A SPIKING NEURAL NETWORK APPROACH TO AUDITORY SOURCE LATERALISATION

Robert Luke [1,2]

[1] Macquarie University
Department of Linguistics
Australian Hearing Hub, Sydney, Australia

David McAlpine [1]

[2] The Bionics Institute
East Melbourne, Australia

ABSTRACT

A novel approach to multi-microphone acoustic source localisation based on spiking neural networks is presented. We demonstrate that a two microphone system connected to a spiking neural network can be used to localise acoustic sources based purely on inter microphone timing differences, with no need for manually configured delay lines. A two sensor example is provided which includes 1) a front end which converts the acoustic signal to a series of spikes, 2) a hidden layer of spiking neurons, 3) an output layer of spiking neurons which represents the location of the acoustic source. We present details on training the network, and evaluation of its performance in quiet and noisy conditions. The system is trained on two locations, and we show that the lateralisation accuracy is 100% when presented with previously unseen data in quiet conditions. We also demonstrate the network generalises to modulation rates and background noise on which it was not trained.

Index Terms— spiking neural networks, binaural localisation algorithms, acoustic source localisation, machine learning

1. INTRODUCTION

Determining the location of a sound source is an important skill that has evolved to be simultaneously accurate and energy efficient. Many species can localise sources using two ears to an accuracy of just a few degrees [1]. Determining the location of an acoustic source is also an essential element of many digital signal processing (DSP) platforms. Acoustic source localisation in DSP systems is used for a variety of purposes including scene analysis and noise reduction. Acoustic source localisation is often at the front end of DSP pipelines and used to gate when other algorithms are activated. For this reason it is desirable to design an acoustic localisation system that can run continuously with minimal memory and energy requirements. In this paper we investigate whether bio-inspired spiking neural networks can be used to localise acoustic sources.

Animals primarily use two cues to determine the location of an acoustic source in the horizontal plane. The interaural time difference (ITD) is the difference in arrival time of sounds across the two ears, it is generated by the difference in distance from each ear to the acoustic source. ITD sensitivity in humans is dominated by low frequencies ($\lessapprox$ 2 kHz). The interaural level difference (ILD) is the difference in intensity of sounds across the two ears, ILD sensitivity is dominated by high frequencies ($\gtrapprox$ 2 kHz). In this paper we demonstrate the performance of a system designed to utilise low frequency ITD cues.

A large number of binaural localisation algorithms have been developed that rely on ITD and ILD cues, see Courtois et al [2] for a review. Most methods consist of a front end that roughly models the basilar membrane, and a computational or statistical based backend. Computational approaches are based on inverse HRTF filters, optimising cross correlation lags, and other similar methods. Statistical approaches use Bayseian estimates, perceptual models, ITD and ILD classifiers, and other similar methods. These techniques perform well in quiet conditions, but often fail in loud or reverberant environments. Methods that are effective in challenging acoustic environments require large amounts of memory, processing, or knowledge of the room acoustic parameters a-priori [2]. There is a need for computationally efficient binaural localisation algorithms that work in challenging acoustic environments.

Recently, spiking neural network approaches have been proposed as a potential mechanism for binaural localisation. These networks are described as the third generation of neural networks, and employ spiking neurons as the computational unit, this allows the system to inherently encode spatiotemporal information [3]. Spiking neural networks can be trained to perform tasks by varying the weight of synaptic connections between neurons. Spiking neural networks have been used to perform acoustic localisation tasks. For example, Wall et al (2007) [4] exploited the ITD cue by inserting delay lines to imitate the Jeffress model [5]. Goodman & Brette (2010) [6] exploited both ILD and ITD cues to train their system of around 1 million neurons that manually included delay lines and gains. Both these approaches showed good localisation performance in quiet but did not report whether performance generalised to conditions with background noise. Wall et al

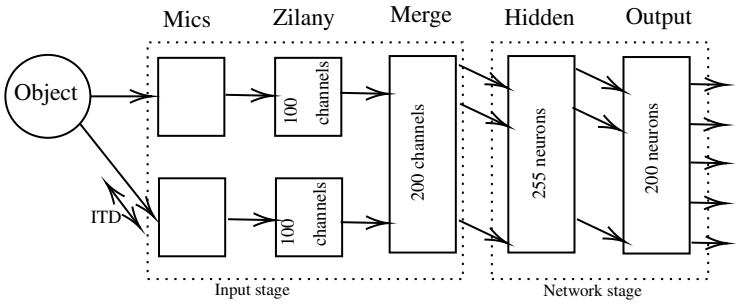

Fig. 1. System Architecture. Sound arrives at the two microphones. The Zilany model is used to generate 100 spiking channels per microphone, these are concatenated in the merge stage. The 200 channels are connected to the hidden layer of 256 neurons, which is then connected to the 200 neuron output layer.

(2012) [7] expanded on previous studies by demonstrating that a delay-line based ILD approach with topology based on the lateral superior olive can generalise to accurately localise sources in noise down to 0.1 dB SNR. The existing literature utilised manually configured delay-lines to achieve ITD sensitivity. However, delay line systems may not be the mechanism by which biological systems generate ITD sensitivity [8]. We present a spiking neural network with no explicit delay lines that accurately estimates the lateralisation of an incoming sound source. We do not attempt to create an exact model of the auditory system, but endeavour to mimic just a few aspects of the biological system that may improve localisation performance in noise.

In this article we present a system for acoustic binaural localisation based purely on ITD cues. We do not attempt to optimise the memory or energy requirements of the system, but aim to establish whether spiking neural networks are a suitable tool to perform acoustic localisation. For this reason we start with the simplest localisation task of lateralisation (determing if a signal came from the left or right). We evaluate the system performance for a range of SNRs and input signals on which the system was not trained. The remainder of the paper is arranged as follows. In section 2 the system architecture, training, and evaluation is described. Experimental results and system performance are discussed in section 3. Finally a brief conclusion is presented in section 4.

2. METHODS

This section describes the design and evaluation of the acoustic localisation system. The architecture of the system is described in section 2.1. Training and stimuli are described in section 2.2. System evaluation is described in section 2.3.

2.1. System Architecture

The system architecture is illustrated in figure 1. Sound waves arrive at the two sensors from an acoustic source. The waveform arriving at each sensor has an ITD generated by the location of the source. The acoustic waveforms are then converted to a series of spikes using the Cochlea toolbox [9] which implements the Zilany model [10], this model gives a realistic approximation of auditory nerve activity given an acoustic input. We parameterised the Zilany model to have 20 centre frequencies between 125 and 1000 Hz. For each centre frequency 3 high, 1 mid, and 1 low spontaneous rate neuron was simulated, in total 100 neuron channels were simulated per microphone. In this work we have used a software model, for implementation an analog model of cochlea could be used [11]. The spiking output from each microphone was concatenated to create a 200 neuron input to the spiking neural network.

The 200 neurons from the input stage were fed to a two layer spiking neural network of leaky integrate-and-fire neurons. The first (hidden) layer contains 256 neurons, the second (output) layer contains 200 neurons. The output of the spiking network was 200 channels, each channel was designed to correspond to an ITD. The neuron to ITD mapping increased monotonically from -2 ms through to +2 ms. For example, neuron 50 corresponds to -1 ms, neuron 100 corresponds to 0 ms, and neuron 150 corresponds to +1 ms.

To estimate the ITD of the incoming signal, the output spikes were accumulated into bins using a histogram approach. The ITD was estimated as the histogram bin with the most counts. Figure 3 illustrates the output of a trained network on unseen data. Each point on the central plot illustrates when an output neuron fired, the histogram shows the distribution of firing neurons over the 1.0 second example segment.

2.2. Network Training

The spiking neural network was trained using the SuperSpike software [12]. Audio was presented to the system to train the network, and the synaptic weights were optimised to produce the desired target output using symmetric feedback. The input audio was 6 Hz amplitude modulated, band-pass noise (200 - 800 Hz). The same audio was applied to both microphones and one channel had a known ITD applied. The target output was a sequence of spikes that fired with greatest probability at the associated ITD neuron, and with decreasing probability for neurons further from the true ITD. The target spike sequence was generated by setting the firing rate of each neuron based on a gaussian centred at the true ITD with variance of 20 neurons. Figure 2 shows the input and target data for a 10 ms ITD (an exaggerated value for visualisation). The input neurons 1-100 are from the left channel output of the Zilany model, the 101-200 neurons are from the right channel output of the Zilany model, the right channel neurons have

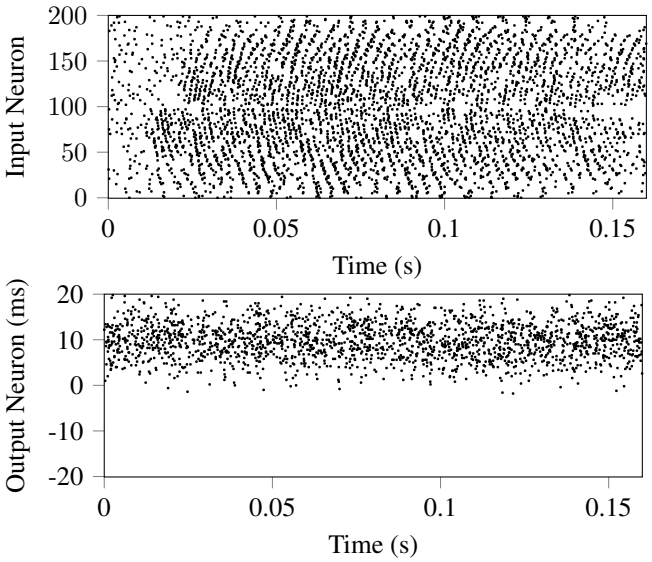

Fig. 2. Input and target signals for training the spiking neural network (network stage in Fig 1). The upper plot shows the input to the spiking network generated from audio with a delay of 10 ms (exaggerated for visualisation), note the shift between first and second 100 neurons. The lower plot shows the target output with peak spiking rate at 10 ms.

been flipped so it is easier to observe the time delay. Note that the top 100 neurons are roughly a delayed version of the bottom 100 neurons, by 10 ms. The output neurons are most densely firing at the neurons associated with the ITD of 10 ms, and there is decreasing firing rate at neurons further from the specified ITD.

To train the network pairs of known input and output spikes were presented to the optimisation software. Each training pair consisted of 1.8 seconds of audio presented to the system for 25 training blocks. 90 unique pairs of input and output were presented to the system. The initial network weights were set to 0.05, with a learning rate of 1e3, and with no spurious spiking. In each training set the source location randomly jumped between ± 1 ms at the minima of the amplitude modulation. All training stimuli were applied to the Zilany model at 70 dB SPL with no interfering noise.

2.3. Performance Evaluation

Once trained, the system was evaluated using audio on which the network was not trained (new generation of noise for acoustic signal, random initialisation of Zilany model, etc). The novel input was presented to the system and the output spikes were recorded. The distribution of the output spikes as a factor of neuron were calculated using a histogram procedure with 20 bins. The estimated location of the source was defined as the ITD associated with the largest histogram bin. The location was determined as correct if it had the same sign as the known input ITD, essentially judging if the system correctly lateralised the source.

The system was evaluated in quiet and in noisy situations with different SNRs. Different SNRs were generated by adding white noise with a bandwidth of 0-1000 Hz to the acoustic source (amplitude modulated bandpass noise with ITD applied). The noise added to each microphone was independent. A range of SNRs were tested from -60 to 40 dB in steps of 20 dB. 20 realisations were computed for each SNR and location (± 1 ms), and the percentage of correct lateralisation estimates was calculated.

The ability of the system to generalise to unseen data was tested by evaluating performance for previously unseen modulation rates. Sounds with modulation rates of 5, 6.4 (the training rate), 7, 8 and 12.8 Hz were presented to the system and the localisation performance was calculated.

3. RESULTS AND DISCUSSION

The system was trained as described in section 2. Figure 3 shows an example output from the network for a previously unseen acoustic signal arriving with an ITD of -1 ms in noise with an SNR of 10 dB. The central plot shows the spiking activity per neuron as a function of time. The histogram indicates that peak neural firing occurs at the neurons associated with -1 ms, as indicated by the red line. In this example the system has correctly identified the input ITD.

The spiking neural network was trained to discriminate whether a sound was arriving at two microphones from the left or the right with an ITD of 1 ms. The network was tested on 20 previously unseen acoustic signals. When the audio was presented in quiet (no additive noise) the lateralisation accuracy was 100%.

The network was tested with previously unseen realisations of sounds arriving from the left or right with an ITD of 1 ms in noise. The signal was always presented at 70 dB SPL, and the noise was varied to achieve the desired SNR levels. The red line in figure 4 shows the performance of the network as a function of SNR for an input signal with the same modulation rate as the network was trained on. As the noise level increased, the lateralisation accuracy decreased, but still achieved over 80% accuracy at 0 dB SNR. The performance vs SNR curve of the system mimics that of typical psychometric function from psychoacoustic experiments [13], and does not have a catastrophic failure when presented with previously unseen conditions.

The ability of the system to generalise to novel acoustic sources was examined by modifying the modulation rate of the input. The blue line in figure 4 illustrates the performance per SNR for an 8 Hz amplitude-modulated signal. The performance at low SNRs was worse than for the trained modulation rate, but again the system did not completely fail. The inset shows the performance at 20 dB SNR for a range of modulation rates. Generally the system was robust to data that

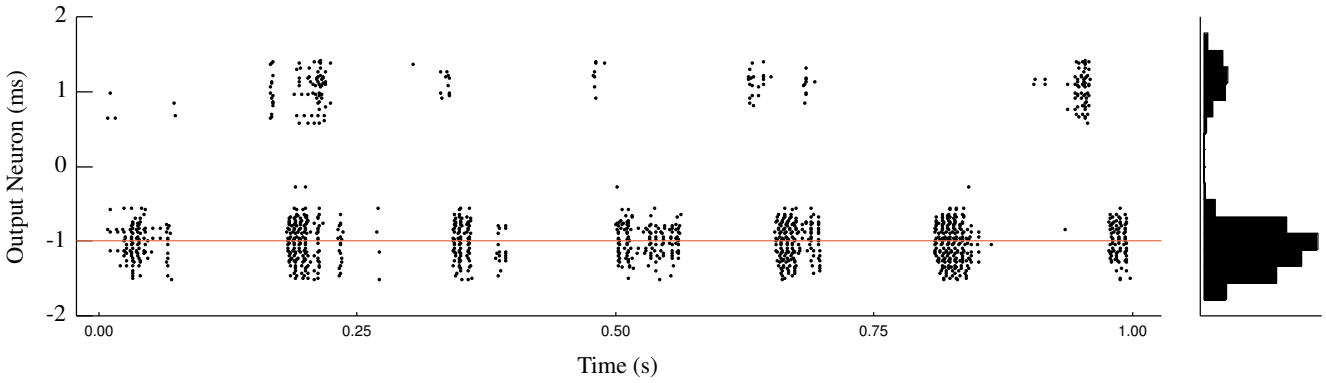

Fig. 3. Localisation system output for previously unseen data (with an ITD of -1 ms) presented in noise with SNR of 10 dB. Each black dot represents a neuron spike. The histogram on the right illustrates the frequency of spikes for each time range. The red line indicates the ITD bin where the highest rate of spiking occurred.

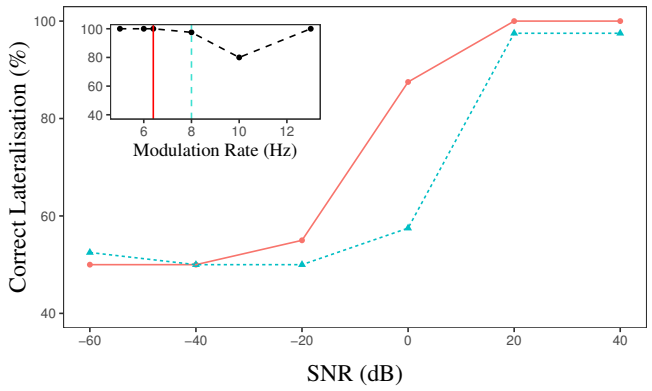

Fig. 4. Lateralisation performance as a function of SNR for sources with a modulation rate of 6.4 Hz (solid red) and 8 Hz (dashed blue). The system was trained using only data with a 6.4 Hz modulation rate. The inset shows the performance for additional modulation rates evaluated at 20 dB SNR.

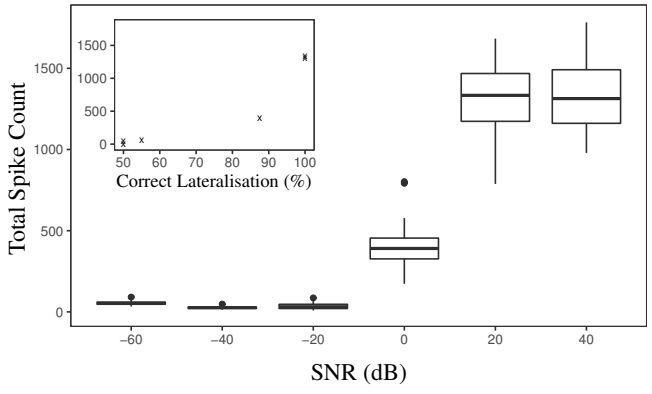

Fig. 5. Total spike rate as a function of SNR. The number of spikes at the output of the system is monotonically increasing with SNR. The inset shows the total spike rate as a function of system accuracy.

differed from what it was trained on. The system performed well in noise that had not been used in training, and for modulation rates other than was used in training.

Figure 5 illustrates the total number of output spikes per second for different SNRs. The spiking rate decreases with decreasing SNR. This was unexpected as the acoustic input level is higher for low SNRs (the signal was held constant at 70 dB SPL and the noise varied).

Assuming a fully connected network, this system would require 51712 synaptic connections. Future work to reduce the network size may include synaptic pruning stages, and smaller hidden and output layers. The number of output neurons was overspecified to support future work to generalise the lateralisation experiment to a localisation task. Similarly the size of the hidden layer may be optimised, a much smaller hidden layer may achieve similar performance in this task.

4. CONCLUSION

A spiking neural network based binaural lateralisation system is presented. The system architecture is described along with methods for training the network. The system achieves 100% lateralisation accuracy in quiet. The system was found to generalise to perform in noise, and to accurately lateralise signals with modulation rates on which it was not trained. To the best of this author's knowledge, this is the first example of a spiking neural network that can localise based purely on ITDs with no pre-defined delay-lines, instead exploiting the spatiotemporal properties of the spiking computational unit.

5. REFERENCES

[1] Benedikt Grothe, Michael Pecka, and David McAlpine, "Mechanisms of sound localization in mammals," *Physiological reviews*, vol. 90, no. 3, pp. 983–1012, 2010.

[2] Gilles Courtois, Patrick Marmaroli, Morten Lindberg, Yves Oesch, and William Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*. Audio Engineering Society, 2014.

[3] Wolfgang Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[4] Julie A Wall, Liam J McDaid, Liam P Maguire, and Thomas M McGinnity, "A spiking neural network implementation of sound localisation," *Proceedings of the IET Irish Signals and Systems*, pp. 19–23, 2007.

[5] L. A Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[6] Dan F M Goodman and Romain Brette, "Learning to localise sounds with spiking neural networks," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds., 2010, pp. 784–792.

[7] Julie A Wall, Liam J McDaid, Liam P Maguire, Thomas M McGinnity, and Senior Member, "Spiking Neural Network Model of Sound Localization Using the Interaural Intensity Difference," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 4, pp. 574–586, 2012.

[8] David McAlpine and Benedikt Grothe, "Sound localization and delay lines - Do mammals fit the model?," *Trends in Neurosciences*, vol. 26, no. 7, pp. 347–350, 2003.

[9] Marek Rudnicki, Oliver Schoppe, Michael Isik, Florian Völk, and Werner Hemmert, "Modeling auditory coding: from sound to spikes," *Cell and Tissue Research*, vol. 361, no. 1, pp. 159–175, 2015.

[10] Muhammad S A Zilany, Ian C Bruce, and Laurel H Carney, "Updated parameters and expanded simulation options for a model of the auditory periphery.," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 283–6, 2014.

[11] Vincent Chan, Shih Chii Liu, and André van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 1, pp. 48–59, 2007.

[12] Friedemann Zenke and Surya Ganguli, "SuperSpike: Supervised Learning in Multilayer Spiking Neural Networks," *Neural Computation*, vol. 30, no. 6, pp. 1514–1541, 2018.

[13] Robert Luke, Lieselot Van Deun, Michael Hofmann, Astrid van Wieringen, and Jan Wouters, "Assessing temporal modulation sensitivity using electrically evoked auditory steady state responses," *Hearing Research*, vol. 324, pp. 37–45, 2015.