Neural and Statistical Processing of Spatial Cues for Sound Source Localisation

Jorge Davila-Chacon[1], Sven Magg[1], Jindong Liu[2] and Stefan Wermter[1]

***Abstract*—When confronting binaural sound source localisation (SSL) algorithms with different environments and robotic platforms, there is an increasing need for non-linear integration methods of spatial cues. Based on interaural time and level differences, we compare the performance of several SSL systems. The architecture has three degrees of freedom, i.e. each tested architecture employs a different combination of *representation* of binaural cues, *clustering* and *classification* algorithms. The heuristic for the selection of methods is the same at each degree of freedom: to compare the impact of traditional statistical techniques versus machine learning algorithms with different degrees of biological inspiration. The overall performance is evaluated in the analysis of each system, including the accuracy of its output, training time and adequateness for life-long learning. The results support the use of hybrid systems, consisting different kinds of artificial neural networks, as they present an effective compromise between the characteristics evaluated.**

I. Introduction

AUDITION is a complex perceptual capability of most vertebrates that remains unparalleled by machines. More specifically, one important component of audition is the localisation of sound sources in space. This ability to pinpoint sound sources around us is crucial for a safe interaction with the environment and for improving communication with other individuals [1]. For both reasons, SSL is a desired capability for robotic systems.

The location of sound sources can be specified with its components in the azimuth, elevation and depth planes. In this paper we will only address the problem of SSL along the azimuth plane on the frontal 180°. A common approach for SSL systems are microphone arrays of different size and geometry. The system of Valin et al. [2] achieves an angular resolution of 3° with an array of 8 microphones. Similarly, the array of 32 microphones of Tamai et al. [3] reaches 5° accuracy on the azimuth and elevation. The drawback of these kind of approaches is that they only use the time-difference-of-arrival (TDOA) between microphones. Therefore, it remains an open issue to take advantage of level differences between sensors as a cue for SSL.

An alternative paradigm to multiple microphone arrays is binaural SSL. Humans are a clear example that it is possible to achieve accurate SSL using only two sound sensors or ears. Additionally, binaural SSL also relies on the effect of our pinnae, head and torso on the sound frequency components (FC), and on the capacity to move our head [4].

With only one pair of microphones separated by a head-like structure, an SSL system can use different binaural cues to locate sound sources in space. This configuration allows the estimation of interaural time (ITD) and level (ILD) differences. Both spatial cues are important, as ITDs convey more accurate information in low FCs and ILDs in high FCs. This effect is known as the *duplex* theory of SSL, and it places the boundary between low and high frequencies around 1500-3000 Hz [4]

Voutsas and Adamy [5] used spiking neural networks (SNN) and a multiple-delays model to estimate ITDs. Their system can localise low frequency sounds with 30° accuracy and its performance decreases for sounds with high FCs. Nevertheless, across-frequency integration keeps the system accuracy high for broadband stimuli. High frequencies also contain useful spatial information that could improve SSL systems. Making use of ITDs, ILDs and interaural envelop differences (IED), Rodemann et al. [6] developed a 10° resolution system. However, the model is sensitive to noise and reverberation, and less accurate for high FCs.

It has been shown that bio-inspired methods can achieve near-optimal integration of multimodal information [7]. In the case of SSL, the modalities we are interested in are time and level differences. In the inferior colliculus (IC) mammals integrate ITDs and ILDs, encoded in the medial (MSO) and lateral (LSO) superior olive respectively [8].

Making use of such neurophysiological principles, Willert et al. [9] and Nix and Hohmann [10] integrated probabilistic models of the MSO, LSO and IC. Using probabilistic models, both systems can reach a resolution of 15°. A possible extension of this work is its implementation with SNNs in order to exploit the dynamics of neural populations providing robustness to noise. Liu et al. [11] proposed a model of the MSO, LSO and IC using SNNs and Bayesian inference. The system achieves 30° resolution under reverberation and low noise conditions. Afterwards, Davila-Chacon et al. [12] adapted this approach [11] to a robotic platform subject to high levels of ego-noise and were able to increase the resolution to 15°.

The objectives of the current study are to extend our previous work in [11] and [12]. The first novel contribution is to provide the system with alternative representations of spatial cues. For this purpose we explore the system's performance with different spatial cues being represented with either a sta-

[1] Knowledge Technology Group, Department of Informatics, University of Hamburg. Vogt-Kölln-Straße 30, D-22527, Hamburg, Germany (email: {davila, magg, wermter}@informatik.uni-hamburg.de)

[2] Department of Computing, Imperial College London. Huxley Building, South Kensington Campus, SW7 2AZ, London, UK (email: j.liu@imperial.ac.uk)

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) and project RobotDoc under 235065 ROBOT-DOC from FP7, Marie Curie Action ITN.

tistical technique or with bio-inspired methods. The second contribution is to improve the system's robustness to ego-noise with a non-linear classification layer. Therefore, each representation is used as input to either statistical or neural classifiers in order to compare their accuracy, robustness to ego-noise and computational cost.

II. Basis Methodologies

Fig. 1 depicts our experimental setup. It consists of a humanoid robotic head immersed in a virtual reality (VR) setup designed for audio-visual integration [13]. The iCub is a platform designed for studies in embodied cognition and cognitive developmental robotics [14]. The iCub head has a geometry similar to an average 3.5-year-old child and is equipped with a pair of microphones surrounded by pinnae.

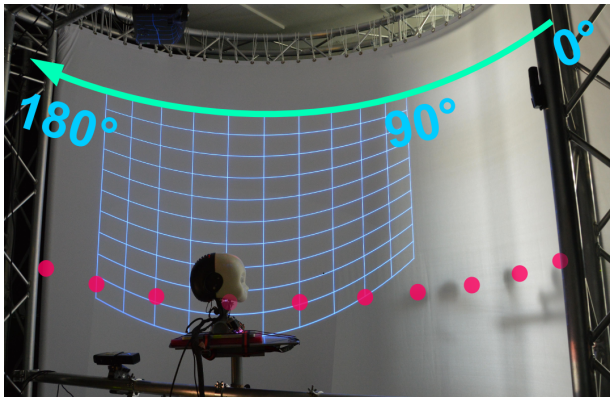


Figure 1: Audio-visual VR experimental setup. The grid shows the curvature of the projection screen surrounding the iCub humanoid robot head and the dots represent the location of the sound sources behind the screen.

During the experiments the position of the head is fixed. The iCub head produces ∼60 Hz of ego-noise and reverberation is reduced by damping curtains. The stimuli consist of $0.25ms$ segments of white noise (WN) and the words *hello, look, fish, coffee* and *tea* recorded from male and female subjects. The WN class consisted 12 instances and the speech class consisted of 40 instances of each word. Each instance of both classes is presented once to the iCub between 0° and 180° at 15° steps along the azimuth plane, at the same elevation angle and at $\sim 1.3m$ distance.

The system is tested with four *training / testing* configurations: *Speech / Speech*, *WN / WN*, *WN / Speech* and *Speech / WN*. As can be expected, the highest performance was obtained when training and testing with different instances of the same class of sounds, i.e. with the *Speech / Speech*, *WN / WN* configurations. The lowest performance came from the *WN / Speech* configuration. However, some architectures were able to generalise between classes in the *Speech / WN* configuration. For this reason, in this paper we focus on the results obtained with the *Speech / WN* configuration as it is interesting to analyse the generalization achieved by the learning process.

We implement an architecture with three degrees of freedom in order to compare different SSL systems. The architecture is depicted in Fig. 2. Each degree of freedom represents a layer, or processing step, that can be accomplished by alternative methods. The total architecture layers consist of *preprocessing*, *representation*, *clustering* and *classification* of binaural sound input. From these layers, only the preprocessing is performed by a fixed algorithm and therefore not considered a degree of freedom. During this step sound input is decomposed in several FCs with the Patterson-Holdsworth filter bank (PHFB) [15].

The representation layer is in charge of numerically characterising ITDs and/or ILDs. The clustering layer is an intermediate step that can potentially improve the performance of classifiers, as it can distribute a large number of prototype vectors similarly to the underlying distribution of the training data. The clustering layer is not present in some of the tested systems, as it is also possible to directly classify the output of the representation layer. Finally, the classification layer generates an output angle that can be used for motor control. In the following subsections we detail further each of the processing layers in the architecture.

A. Preprocessing of Sound Signals

The first stage in our SSL system is the PHFB. This filter decomposes the left (L) and right (R) sound recordings in frequency components $f \in \{1, 2 \dots F\}$, where $F = 20$. All f are equally separated on a logarithmic scale between 200 Hz and 4000 Hz, with an increase in bandwidth resembling the human cochlea. Afterwards, only the corresponding f from L-R signals are compared for the extraction of spatial cues. This step is used by all classification methods we describe in this paper (see Fig. 2).

B. Representation of Spatial Cues

The basis of SSL algorithms is the set of localisation cues chosen as input. Therefore, the method used to represent, or characterise, spatial cues could influence the accuracy of the system's output. We want to compare the performance of our SSL system when representing spatial cues with traditional signal processing techniques against bio-inspired methods.

For this reason we choose two of the most representative methods in binaural SSL research for representing ITDs: Cross-correlation (CCR) [16] and MSO Jeffress coincidence detector [11], [17]. We also make use of ILD cue and represent it with an LSO model previously presented by the authors [11]. Furthermore, we compare two integration methods for the MSO and LSO outputs. The first method (MSO-LSO) simply appends the output of the MSO and LSO models, and the second method (Bayes IC) integrates the output of both models using bayesian inference. In Fig. 3 are shown further details on the MSO, LSO and IC models. In the following sub-subsections we detail further each of the representation methods.

1) CCR – Cross-Correlation: The CCR technique is used to estimate the cross-correlation sequence $CCR_{L,R}$ between

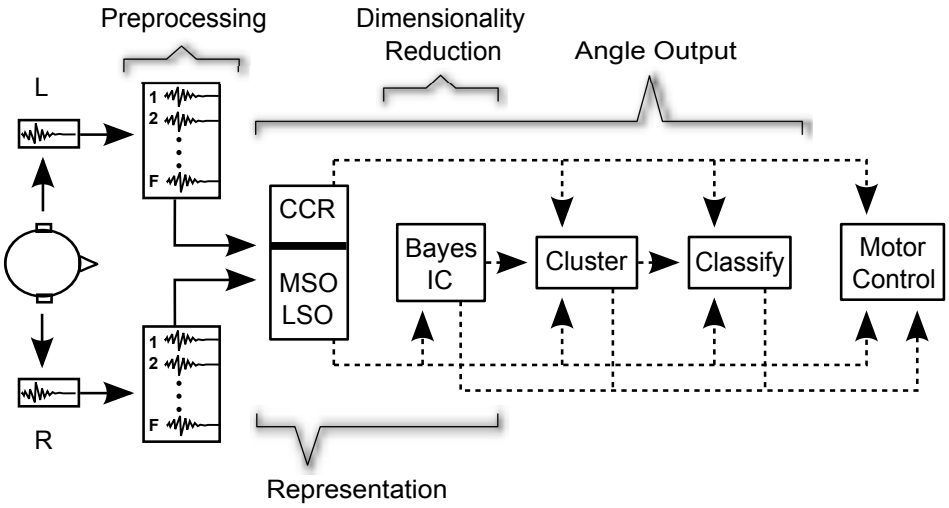


Figure 2: Testing architecture. Solid lines represent fixed steps and dotted lines represent different systems that can be used for motor control. The shadowed brackets indicate the processing layers in the architecture. The *preprocessing* consists on decomposing the sound input in several FCs with the Patterson-Holdsworth Filter Bank (PHFB) [15]. Then, the *representation* layer numerically characterises different spatial cues. Alternatively, the representation provided by the Bayes IC integrates output from the MSO and LSO in vectors with reduced dimensionality. The *clustering* layer distributes a larger number of prototype vectors in the space of represented cues in order to test for a possible improvement in SSL accuracy. All systems were tested with and without this intermediate layer. Finally, the *classification* layer produces an output angle that can be used for motor control.

L and R input signals, assuming them to be random stationary processes sampled from time window Δt.

$$CCR_{L,R}(j,f,\Delta t) = \begin{cases} \sum_{i=0}^{J-1} L_{i,f,\Delta t} \cdot R_{i-j,f,\Delta t}, & \text{for} \quad 0 \leq j \leq J \\ CCR_{L,R}(-j,f,\Delta t), & \text{for} \quad -J \leq j < 0 \end{cases}, \tag{1}$$

where i represents sampled values from the input signals, j are the ITD shifts made when computing the correlation sequence and J is the length of the input signals.

We use the correlation sequences of all f as input to the clustering or to the classification algorithms. However, the output angle Θ can also be estimated directly from the j that maximizes the correlation over all f with the *winner-takes-all* (WTA) rule:

$$ITD_{win} = \arg\max_{j} \left(\sum_{f} CCR_{L,R}(j,f,\Delta t) \right) . \tag{2}$$

We are interested in using WTA for benchmarking, as it is the classification technique the authors previously used in the MSO, LSO and IC models [11], [12]. Due to the geometry of the head, ITDs vary non-linearly as a sound source moves around us. Therefore, the output angle is computed as follows:

$$\Theta = \sin^{-1} \left(\frac{ITD_{win} - ITD_{max} + 1}{ITD_{max}} \right) , \tag{3}$$

where ITD_{max} is the maximum possible ITD that occurs when the sound source is aligned with the interaural axis.

2) MSO – Jeffress Coincidence Detector: One of the methods we use for extracting ITDs is Liu et al. [11] SNN model of the MSO. It was developed by some of the authors and we want to test it in a different anthropomorphic head with ego-noise. This method is inspired by neurophysiological theories describing the underlying mechanisms of the MSO [18]. After decomposing the sound signals with the PHFB, each frequency component f is phase-locked to its positive values. This means that the greatest probability of a spike being produced by a hair cell in the organ of Corti occurs when the amplitude of vibrations in the basilar membrane is maximal.

Afterwards, all maximum positive values in time window Δt are compared and phase shifts between these maximums are used to estimate ITDs. As a last step, neurons $k \in \{1, 2 \dots K\}$ in the MSO respond to different ITDs and for every time window Δt generate a spikes matrix $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$ of size $F \times K$.

We can feed the classification algorithms with $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$, or directly compute the output angle Θ from the k with maximal neural activity over all f. For the latter case, ITD_{win} could be estimated using the WTA rule as in eq. (4) and Θ as in eq. (3).

$$ITD_{win} = \arg\max_{k} \left(\sum_{f} \mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}} \right) . \tag{4}$$

3) LSO – Representation of ILDs: For estimating ILDs we use Liu et al. [11] SNN model of the LSO. It also was developed by some of the authors and our current objective is to test it in a different anthropomorphic head with ego-noise.

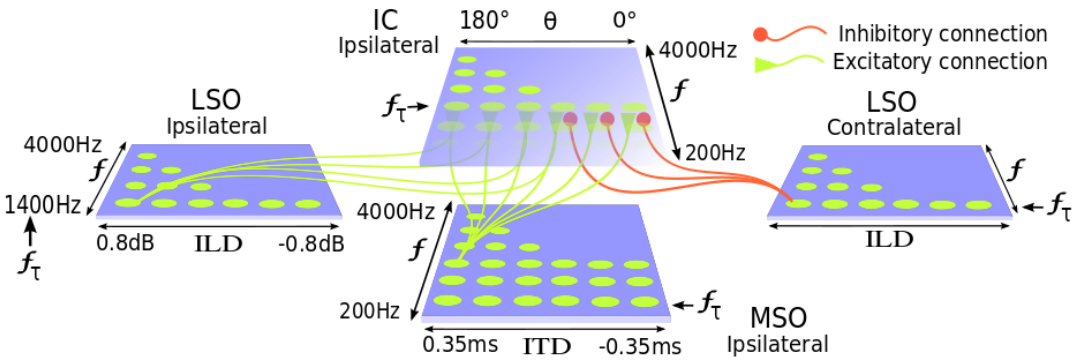


Figure 3: Multiple delay lines deliver spike-trains to MSO cells according to the Jeffress model [8]. MSO neurons respond to frequencies between 200 Hz and 4000 Hz. The difference of the wave amplitudes that produced a spike in the MSO is used to generate a spike in the LSO. LSO neurons respond to frequencies between ∼1000 and 4000 Hz. The MSO has excitatory connections to the IC in all frequencies. The LSO has excitatory and inhibitory connections to the IC in frequencies between ∼1000 and 4000 Hz.

In the LSO model neurons $k \in \{1, 2 \dots K\}$ fire depending on differences in L-R amplitudes for each f. Using the same pairs of L-R values from which ITDs are measured, ILDs are computed as $\log(L_{f,t}/R_{f,t})$. Therefore, at every time step Δt a spikes matrix $\mathbf{S}^{\mathbf{LSO}}_{\mathbf{\Delta t}}$ of size $F \times K$ is generated. Afterwards, output angles can be obtained following the same procedures applied to $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$.

4) IC – Bayesian Dimensionality Reduction: Reducing efficiently the dimensionality of input vectors can decrease the amount of data and time required for training machine learning algorithms. For this reason we also test the clustering and classification algorithms with an integrated version of the MSO and LSO output vectors. Such integrated vectors are constructed using Bayesian inference in a model of the inferior colliculus (IC) [11]. An important computational advantage comes from the IC dimensionality reduction, as IC output vectors are more than six times smaller than the MSO and LSO output vectors together. More details of the IC integration architecture are shown in Fig. 3.

Additional benefit from this integration process comes from the overlap of MSO excitatory connections and LSO inhibitory connections. The LSO captures the useful information for SSL contained in high frequencies, but generates ambiguous information from low frequencies. The MSO captures the useful information for SSL contained in all frequencies, but also generates ambiguous information from high frequencies. For this reason, LSO inhibitory connections can help to remove misleading information generated by the MSO at high frequencies. Therefore, the IC provides a more accurate representation of auditory cues along all f.

Similar to the previous cues, the IC model generates a spikes matrix $\mathbf{S}^{\mathbf{IC}}_{\mathbf{\Delta t}}$ at every time step Δt. Again, output angles can be computed with the same procedures applied to $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$. Further details on the architecture of the MSO, LSO and IC models can be found in [12]. Now we proceed to introduce and justify the selection of clustering methods.

C. Clustering of Spatial Cues

Clustering algorithms can be used directly for classification when having the same number of prototypes $p \in \{1, 2 \dots P\}$ and target classes $c \in \{1, 2 \dots C\}$. However, with a larger P it is possible to cover more closely the distribution underlying the training data, hence, improving the overall performance of the system. In the case of SSL, the distribution of auditory cues in each representation space can be highly convoluted. Therefore, using $P \gg C$ can spread the trained prototypes closer to the distribution of the characterised cue.

Due to the fact that several p can belong to a single c, an additional requirement is the inclusion of another layer in the architecture for classifying the winning p. Again, the criteria for selecting clustering algorithms is to compare a common statistical technique against a neural method, for which we choose K-Means (KM) [19] and Self Organizing Feature Maps (SOM) [20].

1) K-Means Clustering: Due to its simplicity and speed relative to other clustering techniques, KM is included as a benchmark against the more sophisticated SOM. The best results are achieved with $K = 26$ and using a randomly chosen sample of the training data as starting positions for the prototypes. The Euclidean distance is used as the standard metric for all mathematical procedures described in this paper.

2) Self Organizing Map: Due to its topology-preserving property SOMs facilitate visualisation of the data structure in lower dimensions. We use the SOM in two different configurations. In the first one $P = C$ and its output can be directly used for motor control. In the second configuration $P = C^2$ and a classification layer is added on top of it. In both cases the ordering phase consists of 1000 steps, has a learning rate $\eta = 0.9$ and the neighbourhood distance (*ND*) decreases from the furthest neuron to 1. The tuning phase consists of additional 4000 steps where $\eta = 0.02$ and *ND* $= 1$.

D. Classification of Spatial Cues

In our testing architecture, the classification layer receives input from the representation layer or from an intermediate clustering layer. Following the same heuristic, we compare a common statistical technique for benchmarking against a pair of artificial neural networks (ANN). K-Nearest Neighbours (KNN) [21] is the chosen statistical technique and the selected ANNs are the Multilayer Perceptron (MLP) [22] and Radial Basis Functions network (RBF) [23].

1) K-Nearest Neighbours: KNN is a relatively simple, yet powerful, classification technique. Instead of exhaustive search, we use a KD-Tree to reduce the cost of finding a nearest neighbour from $O(N^2)$, to $O(NlogN)$ for N data points [24]. The best performance is obtained with $K = 4$.

2) Radial Basis Functions Network: An important advantage of RBF networks over other ANNs is its much faster training procedure. The number of neurons in the hidden layer is equal to the number of training instances and the network shows best overall performance with a spread $\sigma = 10$.

3) Multilayer Perceptron: The MLP is a universal function-approximator robust to noise, whose internal dynamics are one of the best understood in the field of ANNs. During training we use the following data ratios: *training* $= 0.8$, *validation* $= 0.1$ and *testing* $= 0.1$. We use hyperbolic tangent as activation function and, due to its increased speed for large networks, we use the scaled conjugate gradient [25] method for training the MLP. The network parameters are set to the standard values $\sigma = 5 \times 10^{-5}$ and $\lambda = 5 \times 10^{-7}$ according to [25].

The number of hidden neurons (*HN*) changes depending of the architecture being tested. When the MLP receives input from the representation layer $HN = \lfloor \mathbf{v}_{\dim}/2 \rceil$, where $\mathbf{v}_{\dim}$ is the dimensionality of input vector $\mathbf{v}$. When the MLP receives input from the clustering layer $HN = C \times 2$.

III. Results and Discussion

The system's output is analysed using measures from information retrieval theory [26]: *recall* (Re), *precision* (Pr), *specificity* (Sp), *accuracy* (Ac) and *F-measure* (Fm). The value of each measure is computed from information contained in the confusion matrices of the output angles. Specifically from the *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN):

$$Pr = \frac{TP}{TP + FP} \ , \tag{5a}$$

$$Re = \frac{TP}{TP + FN} \ , \tag{5b}$$

$$Sp = \frac{TN}{TN + FP} \ , \tag{5c}$$

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \ , \tag{5d}$$

$$Fm = 2 \times \frac{Pr \times Re}{Pr + Re} \ . \tag{5e}$$

The performance of all representation and classification algorithms is displayed in recall-precision plots in Fig. 4. In all representations the top performance is achieved when *training / testing* with *Speech / WN*. From all the tested representation methods, three lead to much more accurate results: MSO, MSO-LSO and Bayes IC. Furthermore, within those three representations, three classification algorithms perform significantly better than the rest with $Re > 0.98$ and $Pr \geq 0.89$: KNN, MLP and RBF.

The performance measures of the three best, or winning, classifiers is shown in Tables I, II and III. In order to show the considerable increase in performance of the winning classifiers, Table IV shows the performance results of the *second best* classifiers with $Pr > 0.7$. These *second best* systems achieved higher performance when clustering and classifying input from the MSO representation.

Table I: K-NN - Classification performance with each representation method. Best results are highlighted in bold.

	Pr	Re	Sp	Ac	Fm
CCR	0.21	0.75	0.59	0.61	0.33
MSO	**0.96**	**1.00**	**0.96**	**0.98**	**0.98**
LSO	0.15	0.65	0.54	0.55	0.25
MSO-LSO	**0.89**	**0.99**	**0.90**	**0.94**	**0.93**
Bayes IC	0.18	0.75	0.53	0.56	0.29

Table II: MLP - Classification performance with each representation method. Best results are highlighted in bold.

	Pr	Re	Sp	Ac	Fm
CCR	0.55	0.89	0.75	0.79	0.68
MSO	**0.98**	**1.00**	**0.98**	**0.99**	**0.99**
LSO	0.21	0.76	0.56	0.59	0.33
MSO-LSO	**0.95**	**0.99**	**0.95**	**0.97**	**0.97**
Bayes IC	**0.93**	**0.99**	**0.94**	**0.96**	**0.96**

Table III: RBF - Classification performance with each representation method. Best results are highlighted in bold.

	Pr	Re	Sp	Ac	Fm
CCR	0.20	0.75	0.56	0.58	0.32
MSO	**0.97**	**1.00**	**0.97**	**0.98**	**0.98**
LSO	0.42	0.83	0.70	0.73	0.56
MSO-LSO	**0.93**	**0.99**	**0.93**	**0.96**	**0.96**
Bayes IC	0.27	0.73	0.68	0.68	0.40

Table IV: Performance of the *second best* systems with $Pr > 0.7$. The same classifier has better performance when clustering input from the MSO.

	Pr	Re	Sp	Ac	Fm
MSO: KM-RBF	0.75	0.94	0.85	0.88	0.84
MSO: SOM-RBF	0.75	0.94	0.85	0.88	0.83

In the following subsections we detail the performance of the best classifiers with each of the cue representations, and compare them against the WTA classification rule we applied in our previous work with a different robotic platform [12]. In all cases the *training / testing* configuration is *Speech / WN*.

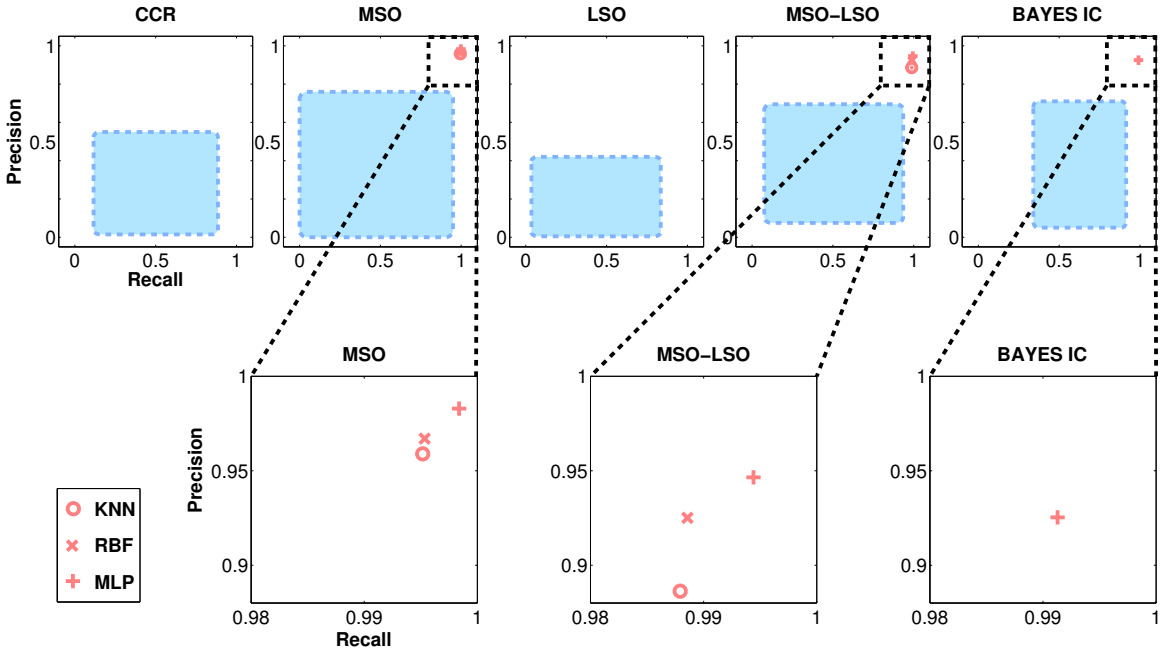


Figure 4: Recall - Precision. The markers show the performance of the best representation methods. For clarity, the area where performance of the non-winning classifiers fall is represented by shaded squares in the top five plots. In all representations the best performance is achieved by the *training / testing* configuration *Speech / WN*. Three representation methods lead to a significantly better performance: MSO, MSO-LSO and IC. Within these representations, three classification algorithms obtain best results with a recall $Re > 0.98$ and precision $Pr \geq 0.89$: KNN, MLP and RBF.

A. CCR: Representation of ITDs

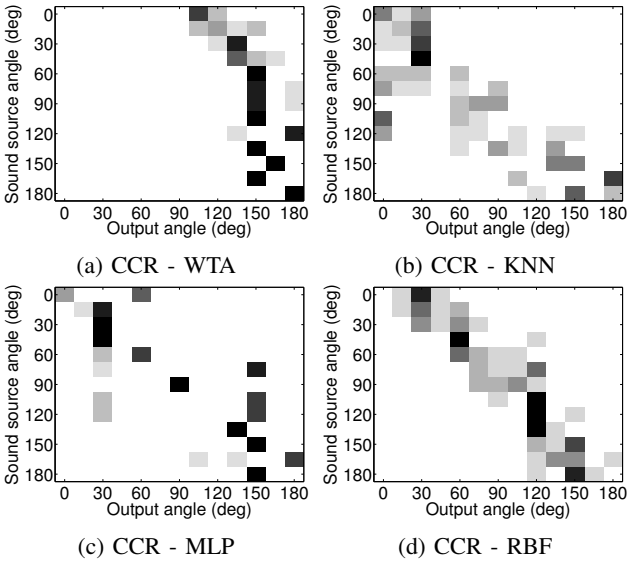
This statistical method shows lower performance than the MSO model, its bio-inspired counterpart. However, the confusion matrices in Fig. 5 show that the angle deviation from ground truth of KNN and RBF outputs is small for practical purposes when using CCR as input. It remains an open possibility to improve the performance of this method when adding a noise cancelling layer to the system. This enhancement is desirable for online applications as CCR provides a faster characterisation of ITDs than the MSO.

B. MSO: Representation of ITDs

The MSO allows the system to reach the highest accuracy relative to all other representations. Also it is the only representation that allows the three winning classification methods to perform almost perfectly. Figure 6 clearly shows the improvement of the winning classifiers with respect to the baseline method WTA. The MSO performed robustly under high levels of ego-noise, even when the noise frequency components were overlapping with the f provided by the PHFB.

C. LSO: Representation of ILDs

This bio-inspired method is the only one we used for representing ILDs, as there are no standard statistical techniques for benchmarking. The extraction of ILDs is dependant

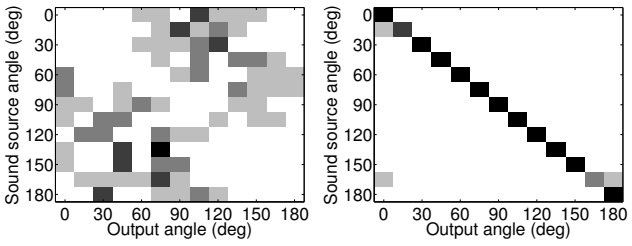


(a) CCR - WTA (b) CCR - KNN

(c) CCR - MLP (d) CCR - RBF

Figure 5: Confusion matrices when using CCR representation as input for WTA and the winning classification methods.

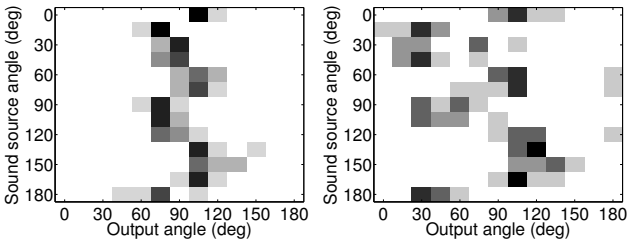
on the geometrical and material properties of the robotic head being used. In previous work the authors successfully used ILDs for SSL [11] with a styrofoam humanoid head. Nevertheless, Fig. 7 shows that the classification techniques can not infer correctly the location of sound sources from



(a) MSO - WTA (b) MSO - KNN

(c) MSO - MLP (d) MSO - RBF

Figure 6: Confusion matrices when using MSO representation as input for WTA and the winning classification methods.



(a) LSO - WTA (b) LSO - KNN

(c) LSO - MLP (d) LSO - RBF

Figure 7: Confusion matrices when using LSO representation as input for WTA and the winning classification methods.

ILDs extracted with the iCub head.

One possible reason for this low performance is the presence of high levels of ego-noise. After inspecting a spectrogram of the iCub's ego-noise, we found that 15 of the 20 frequency components f provided by the PHFB cochlear model are located in the spectral region with most intense ego-noise. Therefore, noise in these frequencies can significantly reduce the SNR of incoming stimuli and impede the use of ILDs for SSL. More details on the PHFB preprocessing step are given in subsection II-A.

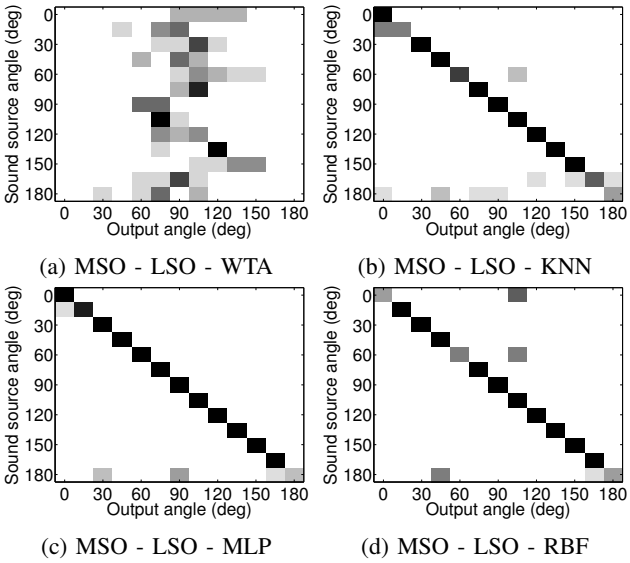
Another possibility is that the inaccuracy of the system when using ILDs is due to the material properties of the robotic head. Differently from [11], the iCub head is hollow and has openings in the back, reducing in this way the shadowing effect needed to effectively use ILDs for SSL.

D. MSO - LSO: Linear Integration of ITDs and ILDs

This integration of ITDs and ILDs, represented by the MSO and LSO models, is much simpler than the IC Bayesian integration. In this case the MSO and LSO activation matrices for Δt are simply appended and used as input for the next system layer. It is interesting to see in Fig. 8 that the performance of the three winning algorithms dramatically increase with respect to the IC method, even though the complexity of the characterisation procedure decreases. However, it is also important to keep in mind that the training procedure also becomes more demanding as the dimensionality of the input vectors to the classification layer grows by a factor of $\sim$7.

E. IC: Bayesian Integration of ITDs and ILDs

This representation is the most biologically plausible from the set we describe in this paper, but it is also the most computationally expensive. On the other hand, the dimensionality reduction provided by this method speeds up considerably the training procedure of the classification algorithms.

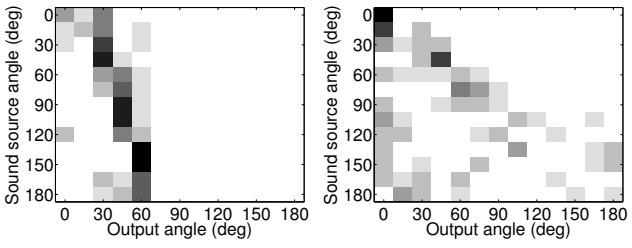


(a) MSO - LSO - WTA (b) MSO - LSO - KNN

(c) MSO - LSO - MLP (d) MSO - LSO - RBF

Figure 8: Confusion matrices when using MSO and LSO representations as input for WTA and the winning classification methods.

Fig. 9 shows the confusion matrices when simply using WTA for classification, versus the performance of the three winning classifiers. The output of WTA is strongly biased towards a small range of angles on the 0° quadrant, possibly due to the non-linear encoding of information across the IC neurons. Also KNN and RBF show a bias, albeit smaller, towards the same region. In contrast, the MLP is capable of correctly encoding the spiking activity of the IC.



(a) Bayes IC - WTA (b) Bayes IC - KNN

(c) Bayes IC - MLP (d) Bayes IC - RBF

Figure 9: Confusion matrices when using Bayes IC representation as input for WTA and the winning classification methods.

IV. Conclusion and Future Work

In this study we compare different methods for the representation and classification of spatial cues for SSL. We found three best representation methods: MSO, MSO-LSO and Bayes IC. There are also three winners from our set of classifiers: KNN, MLP and RBF.

The fastest method for representation of ITDs is the MSO model alone. Nevertheless, MSO-LSO and Bayes IC methods can be more robust when classifying sounds richer in high frequency components.

It has been shown that the LSO model performs well under lower levels of ego-noise [12], more precisely, with levels of $\sim$40 Hz instead of $\sim$60 Hz. In future work we will test the system when using ILDs in combination with a noise cancelling module, as we expect this configuration will improve the accuracy of SSL with the iCub head.

With respect to training speed, the fastest classification method is KNN. However, for life-long learning the standard KNN method would become computationally expensive, i.e. slow, as the system would need to store a very large number of prototypes from possibly several environments. Therefore, for practicality the MLP and RBF networks represent a better option in terms of online speed.

Finally, one aim of our further work is the propagation of probabilities in time and the use of vision. We expect that both additions will improve the confidence of the classification algorithms and their robustness against higher levels of reverberation.

References

[1] N. Roman, D.L. Wang, and G.J. Brown. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114:2236–2252, 2003.

[2] J.M. Valin, F. Michaud, J. Rouat, and D. Létourneau. Robust sound source localization using a microphone array on a mobile robot. In *International Conference on Intelligent Robots and Systems*, volume 2, pages 1228–1233. IEEE, 2003.

[3] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi. Three ring microphone array for 3D sound localization and separation for mobile robot audition. In *International Conference on Intelligent Robots and Systems*, pages 4172–4177. IEEE, 2005.

[4] J. C. Middlebrooks and D. M. Green. Sound localization by human listeners. *Annual review of psychology*, 42(1):135–159, 1991.

[5] K. Voutsas and J. Adamy. A biologically inspired spiking neural network for sound source lateralization. *Transactions on Neural Networks*, 18(6):1785–1799, 2007.

[6] T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, and B. Scholling. Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. In *International Conference on Intelligent Robots and Systems*, pages 860–865. IEEE, 2006.

[7] J. Bauer, C. Weber, and S. Wermter. A som-based model for multi-sensory integration in the superior colliculus. In *International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 2012*, pages 1–8. IEEE, 2012.

[8] J. Schnupp, I. Nelken, and A. King. *Auditory neuroscience: Making sense of sound*. The MIT Press, 2011.

[9] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner. A probabilistic model for binaural sound localization. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(5):982–994, 2006.

[10] J. Nix and V. Hohmann. 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.

[11] J. Liu, D. Perez-Gonzalez, A. Rees, H. Erwin, and S. Wermter. A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation. *Neurocomputing*, 74(1-3):129–139, 2010.

[12] J. Davila-Chacon, S. Heinrich, J. Liu, and S. Wermter. Biomimetic binaural sound source localisation with ego-noise cancellation. *Artificial Neural Networks and Machine Learning (ICANN), Lausanne, Switzerland, September 2012*, pages 239–246, 2012.

[13] J. Bauer, J. Davila-Chacon, E. Strahl, and S. Wermter. Smoke and mirrorsvirtual realities for sensor fusion experiments in biomimetic robotics. In *International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Hamburg, Germany, September 2012*, pages 114–119. IEEE, 2012.

[14] R. Beira, M. Lopes, M. Praga, J. Santos-Victor, A. Bernardino, G. Metta, F. Becchi, and R. Saltarén. Design of the robot-cub (icub) head. In *International Conference on Robotics and Automation (ICRA), 2006.*, pages 94–100. IEEE, 2006.

[15] M. Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. Technical report, Apple Computer, Perception Group, 1993.

[16] J. Benesty, M.M. Sondhi, and Y.A. Huang. *Springer handbook of speech processing*. Springer, 2007.

[17] P.X. Joris, P.H. Smith, and T.C.T. Yin. Coincidence detection minireview in the auditory system: 50 years after Jeffress. *Neuron*, 21:1235–1238, 1998.

[18] G. Ashida and C. E. Carr. Sound localization: Jeffress and beyond. *Current opinion in neurobiology*, 21(5):745–751, 2011.

[19] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[20] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[21] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[22] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[23] J. Park and I. W. Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.

[24] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[25] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4):525–533, 1993.

[26] C.J. Van Rijsbergen. Evaluation. In *Information Retrieval*. Butterworths, London, 1979.