In: Proceedings of the 22nd International Conference on Artificial Neural Networks (ICANN 2012), Part I, LNCS 7552, pp. 239-246, Springer Heidelberg, 2012.

Biomimetic Binaural Sound Source Localisation with Ego-Noise Cancellation

Jorge Dávila-Chacón[1], Stefan Heinrich[1], Jindong Liu[2], and Stefan Wermter[1]

[1] University of Hamburg, Department of Informatics, Knowledge Technology Group
Vogt-Kölln-Straße 30, D-22527 Hamburg, Germany
[2] Imperial College London, Department of Computing
Huxley Building, South Kensington Campus, London, SW7 2AZ, UK
http://www.informatik.uni-hamburg.de/WTM/

Abstract. This paper presents a spiking neural network (SNN) for binaural sound source localisation (SSL). The cues used for SSL were the interaural time (ITD) and level (ILD) differences. ITDs and ILDs were extracted with models of the medial superior olive (MSO) and the lateral superior olive (LSO). The MSO and LSO outputs were integrated in a model of the inferior colliculus (IC). The connection weights between the MSO and LSO neurons to the IC neurons were estimated using Bayesian inference. This inference process allowed the algorithm to perform robustly on a robot with ∼40 dB of ego-noise. The results showed that the algorithm is capable of differentiating sounds with an accuracy of 15°.

Keywords: Binaural sound source localisation, Spiking neural networks, Bayesian inference, Inferior colliculus.

1 Introduction

Audition can inform us about the spatial location of distant events. Sounds can provide information comparable to visual stimuli in scenarios where vision is impeded. SSL can help robots to cope with environment hazards and to communicate [6]. A meta-objective of artificial SSL systems is their portability to different robots. This paper describes the design and implementation of Liu et al. [2] algorithm for sound localisation on a Nao robotic platform with ∼40 dB of ego-noise.

Voutsas and Adamy [9] created a multiple delay-lines model using SNNs. Their model has 30° resolution, and uses ITDs with good results only for sounds with low fundamental frequencies. However, integration across-frequencies kept localisation accuracy high for broadband stimuli.

Rodemann et al. [5] developed a model with 10° resolution based on ITDs, ILDs and the interaural envelop difference (IED). Different localisation cues are computed in parallel and a weak winner-takes-all strategy integrates different cues. In all testing conditions the higher frequencies had greater localisation errors, and merging the cues in an IC model remained an open improvement.

Willert et al. [10] and Nix and Hohmann [3] presented probabilistic models (both with 15° resolution) of the MSO, the LSO and the IC. They used Bayesian inference to estimate the connections between modules. In both cases, a more realistic neural processing could be obtained through the implementation of SNNs. The results from these studies have provided valuable insights on the design requirements of a biomimetic SSL system.

Finally, Liu et al. [2] implemented models of the MSO, LSO and IC using SNNs. Connections from the MSO and LSO to the IC were estimated using Bayesian inference. The algorithm had good performance at 30° resolution, using a human-shaped foam head, and under low levels of noise. Whether the system would perform quite as well on a non-humanoid robot head with strong ego-noise was open research.

For biological plausibility our model was implemented with artificial SNNs. The algorithm simulates the functioning of the human cochlea, the cochlear nucleus (CN), the medial nucleus of the trapezoid body (MNTB), the superior olivary complex (SOC) and the inferior colliculus (IC).

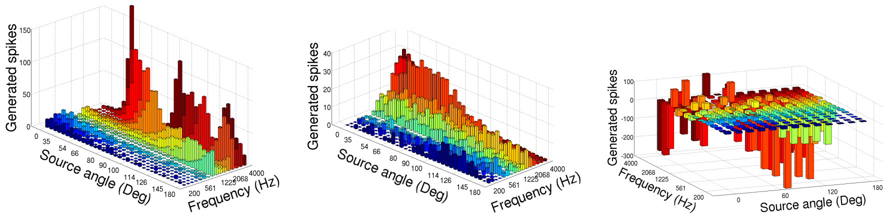
First the cochlea is emulated with the Patterson-Holdsworth filter bank [8] separating the sound wave in several centre frequency channels. Hair-cells are simulated by phase-locking the sound waves. Interaction between the CN and the MNTB is represented in the topology of connections arriving at the medial superior olive (MSO) and the lateral superior olive (LSO) [7]. Interaural time differences (ITD) were extracted in the MSO, and interaural level differences (ILD) in the LSO. Finally, MSO and LSO outputs were merged in the IC. The topology of connections was based on anatomical findings in mammals [4].

2 Computational Model

First sound is decomposed with the Patterson-Holdsworth filter bank [8] in $f \in \{1 \dots n_f\}$ frequency components equally separated on a logarithmic scale. Each of the n_f frequency components is analysed separately. The maximum amplitudes of the sound waves are used to generate spikes directed to the MSO and LSO. See Fig. 1 for MSO, LSO and IC activation examples.

Figure 3 shows the MSO connectivity scheme. In the model the MSO has $i \in \{1 \dots n_{MSO}\}$ neurons for each of the n_f frequency components. $S_{i,f}^{MSO}$ is the number of spikes produced by neuron $MSO_{i,f}$ for a given sound. The value of n_{MSO} depends on the smallest ITD the system is able to detect. Such sensitivity is limited by the distance between the robot microphones, and by the sample rate of the sound card used for recording the sounds. Each neuron $MSO_{i,f}$ is maximally sensitive to sounds produced at angle α_i. The MSO output angle $O^{MSO} = \alpha_i$ for the i that maximizes $\sum_i S_{i,f}^{MSO}$, following the *winner-takes-all* rule.

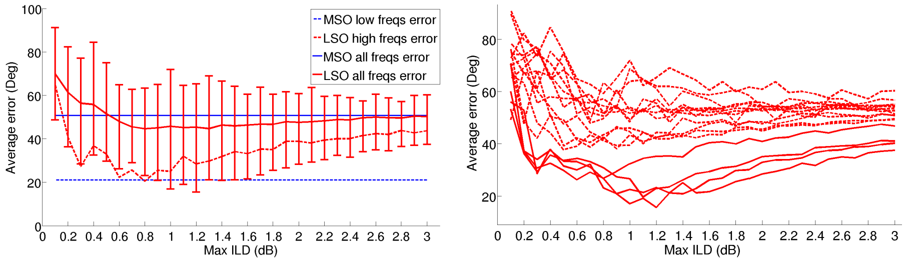
The maximum ILD resolution $\max_{ILD}$ that can be achieved depends on the geometry of the robot's head, and it is essential to know its value to perform efficient SSL based on ILDs. For SSL with humanoid dummy heads [2] it is possible to obtain the value of $\max_{ILD}$ from literature. However, for Nao's head it



(a) MSO (b) LSO (c) IC

Fig. 1. Activation for a sound produced at 15°. Notice that lower and higher frequencies are more informative in the MSO and LSO respectively and that the IC has a more coherent spatial representation across frequencies.

was necessary to estimate the $\max_{ILD}$ from the statistics of the LSO activation. We estimated the best LSO performance for all and each of the n_f frequency components from a range of $\max_{ILD}$ values. Figure 2 shows the MSO and LSO output errors plotted against $\max_{ILD}$ values between 0.1 and 3 dB.



(a) (b)

Fig. 2. (a) MSO and LSO output errors estimated from their best frequency components (dotted lines), and average error from all frequencies (solid lines). In both cases the MSO error (blue lines) is constant for all $\max_{ILD}$ values and the best performance was reached at ∼0.8 dB. (b) LSO errors for each frequency component. Higher frequencies (solid lines) have better performance than lower ones (dotted lines).

Figure 3 shows the LSO connectivity scheme. In the model the LSO has $i \in \{1 \dots n_{LSO}\}$ neurons for each of the n_f frequency components. The maximum value of n_{LSO} is limited by the bits of the sound data. Therefore, it is possible to have many more neurons in the LSO than in the MSO. For the sake of simplicity, n_{LSO} was chosen to be the same as n_{MSO}.

The computation of the connection weights from the MSO and LSO to the IC was inspired by the work of Willert et al. [10] using Bayesian inference. No connections were generated between neurons sensitive to different frequencies. Figure 3 shows the IC connectivity scheme.

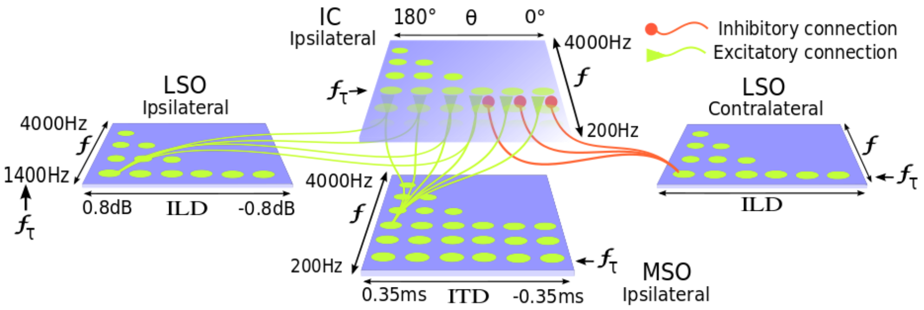


Fig. 3. Multiple delay lines deliver spike-trains to MSO cells according to Jeffress model [7]. MSO neurons respond to frequencies between 200-4000 Hz. The difference of the wave amplitudes that produced a spike in the MSO is used to generate a spike in the LSO. LSO neurons respond to frequencies between ∼1-4 kHz. The MSO has excitatory connections to the IC in all frequencies. The LSO has excitatory and inhibitory connections to the IC in frequencies >1 kHz.

The benefit from this integration is related to the overlap at the higher frequencies of MSO excitatory connections and LSO inhibitory connections. The LSO provides misleading information for the lower frequencies and useful information for the higher frequencies. The MSO provides useful information in all the frequencies, but also potentially misleading information in the higher frequencies. Therefore, the LSO inhibitory connections can help to keep only useful information given by the MSO at higher frequencies. Thus, the system makes more efficient use of auditory cues along the audible frequency range.

In the model the IC has $j \in \{1 \ldots n_{IC}\}$ neurons for each of the n_f frequency components. $S^{IC}_{j,f}$ is the number of spikes produced by neuron $IC_{j,f}$ for a given sound. The value of n_{IC} equals the total number of azimuth angles θ in half circle in front of the robot where a sound is produced. $E^{MSO}_{i,j,f}$ and $E^{LSO}_{i,j,f}$ are the MSO and LSO excitatory connection weights from the ipsilateral neurons $MSO_{i,f}$ and $LSO_{i,f}$ to neuron $IC_{j,f}$. $I^{LSO}_{i,j,f}$ is the LSO inhibitory connection weight from the contralateral neuron $LSO_{i,f}$ to neuron $IC_{j,f}$. The IC uses the weighted input from the MSO and LSO to compute its output angle $O^{IC} = \theta_j$ for the j that maximizes $S^{IC}_{j,f}$, where

$$S^{IC}_{j,f} = \sum_i \left(S^{MSO}_{i,f} \cdot E^{MSO}_{i,j,f} + S^{LSO}_{i,f} \cdot E^{LSO}_{i,j,f} - S^{LSO}_{i,f} \cdot I^{LSO}_{i,j,f} \right) . \tag{1}$$

2.1 Bayesian Framework

Let $p(S^{MSO}_{i,f}|\theta_j)$ be the likelihood that neuron $MSO_{i,f}$ produces a spike in the time frame $\varDelta t$ for a sound produced at angle θ_j:

$$p\left(S^{MSO}_{i,f}|\theta_j\right) = \frac{S^{MSO}_{i,f}}{\sum_{i'} S_{i',f}} . \tag{2}$$

Spikes are summed over all MSO neurons sensitive to frequency component f. The prior $p\left(\theta_j\right) = 1 \ / \ n_{IC}$ is the probability of a sound to be produced at angle θ_j and it is the same for all angles. Let $p(S_{i,f}^{MSO})$ be the evidence that neuron $MSO_{i,f}$ produces a spike in the time frame $\varDelta t$:

$$p\left(S_{i,f}^{MSO}\right) = \sum_j p\left(S_{i,f}^{MSO}|\theta_j\right) p\left(\theta_j\right). \tag{3}$$

Once the likelihood, prior and evidence are calculated, the posterior $p\left(\theta_j|S_{i,f}^{MSO}\right)$ for the same time frame $\varDelta t$ can be computed from Bayes rule:

$$p\left(\theta_j|S_{i,f}^{MSO}\right) = \frac{p\left(S_{i,f}^{MSO}|\theta_j\right) p\left(\theta_j\right)}{p\left(S_{i,f}^{MSO}\right)} = P^{MSO}. \tag{4}$$

The Bayesian inference described so far is also used for computing the LSO inhibitory and excitatory connections. Finally, the connection weights are set according to the following functions:

$$E_{i,j,f}^{MSO} = \begin{cases} P^{MSO}, & \text{if } \ P^{MSO} > \left(\omega_{E,f}^{MSO}.\arg\max_{\theta_j}\left(P^{MSO}\right)\right) \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

$$E_{i,j,f}^{LSO} = \begin{cases} P^{LSO}, & \text{if } \ P^{LSO} > \left(\omega_{E,f}^{LSO}.\arg\max_{\theta_j}\left(P^{LSO}\right)\right) \ \wedge \ f \geqq f_\tau \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

$$I_{i,j,f}^{LSO} = \begin{cases} 1 - P^{LSO}, & \text{if } \ P^{LSO} < \left(\omega_{I,f}^{LSO}.\arg\max_{\theta_j}\left(P^{LSO}\right)\right) \ \wedge \ f \geqq f_\tau \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

The maximum posterior probability for every frequency f is thresholded by $\omega_{E,f}^{MSO}$, $\omega_{E,f}^{LSO}$ and $\omega_{I,f}^{LSO}$. These weights are real-valued numbers from the closed interval [0 1] and determine which connections will be pruned. The value of f_τ marks the transition between the lower and higher frequency spectrum.

3 Experimental Results

The experimental setup is displayed in Fig. 4. According to specifications, Nao's distance between left and right microphones is ∼0.12 m. Therefore, the highest frequency that does not generate ITD ambiguities is $f_\tau \approx 1400\,\text{Hz}$. Background noise was 44.6 dBA at the right microphone and 41.6 dBA at the left microphone.

3.1 Preliminary Study

In a preliminary study, we tested Nao with a SSL system based on the cross correlation between two microphone pairs (left-front and right-front). A feedforward network was trained with 4 speech recordings from 24 directions equally

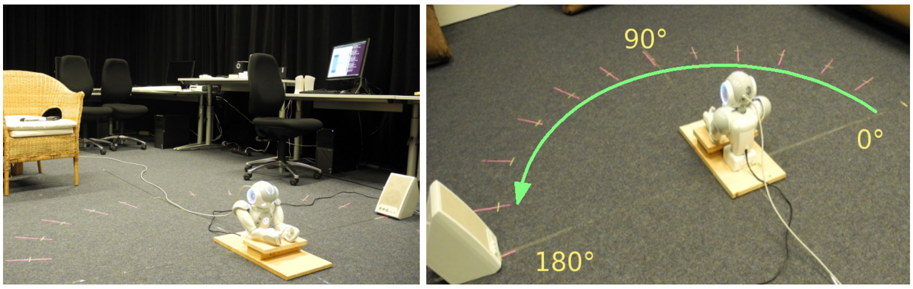
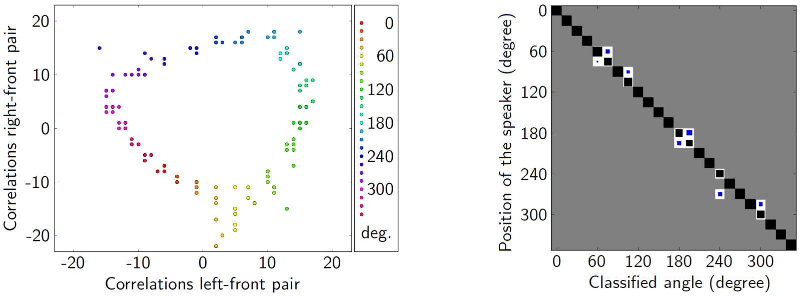


Fig. 4. Sounds were played around the Nao in half circle ∅2 m, from 0° to 180° in 15° steps. 13 recordings were made in a room with reverberation damped by curtains.



(a)

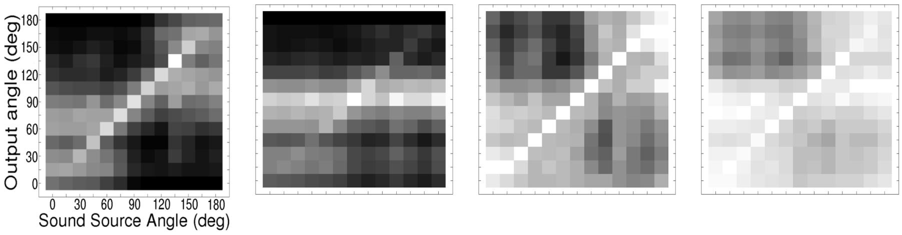
(b)

Fig. 5. Results of the preliminary study on SSL with Nao robot. (a) Cross correlation of ITD pairs. (b) Confusion matrix of the network output.

spaced around the robot. The differences in the correlation (see Fig. 5) for every ITD pair were fed to a network with $|I_I|= 2$ input, $|I_H|= 6\ldots72$ hidden, and $|I_O|= 24$ output neurons. The network test showed very good performance with 91% accuracy rate. The study showed that SSL can achieve good rates of accuracy with Nao's basic microphones. However, using more than two microphones was avoided in the following experiments for the sake of biological plausibility.

3.2 Biomimetic Computation Results

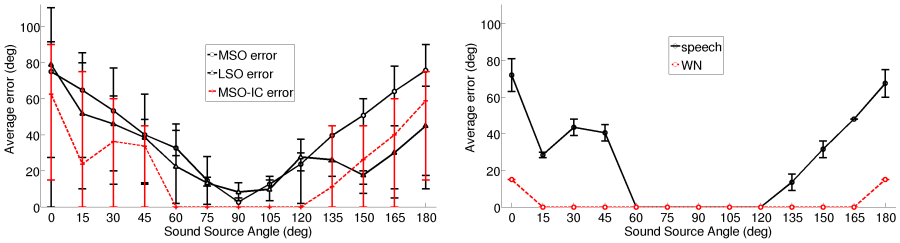
In the first of two experiments with the SNN model, the robot was trained with 1 s of uniform white noise (WN) and in the second one with speech. The recordings were split in 16 frequency components between 200-4000 Hz. Speech consisted of 4 instances of the words *hello, look, fish, coffee* and *tea*. In both experiments the testing sounds were different instances of the same words and 0.25 s samples of WN. The system had good performance when trained with speech and tested with WN even if the system was not able to generalise from WN to speech. Interestingly, the best performance for both experiments was when testing with WN (see Fig. 6).



(a) WN / WN.
(b) WN / S.
(c) S / WN.
(d) S / S.

Fig. 6. IC output confusion matrices when the system was trained / tested with white noise (WN) and speech (S). The speech output is for the word *fish*. Lighter areas indicate higher values.

Figure 7 shows the results of the second experiment. The IC lower boundaries in Fig. 7(a) indicate perfect localisation performance for WN except for 0° and 180°. Figure 7(b) details further the IC output. The IC performance highly improved for all angles and all sound classes with respect to the first experiment, even though the MSO and LSO outputs did not change substantially. The error dropped to zero between 60° and 120° for all sounds.



(a) MSO, LSO and IC average errors
(b) IC error for WN and speech

Fig. 7. (a) Average errors for all testing sounds when training with speech. Notice that the IC has higher accuracy than the MSO and LSO. (b) IC error for each testing sound. It can be seen that WN localisation is always zero for most angles.

4 Conclusion and Future Work

In this paper we confirm the robustness of a biomimetic approach to SSL. Integration of auditory cues in the IC showed higher performance than the MSO and LSO, with no error in the 60° in front of the robot and near perfect localisation accuracy for WN. The optimised algorithm proved capable of segregating sound sources with similar precision to state-of-the-art algorithms [3,10,9].

Estimating the optimal $\max_{ILD}$ value for Nao's head allowed to double the resolution for localisation in comparison to Liu et al. [2]. The $\max_{ILD}$ was found through a statistical analysis of the LSO activation across its frequency components. Frequency decomposition opens the possibility of localising concurrent and dynamic sound sources [2]. Such advantage lacks in networks learning ITD pairs directly extracted from the sound wave cross correlation.

The Bayesian inference process allowed the system to perform correctly under high levels of ego-noise. When the MSO and LSO are presented only with the robot's ego-noise, their output is a fixed direction. However, ego-noise activation is evenly distributed among IC neurons and their output equals the front angle.

The processes underlying spatial hearing can be used for the segregation of speech by increasing its SNR [6]. Part of our future work will be directed towards the enhancement of speech recognition systems with the aid of SSL. Ultimately we pursue a multimodal approach to the long standing *Cocktail Party Problem*, and SSL is an essential ingredient in such enterprise [1].

Acknowledgements. The authors would like to thank Cornelius Weber for valuable manuscript reviews and Robert Keßler for the initial study with the Nao robot. This work was supported by the DFG German Research Foundation (grant #1247) - International Research Training Group CINACS (Cross-modal Interaction in Natural and Artificial Cognitive Systems).

References

1. Even, J., Heracleous, P., Ishi, C., Hagita, N.: Multi-modal front-end for speaker activity detection in small meetings. In: International Conference on Intelligent Robots and Systems. pp. 536–541. IEEE (2011)
2. Liu, J., Perez-Gonzalez, D., Rees, A., Erwin, H., Wermter, S.: A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation. Neurocomputing 74(1-3), 129–139 (2010)
3. Nix, J., Hohmann, V.: Sound source localization in real sound fields based on empirical statistics of interaural parameters. The Journal of the Acoustical Society of America 119, 463 (2006)
4. Recanzone, G., Sutter, M.: The biological basis of audition. Annual Review of Psychology 59, 119–142 (2008)
5. Rodemann, T., Heckmann, M., Joublin, F., Goerick, C., Scholling, B.: Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. In: International Conference on Intelligent Robots and Systems. pp. 860–865. IEEE (2006)
6. Roman, N., Wang, D., Brown, G.: Speech segregation based on sound localization. The Journal of the Acoustical Society of America 114, 2236–2252 (2003)
7. Schnupp, J., Nelken, I., King, A.: Auditory neuroscience: Making sense of sound. The MIT Press (2011)
8. Slaney, M.: An efficient implementation of the patterson-holdsworth auditory filter bank. Tech. rep., Apple Computer, Perception Group (1993)
9. Voutsas, K., Adamy, J.: A biologically inspired spiking neural network for sound source lateralization. Transactions on Neural Networks 18(6), 1785–1799 (2007)
10. Willert, V., Eggert, J., Adamy, J., Stahl, R., Körner, E.: A probabilistic model for binaural sound localization. Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 36(5), 982–994 (2006)