Real-time neuro-inspired sound source localization and tracking architecture applied to a robotic platform

Elena Cerezuela Escudero[a], Fernando Pérez Peña[b], Rafael Paz Vicente[a], Angel Jimenez-Fernandez[a], Gabriel Jimenez Moreno[a], Arturo Morgado-Estevez[b]

[a] Robotics and Computer Technology Lab (RTC), Universidad de Sevilla, ETSI Informática, Avd. Reina Mercedes s/n, Seville 41012, Spain
[b] Applied Robotics Research Lab, Universidad de Cádiz, Faculty of Engineering, Avda. Universidad, 10, Puerto Real, Cadiz 11519, Spain

Keywords:
Sound localization
Interaural intensity difference
Spike signal processing
Neuromorphic auditory sensor
Neurorobotics
FPGA

A B S T R A C T

This paper proposes a real-time sound source localization and tracking architecture based on the ability of the mammalian auditory system using the interaural intensity difference. We used an innovative binaural Neuromorphic Auditory Sensor to obtain spike rates similar to those generated by the inner hair cells of the human auditory system. The design of the component that obtains the interaural intensity difference is inspired by the lateral superior olive. The spike stream that represents the IID is used to turn a robotic platform towards the sound source direction. The architecture was implemented on FPGA devices using general purpose FPGA resources and was tested with pure tones (1-kHz, 2.5-kHz and 5-kHz sounds) with an average error of 2.32°. Our architecture demonstrates a potential practical application of sound localization for robots, and can be used to test paradigms for sound localization in the mammalian brain.

1. Introduction

Sound localization is a function that the ears, auditory pathways and auditory cortex of the brain perform together to determine the source of a sound. It is a powerful feature of mammalian perception that allows the animal to be aware of the environment and to locate prey and predators. This has inspired researchers to develop new computational models of the auditory pathways and biological mechanisms that underlie sound localization in the brain.

The ability to model the ways in which mammals locate a sound source can improve the perception and navigation of mobile robots, allow the development of better virtual realities, improve teleconferencing, provide surveillance systems with omnidirectional sensitivity, and enhance hearing aids.

During the last decades, the structure and function of pathways in the auditory brainstem for sound localization have been extensively studied [1–4]. The direction of a sound in the horizontal plane is determined by a combination of binaural cues derived from the incident acoustic waves arriving at the ear from different angles: interaural time difference (ITD) and interaural intensity, or level, difference (IID or ILD, respectively). Sounds that do not generate directly in front of or behind the receptor arrive earlier at one ear than at the other, creating an ITD. For wavelengths roughly equal to, or shorter than, the diameter of the head, a shadowing effect is produced at the ear that is further away from the source, creating an IID [1,2].

For example, in general terms, if a pure tone sound source is positioned on the left side, the sound signal at the left ear is represented by the equation:

$$Left_{signal} = a \times \sin(2\pi f t) \quad (1)$$

where a is the sound amplitude, f the sound frequency and t the time. The sound at the right ear is represented by the equation:

$$Right_{signal} = (a/\Delta a) \times \sin(2\pi f(t - \Delta t)) \quad (2)$$

where Δa and Δt are related to, respectively, the intensity difference (IID) caused by the shadowing effect of the head, and the additional time (ITD) required for the sound wave to travel the further distance to the right ear.

Due to the head size, the ITD cue in humans is effective for locating low frequency sounds (20 Hz - 1 kHz). However, the information it provides becomes ambiguous for frequencies above 1 kHz. In contrast, the IID cue is not useful for locating sounds below 1 kHz, but it is more efficient than the ITD cue for mid-and high-frequency (<1 kHz) sound localization.

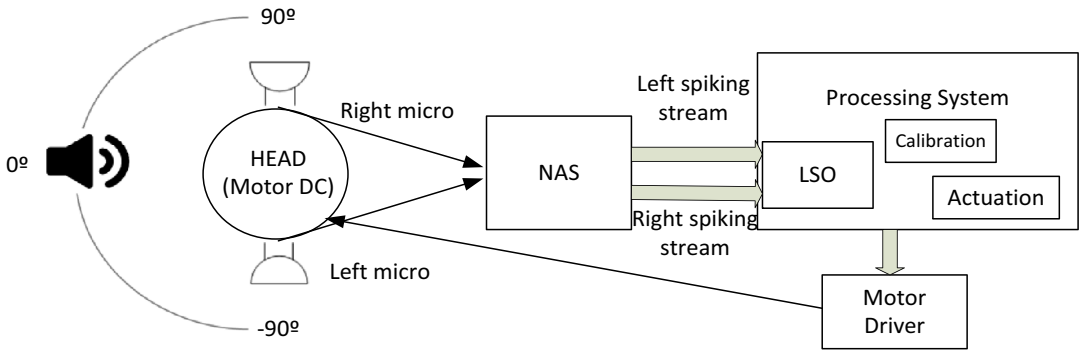

Fig. 1. Tracking Sound system architecture.

The ITD and IID cues are extracted in the medial and lateral superior olive, respectively (MSO and LSO) which are located within an area of the auditory system called the superior olivary complex [2]. The LSO has a tonotopical organization: high frequencies are represented in the middle of the LSO and continually decreasing frequencies to both sides [5]. LSO neurons are inhibited by sounds to the contralateral ear and excited by sounds to the ipsilateral ear, resulting in a neural form of subtraction [2].

There is an increasing demand for the development of real-time and low-power sound localization techniques in the industry of hearing aid [6–8] and robotic applications [9–11]. Currently, various digital processing techniques based on Fast Fourier Transform (FFT) have been proposed to determine the source of a sound signal [12,13]. However, these techniques require high power consuming devices, such as Digital Signal Processors (DSP), and memory to perform such complex signal processing. Traditional Digital Signal Processing (DSP) techniques commonly apply Multiply-Accumulate (MAC) operations over a collection of discrete samples codified as fixed or floating point representations. MAC operations often require dedicated and complex resources, i.e. float-point multipliers, which are available in FPGAs as dedicated expensive resources in relatively small quantities. Therefore, applying a sequence of MAC operations over a dataset with these units requires multiplexing them in time. So they are reused with different input data and output results, which are stored in a global memory. It often requires high frequency clock signals to achieve a competitive data throughput. Furthermore, large memory depths to store intermediate data and results are needed. These facts are reflected in the power consumption and circuitry complexity. On the other hand, Spike Signal Processing (SSP) implements the basic operations that commonly are performed in DSP, but over spike rate coded signals [14]. Thus, operations are performed directly over spike streams, being equivalent to simply adding or removing spikes at the right moment (although it is not evident which). The circuits that implement SSP operations use general purpose FPGA resources, as counters, comparators and logic gates. This allows the building of large scale dedicated systems in hardware, which process spike coded signals in real time using low frequency clocks in a fully parallel way for (low cost) FPGAs, for example, the auditory sensor used in this work demands 29.7 mW for 64 channels in stereo operation [15].

The ability to replicate the ways in which mammals locate a sound source could allow the development of better virtual realities. In addition, the performance of robotics with lower power consumption will be increased. Furthermore, hearing aids could be enhanced by improving the localization of individual sounds. These improvements, which are enabled by the ability to understand and mimic mammalian sound localization, are the main reasons for the research carried out in this paper.

The aim of this research involves the development of a spike-based system that processes and extracts the binaural cue of IID with a topology inspired by the mammalian auditory pathways, specifically the LSO. Using the IID cue, the system performs the task of tracking a sound in real time, in a biologically inspired way.

In this paper, to obtain spike rates similar to those generated by the inner hair cells of the human auditory system, we used an innovative binaural Neuromorphic Auditory Sensor (NAS). This decomposes an audio signal into different frequency bands where the audio information is encoded in the spike rates [15]. Using the out coming spike rates from the NAS as the stimulus to the LSO model we propose, the whole architecture deals with biologically inspired data. The NAS, the LSO model and the actuation system have been implemented on FPGA devices using general purpose FPGA resources. These models are developed using SSP techniques [14].

There are previous works that propose audio localization systems inspired by the mammalian auditory system: the papers by [8] and [11] reported that a neuromorphic silicon cochlea can be used for spatial audition and auditory scene analysis; both papers were based on ITD. In [8], the sound localization circuit was devised by mimicking the neuronal organization of barn owl's auditory pathway to obtain ITD.

The works of [16] and [17] proposed a Spiking Neural Network (SNN) to partially simulate the superior olivary complex, but they did not use a neuromorphic device to obtain the spike streams that represent sound. In the system proposed in [16], the input sound passes through a Gammatone filterbank and is then encoded into phase-locked spikes using a model of the half-wave rectified receptor potential of inner hair cells. ITD processing uses a series of delays and a leaky integrate-and-fire neuron model; the ITD is calculated for all frequency channels to form a full map of ITD processing. IID processing does not use a neuron model; instead, a logarithmic ratio computes the intensity difference. The model classifies the sound source between 7 discrete azimuthal angles (from $-90°$ to $90°$ in steps of $30°$). The model was tested using a robotic head on broadband sounds, both noise and speech, and it achieved overall localization accuracies of 80%. The paper by [17] presented a SNN architecture to simulate the sound localization ability of the mammalian auditory pathways using the IID. To train and validate the localization ability of the architecture, experimentally derived head-related transfer function acoustical data from adult domestic cats were employed; the supervised learning algorithm known as "remote supervision method" was used for the training to determine the azimuthal angles. The SNN classified the sound source between 13 discrete azimuthal angles (from $-60°$ to $60°$ in steps of $10°$). The experimental results using the same sound frequency used for the training were 52% for 5-kHz sounds, 83% for 15-kHz sounds and 40% for 25-kHz sounds. Reference [18]

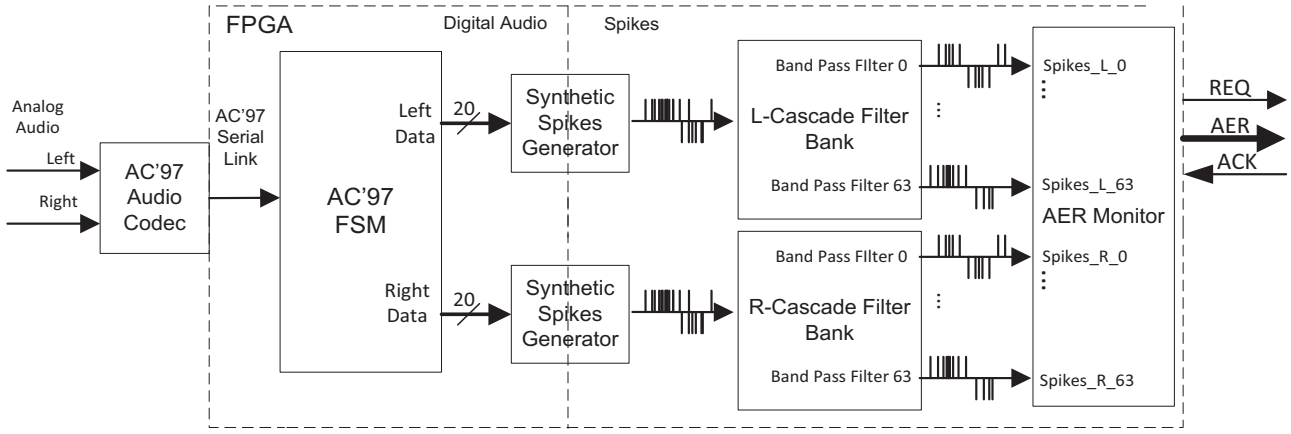

Fig. 2. NAS Architecture.

proposed an FPGA implementation of a sound localization, based on pulsed neural network. The pulsed neural network extracts the ITD to classify between 7 angle classes with a resolution of 30°. In reference [19], the authors obtain the ILD from the energy levels of multiple microphone pairs and propose an algorithm to extract the bearing angle.

The present work is structured as follows. Section 2 presents the global architecture of the sound localization and tracking system. In this section, the LSO model that obtains the IID cue and the processing system that implements that model to track a sound are described in detail. In Section 3, the experimental results are presented to show the feasibility and performance of the sound localization and tracking system. Finally, a conclusion section is presented in section 4.

2. Architecture of real-time sound localization and tracking system

In this work we present a neuromorphic real-time sound tracking system which consists of a neuromorphic auditory system. The aim is to model the functionality of the LSO to locate and track high-frequency sounds in a biologically inspired way. The system proposed obtains the IID auditory cue in the same way the LSO does. Then, the IID auditory cue is used as the input for the spike-based processing system that tracks the sound. Fig. 1 shows the experimental setup to locate and track the sound source using the NAS, and a Processing System based on LSO and Actuation components. The experimental setup consists of a robotic platform based on a cork porexpan head with one microphone on each side of the head. The sound signals from the microphones are sent to the NAS where they are decomposed into different frequency bands. Then, the NAS produces spikes that represent the spectral information of the original audio signal; the final output of the NAS are the binaural spiking stream flows. These binaural cues are sent to the LSO system, which extracts the IID. Finally, the IID spiking streams are used by the calibration system and actuation system to track the sound source.

This section briefly describes the NAS, the LSO model and the processing system.

2.1. Neuromorphic auditory system

To generate the spike streams inspired by the human auditory system, a neuromorphic device proposed in a previous work [15] was used, which decomposes an audio signal into different frequency bands of spiking information, in the same way a biological cochlea sends the audio information to the brain. The biological cochlea performs the transduction between the pressure signal representing the acoustic input and the neural signals that carry information to the brain. Due to the physical characteristics of a part of the cochlea, the basilar membrane, the cochlea divides an input signal into its frequency components. Thousands of hair cells on the membrane generate action potentials, or spikes, that travel along nerve fibres to higher-order auditory brain areas.

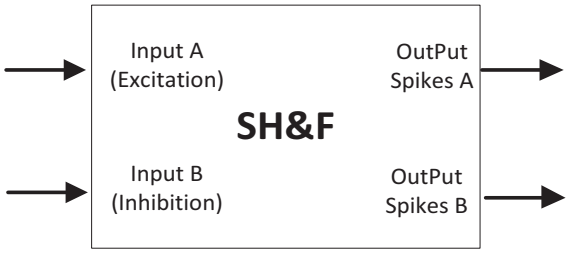

Fig. 3. Spike Hold & Fire block.

The architecture of the NAS is shown in Fig. 2. The system's inputs are the digitalized audio streams (the left and right audio signals), which represent the audio signals of a binaural system. A Synthetic Spike Generator [14,20] converts these digital audio sources into two spike streams. Then, the cascade band pass filter bank splits the spike streams into 64 frequency bands using 64 different spiking outputs. These outputs are combined by a monitor block which encodes each spike according to Address-Event Representation (AER) and transmits this information to the classification layers [21]. All the elements required for designing the NAS components (Synthetic Spike Generators, cascade filter bank and the AER monitor) have been implemented in VHDL and designed as spike-based building blocks [14]. The L-Cascade Filter bank and R-Cascade Filter bank have identical architecture. Table 1 shows the NAS features. The NAS has been used before in [22] to measure the speed of DC motor and in [23–25] for audio sound classification. In [26] a software tool (NAVIS) to develop the first post-processing layer using the NAS information is proposed.

2.2. Lateral superior olive model

We present a LSO model which decodes the IID from the binaural spiking streams generated by the NAS.

For sound waves that have similar or smaller wavelengths than the diameter of the human head, a shadowing effect is produced at the ear further from the source, creating an IID [1]. Processing in the LSO involves taking as input the two sound signals in the form of a neural stimulus from each ear. The ipsilateral stimulus takes an excitatory form and the contralateral is inhibitory, after

Table 1
NAS characteristics.

Number of bands	Frequency range	Dynamic range	Max. Event rate	Clock frequency	Hardware resources
64×2	9.6 Hz – 14.06kHz	75dB	2.19Mevents/s	27 MHz	11,141 slices

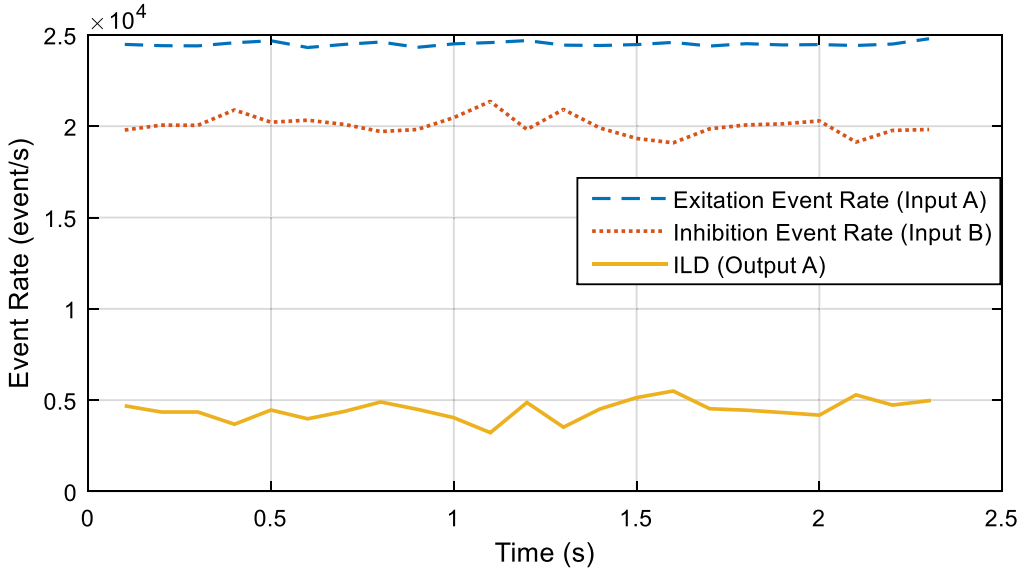

Fig. 4. ILD (encoded as an event-rate) measured over time obtained from the input data that generates the left and right NAS when the sound source is placed at 45° from the head.

passing through the medial nucleus trapezoid body (MNTB). The convergence of excitatory inputs from the ipsilateral ear and inhibitory inputs from the opposite ear resembles a relatively simple subtraction process that creates the well-described IID sensitivity of LSO neurons. These neurons produce an output related to the IID [2].

The architecture of the system is shown in Fig. 3. It is based on the Spike Hold & Fire (SH&F) block that transforms two input spike rates into two output spike rates. The outputs are proportional to the subtraction between the two input rates. Subtracting a spike-based input signal from another will yield a new spike signal whose spike rate will be proportional to the difference between both input spike rates. If that subtraction is positive, the spike is fired by output port A, and if the subtraction is negative, the spike is fired by output port B [27,28]. This block was successfully used in previous works for spike filtering design [14,15,29] and for implementing a spike-based closed-loop robotic controller [28].

The function of the SH&F is to hold incoming spikes for a fixed period of time while monitoring the input evolution to determine output spikes. Fig. 3 shows the SH&F block, which has two inputs: A (Excitation input) and B (Inhibition input). If a spike is received at input port A, an "A" state is held internally. In the case that no spikes are received, nothing is done. When a new spike arrives, it behaves in one way or another, according to the spike input port. If SH&F receives an excitatory spike (port A), the held spike is fired, and a new "A" state is held internally. If a spike is received in port B, the held spike is cancelled and no output spike is fired at any output port. Similar SH&F behavior can be extended to input spikes in port B using the same logic: hold, cancel, and fire spikes according to input spikes ports. With this implementation of LSO, the MNTB function, in order to obtain inhibitory input, is carried out by input B of the SH&F block. The output Spikes A firing rate is proportional to the ILD corresponding to the sound originating to the left of the head and the output Spikes B firing rate is proportional to the ILD corresponding to the sound originating to the right. Fig. 4 shows an example of how the SH&F works: the two inputs and output along time are shown to better explain the model functioning, in order to exhibit how the inhibition/excitation principle influences the output.

This model considers the SH&F block to be akin to the subset of biological neurons of the LSO that deal with the narrow frequency bands of sound to track.

2.3. Processing system

The architecture of the processing system is shown in Fig. 5. This system interfaces the NAS system, processes the data from the NAS generating the IID cue and produces the commands to drive the motor.

The NAS will send the events from the 64 bands for both the right and left microphone-ear. To characterize our system, the processing system only samples the bands (left and right) responding to the fundamental frequency of the sound to track.

The architecture consists of a calibration stage where the baseline is established (i.e. ILD = 0 dB), three SH&F blocks where the ILD and the error are computed, two spike generators [20] to generate the reference, a serial communication component and a spike lengthening mechanism.

The behavior of this layer is divided into two steps: initial calibration and normal behavior.

Initial calibration

1. The calibration process is the first step that takes place. There is an initial five second countdown to let us prepare the scenario, i.e. to place the speaker in front of the head (0°) and play a sound with the same fundamental frequency as the sound to track. The rest of the components of the actuation layer will be disabled during this entire calibration phase.
2. After the five second countdown, two components are launched: a four second countdown and the baseline rate component (Fig. 5). This last component receives the addresses from the AER bus interfacing the NAS system. It includes two registers where the events received from each ear are counted (only the channels of interest). Then, at the end of the four

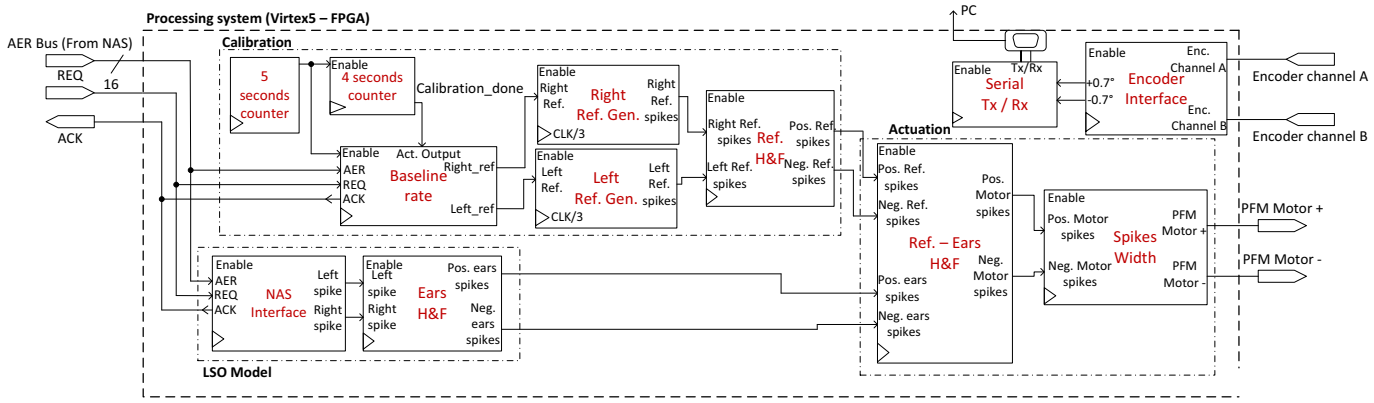

Fig. 5. Block diagram of the processing architecture. For the sake of simplicity, we have avoided drawing the asynchronous reset of all components of the architecture and the way signals are combined. For instance, the ACK out signal can be driven by the NAS interface component and the calibration component, so a multiplexer driven by the "calibration_done" signal, is used to select which component is sending the ACK out of the FPGA. Also, this "calibration_done" signal will enable the rest of the components that are not involved in the calibration stage.

second countdown, we will have the total number of events received from the right and left ear. Then, both registers are shifted two positions to the right (division by four) so the result in the registers are the events per second received by each ear considering the reference condition: the head facing the direction of the sound source. Therefore, these rates will play the role of reference for our control system.

3. The way to generate the rate computed only in the previous stage is to use the spikes generator presented in [20]. There are two of these modules used in our architecture: right reference generator and left reference generator; each will receive the input data from the baseline rate component.

   The rate generated by this block is defined according to the following equation:

   $$Firing\ rate = f_{clk}/2^{(N-1)} \times input = gain \times input$$

   Since we have calculated the desired output rate (firing rate in the above equation) we need to know the input we have to supply to this module to generate such a rate. Then, if the gain of the generator is a power of two values, it is easy to right-shift the computed rate again, in order to have the correct input value for the generators.

   If we consider a clock frequency of 100 MHz and then this data is introduced into the equation, it is not possible to reach a power of two gain by the generator. To achieve it, we have divided the clock frequency by three and used 24 bits to implement the generator, resulting in a gain of $1.98 \sim 2$.

   Implementing the generator in this way allows us to right-shift the reference rate calculated in the baseline rate component and supply the result as the input for the spike generators.

4. The output spiking rates of both generators will be subtracted using the reference Hold & Fire component shown in Fig. 5. The resulting rate of spikes is the reference for our design when the normal tracking behavior is taking place.

Normal behavior

During the normal phase, the NAS interface block is receiving the addresses from the AER bus and detecting the channels of interest, which depend on the frequency of the sound to track/locate. Once an event on any of these channels is received, it is propagated to the Ears SH&F component in the architecture where the rates are subtracted (the channels tuned for the same frequency of each ear are cancelled) to compute the ILD (Fig. 6). This SH&F block implements the LSO model to obtain a spike rate that represents the IID cue of a narrow frequency band of sound. To remove the error caused by gain difference between the microphones, the resulting rate is compared with the reference computed previously during the calibration process (in Ref.-Ears SH&F). The final rate (error if we compare with a conventional control system) is sent to the module responsible for lengthening the time duration of the events. The reason for doing this, is to avoid filtering events by the motor (with a 100 MHz clock frequency, the time length of an event will be 10 ns and this length will be filtered out by the motor).

The event time length can be configured from the PC and this length will be the same for all the events sent to the motor. Using Pulse Frequency Modulation (PFM) removes the delay associated with PWM and makes the direct use of the events available [30].

Finally, the encoder channels are read and this information is serially transmitted to the PC whenever the motor turns 0.7° in any direction.

One of the differences between our architecture and the ones proposed in [16] and [17] is: they did not use a neuromorphic device to obtain the spike streams that represent sound. Instead of this, in [17], they used an experimentally derived, head-related, transfer function acoustical data for each ear and, in [16], a second-order Gammatone filterbank was used. This means that the system in [17] has a high computational cost to be implemented [31,32] and the system proposed in [16] needs high power-hungry devices to be implemented [33]. The SSN architecture presented in [17] is different, depending on if the sound arrives from the left or the right side of the head and it needs a MNTB node to obtain the inhibitory input. The architecture presented in this paper is able to receive the sound from left or right and does not need an inhibitory node because the SH&F is able to subtract signed spikes. The SNN of [17] needs a receptive field layer and an output layer to determine the angle (each angle corresponds to a SNN class). However, in this work we do not need these layers because the IID cue obtained by the processing layer is used to control the motor. All neurons in the architecture presented in [17] used a LIF model and they have been trained by the supervised learning algorithm known as "remote supervision method" . Our architecture does not need training stage but it does need a calibration stage that takes 4 s.

3. Experimental results

The experimental setup is shown in Fig. 1. The distance between the speaker and the head is 40 cm at different azimuthal angles (0° to 90° in steps of 15°). The microphones are on each side of the head (omnidirectional pick-up pattern). The head is placed

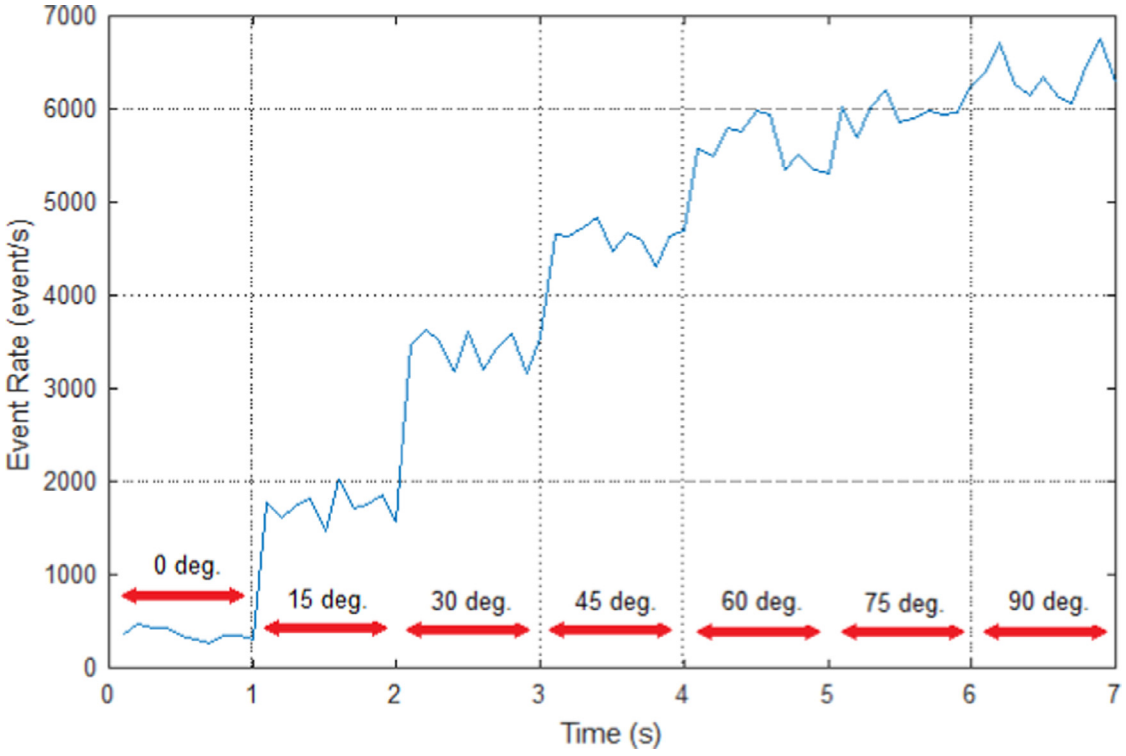

Fig. 6. ILD (encoded as an event-rate) measured over time. The sound source is placed at 0, 15, 30, 45, 60, 75 and 90° from the motionless head during a second for each angle.

Table 2
Data obtained for the first experiment: 1 kHz stimulus.

Target angle (degrees)	Mean angle reached (degrees)	Standard deviation (degrees)	Max. angle (degrees)	Min. angle (degrees)
0	1.97	0.61	2.81	0.70
15	14.16	2.60	16.87	6.33
30	30.67	0.56	31.64	29.53
45	42.83	0.60	44.30	42.19
60	56.25	0.65	58.36	55.55
75	75.84	0.50	76.64	75.23
90	93.21	1.06	94.92	91.41

on top of a platform driven by a DC motor with an encoder (Micromotor Ref. 2224R006SR plus gearhead Ref. 20/1112:1 and encoder Ref. IE2-512 from Faulhaber). The NAS is implemented using a Virtex5 FPGA (XC5VFX70T) and it uses up to 99% of the total slices available (11,141). The FPGA is in a Xilinx development board (ML507), which, among other components, includes the AC' 97 audio codec. The NAS output is connected to the processing system using the AER protocol. The processing system is also implemented using a Virtex5 FPGA (XC5VFX30T), which uses up to 4% of the total slices available on the device (5120). Using the Xilinx XPower tool, the estimated power consumption for the processing system is 28.63 mW and 29.7 mW for the NAS [15]. The test scenario is placed at the center of a classroom with a RT_{60} of 3.06 s. The specifications of the microphone are: transducer principle based on back electret condenser element, the frequency response range is between 20 and 16,000 Hz, the sensitive is -64 dB ± 3 dB and the impedance is 1000 Ohm.

Fig. 6 shows the IID measured at the output of the SH&F which models the LSO (Ears H&F shown in Fig. 5). The experiment consists of measuring the SH&F output rate when the sound source is placed at 7 azimuthal angles, [0–90] degrees within progressive stages of 15°. The robotic platform is not moving to obtain the IID. The sound played is a pure tone of 1 KHz and it lasted one second at each angle. Then, to obtain the event rate, the spike output was sampled at a frequency of 10 Hz (100 ms period). It can be seen how the event rate is higher as the angle increases. This rate is equivalent to the perceived IID when listening to a sound of 1 kHz [34].

The system was tested using three different pure tones: 1 kHz (channel 19 of the NAS), 2.5 kHz (channel 13 of the NAS) and 5 kHz (channel 7 of the NAS). For the three tests, the following identical sequence was followed: first, we placed the head in front of the speaker where the calibration phase takes place (4 s); then, we moved the speaker to each target angle (0° to 90° in steps of 15°) and recorded the angle reached by the robotic platform (10 s each target). The real angle reached by the head is measured by the magnetic encoder included within the motor. The encoder channels are sent to the FPGA where they are processed and, finally, the FPGA serially transmits to the PC whenever the motor turns 0.7° in any direction. This let us measure the real angle reached by the head. We then compared this measurement with the target angle (where the speaker was positioned). We repeated each test 10 times. The tests can be extended to a wider range without any recalibration.[1]

Fig. 7 shows the behavior of the system when the stimulus is a pure tone of 1 kHz and the measurements are taken every 15°. The

[1] Video with the full range operation: https://youtu.be/xUGVNpbodf8.

Table 3
Data obtained for the first experiment: 2.5 kHz stimulus.

Target angle (degrees)	Mean angle reached (degrees)	Standard deviation (degrees)	Max. angle (degrees)	Min. angle (degrees)
0	2.77	0.56	4.21	1.40
15	14.54	1.23	16.87	11.95
30	30.22	0.48	30.93	28.82
45	45.72	1.31	49.92	44.29
60	59.79	0.80	61.87	58.35
75	84.46	0.75	86.48	82.96
90	94.96	0.42	95.62	94.21

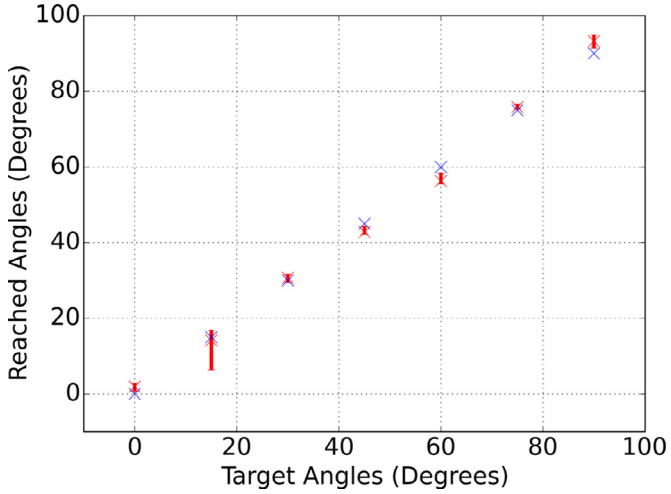

Fig. 7. The system is tested with a pure tone of 1 KHz as the stimulus. The target angles are 0, 15, 30, 45, 60, 75 and 90°, represented in the graph with a blue cross. In red, the mean, maximum and minimum angles reached by the platform. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

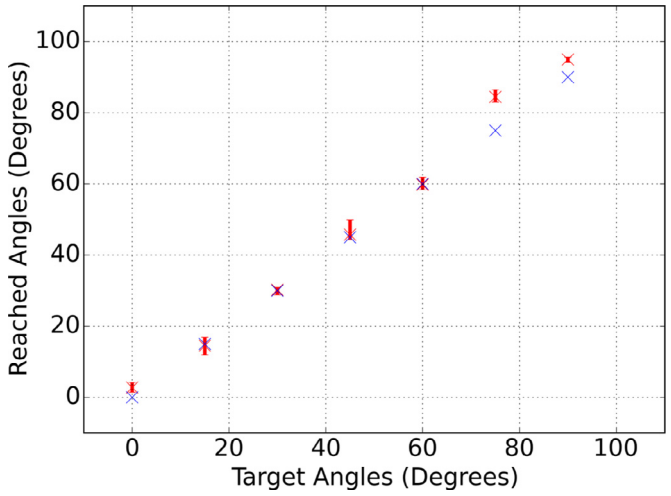

Fig. 8. The system is tested with a pure tone of 2.5 KHz as the stimulus. The target angles are 0, 15, 30, 45, 60, 75 and 90°, represented in the graph with a blue cross. In red, the mean, maximum and minimum angles reached by the platform. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

figure shows the mean angle reached, the maximum and minimum values reached (absolute error from the mean achieved) in red and, with a blue cross, the ideal behavior is represented. Table 2 shows the data for this first experiment.

Fig. 8 shows the behavior of the system when the stimulus is a pure tone of 2.5 kHz and the measurements are taken every 15°. The figure shows the mean angle reached, the maximum and minimum values reached (absolute error from the mean achieved) in red and, with a blue cross, the ideal behavior is represented. Table 3 shows the data for this second experiment.

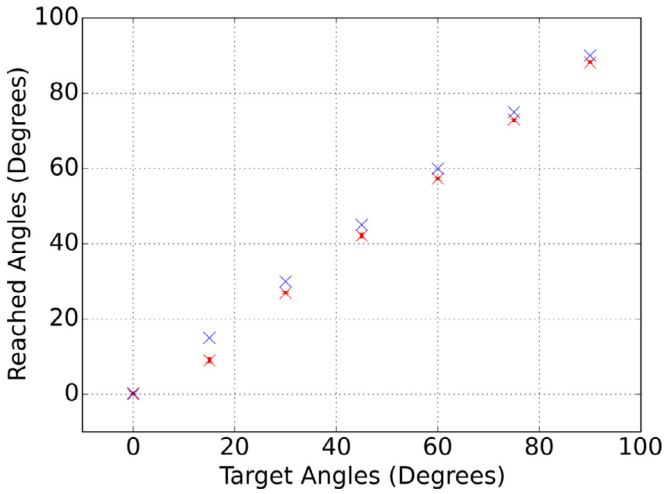

Fig. 9. The system is tested with a pure tone of 5 kHz as the stimulus. The target angles are 0, 15, 30, 45, 60, 75 and 90°, represented in the graph with a blue cross. In red, the mean, maximum and minimum angles reached by the platform. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 9 shows the behavior of the system when the stimulus is a pure tone of 5 kHz and the measurements are taken every 15°. The figure shows the mean angle reached, the maximum and minimum values reached (absolute error from the mean achieved) in red and, with a blue cross, the ideal behavior is represented. Table 4 shows the data for this third experiment.

The average error of the system is 1.92° for the 1 kHz stimulus test, 2.49° for the 2.5 kHz test and 2.57° for the 5 kHz test. Therefore, it results on an average error of 2.32°.

The average error is very low when the stimulus is a pure tone of 1 kHz: 1.92°. This error is higher when the frequency of the tone is increased. It goes up to 2.49° for the 2.5 kHz test and 2.57° for the 5 kHz test. This is most likely due to the fact that the NAS is designed in a way that it generates larger spike rates when the frequency of the stimulus targets low bands of the system, due to the interference of the filters. Therefore, higher rates arrive at the actuation layer. Then, since we are using PFM, there is a trade-off between the spike rate and the time length of a single spike. The higher spike rate produces a mismatch in that relation, creating a larger error.

Considering that the environmental SNR of these experiments is 73 dB, we checked the behavior of the system when higher level of noise is present. We have tested it when a set of SNR is applied: (−5 dB, 0 dB, 4 dB, 14 dB, 28 dB). The SNR is measured when the sound sourced is positioned at azimuthal angle of 0° and the white noise source is equidistantly located between two microphones and above the head. Figs. 10–12 shows the performance

Table 4
Data obtained for the first experiment: 5 kHz stimulus.

Target angle (degrees)	Mean angle reached (degrees)	Standard deviation (degrees)	Max. angle (degrees)	Min. angle (degrees)
0	0.32	0.36	0.70	0.00
15	9.00	0.38	9.84	8.44
30	26.97	0.34	27.42	26.72
45	42.16	0.47	42.89	41.48
60	57.33	0.36	57.66	56.95
75	73.07	0.19	73.12	72.42
90	88.13	0.34	88.59	87.89

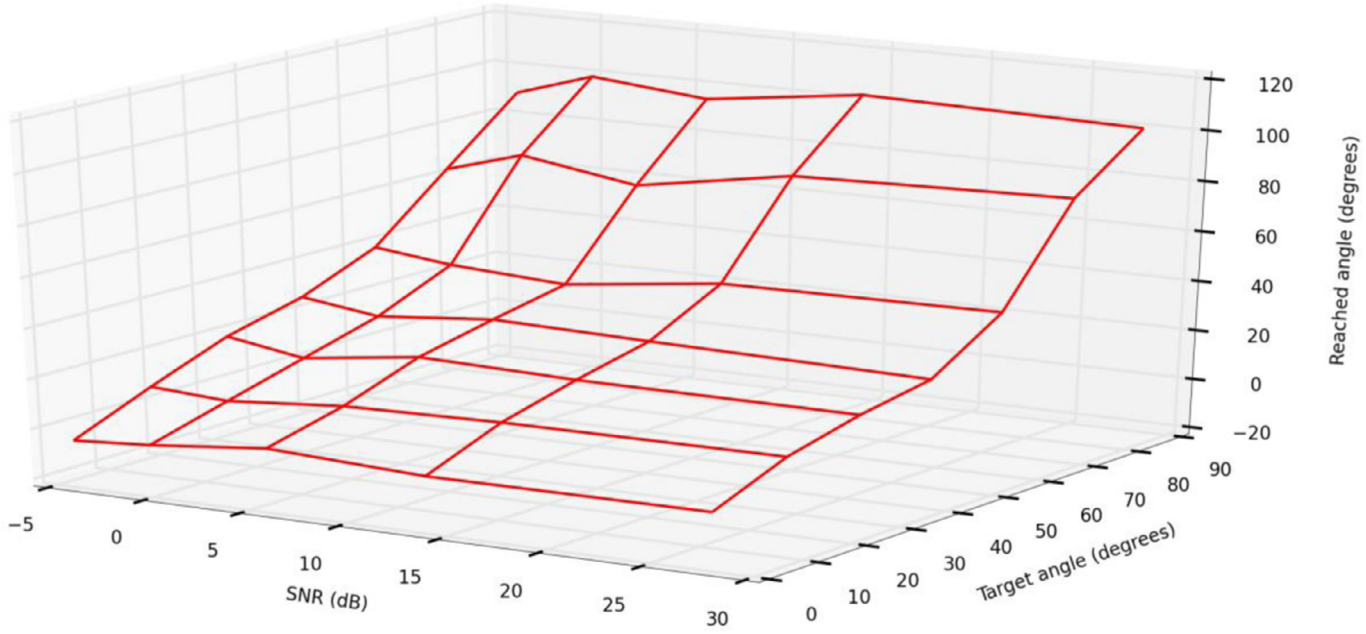

Fig. 10. Angle reached by the robotic platform when the stimulus is a pure tone of 1 KHz and a set of SNR was applied: (−5 dB, 0 dB, 4 dB, 14 dB and 28 dB).

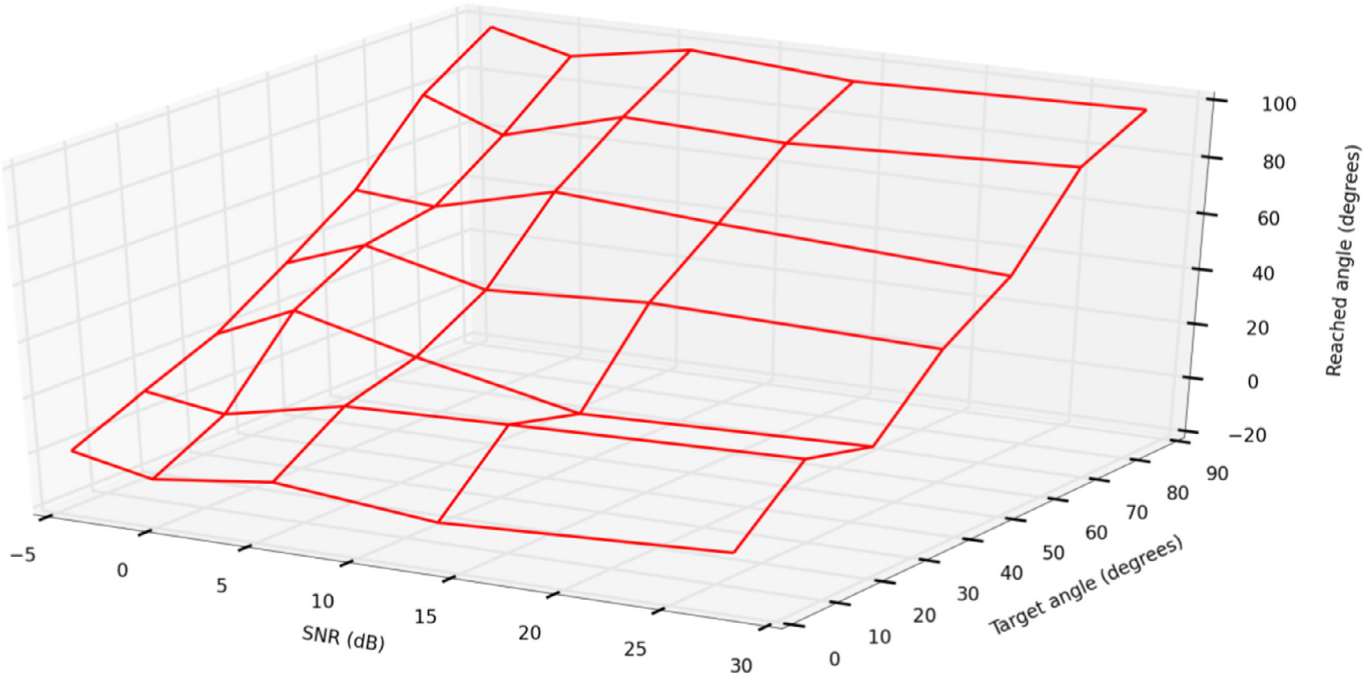

Fig. 11. Angle reached by the robotic platform when the stimulus is a pure tone of 2.5 KHz and a set of SNR was applied: (−5 dB, 0 dB, 4 dB, 14 dB and 28 dB).

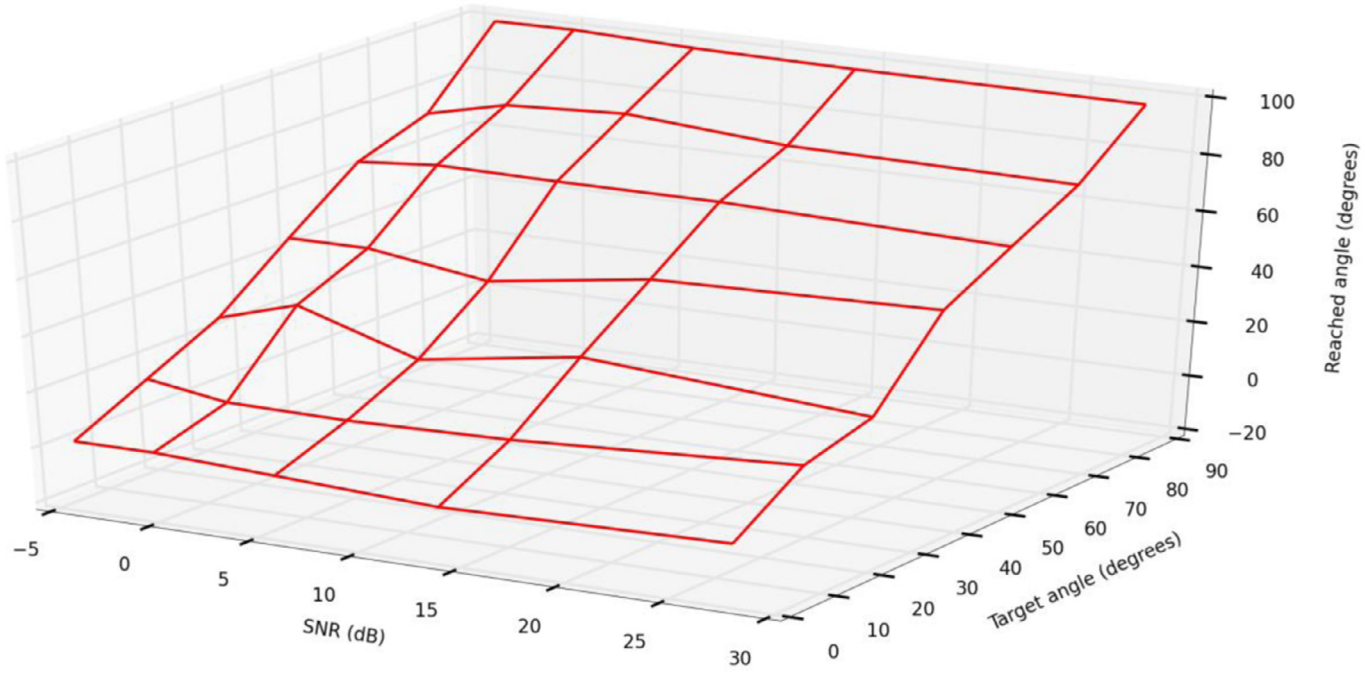

Fig. 12. Angle reached by the robotic platform when the stimulus is a pure tone of 5 KHz and a set of SNR was applied: (−5 dB, 0 dB, 4 dB, 14 dB and 28 dB).

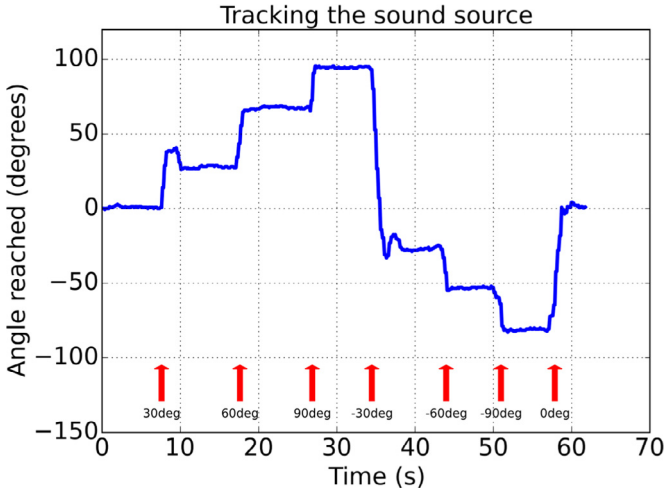

Fig. 13. Trajectory followed by the head when the sound source is moving. The tracking of the sound is accurate. A new target is supplied according to the red arrows on the plot. The real trajectory of the head is shown by the blue trace. The mean error is 4.37° and the time to reach a new supplied target is 125 ms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Mean error (in degrees) obtained for the experiments where the noise is present.

	1 KHz	2.5 KHz	5 KHz
−5 dB	9.61	2.86	4.42
0 dB	8.80	2.66	4.28
4 dB	8.22	2.68	3.09
14 dB	5.22	2.60	2.58
28 dB	5.20	2.63	2.39

of the system when three different pure tones are applied: 1 KHz, 2.5 KHz and 5 KHz as the stimulus.

Table 5 shows the mean error when the noise is present in the experiments. It can be seen how when the SNR is higher the error is lower. However, the mean error is very low (less than ten degrees in the worst case), therefore the system has a high noise tolerance.

Fig. 13 shows the tracking of the sound source by the head when a pure tone of 1 kHz is used. The experiment lasts 61 s and the target to track changes according to the red arrows shown on the figure. The set of targets supplied is (0, 30, 60, 90, −30, −60, −90, 0) and the set of positions reached by the head, on average, is (0, 28, 68, 94, −28, −55, −82, 0) resulting in an error of 3.625° on average. The targets are reached by the head within 125 ms after delivering a new one. From that time, we have to consider that the NAS takes an average of 20 μs to process the digital sound, to generate a spiking output activity and that the processing layer takes 50 ns to process each spike received.

4. Conclusion

This paper describes the design and hardware implementation of a sound localization and tracking system inspired by the mammalian auditory system. The NAS sensor [14] is used to produce a biological cochlea-like output. This output is the stimulus for the processing system where the LSO model is implemented. The architecture proposed for the LSO which performs the subtraction between two input spike rates produces the IID. The IID auditory cue is used as the input for the spike-based actuation stage that tracks the sound.

The model was tested using 1 kHz, 2.5 kHz and 5 kHz pure tones, obtaining a maximum error of less than five degrees, a lower error than those obtained in previous works that proposed a biologically inspired architecture of the LSO, [16] and [17], since the classification accuracy of the SNN architecture proposed by [17] is 52% for 5 kHz sounds, 83% for 15 kHz sounds and 40% for 25 kHz sounds in steps of 10°, and the classification accuracy obtained by the model proposed in [16], using both noise and speech in steps of 30°, is 80%. Furthermore, our system shows a high noise tolerance level when white noise is applied: in the worst condition, the average error is lower than ten degrees.

In comparison to the related work, this paper provides significant novelty and advances in this domain by using topologies that

are faithful to the architecture of the mammalian auditory pathways implemented in hardware to track sound in real time with very high accuracy. Furthermore, the architecture presented in this work can be implemented by using low-cost commercial hardware devices and have a low power consumption because the hardware implementation uses general purpose FPGA resources, such as counters, comparators, logic gates and low clock frequencies. Specifically, the power consumption goes up to 58.33 mW in operation (29.7 mW from the NAS and 28.63 mW from the processing layer).

For these reasons, the proposed system is suitable for the sound localization task in robotics. In addition, the work presented in this paper is a significant step forward in biologically inspired sound localization modeling, because it obtains the IID cue simulating a part of the mammalian auditory system, LSO.

Acknowledgements

This work was supported by the Spanish grant (with support from the European Regional Development Fund) COFNET (TEC2016-77785-P).

References

[1] B. Grothe, M. Pecka, D. McAlpine, Mechanisms of sound localization in mammals, Physiol. Rev. 90 (3) (2010) 983–1012.
[2] D.J. Tollin, The lateral superior olive: a functional role in sound source localization, Neuroscientist 9 (2) (2003) 127–143.
[3] L.A. Jeffress, A place theory of sound localization, J. Comp Physiol. Psychol. 41 (1) (1948) 35.
[4] D. McAlpine, B. Grothe, Sound localization and delay lines–do mammals fit the model? Trends Neurosci. 26 (7) (2003) 347–350.
[5] T.C. Yin, Neural mechanisms of encoding binaural localization cues in the auditory brainstem, Integrative Functions in the Mammalian Auditory Pathway, Springer, New York, 2002, pp. 99–159.
[6] W.C. Wu, C.H. Hsieh, H.C. Huang, O.C. Chen, Hearing aid system with 3D sound localization, in: Proceedings. of the IEEE International Conference on TENCON, 2007, pp. 1–4.
[7] H.J. Simon, Bilateral amplification and sound localization: then and now, J. Rehabil. Res. Dev. 42 (4) (2005) 117.
[8] P.K. Park, H. Ryu, J.H. Lee, C.W. Shin, K.B. Lee, J. Woo, J.S. Kim, B.C. Kang, S.C. Liu, T. Delbruck, Fast neuromorphic sound localization for binaural hearing aids, in: Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 5275–5278.
[9] H. Okuno, K. Nakadai, Real-time sound source localization and separation based on active audio-visual integration, Comput. Methods Neural Model. 2686 (2003) 118–125.
[10] K. Nakadai, D. Matsuura, H. Okuno, H. Kitano, Applying scattering theory to robot audition system: robust sound source localization and extraction, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003, pp. 1147–1152.
[11] A. van Schaik, V. Chan, C. Jin, Sound localisation with a silicon cochlea pair, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 2197–2200.
[12] Y. Tamai, S. Kagami, H. Mizoguchi, K. Sakaya, K. Nagashima, T. Takano, Circular microphone array for meeting system, in: Proceedings of the IEEE Sensors, 2003, pp. 1100–1105.
[13] S.S. Yeom, J.S. Choi, Y.S. Lim, M. Park, M. Kim, DSP implementation of sound source localization with gain control, in: Proceedings of the International Conference on Control, Automation and Systems, 2007, pp. 224–229.
[14] A. Jimenez-Fernandez, A. Linares-Barranco, R. Paz-Vicente, G. Jiménez, A. Civit, Building blocks for spikes signals processing, in: Proceedings of the International Joint Conference Neural Networks, 2010, pp. 1–8.
[15] A. Jiménez-Fernández, E. Cerezuela-Escudero, L. Miró-Amarante, M.J. Domínguez-Morales, F.d.A. Gómez-Rodríguez, A. Linares-Barranco, G. Jiménez-Moreno, in: A binaural neuromorphic auditory sensor for FPGA: a spike signal processing approach, 28, 2017, pp. 804–818.
[16] J. Liu, D. Perez-Gonzalez, A. Rees, H. Erwin, S. Wermter, A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing 74 (1) (2010) 129–139.
[17] J.A. Wall, L.J. McDaid, L.P. Maguire, T.M. McGinnity, Spiking neural network model of sound localization using the interaural intensity difference, IEEE Trans. Neural Netw. Learn. Syst. 23 (4) (2012) 574–586.
[18] Iwasa, K., Kugler, M., Kuroyanagi, S., & Iwata, A.. A sound localization and recognition system using pulsed neural networks on FPGA. in: Proceedings of the International Joint Conference on Neural Networks (pp. 902–907). IEEE.
[19] Birchfield, S.T., & Gangishetty, R.. Acoustic localization by interaural level difference. in: Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05 (Vol. 4, pp. iv–1109). IEEE.
[20] F. Gomez-Rodriguez, R. Paz, L. Miro, A. Linares-Barranco, G. Jiménez, A. Civit, Two hardware implementations of the exhaustive synthetic AER generation method, Lecture Notes in Computer Science, Computational Intelligence and Bioinspired Systems, 3512, Springer, Berlin Heidelberg, 2005, pp. 534–540.
[21] E. Cerezuela-Escudero, M.J. Dominguez-Morales, A. Jimenez-Fernandez, R. Paz-Vicente, A. Linares-Barranco, G. Jimenez-Moreno, Spikes monitors for FPGAs, an experimental comparative study, in: Proceedings of the 12th International Work-Conference Artificial Neural Networks, 2013, pp. 179–188.
[22] A. Rios-Navarro, E. Cerezuela-Escudero, M. Dominguez-Morales, A. Jimenez-Fernandez, G. Jimenez-Moreno, A. Linares-Barranco, Live demonstration: Real-time motor rotation frequency detection by spike-based visual and auditory AER sensory integration for FPGA, in: Proceedings of the IEEE International Symposium on Circuits and Systems, 2015 1907-1907.
[23] E. Cerezuela-Escudero, A. Jimenez-Fernandez, R. Paz-Vicente, M. Dominguez–Morales, A. Linares-Barranco, G. Jimenez-Moreno, Musical notes classification with neuromorphic auditory system using FPGA and a convolutional spiking network, in: Proceedings of the International Joint Conference on Neural Networks, 2015, pp. 1–7.
[24] E. Cerezuela-Escudero, A. Jimenez-Fernandez, R. Paz-Vicente, J.P. Dominguez–Morales, M.J. Dominguez-Morales, A. Linares-Barranco, Sound recognition system using spiking and MLP neural networks, in: Proceedings of the International Conference on Artificial Neural Networks, Springer International Publishing, 2016, pp. 363–371.
[25] J.P. Dominguez-Morales, A. Jimenez-Fernandez, A. Rios-Navarro, E. Cerezuela-Escudero, D. Gutierrez-Galan, M.J. Dominguez-Morales, G. Jimenez-Moreno, Multilayer, spiking neural network for audio samples classification using SpiNNaker, in: Proceedings of the International Conference on Artificial Neural Networks, Springer International Publishing, 2016, pp. 45–53.
[26] J.P. Dominguez-Morales, A. Jimenez-Fernandez, M. Dominguez-Morales, G. Jimenez-Moreno, NAVIS: neuromorphic auditory VISualizer tool, Neurocomputing 237 (2017) 418–422.
[27] Á.F. Jiménez Fernández, Diseño y Evaluación de Sistemas de Control y Procesamiento de Señales Basados en Modelos Neuronales Pulsantes Ph.D.Thesis, University of Seville, Spain, 2010.
[28] A. Jimenez-Fernandez, G. Jimenez-Moreno, A. Linares-Barranco, M.J. Dominguez-Morales, R. Paz-Vicente, A. Civit-Balcells, A neuro-inspired spike-based PID motor controller for multi-motor robots with low cost FPGAs, Sensors 12 (4) (2012) 3831–3856.
[29] M. Domínguez-Morales, A. Jimenez-Fernandez, E. Cerezuela-Escudero, R. Paz-Vicente, A. Linares-Barranco, G. Jimenez, On the designing of spikes band-pass filters for FPGA, in: Proceedings of the International Conference on Artificial Neural Networks and Machine Learning, ICANN, 2011, 2011, pp. 389–396.
[30] F. Pérez-Peña, A. Morgado-Estévez, A. Linares-Barranco, Inter-spikes-intervals exponential and gamma distributions study of neuron firing rate for SVITE motor control model on FPGA, Neurocomputing 149 (2015) 496–504.
[31] J. Mackenzie, J. Huopaniemi, V. Valimaki, I. Kale, Low-order modeling of head-related transfer functions using balanced model truncation, IEEE Signal Process. Lett. 4 (2) (1997) 39–41.
[32] M. Otani, S. Ise, Fast calculation system specialized for head-related transfer function based on boundary element method, J. Acoust. Soc. Am. 119 (5) (2006) 2589–2598.
[33] M. Slaney, An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank, An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank, 35, Apple Computer Perception Group, 1993 Technical Report.
[34] W.E. Feddersen, T.T. Sandel, D.C. Teas, L.A. Jeffress, Localization of high-frequency tones, J. Acoust. Soc. Am. 29 (9) (1957) 988–991.