Closed-loop sound source localization in neuromorphic systems

NEUROMORPHIC
Computing and Engineering

OPEN ACCESS

RECEIVED
23 February 2023

REVISED
26 May 2023

ACCEPTED FOR PUBLICATION
1 June 2023

PUBLISHED
12 June 2023

PAPER

Closed-loop sound source localization in neuromorphic systems

Thorben Schoepe[1,2,*], Daniel Gutierrez-Galan[3,4], Juan P Dominguez-Morales[3,4], Hugh Greatorex[1,2], Angel Jimenez-Fernandez[3,4], Alejandro Linares-Barranco[3,4] and Elisabetta Chicca[1,2]

[1] Bio-Inspired Circuits and Systems Lab, Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands
[2] Groningen Cognitive Systems and Materials Center, University of Groningen, Groningen, The Netherlands
[3] Robotics and Technology of Computers Lab, Universidad de Sevilla, Sevilla, Spain
[4] SCORE Lab, I3US, Universidad de Sevilla, Sevilla, Spain
[*] Author to whom any correspondence should be addressed.

E-mail: t.schoepe@fz-juelich.de

Keywords: event-based sensing, neuromorphic systems, sound source localization, interaural time difference, spiking neural networks

Supplementary material for this article is available online

Abstract

Sound source localization (SSL) is used in various applications such as industrial noise-control, speech detection in mobile phones, speech enhancement in hearing aids and many more. Newest video conferencing setups use SSL. The position of a speaker is detected from the difference in the audio waves received by a microphone array. After detection the camera focuses onto the location of the speaker. The human brain is also able to detect the location of a speaker from auditory signals. It uses, among other cues, the difference in amplitude and arrival time of the sound wave at the two ears, called interaural level and time difference. However, the substrate and computational primitives of our brain are different from classical digital computing. Due to its low power consumption of around 20 W and its performance in real time the human brain has become a great source of inspiration for emerging technologies. One of these technologies is neuromorphic hardware which implements the fundamental principles of brain computing identified until today using complementary metal-oxide-semiconductor technologies and new devices. In this work we propose the first neuromorphic closed-loop robotic system that uses the interaural time difference for SSL in real time. Our system can successfully locate sound sources such as human speech. In a closed-loop experiment, the robotic platform turned immediately into the direction of the sound source with a turning velocity linearly proportional to the angle difference between sound source and binaural microphones. After this initial turn, the robotic platform remains at the direction of the sound source. Even though the system only uses very few resources of the available hardware, consumes around 1 W, and was only tuned by hand, meaning it does not contain any learning at all, it already reaches performances comparable to other neuromorphic approaches. The SSL system presented in this article brings us one step closer towards neuromorphic event-based systems for robotics and embodied computing.

1. Introduction

Large progress has been made in improving algorithms and systems that perform sound source localization (SSL) since the breakthrough of machine learning. Recent approaches are able to precisely detect multiple dynamic sound sources using microphone arrays and deep neural networks [1]. However, most of these approaches use a large number of static microphones, large artificial neural networks and run on classical computing hardware such as central processing units (CPUs) and graphics processing units (GPUs) [1, 2]. Those systems are usually not very well suited for edge computing or robotic tasks due to their high power consumption, high latency or space-consuming hardware. In this paper we develop a compact, low-power and low latency approach for SSL using neuromorphic hardware. This approach is, very much like the vertebrate auditory apparatus, embodied in the real world. Embodied systems can exploit the physical and chemical properties of the environment to actively extract relevant information. For example, an active movement strategy of the robot can improve its SSL performance. The localization error increases for larger sound source angles [3]. By actively turning towards the direction of the sound source the angle and the localization error decreases. Dávila-Chacón *et al* [3] and Chan *et al* [4] have used an active motion strategy to increase the SSL accuracy. However, most computation was performed on conventional hardware leading to a large overhead and high latency. In contrast, we are aiming at a low-power, low-latency hardware implementation of SSL. Neuromorphic hardware and event-based sensors, which are inspired by the working principles of the brain, aim at sparse, fast and power-efficient computation. By using neuromorphic hardware components we can develop a system highly suitable for robotic applications. In contrast to the conventional frame-based approach, event-driven computing only samples changes in the sensory input leading to a much sparser representation of the environment, free of redundancies [5, 6]. Spiking neural networks (SNNs), one building block of neuromorphic computing, extract relevant information from event-based data in a network of spiking neurons and synapses. These local computational units integrate, filter and translate rate, spike-timing and spatio-temporal information in a mostly nonlinear way [7]. One advantage of event-based sensors and SNNs is their ability to extract and process precise spike timing information which is highly relevant for SSL based on interaural time difference (ITD). ITD is the difference in arrival time of an auditory wave at the two microphones of a binaural auditory sensor. A recently proposed model for spatio-temporal computation in SNNs is the time difference encoder (TDE) proposed by Milde *et al* in [8]. This building block translates the time difference between events coming from two different input channels into a burst of output spikes. The applicability of this model for temporal encoding in vision and touch has already been demonstrated in [8–12]. The TDE has been implemented in mixed digital/analog sub-threshold complementary metal-oxide-semiconductor (CMOS) hardware [8], only consuming between 1.4 nW (static) and 500 μW (dynamic) [13]. Its most recent field programmable gate array (FPGA) implementation uses 179 lookup tables and 140 registers, hardware resources comparable to other neuron models on FPGA [13]. We aim at developing a hardware system by combining the neuromorphic auditory sensor (NAS), an FPGA implementation of an artificial cochlea, the TDE on FPGA and a compact SNN on SpiNNaker to perform SSL based on the ITD. A great variety of binaural SSL approaches can be found in the literature. These can be divided between software and hardware implementations, which can be both either open-loop or closed-loop. A detailed review on SSL in robotic systems is given in [14]. A variety of neuromorphic open-loop approaches for event-based SSL have been proposed [15–20]. A neuromorphic software example [4] presents closed-loop audio-visual neuromorphic sensory fusion on a sound-localizing robot with adaptive ITD for robot navigation. In that case, all the processing is performed in Matlab, which limits the real-time capabilities of the system. Focusing on neuromorphic hardware solutions, very few can be found. For example [21], presents an open-loop implementation of a bio-inspired model of SSL on the IBM TrueNorth platform. Regarding neuromorphic closed-loop SSL hardware implementations [22], presents an FPGA-based system using interaural level difference (ILD) on a single NAS channel to drive a robotic head towards pure tone audio cues. The ILD is the difference in amplitude of a sound wave reaching the microphones of a binaural cochlea caused by the accoustic shadow of the robotic head. In this article we propose, to our knowledge, the first neuromorphic closed-loop robotic hardware system that uses the ITD for SSL in real time. Our system is one out of very few approaches with a fully spiking sensor-to-actuator pipeline. The main contributions of this work include the following:

- The ITD based SSL network, a new bio-inspired approach.
- Implementation of the SSL network on an event-based neuromorphic hardware closed-loop system.
- The full characterization of the closed-loop system in a real-world static SSL task, with a standard deviation of 4.2, 7.9 and 16.9° for 250 Hz, 500 Hz and human speech respectively.
- Comparison to other neuromorphic approaches.

The paper is organized as follows: section 2 introduces the different hardware components and the SNN developed in this article. Section 3 gives an overview of the two experiments conducted: an 180 degrees sweep of the sound source and a closed-loop static SSL task. Section 4 presents the results of the two experiments. Finally, in section 5, the results are compared with other approaches and the model's further development as well as its suitability for robotics and embedded systems application is discussed.

2. Methodology

2.1. NAS

The NAS [6] is a spike-based digital audio sensor inspired by Lyon's model of the biological cochlea [23], implemented on FPGA. This sensor decomposes incoming audio signals into their frequency components as the inner hair cells do in the human ear, producing a stream of address events (AEs) [24]. It was implemented using a spike-based low-pass filter (SLPF) bank with a cascade topology [25]. Each SLPF represents a frequency range, and its output consists of a stream of AEs. In this work, we used a 64-channel binaural NAS generated with OpenNAS [26][5] implemented on an AER-Node board, which has a Spartan-6 FPGA [27]. The 64 channels were configured so that they correspond to frequency bands that are distributed along the whole range of the human hearing (20 Hz–20 kHz, approximately [28]).

2.2. Spiking neural network architecture (SpiNNaker)

SpiNNaker [29] is a massively-parallel multicore computing system in which very large SNNs can be deployed and simulated in real time. A 4-node SpiNN-3 machine, which consists of 72 200 Hz ARM968 processor cores, was used. It has a 100 Mbps Ethernet connection for the communication between the computer and the board. A PyNN-based [30] software package called sPyNNaker [31] is used for design and simulation purposes. SpiNN-3 contains two spiNNlinks [32]. The first spiNNlink was used to connect the FPGA, on which the NAS and TDEs are implemented, to the SNN. The second spiNNlink was used to connect the SNN to the FPGA for motor control (see figure 2).

2.3. TDE

The TDE model [8] encodes the time difference between two input events occurring at different input channels in a short burst of output spikes. The time difference is conveyed in the number of spikes as well as the instantaneous firing rate. The model consists of two inputs, the facilitatory gain (fac) and the trigger synapse (trig), as well as one spiking output shown in figure 1(a). Upon the arrival of an event at the facilitatory input, an exponentially decaying facilitatory variable, the gain, is set to its maximum amplitude. The arrival of an event at the trigger synapse shortly after an event at the facilitatory synapse (i.e. small time difference Δt) leads to the generation of an excitatory postsynaptic current (EPSC) (see figure 1(b)). The EPSC amplitude is proportional to the value of the facilitatory variable at the arrival of the trigger event. Hence, the amplitude of the EPSC is inversely proportional to the time difference. A leaky integrate and fire (LIF) neuron integrates the postsynaptic current from the trigger synapse in its membrane potential (V_{mem}). A digital output pulse (also called spike) is generated when V_{mem} reaches the spiking threshold θ_{spike}. As it can be seen in figure 1(e), the number of output pulses is inversely proportional to the time difference between the two input events. When the time difference is much longer than the facilitatory time constant, no EPSC is generated and therefore there are no output spikes. For negative time differences (an event occurs at the trigger synapse shortly before an event at the facilitatory input, as in figure 1(d)) no output spikes occur. The TDE is a direction-selective module.

2.4. SSL system

The SSL system depicted in figure 2 consists of a binaural microphone, three hardware boards and a pan-tilt unit. The 3Dio binaural microphone receives auditory stimuli at its right and left artificial ears which are 150 mm apart. The shape of the ear and the two white disks on each side of the microphone increase the ITD between the two microphones by modulating the minimum distance of a sound wave reaching both ears. The left and right microphone output are sent to the left and right channels of the NAS implementation on the AERNode FPGA board respectively. The NAS converts the auditory input into events and extracts the different frequency components from the events. The NAS output consists of 64 frequency channels per ear. A phase-lock stage was implemented on the same FPGA to retrieve only the timing information of the NAS output. The NAS and phase-lock filter are explained in more detail in section 2.5.1. The phase-lock output of channel 32 is sent to the four left-right and four right-left sensitive TDEs implemented on the same AERNode board. The TDE output is forwarded through a SpiNNlink connector to a SpiNN-3 board which includes the subsequent SNN components. An input population of one-to-one connected LIF neurons receives the events from the TDEs. This population is necessary due to software constraints in the SpiNNaker communication fabric. It simply forwards the TDE spikes from the FPGA to the SpiNNaker board. The TDE input is further processed by the Time to Rate Network, Ring Attractor Network and Center Detector Network in subsequent order. The spikes of the Center Detector Network are sent to the motor control board

[5] https://github.com/RTC-research-group/OpenNAS. Retrieved 26 May 2023.

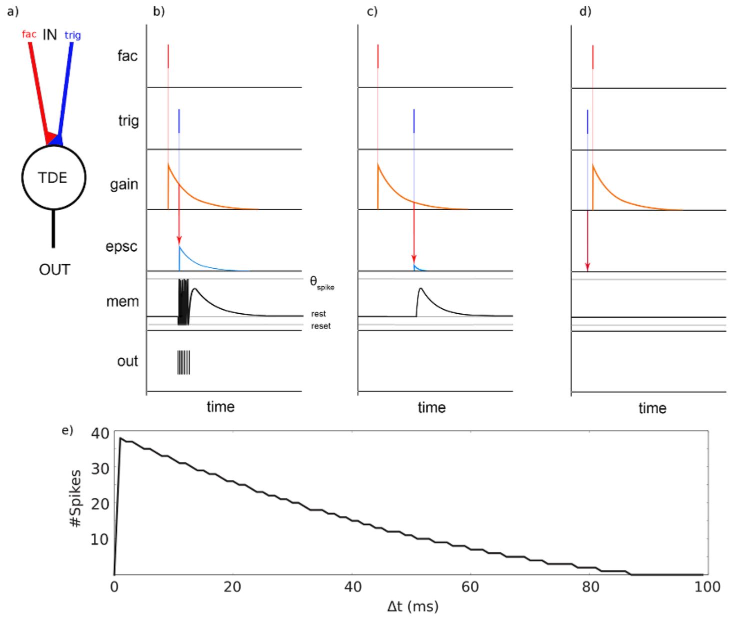

Figure 1. TDE operating principle. (a) TDE basic composition. Input received by the trigger (trig) synapse is gated through the facilitatory (fac) synapse. Gain depends on time difference. (b)–(d) TDE response for small, large and negative time differences. (e) TDE response curve.

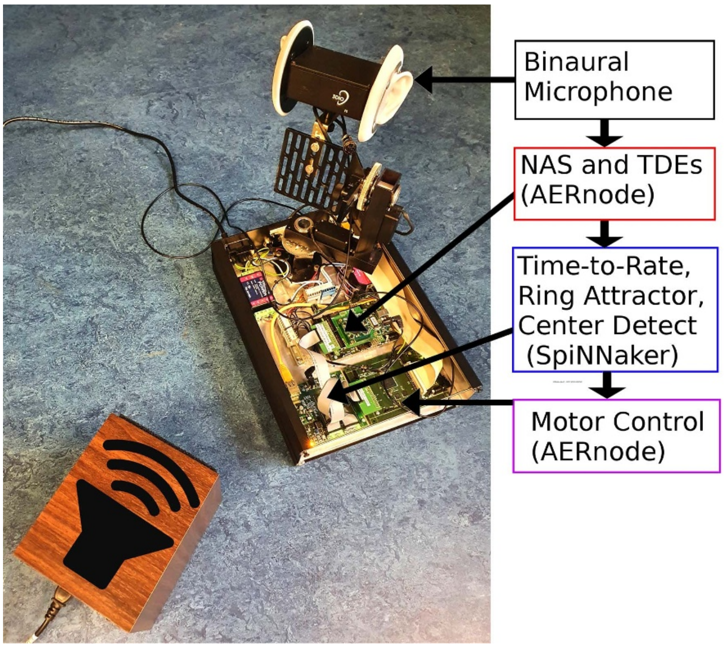

Figure 2. Neuromorphic closed-loop SSL system. Sound waves are received by the two microphones of the binaural microphone. The microphone outputs are connected to the first AERnode FPGA board. On the FPGA board the NAS, phase-lock filter and TDEs encode the ITD information of the auditory signal into a spiking rate. The TDE output is further processed in the Time to Rate Network, Ring Attractor Network and Center Detector Network on a SpiNNaker board. The identifier of the spiking neuron in the Center Detector Network is passed to another AERnode board. This board decodes the identifier into a PWM signal that drives the pan-tilt-unit to the estimated position.

(another AERNode board), which decodes the identifier of the spiking neuron into a pulse width modulation signal. This signal is received by the motor of the pan-tilt unit controlling the yaw angle. The different network components are further elaborated in the following section.

2.5. SSL network

Our SSL network does not contain any learning. The model was engineered and tuned by hand. The network consists of three sub-networks. The first sub-network, the Time to Rate Network (figure 3, section 2.5.1),

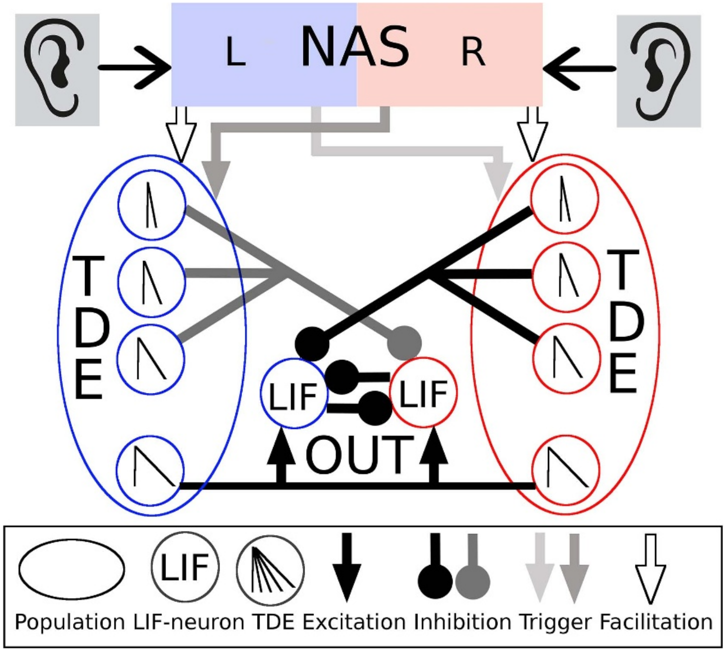
L
NAS
R
TDE
TDE
LIF
LIF
OUT
LIF
Population LIF-neuron TDE Excitation Inhibition Trigger Facilitation

Figure 3. Time to Rate Network. The NAS converts auditory signals from the left and right microphone into events. The events from the left NAS channel excite the facilitatory synapses of the neurons in the left TDE population (white arrow). Events from the right NAS channel project onto the trigger synapses of the left TDE population (grey arrow). For the right TDE population it is the other way around. Each TDE population consists of four TDEs with four different facilitatory time constants. The time constants are depicted as triangles inside the TDE symbols. The TDE of each population with the largest facilitatory time constant excites both output neurons (OUT). The other three TDEs of each population inhibit the opposing output neuron. This mechanism leads to an activation function of the output neurons inversely proportional to the TDEs response.

encodes the ITD in the NAS output into a spiking rate proportional to the angle of the sound source. The left and right output neurons of the Time to Rate network are excitatory all-to-all connected to the left and right input-population of the second sub-network, the Ring Attractor Network (figure 5(a), section 2.5.2). A stronger input to the left input-population drives the activity bump of the Ring Attractor Network to the left and vice versa. Hence, the activity bump in the network moves towards the direction of the sound-source driven by the difference in the TDE activity. The third sub-network, the Center Detector Network (figure 5(b), section 2.5.3), filters the activity of the Ring Attractor Network to remove jitter in the movement-behavior of the pan-tilt unit. The spikes of the Center Detector Network are sent to the motor control board which decodes the identifier of the spiking neuron into a position of the pan-tilt unit. Next, the three sub-networks are explained in detail.

2.5.1. Time to rate network (figure 3)

The two microphones used in this setup encode the sensed auditory signals in electronic signals and send these to the left and right channel of the NAS respectively. The NAS converts the auditory signals into events with an event rate proportional to the amplitude of the input [33]. As part of the NAS an event-based cascade filter bank separates the different frequency components of the event-based signal into 64 frequency channels per ear. We only use one NAS output channel per ear with a maximum response close to 500 Hz. We chose 500 Hz, because ITD-based SSL with a human-inspired binaural microphone only works for frequencies lower than 1 kHz. The NAS cascade filter bank design of the channels also passes through signals with a lower rate than the main frequency, but with a weaker amplitude. Hence, we could use the same frequency channel for all subsequent experiments. Before sending the events into the Time to Rate network a phase-lock filter is applied to the events. Each channel has two different types of events, positive events which are generated by the positive phase of an auditory signal and negative events generated by the negative phase. The phase-lock filter detects the point in time when the signal switches from positive to negative events, similar to an auditory signal crossing the zero threshold. Every time the signal switches from positive to negative events a single spike is elicited by the phase-lock filter. Ambiguities arise in the estimate of the temporal difference when the ITD is larger than half the frequency of the signal, in our case the frequency of the phase-lock filter events. Therefore, we reduce the output frequency to the minimum possible by performing the zero crossing only in one direction. Due to the phase-lock filter the spikes only inherit timing information which is necessary for computing the ITD. The output of the two phase-lock filtered event

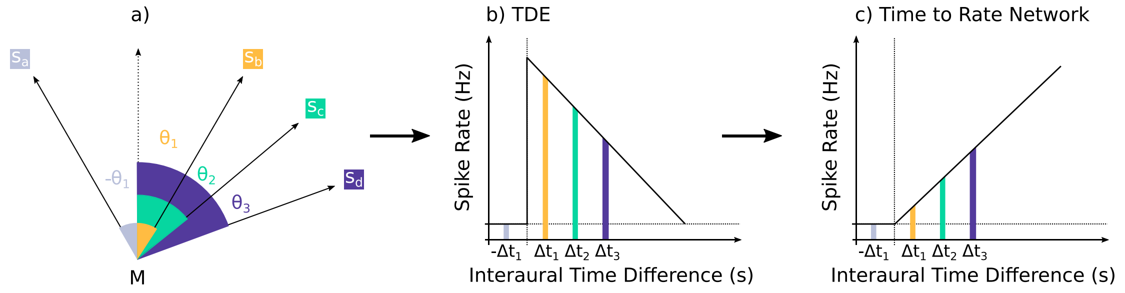

Figure 4. Relation between sound source angle, ITD, TDE output rate and Time to Rate Network output Rate. (a) Angle between a sound source S_{a-d} and the binaural microphone M. (b) A small angle θ_1 between sound source and microphones leads to a small ITD Δt_1 which encodes in a high TDE spike rate. The ITD increases with larger angles while TDE spike rate decreases inversely proportional. (c) The output of the Time to Rate Network is proportional to the angle of the sound source. An angle increasing from θ_1 to θ_3 leads to an increasing ITD from Δt_1 to Δt_3 which leads to an increasing spike rate.

signals is send to two TDE populations. One TDE population is left-right connected, that means it receives facilitatory input from the left cochlea output and trigger input from the right cochlea output. That way only signals which reach first the left and then the right ear elicit a response in the TDE output. The response is maximum at ITDs close to zero (see figure 4). The other TDE population is right-left connected, responding to auditory signals closer to the right ear.

The difference in spiking activity between the two TDE populations reaches its maximum at time differences close to zero (sound source in front of the 3Dio binaural microphone). For example, when a sound source is located to the right of the microphones, first the right microphone and shortly after the left microphone receive an auditory signal (see figure 4(a)). The right-left connected TDE population elicits a response while the left-right connected population is silent. The closer the position of the sound source to the middle between the two microphones, the stronger is the response of the right-left connected TDE population, while the other population remains silent (see figures 4(a) and (b)). Hence, the response difference reaches its maximum at time differences slightly bigger than zero, and decays back to zero for time differences much bigger than the facilitatory time constant τ_{fac}. This response profile is not suitable for the generation of the required motor action. Ideally a big angle should lead to a big difference in TDE left and right response. This proportional translation from angle to rate-difference can easily be converted into a motor action, for example using a ring attractor to turn the actuator into the direction of the sound source. The working principle of a ring attractor is explained in the next section in detail. Because of this unfit transfer function of the TDEs, we designed a Time to Rate Network shown in figure 3, which translates the TDE response into the desired output profile. The desired network output shown in figure 4(c) is exactly opposite to the TDE response profile shown in figure 4(b). The required output is proportional to the ITD, the TDE output is inversely proportional. Therefore, the required spiking activity can be obtained by inverting the TDE response. The easiest way to invert the TDE signal is by subtracting the signal from a fixed offset. In spiking neurons this operation can be done on the synaptic level by subtracting excitatory (fixed offset) and inhibitory (input signal) currents from each other. In our Time to Rate network we perform this synaptic operation on two output LIF neurons (see figure 3). The two neurons receive excitatory input from one left-right connected and one right-left connected $\mathrm{TDE}_{(l3,r3)}$ with a large τ_{fac} (440 us on FPGA). This excitatory input drives the neurons into an active state which serves as baseline activity. Three left-right connected $\mathrm{TDE}_{(l0-l2)}$ and three right-left connected $\mathrm{TDE}_{(r0-r2)}$ with smaller τ_{fac} (68us, 190us and 320us on FPGA) are inhibiting the right and left output neuron respectively. This operation results in a nonlinear inversion of the output neuron activity. Inhibition between the two output neurons silences the activity of the less excited neuron. The spiking response of the output neurons is proportional to the azimuth angle of the sound source as shown in the closed-loop experiment and therefore well suited for motor control as explained in the following.

2.5.2. Ring attractor network (figure 5(a))

The ring attractor is a model of neural computation which closely resembles structures found in vertebrates as well as in insects [34, 35]. A typical realization of a ring attractor network is depicted in figure 5(a) and consists of four neural populations all connected in a ring-like structure, called input right ($\mathrm{IN}_{\mathrm{right}}$), input left ($\mathrm{IN}_{\mathrm{left}}$), direction right ($\mathrm{DIR}_{\mathrm{dr}}$) and direction left ($\mathrm{DIR}_{\mathrm{dl}}$). A bump of spiking activity is self-maintained in all four populations due to recurrent excitatory connections (see figure 5(a)). This bump represents the relative angular orientation of the vertebrate or insect in its environment. In this article we use the ring attractor to update the current angular position of a pan-tilt-unit. We are using an adapted version of the fruit fly

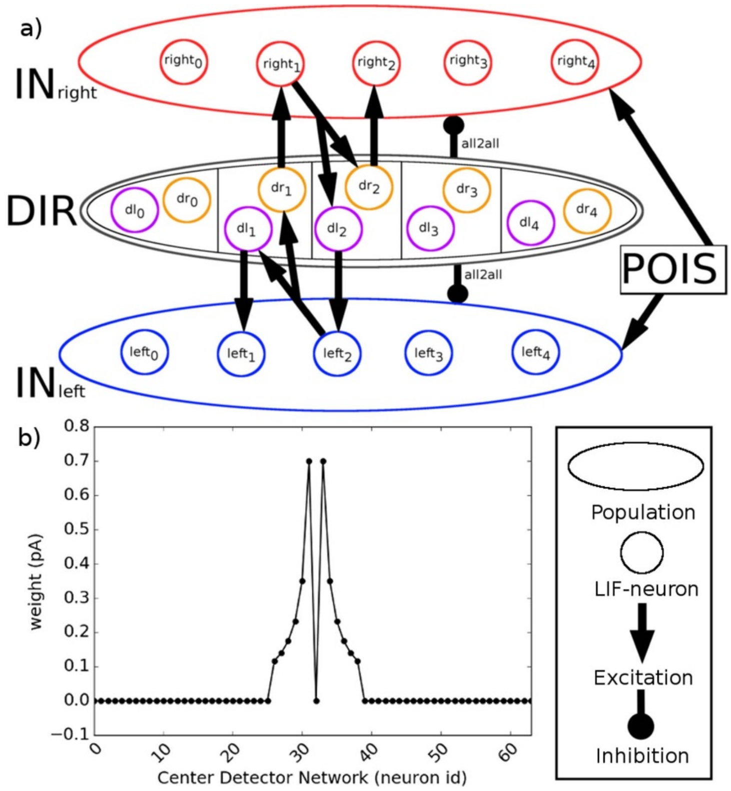

Figure 5. Ring Attractor Network and Center Detector Network. (a) Ring Attractor Network. Direction population neurons (DIR) are one-to-one connected to input neurons. Neurons of right input population ($\mathrm{IN}_{\mathrm{right}}$) are $i+1$ connected to left and right direction population. Left input population neurons ($\mathrm{IN}_{\mathrm{left}}$) are $i-1$ connected to left and right direction population. Only a few connections are shown for clarity. Direction population is all-to-all inhibitory connected to both input populations. Excitatory Poisson input spike trains (POIS) to the input populations maintain the bump in the Network. (b) Excitatory weight kernel from neuron 32 of Ring Attractor Network to all neurons of Center Detector Network. Same weight kernel is repeated for all neurons of the Ring Attractor Network. This filter amplifies spiking activity at the center of the ring attractor bump.

inspired ring attractor from [35] (see figure 5(a)). Our Ring Attractor Network is designed in a semicircle configuration with open ends since it only represents an angular range of 180°. Neurons of the two direction populations $\mathrm{DIR}_{\mathrm{dr}}$ and $\mathrm{DIR}_{\mathrm{dl}}$ are connected in a one-to-one fashion to the corresponding neurons in the two input populations $\mathrm{IN}_{\mathrm{right}}$ and $\mathrm{IN}_{\mathrm{left}}$ respectively. The right/left input neurons project back to the direction neurons of both populations $\mathrm{DIR}_{\mathrm{dr}}$ and $\mathrm{DIR}_{\mathrm{dl}}$ with excitatory connections to the first neighbor on the right/left side. Figure 5(a) exemplifies the pattern of connectivity between the direction neuron population and the input populations. Neuron dr_1 is one-to-one connected to neuron right_1 which is connected to neurons dr_2 and dl_2 in the direction population. This pattern is then repeated across the ring attractor topology, with neuron dr_2 connected to right_2 and so on. This connection profile causes a chain of excitation going to the right side through the right input population and vice versa. All-to-all inhibition from the direction neurons to the input neurons stops the excitatory wave from travelling further. The excitatory wave only travels until the point where the all-to-all inhibition is approximately as strong as the wave of excitation. This configuration leads to a bump of activity in the Ring Attractor Network. The activity is maintained by Poisson spike input trains (POIS) projecting onto the input populations. Hence, every neuron in the input population receives excitatory spike trains with a Poisson distributed spike rate around 10 Hz. As long as the spike rate in all four populations is balanced and no external input is received, the bump of activity remains at the same location in the network. When one of the two input populations receives additional excitatory spike trains from an external source, the target neurons reach a more excited state. The all-to-all inhibition from the direction neurons to the input neurons does not prevent the wave of excitation from travelling anymore. The excitatory bump starts to move, to the left or to the right when either the left or right input population are excited respectively. The movement velocity of the bump is proportional to the difference in input rate received by the left and right input population. The left and right input population receive all to all excitatory input from the left and right Time to Rate network output neuron respectively. Since each step of conversion from sound source angle to rate (Time to Rate network) to velocity (Ring Attractor Network) is proportional, the ring attractor bump moves with a velocity proportional to the angle of the sound source. Hence, the spiking rate of our Time to Rate Network is converted into a well defined movement towards the sound-source. While no sound is perceived by the system, the input to the Ring Attractor Network is close to zero. Therefore, the bump of spiking activity maintains its current position in the ring attractor. When a sound signal causes a rate difference between the left and right Time to Rate Network output neurons, which are connected to the left and right Ring Attractor Network input populations respectively, this difference drives the Ring Attractor Network bump into the direction of the sound-source.

2.5.3. Center detector network (figure 5(b))

Our motor control board (see section 2.4) translates the identifier of a spiking neuron into a position of the pan-tilt unit. However, the bump of activity in a ring attractor is typically spread over several neurons. Its width and movement characteristics are defined by the ratio of recurrent excitation and inhibition. On one side weak inhibition increases the width of the bump. In this case only very little excitation from an external source is needed to overcome the inhibition and move the bump into one direction. The weaker the inhibition the faster the bump can move. However, if the inhibition is too weak and the network receives strong excitatory input, the network tends to completely overcome the inhibition and all neurons remain active all the time, the network is unstable. On the other hand, strong inhibition decreases the width of the bump since the excitatory wave can no longer travel as far. Therefore more excitation is required to overcome the inhibition and move the bump. At some point the inhibition is too strong so that even the highest biologically reasonable input frequency of around 500 Hz cannot move the bump anymore. Hence, in this work we tuned the network to find the right trade off between movement velocity and network stability. This results in a bump which is typically spread over five to ten neurons. Due to these fluctuations in the spiking id, the pan-tilt unit tends to shake when the Ring Attractor Network is directly connected to the motor control board. To overcome this problem we designed a neuron-based filter called the Center Detector Network which extracts the center of the bump from the Ring Attractor Network output. Additionally, the activity is stabilized by this network as it acts as a low pass filter on the input. The neurons of the Ring Attractor Network are connected to the neurons of the Center Detector Network with the weight distribution shown in figure 5(b). This connectivity profile enhances the activity of the neuron in the center of an activity bump through lateral excitation. We connected the Center Detector Network population in a winner-take-all (WTA) fashion in order to maintain only the activity of the neurons at the center of the bump. The neurons are recurrently all-to-all connected to a global inhibitory neuron. This neuron reduces the activity of all neurons so that only the winning neuron at the center of the bump remains active. The spiking activity of this neuron updates the position of the pan-tilt unit through the motor control board. Each neuron in the Center Detector Network corresponds to one specific angular orientation of the pan-tilt unit. Hence, the number of neurons in the Ring Attractor Network and subsequent Center Detector Network determine the spatial resolution of the whole system. We chose a population size of 64 neurons in the Center Detector Network corresponding to an angular range of 180°. This leads to a spatial resolution of $\sim$2.8° per neuron. A decrease in the number of neurons would lead to a lower spatial resolution and therefore lower precision of the system, a faster angular velocity of the bump and a reduced number of required computational units. Depending on the application the right trade off between these factors can be chosen.

3. Experiments

We conducted two different types of experiments. The system was tested with an open-loop sound source sweep and a closed-loop static SSL task.

3.1. Pure-tone response

In this section the open-loop response of the Time to Rate network was evaluated using synthesized and real world data. In the first experiment, two POIS trains representing the left and right output of the NAS were synthesized. The time difference between each spike pair was sampled from a Gaussian distribution with a mean similar to the ITD ranging from -800 to 800 μs and a standard deviation of 40 μs. To replicate the rather noisy nature of the sensory system we added random spikes with a mean frequency of 500 Hz to the synthesized POIS trains. The Time to Rate Network was implemented in the NEST simulation platform. The TDE is implemented as a current-based LIF neuron with linear currents and linearly decaying membrane potential. Four left-right connected and four right-left connected TDEs with varying τ_{fac} (440, 315, 190 and 63 μs, respectively) received the synthesized spike-trains.

In a second experiment we characterized the performance of the Time to Rate Network using neuromorphic hardware in a physical setup (see figure 2). The NAS receives binaural auditory input from a 3Dio microphone placed on a pan-tilt unit. The left and right 500 Hz channels of the NAS are sending events to four right-left connected and four left-right connected TDEs on FPGA with different time constants

(440, 315, 190 and 63 μs, respectively). The digital TDE variables change linearly and the model does not include a membrane potential. The TDE spikes are sent to the SpiNNaker board. The TDEs are not directly implemented on the SpiNNaker processor for several reasons. Since the ITD lies typically in the range of tens of microseconds and the smallest time step supported by the SpiNNaker board is one millisecond, the processor cannot detect such small time differences. Furthermore, having the NAS and TDEs on the same FPGA enables a much faster and more parallel communication, which increases the response speed of the system and keeps the error in timing low. The SpiNNaker board includes the remaining parts of the Time to Rate Network. In this open-loop experiment, we perform a 180° horizontal, anti-clockwise turn of the pan-tilt unit. A speaker playing a 500 Hz pure tone was placed at the distance of 50 cm from the 3Dio microphone.

3.2. Closed-loop localization

In this experiment we investigate the movement velocity and precision of the SSL system in a closed-loop setup. A speaker playing different pure tones and words from the Google speech command dataset [36] is placed in front of the robotic binaural cochlea depicted in figure 2, 50 cm distant. The pan-tilt unit is initialized at nine different angles from -100 to 80°. A constant POIS train holds the initial position of the pan-tilt unit for the first ten seconds. After this initialization time, the difference in TDE activity moves the pan-tilt unit towards the sound source direction. The movement characteristics in dependency of the angular difference and the localization error are measured for the following 110 s. An HD webcam is placed above the pan-tilt unit and a red stripe is mounted onto the hardware setup to track the angular position of the pan-tilt unit through color tracking and shape detection using Python's OpenCV library.

4. Results

4.1. Pure-tone response

4.1.1. TDE

The TDE encodes the time difference between events anti-proportionally in the number of output spikes. Figure 6 shows how the TDE response profile looks like when using synthesized POIS trains and real auditory input. In figure 6(a) the response of four simulated left-right sensitive TDEs with varying τ_{fac} to synthesized POIS trains is shown. Data points with brighter color refer to longer τ_{fac}. The TDEs show no response for a stimulus with anti-preferred time difference. They jump to a maximum of close to 500 Hz at a time difference of zero microseconds. The response decays for larger positive time differences. The decay is slower for larger τ_{fac}. This response profile shows the same characteristics than the theoretical TDE response curve shown in figure 1. Figure 6(b) is similar to (a) only that we added 500 Hz noise to the input signal. The noise increases the overall response of the TDEs. Figure 6(c) shows the response of left-right sensitive (red) and right-left sensitive (blue) TDEs to synthesized POIS trains with Gaussian distributed time differences with a 40 ms standard deviation and 500 Hz noise. The Gaussian distribution of time differences leads to a smoother transition from anti-preferred to preferred time differences. While the response maximum still stays at the preferred time difference the response profile almost resembles a Gaussian distribution. The response maximum shifts to higher time differences for larger τ_{fac}. This simulated behavior looks very similar to the response of the TDEs on FPGA to a sound source sweep shown in figure 6(d). A 500 Hz pure tone was played by a speaker in front of the setup shown in figure 2, 50 cm distant. The pan-tilt unit performed a 180° anti-clockwise sweep of the 3Dio microphones in steps of 22.5°. Starting at two seconds run time, the pan tilt unit is moving to its next location every four seconds. During the occurrence of a turn every four seconds, starting at second two, the response of all TDEs increases suddenly and the difference between left and right sensitive TDEs drops almost to zero. This response can be explained by the sound of the motor placed below the 3Dio microphones. The motor sound, which arrives at the same time at both ears, causes a strong response of the TDEs due to a time difference close to zero. This motor noise causes a kind of saccadic movement scheme in closed-loop configuration.

4.1.2. Time to rate network

As stated in section 2.5.1 the response profile of the TDE is not very well suited to drive a motor towards a sound source location since the difference between left and right TDE response has its maximum close to an ITD of zero. Hence, we developed a compact Time to Rate Network which translates the TDE response into a profile proportional to the angle of the sound source. The architecture of this network is elaborated in section 2.5.1.

The response of the Time to Rate Network to synthesized spike trains and a sound source sweep with a 500 Hz sound input are depicted in figures 7(a)–(d) respectively. The response of the right output to POIS

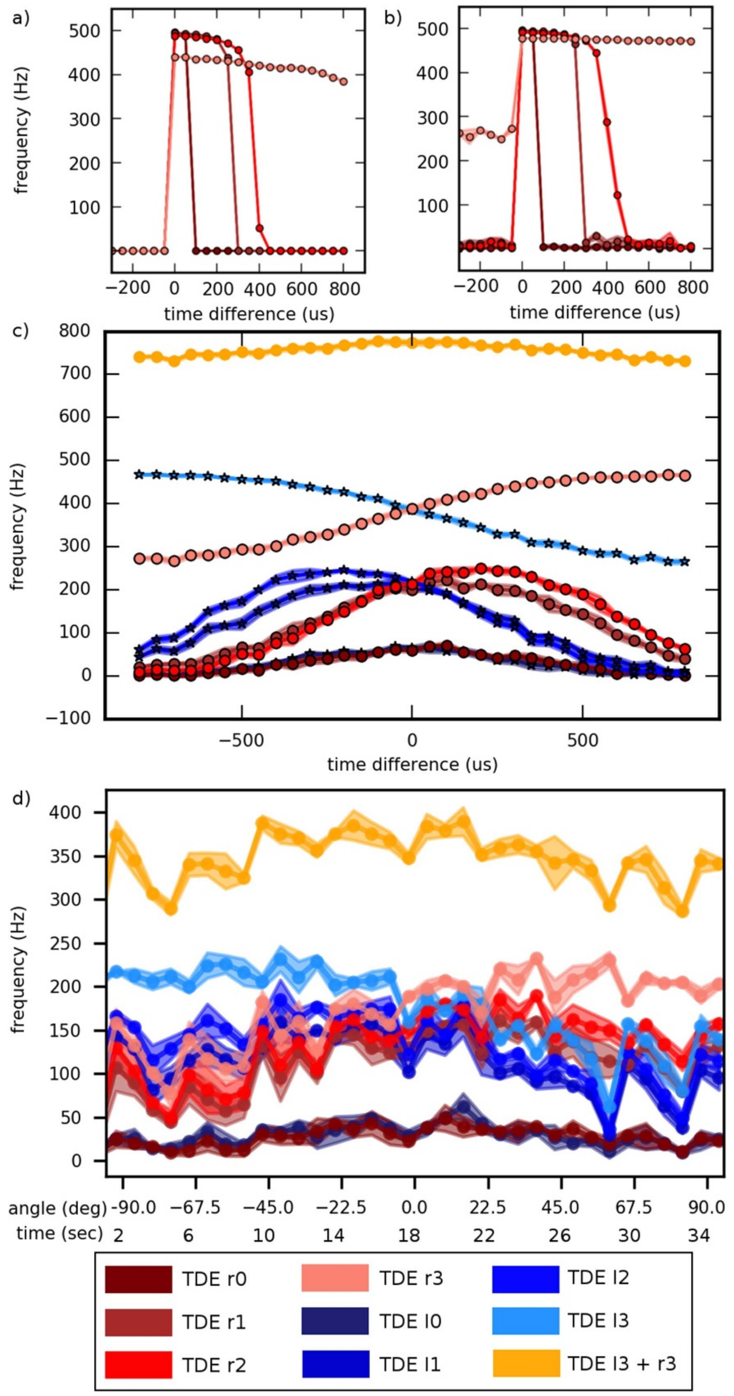

Figure 6. TDEs response to different ITDs. (a) Response of right-left connected leaky integrate and fire TDEs with linear decaying current and membrane potential simulated in Nest receiving computer generated Poisson spike trains with ITDs between −800 and 800 μs. (b) Similar to (a) with 500 Hz noise. (c) Similar to (a) with left-right and right-left connected TDEs with 500 Hz noise and Gaussian distributed time differences with a standard deviation of 40 μs. (d) Response of linear TDEs on AER-Node FPGA board to an anti-clockwise 180° turn of the binaural microphones around the yaw axis (mediated over three trials). A speaker is placed 50 cm distant in front of the microphones playing a 500 Hz constant pure tone. Starting from −90° (two seconds) the pan-tilt unit turns the microphones to the next position 22.5° apart every four seconds.

trains with precise time differences is shown in figure 7(a). The output is zero for negative time differences and jumps to its maximum at time differences larger zero. It slightly decays for larger time differences. The jump at around 500 μs is caused by the WTA mechanism between the left and right output neuron. From 0 to 400 μs the right output neuron is more active and suppresses the activity of the left neuron through inhibition (see supplementary material figure A.1). For time differences larger 400 μs both output neurons receive exactly the same input so that the WTA mechanism does not apply. Both neurons stay active at an intermediate level of activity reducing each other's activity due to slight inhibition. When adding random

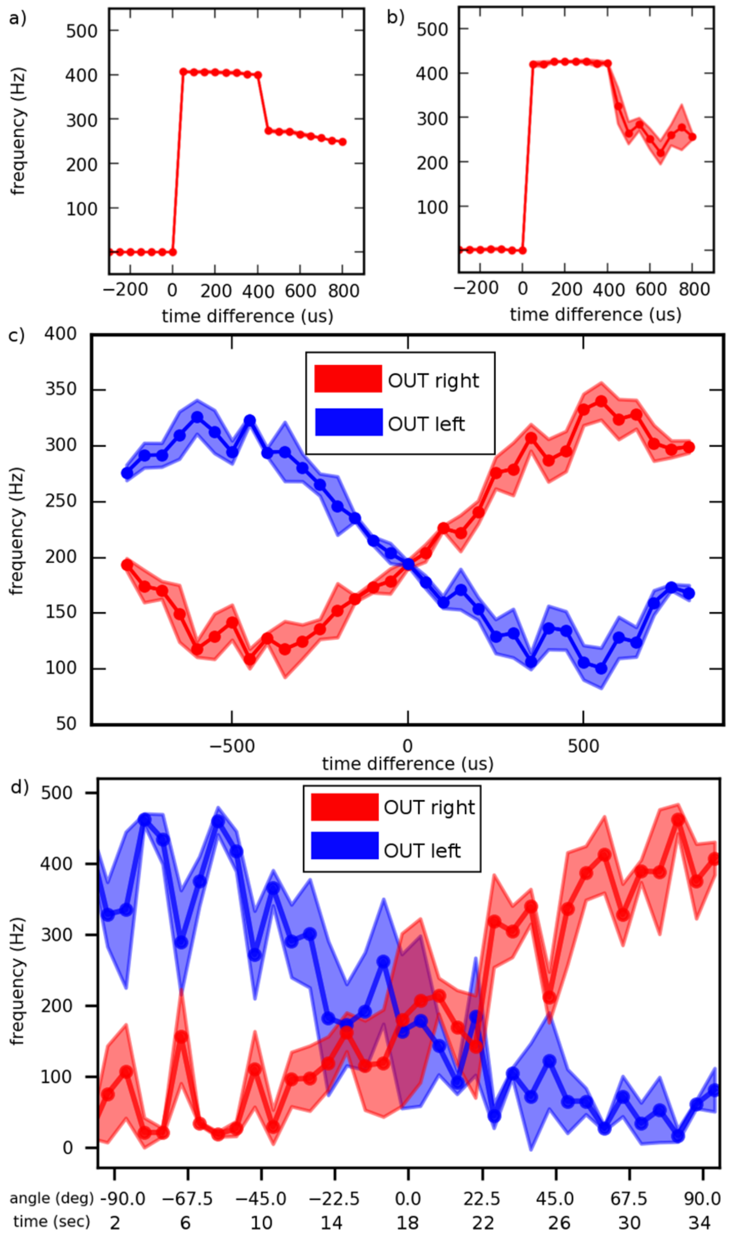

Figure 7. Response of the Time to Rate Network output to different ITDs inputs with different amounts and types of noise. (a) Response of right output LIF neuron in the Nest simulator to computer generated Poisson spike trains with ITDs between −800 and 800 μs. (b) Similar to (a) with 500 Hz noise. (c) Similar to (a) with right and left Time to Rate Network output with 500 Hz noise and Gaussian distributed time differences with a standard deviation of 40 μs. (d) Response of Time to Rate Network output on a SpiNN-3 board to shift of sound source location as described in figure 6.

spikes with a frequency of 500 Hz as shown in figure 7(b) the response becomes more irregular especially for large time differences. The response of the right and left output to POIS trains with 500 Hz noise and Gaussian distributed time differences is shown in figure 7(c). The response of the simulated output neurons increases linearly until a time difference of approximately 600 μs. For larger time differences, the output frequency decays slightly. The response of the Time to Rate Network on SpiNNaker is similar, only that activity increases linearly over the whole 180° range (figure 7(d)). A decrease in response difference during the motor turns can be observed. The output of the Time to Rate Network is very well suited to drive an actuator into the direction of a sound source since its response is proportional to the sound source angle. The performance of the whole closed-loop system is evaluated in the subsequent experiment.

4.2. Closed-loop localization

We evaluated the response of our closed-loop system shown in figure 2 in a real world SSL task. A speaker playing a constant 250 Hz beep, 500 Hz beep or random words from the Google speech command dataset

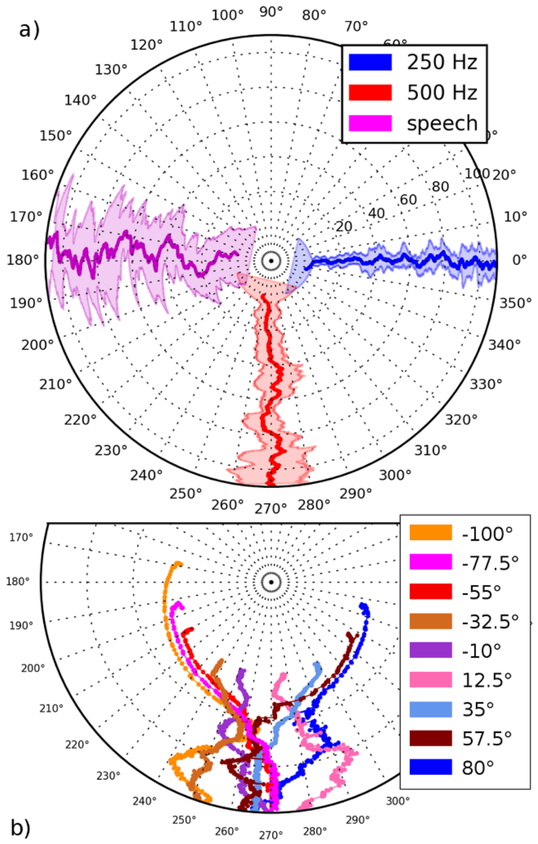

Figure 8. Azimuth angle of robotic platform. (a) Angle of robotic platform when playing a 250 Hz or 500 Hz pure tone or speech from a speaker 50 cm distant. Speaker location are zero, 270 and 180°. (b) Turning response with different starting angles for the first 15 s playing a 500 Hz input sound located around 270°. Legend shows deviation of start angle from sound source position.

was placed at 50 cm distance from the pan-tilt unit. The pan-tilt unit was initialized at nine different angles relative to the sound source angle (−100, −77.5, −55, −32.5, −10, 12.5, 35, 57.5 and 80°). In a first initial turn the pan-tilt unit moves towards the direction of the sound source. After this turn it oscillates around the sound source direction. The full duration of the experiment lasting 110 s is depicted in figure 8(a). The angle of the robotic platform and its standard deviation averaged over nine repetitions is displayed for 250 Hz, 500 Hz and human speech. The precision with which the pan-tilt unit faces the direction of the sound source decreases with increasing frequency and it is lowest for the speech command dataset (see figure 8(a): increase of standard deviation from 250 Hz to 500 Hz to speech). We calculated two values, the standard deviation of the rolling mean and the average of the rolling standard deviation. For 250 Hz, the standard deviation of the rolling mean starting at five seconds run time amounts 1.24° with an average rolling standard deviation of 4.2; for 500 Hz, 2.6° with a standard deviation of 7.9; and, for the speech dataset, 5.5° with a standard deviation of 16.9. Figure 8(b) displays single runs of the same experiment shown in figure 8(a) for a frequency of 500 Hz. The different starting angles of the pan-tilt unit and the first initial turn are visible. The sound source is located at an angle of 270°. In all nine cases the pan-tilt unit finishes its first initial turn towards the sound source during the first five seconds. The turning velocity of the pan-tilt unit stands in a clear linear relationship to the starting angle of the actuator as can be seen in figure 9. A large azimuth angle leads to a large Time to Rate Network output frequency. The high frequency difference causes the ring attractor bump and the corresponding Center Detector Network spikes to move along the ring of neurons with a high velocity. This leads to a fast turn of the pan-tilt unit into the coarse direction of the sound. For smaller angles a small output frequency leads to a slow movement of the ring attractor bump, Center Detector Network spikes and pan-tilt unit which corrects the angular deviation in a more precise manner without a lot of strong movements over the target location. This linear translation from sound source angle to Time to Rate Network output (see figure 10) to velocity of the ring attractor bump and Center Detector Network to velocity of the robotic platform (see figure 9) leads to a robust performance of the neuromorphic system. In figure 11 we analyze the closed-loop network dynamics of the system for a single run. At the

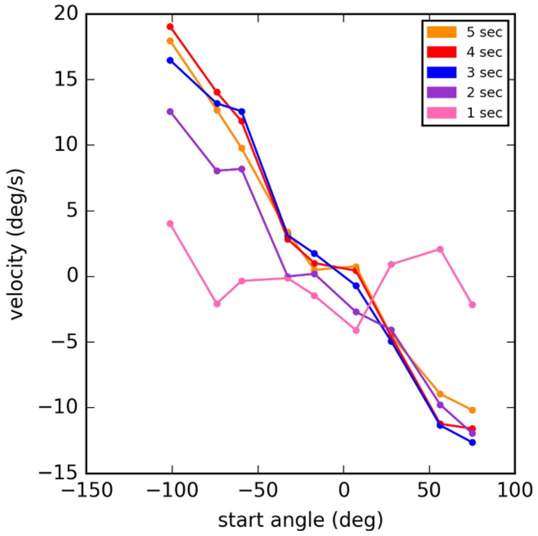

Figure 9. Start angle dependent turning velocity of pan-tilt unit for 500 Hz sound computed from data in figure 8(a). Mean velocity is calculated from a time window starting at time point zero with a duration between one and five seconds. Different colors refer to different time window size.

beginning of the experiment, the binaural cochlea is placed approximately −100° to the right of the speaker. $TDE_{(l0-l3)}$ (b) show a strong spiking response while $TDE_{(r0-r3)}$ (a) are almost silent. The left Time to Rate network output neuron is strongly active, close to its maximum spiking activity of 500 Hz (h), since it receives excitatory input from $TDE_{(r3,l3)}$ but very little inhibitory input from $TDE_{(r0-r2)}$. In contrast, the right Time to Rate network output (g) is inhibited by $TDE_{(l0-l2)}$. This rate difference in the output starts moving the Ring Attractor Network bump (c) to the left. The Center Detector Network (f) moves with the bump but with a much lower spiking activity at the center of the Ring Attractor Network bump. The pan-tilt unit (i) starts moving towards the sound-source-angle. When the pan-tilt unit reaches an angle of approximately −50° the activity of $TDE_{(r0-r3)}$ starts rising (a). This behavior matches the open-loop response profile in figure 6(d). Similarly the response of $TDE_{(r0-r3)}$ starts rising at approximately −45° reaching its maximum response between 22.5 and 67.5°. The increase in right TDE response leads to a slow almost linear decrease in the left Time to Rate Network output from 50 until zero degrees azimuth angle (h). The same response profile can be observed in the open-loop experiment in figure 7(d). Due to the decrease in spiking activity the movement velocity of the pan-tilt unit (i) slows down close to the angle of the sound source. This causes a smooth convergence of the pan-tilt unit towards the speaker direction. Therefore, our approach successfully encodes ITD into rate difference and rate difference into velocity. The system only requires very limited hardware resources so that it could be implemented on compact and energy efficient extreme edge platforms. The system can align itself with the direction of the sound source within seconds using only the information provided by the ITD. The ring attractor architecture is able to transform this information into the appropriate motor action. The physical system is already operating at its maximum speed. In figure 12 the velocity profile of the pan-tilt unit from the run in figure 11 is shown. The figure includes the initialization phase of the experiment (negative time). From second −12 until −8 the robotic platform turns to its initial position of −110° with its maximum velocity of around ±38° per second. From −8 until 0 s the robot is fixed at its start position. At second 0 the pan-tilt unit is released and starts moving towards the direction of the sound source reaching a velocity of up to 33° per second. This velocity is very close to the maximum velocity of ±38° per second. This demonstrates that the approach uses almost the whole dynamic range of the physical system leading to a latency of around five seconds. The movement velocity of the pan-tilt unit in the range of seconds is the limiting factor of the closed loop system. Only by moving to a different pan-tilt unit with higher velocities we can further improve the response time, enabling real-time capable SSL for robotics. Finally this work sets an important cornerstone in the context of embodied computation in fully SNNs. The system is one out of a handful of approaches which use events and spikes from the sensors to the actuators to perform closed loop tasks (for detailed reviews see [37, 38]). These approaches enable us to better understand the full potential of spike based computing for embodied systems.

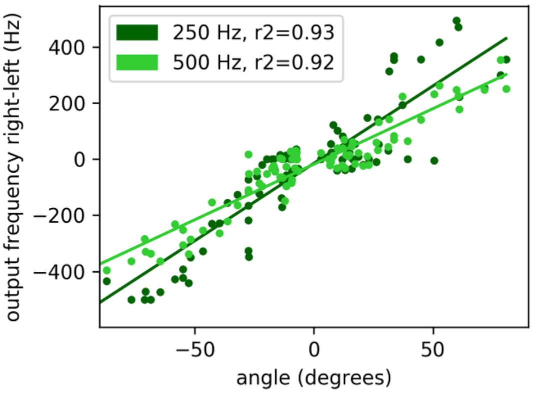

Figure 10. Response of the Time to Rate Network output neurons right minus left for different angles of the pan-tilt unit extracted from the closed-loop experiments.

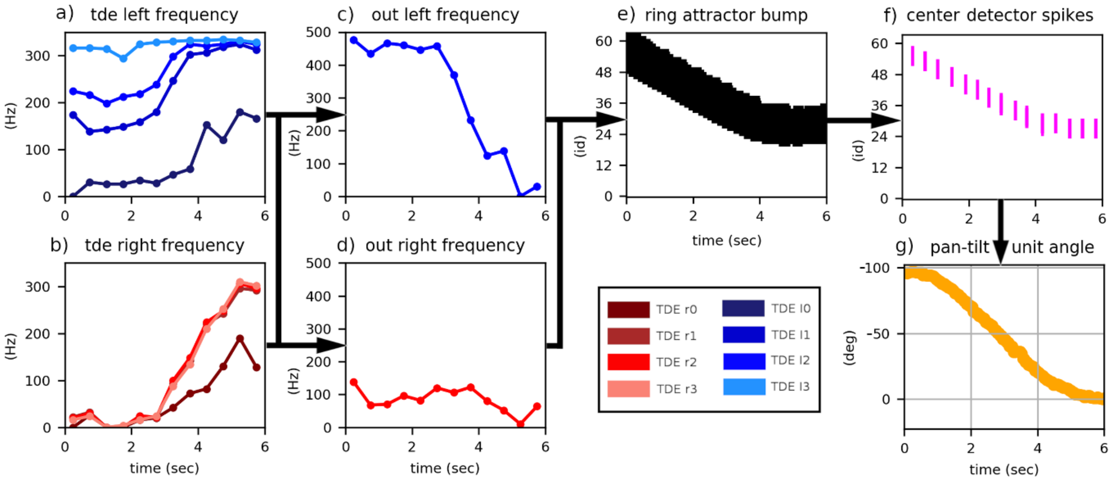

Figure 11. Network response for an example of closed-loop localization. Sound-source position is initialized 100° to the left of the system. (a), (b) Firing rate of four right and left TDEs. Response of left TDEs is much stronger at the beginning. (c), (d) Spiking rate of output neurons. While activity of the right output neurons stays low left output neuron starts close to 500 Hz, the possible maximum spiking frequency. At angles smaller than 50° it starts decaying almost linearly. (e) Rasterplot of one population of the Ring Attractor Network. Bump of spiking activity moves from top to center over time. (f) Rasterplot of Center Detector Network. Each bar is a single event. (g) Azimuth angle between robotic platform and sound source.

5. Discussion and conclusions

Our SSL system can successfully locate and track sound sources with low frequency components such as human speech. In a closed-loop experiment, the binaural cochlea turned into the direction of the sound source with a turning velocity linearly proportional to the angle difference between sound source and pan-tilt unit. After this initial turn, the binaural cochlea stays at the direction of the sound source.

The robotic system developed by Escudero *et al* [22] reaches in closed-loop a mean error of 1.9, 2.5 and 2.6° for 1, 2.5 and 5 kHz. The software implementation of Chan *et al* [4] performs with an error between 4 and 5° after training. Our closed-loop system estimates the position of a sound source with a mean error of $1.24 \pm 4.2, 2.6 \pm 7.9, 5.5 \pm 16.9°$ for a 250 Hz pure tone, 500 Hz pure tone and the Google speech command dataset, respectively. A detailed comparison of our system with other neuromorphic approaches is given in table 1. When using noise or pure tones, all mentioned neuromorphic approaches perform in an error range smaller than 5° (see table 1), comparable to human performance with a precision between 3 (frontal position) and 10° (lateral position) [39]. Our approach has mean errors comparable to the two others. However, there are significant differences in the working principles. Since Chan *et al* computes the location on Matlab software, the problem can be solved purely mathematically reaching a high precision. At the same time the system is limited in its real time capabilities and requires a lot of power. It uses a PC with a power consumption of more than 100 W and sequential processing. Our approach runs on neuromorphic hardware using approximately 1 W and parallel in memory computing. Furthermore, our system is the most compact neuromorphic solution for SSL using the ITD. Most animal-inspired approaches for SSL use arrays of

Table 1. Comparison of different neuromorphic sound source localization approaches. Our approach without the ring attractor requires the lowest number of computational units and works with the lowest spiking rate. Few computational units enable a compact ASIC implementation while a low spiking rate reduce the dynamic computation, hence the dynamic power consumption, of the system. Acronyms: neuron (neu), correlator (corr), hardware (HW), software (SW), open loop (OL), closed loop (CL), interaural time difference (ITD), interaural level difference (ILD), lateral superior olive (LSO).

Implementation	Type	Signal	# neu/corr	# synapse	Cochlea rate (Hz)	Detector rate (Hz)	Power (W)	Mean error (°)
OURS without ring attractor	HW OL	ITD	6	14	⩽250	⩽500	See below	Unknown
OURS with ring attractor	HW CL	ITD	326	∼400	⩽250	⩽500	NAS: 0.0297 TDE: 0.012 SpiNNaker chip: 0.93	250 Hz pure tone: 1.24 ± 4.2 500 Hz pure tone: 2.6 ± 7.9 Speech: 5.5 ± 16.9
[4]	HW/SW CL	ITD	∼3000	117.000	⩽6 k	Unknown	> 100	Pink noise 200–3 kHz: 5.0 White noise 3 kHz: 4.4
[22]	HW CL	ILD	128 LSO	0	⩽6 k	⩽6 k	0.058	1 kHz pure tone: 1.92 ± 0.94 2.5 kHz pure tone: 2.49 ± 0.79 5 kHz pure tone: 2.57 ± 0.35
[21]	HW OL	ILD	1024	∼1216	Unknown	Unknown	Unknown	Unknown

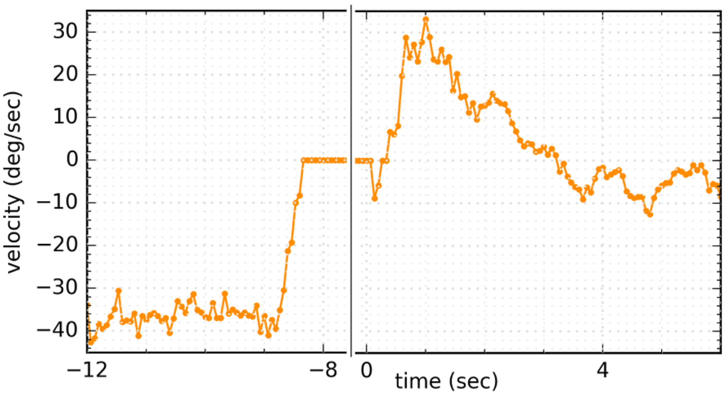

Figure 12. Velocity of the pan-tilt unit for the experiment shown in figure 11 including initialization before real experiment. From −12 until −8 s the pan-tilt unit turns to its initial position of −110°. This turn is performed with the maximum velocity of the pan tilt unit of around ±38° per second. At 0 s the experiment begins and the pan-tilt unit turns towards the direction of the sound source. At around 1 s the pan-tilt unit reaches a velocity of 33° per seconds and then slows down.

coincidence detectors which means that each frequency channel pair requires various different coincidence detectors with different delay lines. The time difference is encoded in a time difference map. In contrast, our approach converts the time difference into a rate which requires only a few TDEs leading to a much more compact solution. Also, our hardware implementation works already very well with only one frequency channel which further reduces the number of required computational units. Even when including the ring attractor our system is more compact than the approach by [4] (see table 1). Approaches using the ILD are usually more compact since they only need to convert the frequency of each channel into an angular location. However, while ITD works only with low sound frequencies ILD works only with high sound frequencies. Hence, by developing an approach which estimates both the ITD and ILD the best performance can be reached covering the full frequency range. In a future implementation we will combine [22] and our approach to develop such a system. Further improvements can be obtained by moving to a different pan-tilt unit. Our system is constrained by the physics of the robotic platform. The maximum speed of the pan-tilt unit lies at 38° per second (see initialization phase in figure 12) leading to response times in the range of seconds. By replacing the pan-tilt unit with a much faster model we could easily move the performance into the range of hundreds of milliseconds. Low power consumption, small network size and low latency make this implementation very well suited for real world robotic applications and computing on the edge.

While other approaches mentioned above only test with noise and pure tones, we evaluated the performance of the proposed system also with speech commands in order to test it in a more realistic scenario.

In this work we used a pan-tilt unit with a positional motor control scheme. Hence, we require a ring-attractor network to store the angular position of the robotic platform and update the position based on the output of the Time to Rate network. Such an implementation can be useful for robotic tasks in which the agent has to be aware of its own joint positions, for example an assistant robot which turns its head towards a human speaker to focus its attention. The approach could be deployed in the iCub humanoid robotic platform, similar to the Jeffress model implementation by [40]. In case of a differential motor control scheme the ring-attractor is not required. The output of the Time to Rate Network can directly be used for differential motor control, e.g. to move a mobile agent towards a human speaker. In that case, the movement velocity of the two motors is controlled by the spiking rate of the two Time to Rate Network output neurons using, for example, pulse frequency modulation [41]. Such a system can be used to detect and approach different types of sound sources. For example a rescue robot equipped with this system could move towards the direction of human voices to find possible endangered subjects.

The system used in this paper consists of two FPGAs, a SpiNNaker board, 3Dio microphones and a pan-tilt-unit. Only very few resources on the two FPGAs and the SpiNNaker board are used in our current implementation. The full binaural NAS implementation with two times (left and right) 64 frequency channels requires approximately eleven thousand slices on a Xilinx Virtex-5 FPGA with a power consumption of 29.7 mW [6]. Our current system only uses two out of the 128 frequency channels. The eight subsequent TDEs on FPGA require approximately 1120 registers and 1440 lookup tables with a power consumption of around 12 mW [13]. On SpiNNaker, the TDE input population of 8 neurons, the two Time to Rate Network output neurons, the 256 Ring Attractor Neurons and the 64 Center Detector Network neurons require two out of 18 cores on a single ARM chip. Each chip consumes 255 up to 930 mW depending on the intensity of computation [42]. Over all, our SNN uses eleven thousand slices, 1120 registers and 1440 lookup tables and one SpiNNaker chip with a power consumption of up to one Watt. Since the network is relatively small with only 326 neurons and around 400 synapses it could also be implemented on a standard micro-controller. For example Dabbous and colleagues [43] implemented an object-detection network of comparable size on a Raspberry Pi 4. However, their classification network reaches a high latency of 0.42 s when operating on a 10 μs time step. Furthermore, the power consumption of the Raspberry Pi 4 with 3 up to 7 W is much higher than the power usage of the system presented. While in this article we use a relatively large prototyping system consisting of three boards to evaluate the functionality of our SSL approach, our final aim is to implement the network onto a single mixed analog-digital asynchronous CMOS ASIC. In comparison to other neuromorphic approaches we use few computational units which allows for a compact design (see table 1 columns 4 and 5). We use low spiking frequencies which reduces the dynamic power consumption (see table 1 columns 6 and 7). At the same time, the accuracy of our system is comparable to other approaches (see table 1 column 9). Hence, our network is a very promising candidate for an ASIC implementation for edge computing and low power robotics.

Using the TDE on CMOS (1.4 nW–500 μW [13]), and the low-power LIF neuron on CMOS (20 μW–100 μW for 100 Hz [44]) we can aim at the design of a single ASIC with an overall power consumption in the double digit mW range or below. A comparison regarding power consumption to closed-loop deep learning approaches is not possible since, to our knowledge, there is no deep learning closed-loop real-time hardware implementation.

The implementation proposed in this article is, to the best of the authors' knowledge, the first neuromorphic hardware closed-loop SSL system using ITD capable of working in real time. Based on the current results using ITD for SSL, we aim to add an ILD and head-related transfer function part to the system to further increase the precision and also detect elevation angles. Adding more microphones, as done in many SSL implementations, can also increase the precision of the system. At the same time more processing and hence more power is needed to compute data from the additional microphone inputs. Therefore, the right trade off between sufficient precision and number of microphones needs to be found dependending on the application. We will explore this trade off in future applications. This work serves as a first approach towards achieving a closed-loop 360° neuromorphic SSL system. Our SSL system advances the state-of-the-art of neuromorphic event-based systems for robotics and embedded systems. Neuromorphic closed-loop hardware is a promising candidate for robotic systems due to its compactness, low power consumption and real time performance.

Data availability statement

The data that support the findings of this study are available at https://github.com/thorschoepe/SoundSourceLocalization [45].

Acknowledgment

This research was partially supported by the Spanish grant MINDROB (PID2019-105556GB-C33/AEI/10.13039/501100011033) and by the Cluster of Excellence Cognitive Interaction Technology (EXC 277), Bielefeld University, funded by the German Research Foundation. The authors would like to acknowledge the financial support of the CogniGron research center and the Ubbo Emmius Funds (Univ. of Groningen).

The authors would like to thank Terry Stewart, Herman Adema and Antonio Rios-Navarro for their technical support. We are thankful to Nicoletta Risi for her feedback on the manuscript.

ORCID iDs

Thorben Schoepe https://orcid.org/0000-0002-7048-7358
Angel Jimenez-Fernandez https://orcid.org/0000-0003-3061-5922

References

[1] Grumiaux P-A, Kitić S, Girin L and Guérin A 2022 A survey of sound source localization with deep learning methods *J. Acoust. Soc. Am.* **152** 107–51

[2] Evers C, Lollmann H W, Mellmann H, Schmidt A, Barfuss H, Naylor P A and Kellermann W 2020 The locata challenge: acoustic source localization and tracking *IEEE/ACM Trans. Audio, Speech Lang. Process.* **28** 1620–43

[3] Dávila-Chacón J, Liu J and Wermter S 2019 Enhanced robot speech recognition using biomimetic binaural sound source localization *IEEE Trans. Neural Netw. Learn. Syst.* **30** 138–50

[4] Chan V Y-S, Jin C T and van Schaik A 2012 Neuromorphic audio-visual sensor fusion on a sound-localising robot *Front. Neurosci.* **6** 21

[5] Gallego G *et al* 2022 Event-based vision: a survey *IEEE Trans. Pattern Anal. Mach. Intell.* **44** 154–80

[6] Jiménez-Fernández A, Cerezuela-Escudero E, Miro-Amarante L, Dominguez-Moralse M J, de Asis Gomez-Rodriguez F, Linares-Barranco A and Jimenez-Moreno G 2017 A binaural neuromorphic auditory sensor for FPGA: a spike signal processing approach *IEEE Trans. Neural Netw. Learn. Syst.* **28** 804–18

[7] Thakur C S *et al* 2018 Large-scale neuromorphic spiking array processors: a quest to mimic the brain *Front. Neurosci.* **12** 891

[8] Milde M B, Bertrand O J N, Ramachandran H, Egelhaaf M and Chicca E 2018 Spiking elementary motion detector in neuromorphic systems *Neural Comput.* **30** 2384–417

[9] Schoepe T, Gutierrez-Galan D, Dominguez-Morales J P, Jimenez-Fernandez A, Linares-Barranco A and Chicca E 2019 Neuromorphic sensory integration for combining sound source localization and collision avoidance *2019 IEEE Biomedical Circuits and Systems Conf. (BioCAS)* pp 1–4

[10] D'Angelo G, Janotte E, Schoepe T, O'Keeffe J, Milde M B, Chicca E and Bartolozzi C 2020 Event-based eccentric motion detection exploiting time difference encoding *Front. Neurosci.* **14** 451

[11] Haessig G, Milde M B, Aceituno P V, Oubari O, Knight J C, van Schaik A, Benosman R B and Indiveri G 2020 Event-based computation for touch localization based on precise spike timing *Front. Neurosci.* **14** 420

[12] Mastella M and Chicca E 2021 A hardware-friendly neuromorphic spiking neural network for frequency detection and fine texture decoding *2021 IEEE Int. Symp. on Circuits and Systems (ISCAS)* pp 1–5

[13] Gutierrez-Galan D, Schoepe T, Dominguez-Morales J P, Jimenez-Fernandez A, Chicca E and Linares-Barranco A 2021 An event-based digital time difference encoder model implementation for neuromorphic systems *IEEE Trans. Neural Netw. Learn. Syst.* **33** 1–15

[14] Rascon C, Meza I 2017 Localization of sound sources in robotics: a review *Robot. Auton. Syst.* **96** 184–210

[15] Faraji M M, Shouraki S B and Iranmehr E 2015 Spiking neural network for sound localization using microphone array *2015 23rd Iranian Conf. on Electrical Engineering* pp 1260–5

[16] Kriener L 2014 Binaural sound localization in spiking neural networks Bachelor Thesis University of Heidelberg

[17] Glackin B, Wall J, McGinnity T, Maguire L and McDaid L 2010 A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization *Front. Comput. Neurosci.* **4** 18

[18] Wall J A, McDaid L J, Maguire L P and McGinnity T M 2012 Spiking neural network model of sound localization using the interaural intensity difference *IEEE Trans. Neural Netw. Learn. Syst.* **23** 574–86

[19] Goodman D and Brette R 2010 Learning to localise sounds with spiking neural networks *Advances in Neural Information Processing Systems* pp 784–92

[20] Pan Z, Zhang M, Wu J, Wang J and Li H 2021 Multi-tone phase coding of interaural time difference for sound source localization with spiking neural networks *IEEE/ACM Trans. Audio, Speech Lang. Process.* **29** 2656–70

[21] Oess T, Löhr M, Jarvers C, Schmid D and Neumann H 2020 A bio-inspired model of sound source localization on neuromorphic hardware *2020 2nd IEEE Int. Conf. on Artificial Intelligence Circuits and Systems (AICAS)* pp 103–7

[22] Escudero E C, Peña F P, Vicente R P, Jimenez-Fernandez A, Moreno G J and Morgado-Estevez A 2018 Real-time neuro-inspired sound source localization and tracking architecture applied to a robotic platform *Neurocomputing* **283** 129–39

[23] Lyon R 1982 A computational model of filtering, detection and compression in the cochlea *ICASSP 1982. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* vol 7 (IEEE) pp 1282–5

[24] The address-event representation communication protocol (available at: www.ini.uzh.ch/~amw/scx/std002.pdf)

[25] Jimenez-Fernandez A, Linares-Barranco A, Paz-Vicente R, Jiménez G and Civit A 2010 Building blocks for spikes signals processing *The 2010 Int. Joint Conf. on Neural Networks (IJCNN)* (IEEE) pp 1–8

[26] Gutierrez-Galan D, Dominguez-Morales J P, Jimenez-Fernandez A, Linares-Barranco A and Jimenez-Moreno G 2021 Opennas: open source neuromorphic auditory sensor HDL code generator for FPGA implementations *Neurocomputing* **436** 35–38

[27] Iakymchuk T, Muñoz A R, Serrano-Gotarredona T, Linares-Barranco B, Jiménez-Fernandez A, Linares-Barranco A and Jiménez-Moreno G 2014 An AER handshake-less modular infrastructure PCB with x8 2.5Gbps LVDS serial links *2014 IEEE Int. Symp. on Circuits and Systems (ISCAS)* pp 1556–9

[28] Rosen S and Howell P 2011 *Signals and Systems for Speech and Hearing* vol 29 (Leiden: Brill)

[29] Painkras E, Plana L A, Garside J, Temple S, Galluppi F, Patterson C, Lester D R, Brown A D and Furber S B 2013 SpiNNaker: a 1-W 18-core system-on-chip for massively-parallel neural network simulation *IEEE J. Solid-State Circuits* **48** 1943–53

[30] Davison A P, Brüderle D, Eppler J, Kremkow J, Muller E, Pecevski D, Perrinet L and Yger P 2009 PyNN: a common interface for neuronal network simulators *Front. Neuroinform.* **2** 11

[31] Rhodes O *et al* 2018 sPyNNaker: a software package for running PyNN simulations on SpiNNaker *Front. Neurosci.* **12** 816

[32] Plana L A, Garside J, Heathcote J, Pepper J, Temple S, Davidson S, Luján M, Furber S 2020 spiNNlink: FPGA-based interconnect for the million-core SpiNNaker system *IEEE Access* **8** 84918–28

[33] Linares-Barranco A, Jimenez-Moreno G, Linares-Barranco B and Civit-Balcells A 2006 On algorithmic rate-coded AER generation *IEEE Trans. Neural Netw.* **17** 771–88

[34] Zhang K 1996 Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory *J. Neurosci.* **16** 2112–26

[35] Turner-Evans D, Wegener S, Rouault H, Franconville R, Wolff T, Seelig J D, Druckmann S and Jayaraman V 2017 Angular velocity integration in a fly heading circuit *eLife* **6** e23496

[36] Warden P 2018 Speech commands: a dataset for limited-vocabulary speech recognition (arXiv:1804.03209)

[37] Bartolozzi C, Indiveri G and Donati E 2022 Embodied neuromorphic intelligence *Nat. Commun.* **13** 1024

[38] Sandamirskaya Y, Kaboli M, Conradt J and Celikel T 2022 Neuromorphic computing hardware and neural architectures for robotics *Sci. Robot.* **7** eabl8419

[39] Risoud M R, Hanson J-N, Gauvrit F, Renard C, Lemesre P-E, Bonne N-X and Vincent C 2018 Sound source localization *Eur. Ann. Otorhinolaryngol. Head Neck Dis.* **135** 259–64

[40] Gutierrez-Galan D, Bartolozzi C, Dominguez-Morales J P, Jimenez-Fernandez A and Linares-Barranco A 2022 Towards the neuromorphic implementation of the auditory perception in the iCub robotic platform *Neuro-Inspired Computational Elements Conf. (NICE 2022)* (New York: Association for Computing Machinery) pp 11–12

[41] Jiménez-Fernandez A, Jiménez-Moreno G, Linares-Barranco A, Domínguez-Morales M J, Paz-Vicente R and Balcells A A C 2012 A neuro-inspired spike-based PID motor controller for multi-motor robots with low cost FPGAs *Sensors* **12** 3831–56

[42] Sugiarto I, Liu G, Davidson S, Plana L A and Furber S B 2016 High performance computing on spinnaker neuromorphic platform: a case study for energy efficient image processing *2016 IEEE 35th Int. Performance Computing and Communications Conf. (IPCCC)* pp 1–8

[43] Dabbous A, Ibrahim A, Alameh M, Valle M and Bartolozzi C 2022 Object contact shape classification using neuromorphic spiking neural network with STDP learning *2022 IEEE Int. Symp. on Circuits and Systems (ISCAS)* pp 1052–6

[44] Indiveri G, Chicca E and Douglas R 2006 A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity *IEEE Trans. Neural Netw.* **17** 211–21

[45] Schoepe T 2023 *Sound Source Localization, Github* (available at: https://github.com/thorschoepe/SoundSourceLocalization)