SOUND LOCALISATION WITH A SILICON COCHLEA PAIR

André van Schaik, Vincent Chan, and Craig Jin

School of Electrical and Information Engineering
University of Sydney
Sydney, NSW 2006, Australia

ABSTRACT

A neuromorphic sound localisation system is proposed. It employs two microphones and a pair of silicon cochleae with address event interface for front-end processing. This allows subsequent processing to be implemented with spike-based algorithms. The system is adaptive and supports online learning. Its localisation capability was tested with white noise and pure tone stimuli, with an average error of around 3° in the -45° to 45° range.

Index Terms—*Auditory neuroscience, neuromorphic engineering, sound localisation*

1. INTRODUCTION

Sound localisation is the ability to identify the direction of a sound and is a key to survival in the animal world. It is used by many predators to hunt effectively, sometimes in complete darkness. On the other hand, preys capable of locating sound accurately will have a higher chance to escape. In robotics, however, sound localisation has received much less focus compared to vision. Nevertheless, sound localisation is expected to become more important as robots are required to operate in the real world and must respond to both visual and audio stimuli.

Two important cues for sound localisation in biology are interaural time difference (ITD), and interaural intensity difference (IID), also known as interaural level difference (ILD). These two cues complement each other – ITD is dominant at low frequency where phase difference can be uniquely and accurately identified, while at high frequency IID is more perceivable as wavelength becomes smaller than the size of head. Many existing sound localisation systems use only ITD to localise sound (e.g. [1]-[4]) because it can be measured accurately and is less dependent on the frequency content of the signal. The microphones in some of these systems are suspended in the air or mounted on a plane so that no IID is available.

In this paper we propose a neuromorphic sound localisation system based on ITD extraction. It is biologically realistic as a cochlea chip with spiking output is used at the front-end and all sub-sequent processing is either spike-based, or can be easily mapped to its spike-based counter-part. Unlike previous implementation involving silicon cochleae [2], [3], our system is adaptive and does not require any prior knowledge of the ITD model.

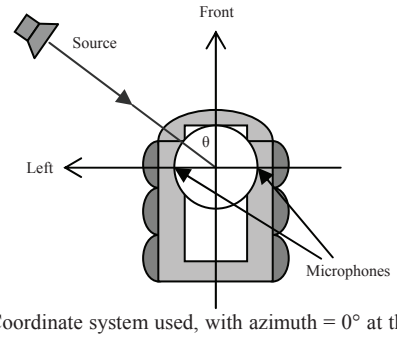

Figure 1. Coordinate system used, with azimuth = 0° at the front, 90° on the left, -90° on the right, and ±180° at the back.

2. THE LOCALISATION ALGORITHM

In previous implementations [2], [3], signals recorded by microphones are filtered by a left and a right cochlea. These cochleae consist of a number of sections and the output of each section is a band-pass filtered version of the input signal, and the centre frequency of the filters decreases exponentially along the cochlea. At each cochlear output section, i, cross-correlation with sampled time delay is performed on the cochlear outputs,

$$R_i[n] = \int x_{left}(t) \cdot x_{right}(t + nT_S)dt \tag{1}$$

where T_S is the resolution of the cross-correlator. ITD is then estimated by:

$$\hat{\tau} = T_S \cdot \arg\max\left(\sum_i R_i[n]\right) \tag{2}$$

It is assumed that ITD is independent of frequency and depends purely on the direction of the source, θ (Figure 1),

$$\tau = f(\theta) \tag{3}$$

Therefore, the estimated direction is computed simply by applying the inverse function,

$$\hat{\theta} = f^{-1}(\hat{\tau}) \tag{4}$$

However, if the microphones are mounted on a head, the introduced diffraction will cause $f(\theta)$ to be frequency dependent [5]. At frequencies under 500 Hz, f can be approximated by a sine function, but this changes gradually to $\sin(\theta) + \theta$ as the frequency increases above 1.5 kHz. Thus, different estimates will be given as the frequency of the source changes. The task is further

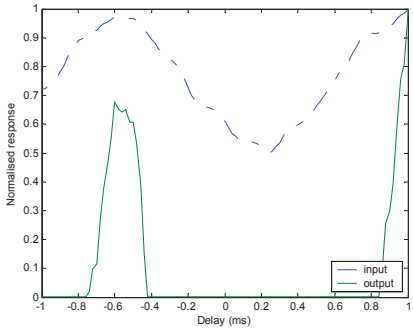

Figure 2. Application of soft-WTA to the result of cross-correlation. The stimulus is a 650-Hz pure tone with an ITD of approximately -0.6ms. This ITD information would have been lost if a normal WTA is used.

complicated when sound localisation is implemented in analogue VLSI as there will be mismatch in the delay lines used in the cross-correlator and phase mismatch between the left and right cochleae, which varies from section to section.

In our proposed system, at each cochlear section, we compute cross-correlation, $R_i[n]$, on the left and right cochlear outputs as before but the results are processed individually so that variations can be compensated using the following steps. The first step is to apply a soft-max function to each of the $R_i[n]$ separately, which captures not only the position of the maximum but also those similar in strength in each channel. This is useful when the input stimulus is a pure tone, resulting in a periodic cross-correlation function, where the result at the true time delay may not be the global maximum. The use of soft-max ensures that the true time delay will not be discarded, as shown in Figure 2. The soft-max function can be obtained using a soft-WTA network [6]. By adjusting the strength of the inhibition in a winner-take-all (WTA) network, one can vary the selectivity. In the example shown in Figure 2, the strength of the inhibition is set to 1 and both peaks are well-preserved.

The next step is to transform the soft-max result $S_i[n]$ to $G_i(\theta)$, a map of auditory activity on the horizontal plane with discrete azimuth angle. We can express the transformation in matrix form,

$$g_i = W_i \cdot s_i \tag{5}$$

where s_i = $[S_i[\text{-k}]\ S_i[\text{-k+1}]\ \ldots\ S_i[k]]^T$ and g_i = $[G_i(\theta_1)\ G_i(\theta_2)\ \ldots\ G_i(\theta_N)]^T$. Such a matrix multiplication is essential in artificial neural networks and has previously been implemented in VLSI [7]. W_i can be thought of as the synaptic connections between the neurons at the output of the soft-WTA and the neurons representing activity in the auditory space found in biology. This representation allows the system to learn sound localisation online by making small incremental changes to the values in W, enabling the system is to slowly adapt to a changing acoustic environment while in operation.

During supervised learning, for each cochlear section and each azimuth position, we present auditory stimuli to the system to produce the soft-max results s_i and $g_i = W_i s_i$. g_i is then compared with a target pattern, t_i, resulting in an error $err = t_i - g_i$. Then we can update W_i with the following rule[i] [8],

$$W_{new} = W_{old} - \varepsilon \cdot err \cdot s_i^T \tag{6}$$

with ε controlling the learning rate. The target functions used for our experiments are Gaussians centred at the known training positions (see section 3 for more detail).

After training, $G_i(\theta)$ will represent the likelihood that the sound arrived from direction θ, with a maximum at the actual source position (for a single source). Therefore, in frequency bands where there is sufficient energy from a single source, the peaks in correlation for that source will align and we can estimate the direction of the source by:

$$\hat{\theta} = \arg\max\left(\sum_i G_i(\theta)\right) \tag{7}$$

3. EXPERIMENTAL SETUP

The experimental setup is shown in Figure 1. A pair of electret microphones is mounted on opposite sides of a sphere, 15 cm in diameter, made of foam. The sphere itself is then fixed atop a robot, 15 cm from the ground. This sphere simulates the effect of head shadowing and diffraction introduced by the head.

A block diagram of the proposed localisation algorithm is shown in Figure 3. For demonstrative purpose, the complete system is simulated in MATLAB, except for the pair of silicon cochleae, which are implemented in hardware. The AER EAR chip [9] contains a matched pair of general purpose silicon cochleae with 32 sections, each having their own inner hair cell circuits and spiking neurons. The cochlea was tuned to cover the range of frequency from 200 Hz to 10 kHz. To simulate many fibres innervating a single cochlear region with our AER cochlea, which has only one output address for this region, we have used a high spike rate. Each channel generates, on average, 6000 spikes per second when a 100 mVpp sine wave of best frequency (BF) is presented.

Due to the limited dynamic range of the silicon cochleae, the signals from the microphones must first be conditioned by automatic gain control (AGC) so that the cochlea can operate under different sound levels without distortion. In biology, due the diminishing phase locking at high frequency in the inner hair cells, only the low frequency channels can be used to extract ITD. In our case, channels with best frequency above 3 kHz are ignored, leaving us with 19 sections. There is also an upper limit of 3 kHz for all stimuli.

[i] The update rule is very similar to that used in the back-propagation algorithm to train multi-layers neural networks.

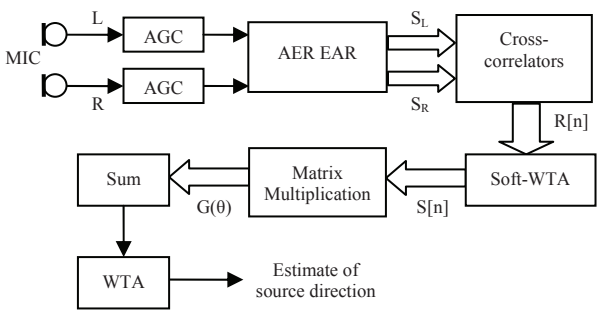

Figure 3. Block diagram of the proposed system. The block arrows represent signals in multiple frequency bands.

We record the impulse responses (IRs) of the microphones in response to a loudspeaker at different azimuth positions in an almost anechoic room. The walls are fitted with sound absorbing material to minimise reflection and the only major reflection comes from the floor, which is covered with thick carpet. The speaker is placed 2.6m away, at the same height as the sphere, and the impulse responses were recorded at 10° steps. These impulse responses allow us to present any stimulus to the AER EAR for both learning and testing, from different directions, by simply convolving the source signal with the appropriate left and right IRs. This method also allows simulated AGC to be applied to the signals before they enter the cochleae.

The outputs from the AER EAR are passed to the cross-correlator. The cross-correlator can be built using either shift registers or silicon axons [10] as delay lines and have neurons spiking if both inputs arrive within a short period of time, as proposed in Jeffress' model [11]. In our simulation, the cross-correlator contains 101 delay positions, from -1ms to 1ms with 20μs resolution.

For the soft-WTA network, although spiking WTA network have been demonstrated [12], for simplicity, WTA dynamics will not be simulated and only the steady-state output will be computed.

For the spatial map of auditory activity, 61 azimuth positions are used, covering the angles from -90° to 90° with 3° resolution. In each frequency channel, the weight matrix W_i is trained with band-limited noise stimuli with a bandwidth of $0.9f_C$ to $1.1f_C$, where f_C is the centre frequency of the channel. For each training example, we set the target t_i to be a Gaussian function centred at the expected position of the source, with $\sigma = 25°$. One of the advantages of choosing a Gaussian function instead of an impulse function is that it updates not only the weights going into the neuron representing the source position, but also those surrounding it. As a result, there is no need to provide training data at every position and the system will be able to interpolate upon successful training.

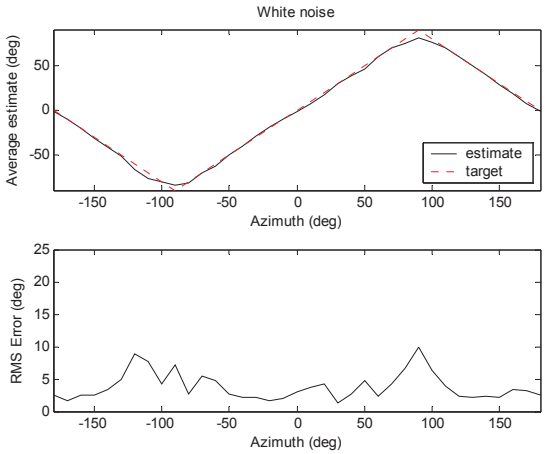

Figure 4. Localisation results and errors at different azimuth, for a white noise stimulus. RMS error is calculated at each position from 5 trials. RMS error across the [-90°, 90°] range is 4.4°.

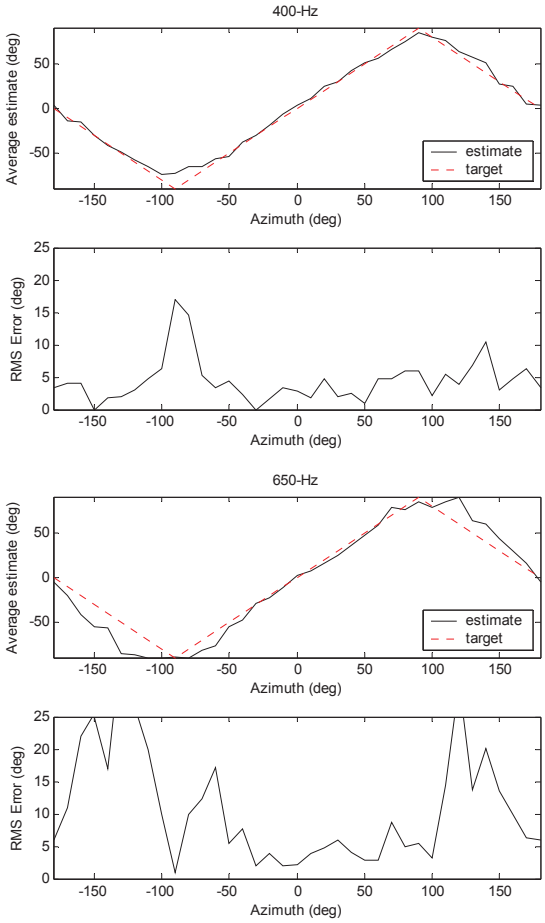

Figure 5. Localisation results and errors at different azimuth, for pure tone stimulus of 400-Hz and 650-Hz. RMS errors across the [-90º,90º] range are 6.2° and 6.9°, respectively. For the 650-Hz, there are big errors outside ±90º due to front-back asymmetry. This is a result of the head being mounted near the front of the robot and interference caused by the body of the robot when the sound arrives from the back.

Table 1. Comparison with other sound localisation systems

Localisation system	Localisation cues used	Stimulus	RMS Error (0°-45° / 45°-90°)
Current work	ITD	Noise	2.7 / 5.5
		Pure tone	3.7 / 8.4
[1]	ITD	Noise, <300Hz	3 / 7
[13]	IPD[i] + IED[ii] + IID + spectral cues	Impulse	5 / 5 (2-D localisation)
[14]	(a) ITD	Noise	2 / 3
	(b) IPD + IID	Noise	1 / 3
[15]	IPD + IID with motion	Noise	1 / 2
[16]	IPD + IID	Speech	3 / 12

4. RESULTS

We tested the performance of our proposed system with three types of stimuli – (a) band-limited white noise (up to 3 kHz), (b) a 400 Hz pure tone, and (c) a 650 Hz pure tone. Each stimulus lasts 100 ms and is multiplied by a Hanning window to remove any apparent onset and offset. These stimuli are convolved with the impulse responses of the target direction and amplified (simulated AGC) before they are played to the cochlea chip.

Figure 4 shows the localisation results when white noise is used. Note that with only 2 microphones, it is not possible to distinguish front-back ambiguity because the resultant ITD's (as well as IID's in the absence of the pinna) are exactly the same. Therefore, sources originating from ±[90°,180°] are mapped to ±[0°,90°]. We also plotted the localisation errors in Figure 4. The RMS error is 2.7° in the range ±[0°,45°] and increases to 5.5° in the range ±[45°,90°]. As expected, errors are largest around ±90° due to ITD being weakly sensitive to changes in θ, and a lower SNR at the far ear.

Figure 5 shows the localisation results for pure tones. Together, RMS error is 3.7° in the range ±[0°,45°] and increases to 8.4° in the range ±[45°,90°]. This drop in performance when a pure tone is used is similar to test results in humans. We compare the performance of our system with other 2-microphone systems in Table 1. The performance of our spike-based system is comparable against some implementations such as [1] and [16].

5. CONCLUSION

An ITD-based neuromorphic sound localisation system has been proposed. It uses a pair of silicon cochleae as a front-end and supports spike-based processing. By individually processing each frequency band, frequency dependent variations are overcome and the final system demonstrates the ability to accurately determine the position of the source for both pure tone and white noise stimuli, in contrast to many existing sound localisations which are tested with only one type of sound. The system can also compensate for systematic offset caused by circuit mismatch. Finally, this new architecture supports online learning, allowing the system to learn while in operation. Future work will concentrate on improving the learning algorithm as well as implementing the full system in hardware to enable real-time operation.

6. REFERENCES

1 P. Julian, A. Andreou, P. Mandolesi and D. Goldberg, "A low-power CMOS Integrated Circuit for bearing estimation," *Proceedings, IEEE International Symposium on Circuits and Systems*, 2003, vol. 5, pp. 305-308.

2 J. Lazzaro and C. Mead, "A silicon model of auditory localization," *Neural Computation*, vol. 1, pp. 41-70, 1989.

3 N. Bhadkamkar and B. Fowler, "Sound localization system based on biological analogy," *Proceedings, IEEE International Conference on Neural Networks*, 1993, vol. 3, pp. 1902-1907.

4 A. van Schaik and S. Shamma, "A neuromorphic sound localizer for a smart MEMS system," *Proceedings, IEEE International Symposium on Circuits and Systems*, 2003, vol. 4, pp. 864-867.

5 G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157-167, 1977.

6 G. Indiveri and T. Delbruck, "Current-mode circuits," in *Analog VLSI: Circuits and Principles*, S.-C. Liu, et al., Eds., MIT Press, Cambridge, MA; London, 2002, pp. 145-175.

7 T. Morie, "Analog VLSI implementation of self-learning neural networks," in *Learning On Silicon*, G. Cauwenberghs and M.A. Bayoumi, Eds., Kluwer Academic Publishers, Boston; Dordrecht; London, 1999, pp. 213-242.

8 J.A. Anderson, "Gradient Descent Algorithms," in *An Introduction to Neural Networks*, MIT Press, Cambridge, MA; London, 1995, pp. 239-279.

9 V. Chan, S.-C. Liu and A. van Schaik, "AER EAR: A Matched Silicon Cochlea Pair With Address Event Representation Interface," *IEEE Transactions on Circuits and Systems I*, vol. 54, no. 1, pp.48-59, 2007.

10 B. Minch, P. Hasler, C.Diorio and C. Mead, "A Silicon Axon," *Advances in Neural Information Processing Systems*, vol. 7, pp. 739-746, 1994.

11 L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology,* vol. 41, pp.35-39, 1948.

12 M. Oster and S.-C. Liu, "A Winner-take-all Spiking Network with Spiking Inputs," *Proceedings, IEEE International Conference on Electronics, Circuits and Systems*, 2004, pp. 203-206.

13 I. Grech, J. Micallef and T. Vladimirova, "Analog CMOS chipset for a 2-D sound localization system," *Analog Integrated Circuits and Signal Processing*, vol. 41, no. 2-3, pp. 167-184, 2004.

14 A. A. Handzel, S. B. Andersson, M. Gebremichael and P. S. Krishnaprasad, "A Biomimetic Apparatus for Sound-source Localization," *Proceesings, IEEE Conference on Decision and Control*, 2003, vol. 6, pp. 5879-5884.

15 S. B. Andersson, A. A. Handzel, V. Shah and P. S. Krishnaprasad, "Robot phonotaxis with dynamic sound localization," *Proceedings, IEEE International Conference on Robotics and Automation*, 2004, vol. 5, pp. 4833-4838.

16 H. G. Okuno and K. Nakadai, "Real-Time Sound Source Localization and Separation based on Active Audio-Visual Integration," *Proceedings, International Work Conference on Artificial and Natural Neural Networks*, 2003, pp. 118-125.

[i] Interaural phase difference

[ii] Interaural envelope difference