Modeling Populations of Spiking Neurons for Fine Timing Sound Localization

Qian Liu*,†, Cameron Patterson†, Steve Furber†, Zhangqin Huang*, Yibin Hou* and Huibing Zhang†

*Embedded System Software Institute, College of Computer Science
Beijing University of Technology, Beijing, 100124, People’s Republic of China.
†Advanced Processor Technologies Group, School of Computer Science,
University of Manchester, Manchester, M13 9PL, United Kingdom.
‡Guangxi Key Laboratory of Trusted Software,
Guilin University of Electronic Technology, Guilin 541004, People’s Republic of China.
Email: {liuq, pattersc}@cs.man.ac.uk, steve.furber@manchester.ac.uk, {zhuang,yhou}@bjut.edu.cn, zhanghuibing@guet.edu.cn

Abstract—When two or more sound detectors are available, interaural time differences may be used to determine the direction of a sound’s origin. This process, known as sound localization, is performed in mammals via the auditory pathways of the head and by computation in the brain. The Jeffress Model successfully describes the mechanism by exploiting coincidence detector neurons in conjunction with delay lines. However, one of the difficulties of using this model on neural simulators is that it requires timing accuracies which are much finer than the typical 1 ms resolution provided by simulation platforms. One solution is clearly to reduce the simulation’s step time, but in this paper we also explore the use of population coding to represent more precise timing information without changing the simulation’s timing resolution. The implementation of both the Jeffress and population coded models are contrasted, together with their results, which show that population coding is indeed able to provide successful sound localization.

I. INTRODUCTION

Localizing sound sources in the horizontal plane is achieved in the lower auditory pathway of the brain by analyzing the physical sound features in high resolution. This is particularly the case for time parameters, including the Interaural Time Differences (ITDs) which lie in the submillisecond range. The ability of mammals to perform sound localization has inspired researchers [1] [2] to propose Spiking Neural Network (SNN) approaches to unlock the brain’s mechanism for electronic exploitation. This research paper therefore develops techniques for SNN simulators that can replicate the fine timing computation that the auditory neural system can achieve.

Sound localization is complex as a means of neural processing [3], but the neural circuit underlying the localization cues is one of the rare cases where the physiological and behavioral function is believed to be well understood [4]. Two dominant binaural localization cues are encoded in the neural circuit: ITDs and Interaural Level Differences (ILDs). The most widely accepted hypothesis is the Jeffress model [5], which detects the ITDs of a sound wave captured by two ears. Although it forms the basis of modern binaural localization, the first evidence was found more than 30 years after the model’s creation, showing that the nucleus laminaris of the barn owl works similarly to the hypothesis [6]. Later practical evidence of the existence of delay lines projecting to the medial superior olive (MSO) was found in cats, but only from the contralateral side [7]. The lateral superior olive (LSO) is where the initial stage of encoding ILDs takes place, and most of the neurons located in the LSO are of excitatory-inhibitory (EI) binaural type [8].

This paper proposes a population coding method for sound localization, where the required fine timing accuracy is achieved using statistics of the constituent spiking neurons which have a coarser time step. The paper is structured as follows: The neurophysiological data and mechanisms of the mammalian auditory pathway which inspire the SNN models are presented in Section II. Section III describes an experimental implementation of the Jeffress model, processing the input spike trains from the cochlea through the auditory pathway to the superior olivary nucleus (SOC). Section IV introduces population coding as a mechanism to achieve more precise ITD timing resolution than the individual constituent neurons of an SNN provide. Results of the Jeffress Model and population coding experiments are compared and contrasted in Section V which details the recognition rates of both models. The paper’s final section draws conclusions from the experimental work, and describes further research to be carried out in the subject area.

II. BIOLOGICAL FUNDAMENTALS

As spatial information is not directly represented in auditory stimuli (unlike visual or somatosensory systems) there is particular interest in understanding the function of sound localization. Further processing is essential to localize the sound source by combining the binaural auditory inputs within the central auditory computing system. The auditory nerve (AN) propagates the sound signal into the cochlear nucleus (CN) of the auditory brainstem, whose neurons propagate information to identical ascending centers on both dorsal and ventral sides of the brain. Auditory processing stations along the central auditory pathway include the SOC, the inferior
colliculus (IC) in the midbrain, the medial geniculate nucleus in the thalamus, and the auditory cortex.

To provide the auditory pathway with stimuli the auditory periphery first has to convert the sound wave signals into spike trains. The vibration of the eardrum generates a wave of displacement travelling along the liquid basilar membrane in the cochlea of the inner ear. The amplitude of the displacement maximizes at the position in the cochlea which responds to the frequency of the wave, with the cochlea's base covering the higher, and the apex the lower frequencies. If the displacement in the cochlea reaches a threshold, the inner hair cells (IHCs) located on the organ of corti will reach the basilar membrane, triggering action potentials in the AN. Different IHCs respond to their own characteristic frequencies (CF) according to their position along the basilar membrane. This spatial arrangement, referred to as tonotopic organization of sound reception is located in the ascending auditory pathway.

For a sinusoidal pure tone, whose frequency is the same as the CF of the IHCs, the basilar membrane always touches the IHCs at a particular phase angle when its displacement reaches the threshold. This results in the corresponding AN fiber firing in a phase locked fashion, encoding the temporal parameters of the input signal (Figure 1). Phase locking forms the basis of detecting time differences of binaural sound inputs. In mammals phase locking of AN is only valid in lower frequency bands, and in frequencies greater than around 1k-1.5k Hz the lock begins to break down, disappearing at frequencies exceeding 3k-4k Hz [9]. As the sound pressure level (SPL) grows from below to well above threshold, the firing rates of most of the AN fibers increase with a sigmoidal relationship (over 20-30dB) and saturate at their maximum firing rate [10].

Much of the processing of sound localization occurs in the SOC, where two computational circuits exist to encode the ITDs and the ILDs. This mammalian auditory brainstem structure is illustrated in Figure 2. The process of binaural cues from both ITDs and ILDs is well known as the ‘duplex theory’ [11]. The maximum possible ITD is about 660 µs according to the typical dimensions of a human skull [12], so the ITD cue is effective for low frequency sounds between 20 Hz-1.5k Hz [13]. Conversely ILDs can be measured most effectively at relatively high frequencies around 2k-3k Hz, where the wavelength is short and the head can act as an effective acoustic shadow [14].

As illustrated in Figure 2 the circuit that encodes ITDs receives excitatory inputs from the anteroventral cochlear nucleus (AVCN) on both ipsilateral and contralateral sides, and projects to the neurons in the medial superior olive (MSO). The neurons in the MSO are sensitive to their own preferred ITDs. In the original Jeffress model input spikes from both sides of AVCN propagate through different delay lines. However Yin and Chan later found [7] that only the contralateral inputs are subjected to multiple parallel delay lines. These MSO neurons perform as coincidence detectors, firing when both inputs from ipsilateral and contralateral AVCNs arrive in the same time interval. By having multiple delay lines of different lengths this MSO model can be described as a two-dimensional matrix of coincidence neurons [4]: ITDs and frequency, see Figure 3. The neurons in the LSO receive excitation from the ipsilateral AVCN and inhibition from the medial nucleus of the trapezoid body (MNTB) on the same side where contralateral excitatory inputs are converted to inhibited connections.

III. MODELING MSO NEURONS

In this section the biologically-inspired SNN model of the MSO for sound localization is presented in detail. The ITD cue is encoded into spike trains by SNNs which enable the sound source to be localized. We start with the overview of the model structure describing the data flow of the system, and follow this with an exploration of timing accuracies which influence the sound localizing performance. The structure of the sound localization model is illustrated in Figure 4. The input of the system is the stereo sound wave as would be received directly by the ears. The data we use is generated by
Listen Project [15] which aims to provide users with intuitive access to personalized audio information spaces. The dataset consists of samples for 24 different azimuths from $0^\circ$ (head-on) to $345^\circ$ with a resolution of $15^\circ$. The front-back confusion is widely accepted to be a result of ITDs, where the same ITD can be produced by opposing sound sources. Therefore in this model only half of the data is used, whose sound source is in front of the head. Furthermore, the ambiguity also occurs at azimuths close to $90^\circ$ and $270^\circ$ ($-90^\circ$), as the ITDs are very close within that range. Figure 5 provides the spike rate statistics of an array of coincidence detection neurons responding to a 200 Hz frequency tone with an accuracy of 10 $\mu$s. The front-back confusion and the ambiguity within the ranges [60°, 120°] and [240°, 300°], are clearly shown in the figure. In this work we therefore focus on the ranges [0, 60°] and [300°, 360°].

The next layer of the model is a series of second-order Gammatone filter banks simulating the tonotopical organization of the cochlea. The bandwidths of the filter banks are the equivalent rectangular bandwidth (ERB) corresponding to the human cochlea along the basilar membrane. The output of the Gammatone filter layer is the decomposed mechanical responses of each ERB frequency channel of the sound wave. There are 40 ERB frequency channels from 200 Hz to 4k Hz in our cochlea model.

The tonotopical movements of the basilar membrane are then fed into the IHC layer where they are converted into spikes. A spike is triggered and propagated to the AN fiber when the displacement of the basilar membrane reaches a threshold. As the previous section described, the phase locking feature of the cochlea is carried by the output spike trains of the IHCs. Each spike train responds to a different ERB frequency channel derived from the tonotopical decomposition of the basilar membrane.

As Figure 3 illustrates, the coincidence detection neurons in the MSO can be considered as a matrix. Each row of neurons receives the spike trains from the corresponding frequency channel of both sides. The coincidence detectors in the same row receive the identical input spike train from the ipsilateral cochlea. The contralateral spikes are propagated through multiple-delay lines, so as to arrive at the coincidence detectors one-by-one in a sequence with defined intervals. A coincidence detection neuron fires when two spikes arrive within the same time step (a relatively short time range), which corresponds to the granularity of the delay line interval. This ensures there is no overlapping time range within adjacent neurons, and only one detector in a row (or possibly two at the boundary of adjacent neurons) fires when a pair of spikes arrive at the MSO. There are two MSOs as illustrated in Figure 4 located on both sides of the SOC respectively. The neurons of the MSO on one side fire when the ipsilateral sound wave arrives later than the contralateral side.

The output layer consists of 9 neurons responding to the azimuth angles from -60° to 60° with a resolution of 15°. The MSO neurons project to the output layer with an all-to-all connection, whose weights are trained by the spike rate statistics of the coincidence detector neurons. Audio output from the LISTEN project is divided into a training set with 216 audio trials, and a test set of 243. The estimated azimuth angle is determined by the most active neuron among the corresponding nine output direction neurons.

The main problem when modeling the coincidence detection neurons in the MSO is the fine time resolution it requires to present ITDs, which are usually less than 660 $\mu$s, while many neural simulators run with a 1 ms resolution. One possible solution is to make the real time simulator run fast enough to handle the time accuracy. One such real-time simulator that has been developed within our group is SpiNNaker [16], and previous experiments show it is able to simulate up
to 36,000,000 synaptic events per chip per second. If the simulation were running with a 1 µs time step, each of SpiNNaker’s 16 application cores could support 2 synapses and a single neuron.

The ITDs cue is effective for low frequency sounds, typically up to 1.5k Hz, therefore only the lower 16 of the 40 ERB frequency channels are involved in the MSO model. Although the ITD precision is always less than 10 µs [17], permitting a human to localize the sound source with an accuracy range of 2°-3°, the initial goal of our research is to classify the sound source into 9 directions with 15° resolution. A MATLAB simulation was created to determine the acceptable tradeoff between the timing precision and the classification accuracy. It requires the adjustment of the input event and the ITD resolution to these timing precisions. If the timing precision is p µs, then the narrowest width of ITD bin and the finest granularity of the delay lines interval will be 2p µs. Every center of an ITD bin is the delayed firing time of the contralateral spike. So, the arrival time of the contralateral spike at the first MSO neuron is always considered to be 0, the number q detector fires only when the ipsilateral spike time locates in the range of [(q - 1)p, (q + 1)p] indicating the ITD is within its range. Figure 6 shows that the classification accuracy remains above 90% when the granularity of the ITDs is as coarse as 200 µs, and we therefore set our model’s timing accuracy to 100 µs. Consequently, there are 6 ITDs bins for MSO on both sides: [0, 100 µs], [100 µs, 300 µs],... , [900 µs, 1100 µs].

The coincidence detectors in the MSO are modeled using Leaky Integrate and Fire (LIF) neurons, which fire when two input spikes arrive with an interval of 100 µs or less. We modeled the neuron using PyNN [18] on the Brian [19] platform using the parameters listed in Table I with two presynaptic neurons located on both sides of the A VCN. Figure 7 is the result of injecting this modeled LIF neuron with two pairs of spikes with intervals of 150 µs and 100 µs respectively. The solid blue line is plotted with data generated using a 1 µs time step simulation while the dashed red line is with 100 µs resolution. The neuron successfully fires when second pair of spikes are within the required interval. Although the simulation using a 100 µs time step is not as accurate as the 1 µs simulation, the neuron performs successfully as a coincidence detector.

| Table I: Parameter Setting for MSO Neurons in PyNN |
|----------------|----------------|----------------|----------------|
| Type           | curr_exp        | cm             | refrac         |
| IF             | -65.0           | 0.02           | 0.0            |
| v_rest         | 0.1             | 5.0            | -68.0          |
| i_offset       | 0.0            | -58.0          |

For neurons in a single MSO there are 6 neurons responding to 6 successive ITDs bins, and it therefore requires 96 neurons to cover the 16 frequency channels for MSO on each side of the brain.

IV. Population Coding Model

Another potential method of representing the precise time resolution is to use population coding, which makes use of populations of imprecise neurons to represent the neural event. In this case, the neural events are the accurate firing time of the AN fibers, with the sound source location encoded in the timing of the spikes. Using populations has precedence, neural function is rarely dependant on single elements, but has resilience to individual neuron mortality for example. Inspired by this redundant biological behaviour, the proposed population coding method works on the basis of the normal distribution. The statistical nature of normal distribution makes the method tolerant to abnormal behaviour of a relative small number of neurons in a population without degrading the performance of the whole network.

Additional inspiration for the proposed normal distribution based population coding originates in the jitter demonstrated...
in the responses of neurons. Although the neurons may be homogeneous and are triggered simultaneously with an identical input, not all neurons fire at the exact same time. The accurate temporal information, the mean of the responses, is instead encoded amongst the normal distribution of the spikes generated by the population of neurons. Figure 8 shows a population of 1000 neurons firing according to a normal distribution with $\mu=5.612$ (red dashed line) and $\sigma=1$. The green solid line highlights the expectation of the normal distribution. In our particular sound localization application, 5.672 ms is the calculated time when a spike event triggered in an AN fiber, an error of 60 $\mu$s when compared with the original time event.

The normal distribution based spike generator requires three parameters: $\mu$, $\sigma$, and the population size (number of neurons). $\mu$ is the input spiking time event of the AN fiber, whereas the other two parameters are the main factors influencing the accuracy of the estimation. The estimation error decreases as the population size grows (see Figure 9(a)), with a better than 100 $\mu$s error at $2^8$ neurons and 10 $\mu$s with $2^{16}$ neurons in the population. As the precision required for our sound localization task with a 15$^\circ$ resolution is less than 100 $\mu$s (from the previous section), we therefore selected $2^8$ neurons for the population size. Next, we varied the target $\sigma$ to determine how the error characteristics change, and were able to minimize the error with a $\sigma$ of 0.45, see Figure 9(b). As $\sigma$ rises, the more similar the normal distribution becomes to an even distribution, causing the error to grow as $\sigma$ increases beyond this point. Below 0.45 the error becomes larger due to the spike time being rounded into 1 ms bins, and with fewer bins the real event time is harder to estimate accurately.

The experiments of Figure 9 are conducted over 1000 trials for every population size, and in the case of $\sigma$ of 0.45 with $2^8$ neurons only three of the 1000 trials exceed the limit of 100 $\mu$s accuracy.

The structure of the MSO model using population coding is shown in figure 10. There is a population of converter neurons inserted between an AN fiber and the MSO neurons transforming a single spike train on that frequency channel to population spike trains. If the population size of converter neurons is $N$, then $N^2$ coincidence detectors are required for the population. The coincidence detection neurons no longer detect sub-ms ITDs, but fire when its combination of two input spikes are within the same 1 ms time slot. If there are $n_i$ spikes input to the population of coincidence detection neurons from the ipsilateral side and $n_c$ from the contralateral side, there will be $n_i \times n_c$ spikes generated in the population. Similar to the original MSO model, contralateral spikes propagate through multiple delay lines, but instead of 200 $\mu$s the granularity of
the delay lines is 1 ms. The correlation of the pair of spikes are encoded in the spike rates of these populations. Figure 11 shows the spike rates of nine populations of coincidence detection neurons against different ITDs. Its X-axis represents the ITDs of pairs of spikes, Y-axis shows the length of delay line, and the Z-axis is the spike rate. The different receptive fields of the populations are clearly identifiable in the figure. The estimation of ITD is worked out with the intrinsic feature of correlation, see Equation 1:

\[
ITD = \frac{\sum (d_i \times s_i)}{\sum s_i}
\]

The population is identified by \(i\), \(1 \leq i \leq M\), \(M\) is the number of populations involved, \(d_i\) is the length of delay line of population \(i\), and \(s_i\) represents the corresponding spike rate of the population. The precision of the estimated ITD is influenced by \(M\), and information is lost when \(M\) is not sufficiently large. This is illustrated in Figure 12 which demonstrates that the errors decrease as \(M\) grows and the errors are approximately identical when \(M\) equals 9, 11 and 13 in a test of 1000 trials for each value of \(M\). Since \(\sigma\) of the normal distribution of the input population is less than 1, only five or six bins are needed to form a distribution, per Figure 8. Thus the spike rate of the population with a delay line longer than 5 ms is close to 0 as the ITDs are within the range [-1 ms, 1 ms], and an indication of this is supplied in Figure 11. Nine populations of coincidence detection neurons are sufficient for the sound localization task.

To model the coincidence detection neurons in MSO using population coding, there are 9,437,184 neurons required (9 populations \(\times 256^2\) neurons \(\times 16\) channels). Although this number of neurons is huge compared to the original model, this is an initial attempt to represent fine timing accuracy using spiking neurons of lower native resolution. In the future work section techniques of reducing this method’s overhead are discussed.

Both the original MSO model and the population coding method were tested using the 243 trials of sound waves from the LISTEN project. A 20 ms sound track captured with an azimuth of 30° and a sample rate of 44100 Hz was taken as an example and its spectrum and sound wave are plotted in Figure 13(a). This binaural sound wave was fed into the gammatone filter bank simulating the mechanism of tonotopical decomposition in the cochlea. The bandwidths of 16 frequency channels were distributed from 200 Hz to 1.5k Hz according to the ERB, with Figure 13(b) illustrating the movement of the basilar membrane with different frequency channels. The wave from the right side (dark blue) was leading the left (light green) and had higher amplitude. The spikes generated within the IHCs, triggered by the movement of the basilar membrane, are displayed in Figure 13(c).

In Figure 5, the ITDs fall into the range of around -300 \(\mu\)s to -500 \(\mu\)s (negative ITDs refer to the right sound wave leading the left wave). It is clearly shown in Figure 14(a) that the MSO neuron responding to the 400 \(\mu\)s ITD bin of the first channel was the only active neuron in that row. The brighter the color in the figure, the more active the neuron. As the frequency increased the ITDs tended to be smaller and noisy because of multiple spikes generated within a 1 ms interval. Compared with the original MSO neurons, the spike rates among the populations of coincidence detection neurons were noisy because of the intrinsic randomness in the population coding method (see Figure 14(b)). Finally, the spikes propagated to the output layer, whose weights were trained according the statistics of the training data, similar to Figure 5. The output neuron for 30° received the most synaptic input from the previous layer which was the best candidate for this classification task, and is shown in Figure 14(c). Similarly, but for the population coding method, Figure 14(d) shows the neural activity at 30° is slightly lower than that of Figure 14(c) but is still the best spiking match. Although the output from the population coding neuronal activity is more noisy, its classification rate was competitive with the
original MSO model. The performance of classification in 10 trials is shown in Figure 15 comparing the population coding technique to the Jeffress model on different timing precisions without changing population settings. In the previous section we targeted parameters of the populations to compete with the performance of the original MSO model, running at a 100 µs time step whose ITD precision is 200 µs. The classification rate of the population coding varied within a range from 89.3% to 93.42% with an average of 91.32%. This is slightly higher than the original model (90.95% on 200 µs ITD precision), as most of the errors generated during the population coding process were less than 100 µs, see Figure 12. Consequently, when the timing accuracy was poorer, the classification rate of population coding was better, while its sound localization ability was worse than the original model when a finer ITD resolution is required. Without changing the time step and any settings of the populations, the population coding method showed feasible and competitive results compared with the original MSO model.

VI. CONCLUSION AND FUTURE WORK

In this paper we have introduced population coding to represent timing information for sound localization. One of the drawbacks of the conventional technique is that it requires neurons with a fast time step in simulation whereas, using populations, neural models with coarser time steps may be used. The results of initial experimentation with the technique show that it is indeed a feasible approach, and methods have been developed for tuning the parameters to be used with the network. In operation, classifications using populations of neurons achieve results that are within a few percentage points of the conventional alternative, albeit requiring many times the number of neurons. The simulated results of the technique are very encouraging, and future work is planned to optimize the number of neurons per population and their structure.

One possible optimization approach is to sample the coincidence spikes to reduce the number of neurons required for the coincidence detectors. Every coincidence detection neuron is randomly connected to two neurons located in the populations of converter neurons. Experiments will be carried out here.

Fig. 13. Sound wave gets through gammatone filter banks converted into spikes.

Fig. 14. Comparing the ITD detection and classification ability of the population coding method and the original model

Fig. 15. Classification rates of 10 trials of tests for the same 243 audio data (boxplot) and the original MSO model (green star) against the ITDs accuracies
to determine the connectivity pattern and how many neurons may be optimized out. When detecting the ITDs, comparing the differences in spike rates of the input neurons is another optimisation opportunity – only one or just a few neurons may be good enough to measure the spike rate differences of two neural populations. These mechanisms need to be explored similarly to the LSO EI type neurons which detect the ILDs.

In the future, this promising technique may also be extended beyond the Jeffress model to other applications where it is not possible to achieve the required level of modelled accuracy with single neurons. The population based coding technique has the potential to be applied to many other problems. We also aim to apply the proposed sound localization models to our mobile robotic sensory platform [20] and explore sound processing of its environment in real time.

In the initial phase of our work into sensor processing, we have succeeded in connecting[1] a hardware SNNs simulation platform, SpiNNaker [16], to an event based silicon cochlea [21]. Spikes generated from the auditory sensor are transmitted to the SpiNNaker ‘brain’ and are displayed with a real-time visualizer. The next step is to permit the ‘brain’ to drive the robot based on its ability to perform sound localization. Moreover, we plan to provide the system with learning algorithms, so it may adapt to new environments dynamically. Concerning the ‘duplex theory’, the ILD cue is also an interesting area we expect to explore in our future work.

ACKNOWLEDGMENT

The SpiNNaker project is supported by the Engineering and Physical Sciences Research Council of the UK, through Grants EP/D07908X/1 and EP/G015740/1, and also by ARM and SiliSistix. Qian Liu and Huibing Zhang are supported by the Guangxi Natural Science Foundation (2013GXNSFAA019324). The authors appreciate the support from all sponsors and industrial partners.

REFERENCES


https://capocaccia.ethz.ch/capo/wiki/2012/csQian