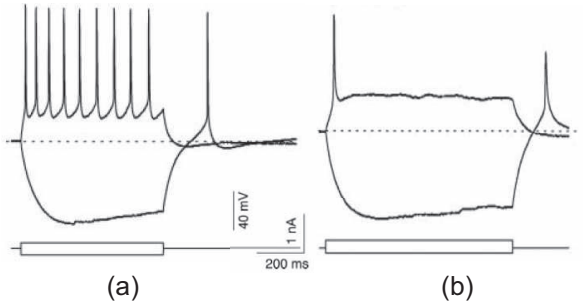


Fig. 4: Two IC cells modelled in this paper: (a) sustained-regular cell (b) onset cell. The square wave at the bottom of each figure indicates two kinds of input current: negative and positive. The corresponding response of the cell is drawn at the top of each figure. In this paper, we only modelled the situation when the cell receives positive input. (Figures adapted from [13].)

[12]. The inferior colliculus (IC) is 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 that range of frequencies.

The cells in the IC can be classified into 6 physiological types [13]: sustained-regular, rebound-regular, onset, rebound-adapting, pause/build and rebound-transient. We tested two of these, the onset and sustained-regular cells in our model. The sustained-regular cell fires with a constant spike rate when driven by a constant inward excitatory current (Figure 4-(a)) and thus provides a measure of the presence of an ongoing sound. In contrast, the onset cell (Figure 4-(b)) only generates a single spike in response to the same inputs.

Taking into account this biological evidence, we propose an IC model for sound source localisation as outlined in Figure 5. It consists of different components according to the frequency range: at low frequencies, as shown in Figure 5a, only the ipsilateral MSO is involved in sound localisation; while in the middle frequency range, shown in Figure 5b, the ipsilateral MSO and both LSOs contribute inputs to the IC. 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. The signs and patterns of these connections are based on the available biological data as discussed above [14].

III. System Model of Sound Localisation

Inspired by the neurophysiological findings and the proposed models presented in Section II, 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 (soma) and several synapses which connect the elements of the network.

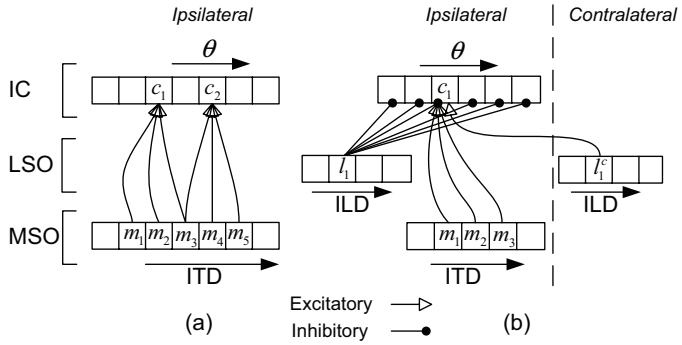


Fig. 5: Schematic diagram of the distribution of inputs to our IC model. We assume there are no connections across frequencies. (a) In the range 200 Hz to 1 kHz, the IC model only receives inputs from the MSO model. (b) From 1 kHz to 4 kHz, the IC model receives inputs from the MSO model and both LSO. The distributions of the projections were calculated using Bayesian statistics (see Section III for details).

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 [15]:

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

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, which will affect the temporal integration of the inputs. In this paper, the value of τ_m is 1.6 ms based on typical biological data. C is the capacitance which is charged by $\sum_k I_k(t)$ from multiple inputs, in order to simulate the postsynaptic current charging the soma. k is the index of synaptic input. The action potential threshold φ controls the firing time t_f. When $u(t) = \varphi$, the soma fires a spike; then $u(t)$ is reset to 0. Afterwards, the soma enters a refractory state for $t_r = 1$ ms during which it is not responsive to any synaptic input. After the refractory period, the soma returns to its resting potential. The difference between the modelled onset cells and the sustained regular cells in IC is the setting of φ. For the sustained regular cell this value is constant, while for the onset cell, after a spike is triggered, φ is set to a positive infinite value to prevent further spiking until there is no more synaptic input for a period t_s=10 ms.

A schematic structure for the sound localisation procedure is shown in Figure 6. The frequency separation occurring in the cochlea is simulated by a bandpass filterbank consisting of 16 discrete second-order Gammatone filters [16], which produces 16 frequency bands between 200Hz and 4kHz.

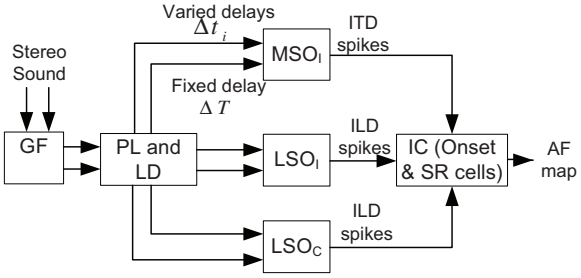


Fig. 6: Flowchart of the biologically inspired sound localisation system. This example only shows one IC; note that there is a symmetric model for the contralateral IC. $\mathrm{MSO_I}$, ipsilateral MSO model; $\mathrm{LSO_I}$, ipsilateral LSO model; $\mathrm{LSO_C}$, contralateral LSO model; GF, Gammatone filterbank; PL, phase locking; LD, level detection; SR, sustained-regular and AF, azimuth-frequency

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 shown in Figure 6, which simulates the halfwave rectified receptor potential of the inner hair cells in the cochlea that leads 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 at the peak phase is detected at the same time (Figure 1) and then 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 2). The spike train of the ipsilateral ear reaches the MSO with a single fixed delay time ΔT. The cells in the MSO are modelled with the parameters in Table I.

The ILD pathway is not modelled directly 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 (Figure 3). When a specific level difference 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 resulting ITD and ILD spikes are fed into the IC model, as shown in Figure 6. The IC model merges the information to obtain a spatial representation of the azimuth of the sound source. According

TABLE I: Parameters for the MSO and IC model

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

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

to the model proposed in Section II, 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 [7], we used an approach based on conditional probability to calculate these gains, as shown in the following functions:

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

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

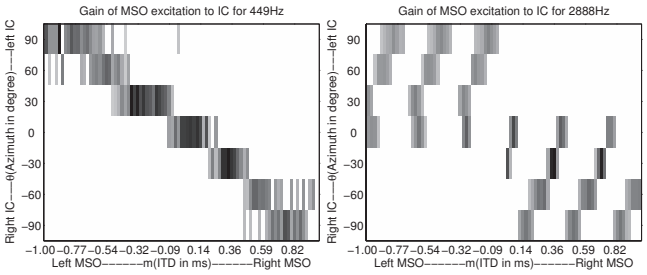
$$c_{l_i\theta_j} = \begin{cases} \text{if} p < 0.6 \max_j(p(\theta_j \,|l_i\ , f)), f >= f_b : \\ 1 - p(\theta_j \,|l_i\ , f) \\ \text{otherwise} : \\ 0 \end{cases} , \quad (4)$$

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 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 (see below), 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. For example, $p(\theta_j \,|m_i\ , f)$ is:

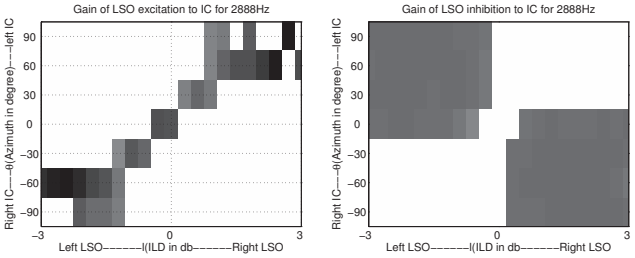
$$p(\theta_j \,|m_i\ , f) = \frac{p(m_i \,|\theta_j\ , f) p(\theta_j \,|f\,)}{p(m_i \,|f\,)} \quad (5)$$

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 azimuth angles (from -90 to 90 degrees in 30 degree steps) using a robot head. The head has 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]



(a) 449Hz (b) 2888Hz

Fig. 7: 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.



(a) 2888Hz, contralateral LSO (b) 2888Hz, ipsilateral LSO

Fig. 8: Gain of the projection from the contralateral (8a, excitatory) and ipsilateral (8b, 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.

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. Figure 7 shows the gain calculated for the MSO projection at two different frequencies, (a) 449Hz and (b) 2888Hz.

A similar procedure is used to calculate the gains of the LSO projection to the IC. Figure 8 shows the gains calculated for this projection, at 2888Hz. Figure 8a shows the excitatory

[1]Sounds were recorded in a low noise environment (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.

contralateral projection, while Figure 8b shows the inhibitory ipsilateral projection.

Equations 2, 3 and 4 used to calculated the gains for the projections between the MSO or LSO and IC have two features: (i) 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 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, Equation 2 implies that there can be multiple m_i that have an active connection to a single IC cell θ_j. For example, in Figure 7 a sound coming from a single azimuth (e.g. 30 degrees) causes multiple MSO cells to respond, to different extents (e.g. cells tuned at -0.54 to -0.09 ms ITD). Furthermore, Equation 3 defines a few-to-one projection from the contralateral LSO to the IC (Figure 8a), while Equation 4 shows a one-to-all projection from ipsilateral LSO to the IC (refer to Figure 8b).

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.

IV. Experimental Results

The model was tested in conjunction with a mobile robot using real sounds. Two groups of experiments were designed: (i) sound source azimuth detection with a stationary robot, and (ii) sound source localisation with a moving robot.

A. Azimuth Detection by a Stationary Robot

In this experiment, four types of sound sources were employed: clicks, white noise, pure tone and speech. They were presented at different azimuths to the stationary robot. The click was 0.04 ms in duration, and the pure tone sounds was 1s and included 500, 1000, 1500, 2000 and 3000 Hz. The speech sounds included five words in English: ”hello”, ”look”, ”fish”, ”coffee” and ”tea”. The robot head is equipped with two omnidirectional microphones and half cones to provide simple pinna.

Figure 9 shows the accuracy of sound source localisation using a model with sustained-regular IC cells. The broadband sound sources such as the click, white noise and speech are generally well localised. In contrast, the localisation performance of pure tones is less accurate. For sounds with more complex spectra and time structures, reverberations are less likely to coincide at the robot's ear with the same frequency as the sound taking the direct path, thus the echo interferes less with the direct sound. With pure tones, however, the frequency of the echo always has the same frequency as the direct sound so resulting in greater interference. As a consequence the sustained-regular cell, whose output reflects the resultant signal does not give an accurate representation of the sound's location.

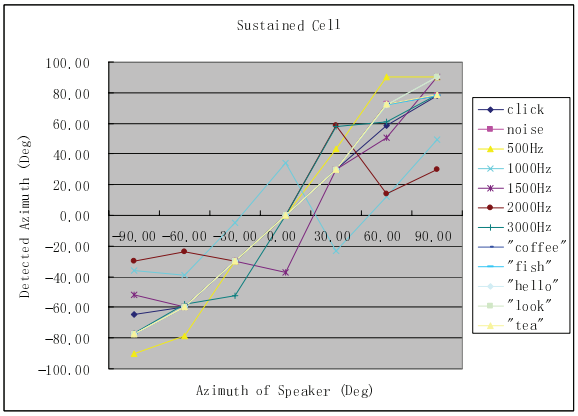


Fig. 9: Accuracy of sound source localisation by a stationary robot, using sustained-regular IC cells

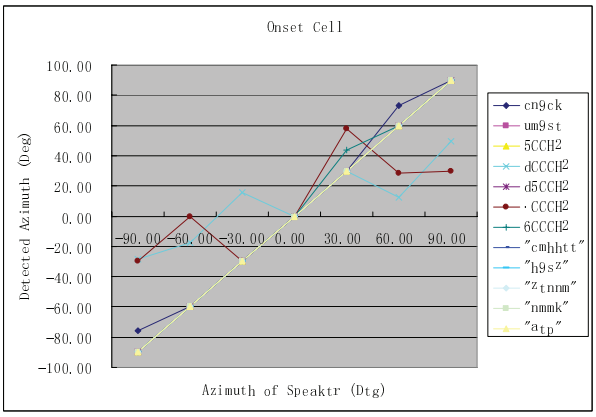


Fig. 10: Accuracy of sound source localisation by the stationary robot, using Onset IC cells. Note that the plot for white noise, pure tones at 500 and 300 Hz, and all speech sounds overlap (yellow line).

The results of localisation using a model that incorporates onset cells are shown in Figure 11. Onset cells provide a more accurate estimate of the location of a sound source than sustained-regular cells for all the sounds tested. However, the performance is poor for pure tones at 1000 and 2000 Hz; possibly because these frequencies coincide with the resonant frequencies of the robot pinnae.

B. Sound Source Location by a Mobile Robot

To test the performance of our model in a moving robot in reverberant environment, we implemented it on a PeopleBot mobile robot which is equipped with the same robot head as in IV-A and a laser scanner. With the assistance of the laser, the mobile robot is expected to localise not only the azimuth but also the distance of the sound source. Figure 11 shows the basic triangular relationship between one sound source and two robot positions. In a 2-dimensional plane, we define four states for the robot performing a sound localisation: (x, y, α, θ), where x, y are the robot position, α is the robot orientation and θ is the azimuth angle of the sound source detected at this position and orientation. When the robot

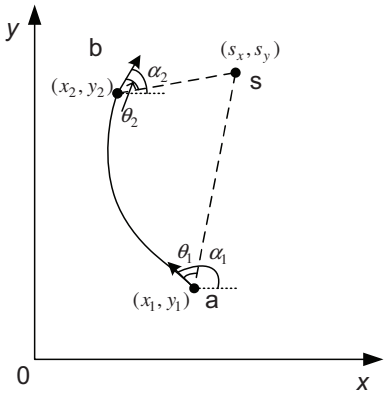


Fig. 11: Localisation of a sound source by a mobile robot. In the figure, the robot moves from point **a** to **b** via a trajectory **ab**. The robot measures the azimuth of the sound source (**s**) at both points **a** and **b**.

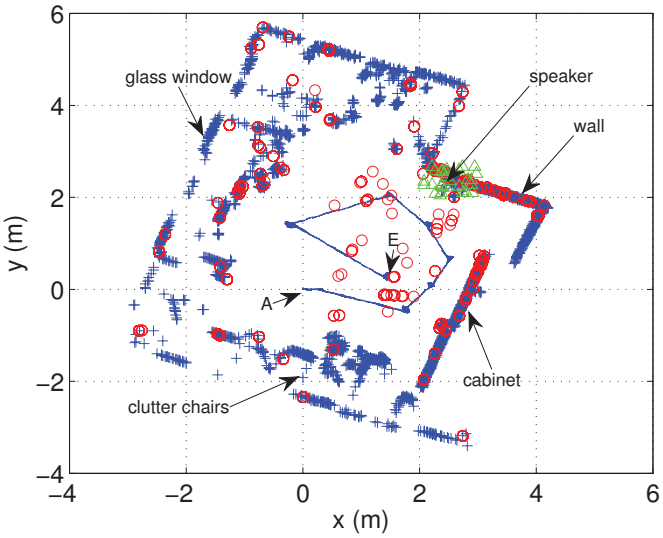


Fig. 12: Sound source localisation by a mobile robot. In the figure, the robot moves from point A to E. Blue points indicate the trajectory of the robot, blue crosses represent the map built using the laser scanner; red circles indicate the sound source localisation calculated using sustained-regular IC cell in the model; and green triangles show the sound source localisation using onset IC cells.

moves from point **a** with status $(x_1, y_1, \alpha_1, \theta_1)$ to point **b** with status $(x_2, y_2, \alpha_2, \theta_2)$, the sound source position (s_x, s_y) can be estimated by triangulation as:

$$\begin{aligned} s_x &= \frac{(x_1 t_1 - x_2 t_2) - (y_1 - y_2)}{t_1 - t_2} \\ s_y &= \frac{(x_1 - x_2) t_1 t_2 - (y_1 t_2 - y_2 t_1)}{t_1 - t_2} \end{aligned} , \qquad (6)$$

where $t_1 = \tan(\alpha_1 - \theta_1)$ and $t_2 = \tan(\alpha_2 - \theta_2)$.

In order to update the position of the robot, we adapted Ryde's algorithm [17] using a laser scanner and assumed that the sound source is inside the room and located on a tripod at the height of the robot head. Currently, the model only detects sound sources located in front of the robot (from -90 degrees left to 90 degrees right). Therefore, if a sound source position calculated from Equation 6 is at the back of the robot the position is ignored. This position checking can be done by calculating $\beta = (s_x - x)\cos(\alpha) + (s_x - y)\sin(\alpha)$. The calculated sound source position is in front of the robot when $\beta \geq 0$. Figure 12 shows the result of a roving sound source localisation task, using a 3000 Hz pure tone (1 second on, 1 second off). The model using onset cells (green triangles) accurately reflects sound source location (with 0.25m standard deviation) despite the effect of echoes, which severely affect localisation when sustained-regular cells are used in the model. (red circles). However, although sustained-regular cells are poorer at localising sounds in this reverberant environment, many of the failed localisations coincide with the position of reverberating objects in the room (compare with the positions detected by the laser, blue crosses), such as the cabinet, wall and the window. In the future, this information can be used as a feedback to detect objects and the boundaries of the room, as well as to adapt our model to localise sound sources outside of the room. In the real auditory system it is likely that such information might contribute to the impression of acoustic ambience that we experience in spaces of different sizes and properties.

V. 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. In this system, both ITD and ILD pathways were modelled based on neurophysiological theories and data. Firing patterns representing ITD and ILD information were computed in models of the MSO and LSO respectively, and, in a similar manner to the biological system, these were projected to the IC where they were merged together to achieve broadband sound localisation. Two types of IC cells were tested in the model. Onset cells were particularly effective in minimising the errors arising from reverberations. The experimental results show that our system can localise a broadband sound source in the range -90 to 90 degrees of azimuth, with sound frequencies in the range 200 to 4000 Hz. Preliminary sound localisation experiments with a mobile robot show that our model is not limited to static situations. The performance of the model suggests its potential application in situations where objects can only be located by their sound, or where sound source detection can offer advantages other methods such as vision, for example in rescue operations in the dark.

In the future, models that incorporate other types of IC cells will be investigated with the aim of implementing sound segregation based on sound localisation to enhance signal recognition in a cluttered environment.

Acknowledgment

This work is 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, Julian Ryde for processing the laser data to generate an accurate map, and Mahmoud Elsaid for assistance on robot experiments.

REFERENCES

[1] T. Yin. Neural mechanisms of encoding binaural localization cues in the auditory brainstem. *Integrative Functions in the Mammalian Auditory Pathway*, vol. I, 2002:pp. 99–159.

[2] L. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol.*, vol. 41, 1948:pp. 35–39.

[3] B. Moore. *An Introduction to the Psychology of Hearing (ed.)*. San Diego: Academic Press, 2003.

[4] P. Smith, P. Joris, and 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. *J. Comp. Neurol.*, vol. 331, 1993:pp. 245–260.

[5] A. Brand, O. Behrend, T. Marquardt, D. McAlpine, and B. Grothe. Precise inhibition is essential for microsecond interaural time difference coding. *Nature*, vol. 417(6888), May 2002:pp. 543–547. ISSN 0028-0836.

[6] N. A. Bhadkamkar. Binaural source localizer chip using subthreshold analog CMOS. In *Proceedings of the 1994 IEEE International Conference on Neural Networks. Part 1 (of 7)*, vol. 3. IEEE, Orlando, FL, USA, 1994, pp. 1866–1870.

[7] V. Willert, J. Eggert, J. Adamy, R. Stahl, and K. E. A probabilistic model for binaural sound localization. *IEEE Trans Syst Man Cybern Part B Cybern*, vol. 36(5), 2006:pp. 982–994. ISSN 10834419.

[8] K. Voutsas and J. Adamy. A biologically inspired spiking neural network for sound source lateralization. *IEEE Trans Neural Networks*, vol. 18(6), 2007:pp. 1785–1799. ISSN 10459227.

[9] T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, and B. Scholling. Real-time Sound Localization With a Binaural Head-system Using a Biologically-inspired Cue-triple Mapping. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. 2006, pp. 860–865.

[10] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *The Journal of the Acoustical Society of America*, vol. 119, 2006:pp. 463–479.

[11] J. J. Guinan, S. S. Guinan, and B. E. Norris. Single Auditory Units in the Superior Olivary Complex: II: Locations of unit categories and tonotopic organization. *International Journal of Neuroscience*, vol. 4(3), 1972:pp. 147–166. ISSN 0020-7454.

[12] K. Glendenning and R. Masterton. Acoustic chiasm: efferent projections of the lateral superior olive. *J. Neurosci.*, vol. 3(8), Aug. 1983:pp. 1521–1537.

[13] S. Sivaramakrishnan and D. Oliver. Distinct K Currents Result in Physiologically Distinct Cell Types in the Inferior Colliculus of the Rat. *Journal of Neuroscience*, vol. 21(8), 2001:p. 2861.

[14] D. L. Oliver. *The Inferior Colliculus*, chap. Neuronal Organization in the Inferior Colliculus. Springer: New York, 2004, pp. 69–114.

[15] W. Gerstner and W. M. Kistler. *Spiking Neuron Models, Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

[16] M. Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer Technical Report*, vol. 35, 1993.

[17] J. Ryde and H. Hu. Mutual localization and 3D mapping by cooperative mobile robots. In *Intelligent Autonomous Systems*. The University of Tokyo, Tokyo, Japan, Mar. 2006.