Building a sound localization system for a robot head *

Oscar Déniz, Jorge Cabrera, Mario Hernández
Universidad de Las Palmas de Gran Canaria
Instituto Universitario de Sistemas Inteligentes
y Aplicaciones Numéricas en Ingeniería
Edificio Central del Parque Científico-Tecnológico
Campus Universitario de Tafira
35017 Las Palmas - España
odeniz@dis.ulpgc.es

Abstract

Sound localization plays a crucial role in the perceptive function of living beings, being vital for the survival of many animal species. It can also play an important role in human-machine interaction and robot interaction with its environment. In the typical setting two audio signals gathered by a pair of microphones mounted on both sides of a head are processed to extract meaningful cues for deriving the approximate azimuthal localization of the sound source. In this paper a new method of feature extraction for sound localization is described. It has been developed for a robot head currently under construction. The proposed method is compared in off-line experiments with another feature extraction scheme developed for the Cog humanoid robot, showing superior performance with all the signals tested.

Keywords: Sound localization, Human computer interaction, Robotics

1 Introduction

Sound localization plays a crucial role in the perceptive function of living beings, being vital for the survival of many animal species. Barn owls, for example, can hunt in total darkness, thanks to their extraordinary sound localization abilities [6]. Humans are no exception. Our ability to localize where a sound comes from warns us of potential danger. Sound localization is also an important attention fixing mechanism, especially in a verbal communication setting.

Our sound localization abilities stem from the fact that we have two ears. Although we could distinguish different sounds with one ear alone, pinpointing where the sounds are coming from requires at least two ears. Reliably localizing a sound source in 3-D space requires even more hearing sensors. Sound differences between the signals gathered in our two ears account for much of our sound localization abilities. In particular, the most important cues used are *Interaural Level Difference* (ILD) and *Interaural*

*The first author is supported by graduate grant *D260/54066308-R* of Universidad de Las Palmas de Gran Canaria. This work was partially funded by Gobierno de Canarias *PI2000/042* and ULPGC *UNI2002/16* research projects.

Time Difference (ITD). ILD cues are based on the intensity difference between the two signals. This intensity difference, which can be of up to 20dB, is caused mostly by the shading effect of the head. ITD cues are based on the fact that sound coming from a source will be picked up earlier by the ear nearest to the sound source. This difference will be maximum when the sound source is directly from one side, and minimum when it is in front of the head. Both ILD and ITD cues are dependent on the sound frequency. ITD cues are reliable for relatively low frequencies (up to 1 Khz, approximately), while ILD cues are better for higher frequencies (see [3] for an explanation of this).

Humans use additional cues for sound localization. The shape of our head and outer ears affect received sounds in a manner dependent on arrival angle and frequency. A model of this process referred to in the literature is the *Head Related Transfer Function* (HRTF). HRTF-based cues allows us to obtain an estimate of the sound source elevation and also to distinguish between sound originating in front of and behind the listener. A more detailed description of sound localization mechanisms can be found in [2, 12, 5, 3].

These and other physiological findings have been emulated in computer-microphone systems with relative success. Sound localization can play an important role in human-machine interaction and robot interaction with its environment. This paper is organized as follows. In Section 2 we briefly describe previous approaches to sound localization that use computer-microphone systems. In Section 3 we introduce new feature (cue) extraction techniques that improve upon those used in previous systems. The advantages of using these new features are analyzed in Section 4. Finally, in Section 5 the most important conclusions are described.

2 Previous work

The first important work on a computer sound localization system is [7]. With a combination of hardware and software the system aims to learn to localize sounds in complex environments. The output of the system can be one three values: frontal, right and left. Both ILD and ITD cues are extracted from signals gathered from two microphones and a pre-amplifier circuit. Signals were previously high-pass filtered to remove background noise and then they were divided into segments. For each segment, the cues extracted are: difference of the two maximum positive values, difference in the positions of these maxima, delay between signals (computed by performing a cross-correlation of both signals), difference in the sum of magnitudes of the signals and filterbank-based cues. Filterbank-based cues are computed by dividing the spectrum of the signals in a number of equally spaced banks, computing the sum of magnitudes in each bank. The cue itself is the difference between the sums of the two signals. 4 banks were used, so the complete feature set used had 8 cues. These cues were fed into a feedforward multi-layer perceptron with three outputs. This network was trained and tested using three sounds (hand clap, spoken "ahh" and door slam). This system is currently working on the *Cog* humanoid robot at MIT [1].

In [1] a similar system is introduced. The input signals were divided into segments. The extracted cues were: difference between maximum values, difference in the positions of the maxima, correlation and difference in the sum of magnitudes. A classifier was not used, the output of the system was programmed (basically by means of comparing values). This can be a disadvantage in certain settings, because many thresholds have to be manually found (think for example that the difference in intensities could not be exactly zero for a perfectly frontal source, because the two microphones and/or pre-amplifier circuits could have different gains).

A work that used only one ITD cue is described in [10]. After performing high-pass and low-pass filtering, a signal level test was performed to discriminate between sound and silence. After that, correlation was performed to obtain the ITD estimate and another threshold test was performed on its result (based on the ratio peak/average correlation values). Finally, in order to discard outliers, many estimates were gathered before giving their median value as a response. Correlation was only computed for the possible range of temporal displacement

[1] Prof. Rodney Brooks, personal communication

values (as the sound speed is finite, there is a maximum delay possible in the ITD cue, and it depends on the distance between microphones), and this in turn allowed for a faster response. The output of the system was an angle, and it was tested with two types of sound (impulsive sound and speech).

In this paper we are mainly interested in studying practical approaches to the problem, as our goal is to integrate a sound localization system in a robot. For examples of simulated auditory models or systems that use more than two microphones or special-purpose hardware see [9, 4]. For the use of sound localization for robot positioning see [8, 11].

3 Feature Extraction

In this work we have used the system in [7] as a base line for comparison, as it uses both ITD and ILD cues and has found practical use. We describe here a new cue extraction procedure that can eliminate some minor errors. The extracted cues for a computer sound localization system are always subject to error because of background noise, electrical equipment noise, and specially echoes and reverberation. Echoes originate when sound signals reflect off planar surfaces. Often the effect of multiple reflective sound paths can be as loud or even louder than the sound travelling a direct path from the source.

An important fact to consider is the effect of using segments of the input signals. All systems described in Section 2 divide the input signal in segments, and extract features from these. However, none of the systems described consider problems that could arise at boundaries. If we consider for example the first extracted cue, difference of maximum positive values, the maximum of signal L (left) could be just at the beginning of the segment. If the source is on the right side, signal L will be delayed with respect to signal R (right). Thus the maximum of signal R is not associated with the maximum in signal L. This in turn affects the second extracted cue, the difference in maximum positions. We propose to extract the first cue as follows. The maximum of signal L is found, be it Ml. Then we search in signal R for the maximum in a zone around the position of Ml. The zone has a length of $2W$, where W is the maximum possible interaural delay. The value of W depends on the distance between microphones and the sound speed. Any (correct) ITD estimate must be equal or lower than W (in absolute value). If Ml falls in the initial zone of the segment, of length W, or in the final zone of the segment, also of length W, it is discarded and we repeat the procedure beginning with signal R. If the maximum of signal R, Mr, also falls in one of these "dangerous" zones, and the zone in which it falls is different from that of Ml, the segment is discarded (no cues are extracted from it). Figure 1 shows the three possible cases. This way, some segments are not used for localization, though the first (and second) cues extracted for other segments should be more reliable.

As for the third cue (temporal displacement as extracted from the maximum of the cross-correlation), we propose to use the option already described in [10], of considering only the maximum in the zone of possible temporal displacement, defined again by W.

Another characteristic of the proposed procedure is related to changes in the volume of the signals. Let us suppose that we want our system to give one of three possible outputs: frontal, left or right. We could fix by hand thresholds for the cue values that define the three regions. Alternatively, these regions could be inferred by training a classifier. In any case, what happens when the input signal has a different intensity (volume) than that used for training/fixing the thresholds? or, what happens when the source gets closer to or farther from the hearing system?. Figures 2 and 3 show that the value of the extracted ILD cues depend on the volume of the sound signal and the distance of the source.

If we consider for example the first cue extracted, difference between maximum values, the obtained value could be incorrectly discriminated by the fixed thresholds/classifier, because the difference is dependent on the intensity of the input signal. Thus, ILD cues (cues 1 and 4) should be normalized. Let Sl and Sr be the sum of magnitudes for the left signal segment and right signal segment, respectively. Then the two ILD cues should be:

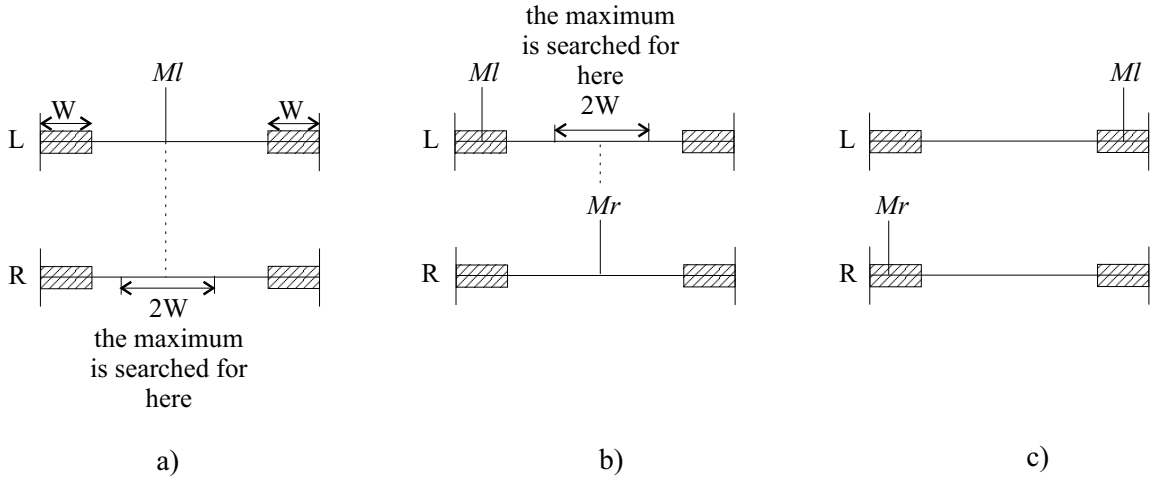

Figure 1: a) Ml does not fall in the initial or final "dangerous" zones, b) Ml falls in the "dangerous" zone, c) both Ml and Mr fall in "dangerous" zones. In the last case the sample is discarded.

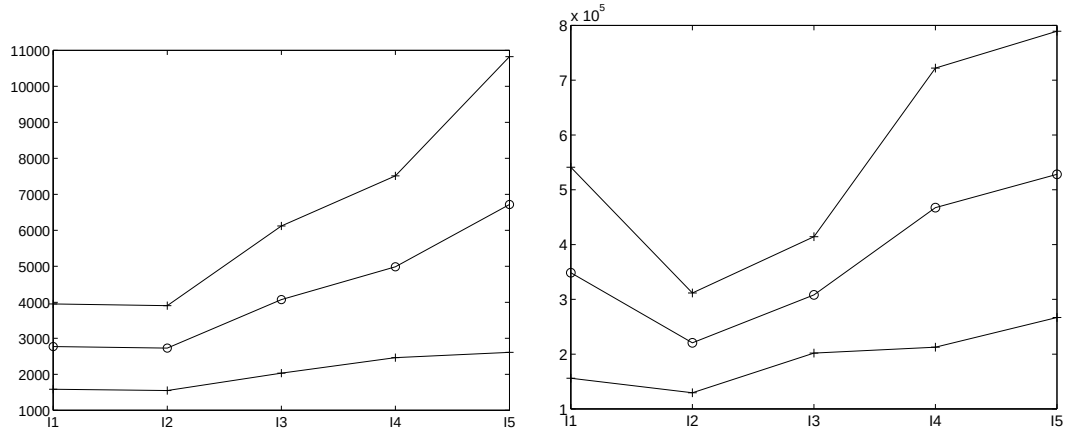

Figure 2: Effect of changes in the intensity of the sound signal. The sound source (a mobile phone) is located on the left side of the head at a constant distance (see Section 4). On the left: mean values obtained for cue 1. On the right: mean values obtained for cue 4. The upper and lower lines are the mean values plus and minus one standard deviation.

$$C_1 = \frac{Ml - Mr}{Ml + Mr} \;\; ; \;\; C_4 = \frac{Sl - Sr}{Sl + Sr} \qquad (1)$$

However, normalization can be of advantage only if the differences in volumes are predominant over the error present in the signals. Otherwise it could be worse for the cues extracted. Let $\mathbf{x}$ be the difference between signals L and R. Any extracted ILD cue can be denoted as $C = f(\mathbf{x}) + \varepsilon_f(\mathbf{x})$, the error in the extracted cue being $e = |\varepsilon_f(\mathbf{x})|$. The normalized cue can be expressed as:

$$C_N = \frac{f(\mathbf{x}) + \varepsilon_f(\mathbf{x})}{g(\mathbf{x}) + \varepsilon_g(\mathbf{x})} \qquad (2)$$

The error is, proportionally, "amplified" if $e_N > \frac{e_y}{|g(\mathbf{x})|}$. From (2), it can be shown that this occurs when:

$$|f(\mathbf{x}) - K(\mathbf{x})f(\mathbf{x}) - K(\mathbf{x})\varepsilon_f(\mathbf{x})| - |\varepsilon_f(\mathbf{x})| > 0, \qquad (3)$$

where

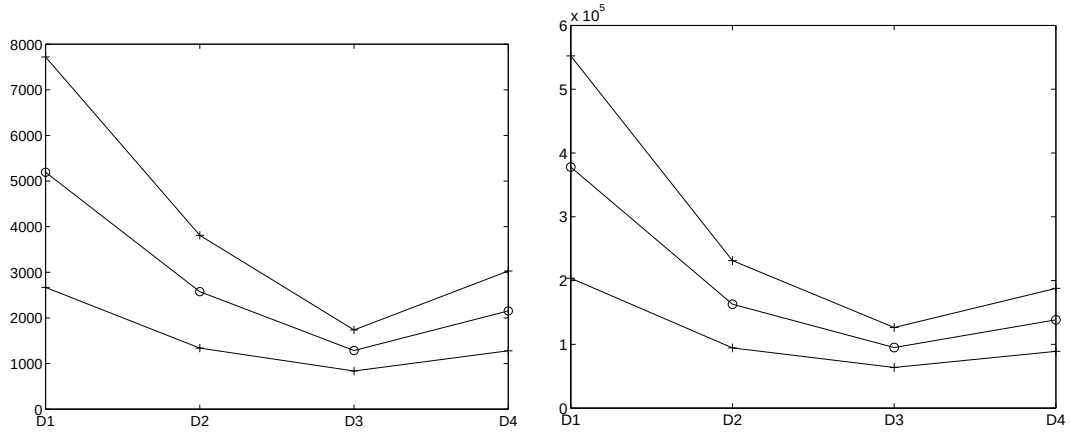

Figure 3: Effect of changes in the distance of the sound source. The sound source (a mobile phone) is located at distances such that Di>Di-1. The sound intensity is constant. On the left: mean values obtained for cue 1. On the right: mean values obtained for cue 4. The upper and lower lines are the mean values plus and minus one standard deviation.

$$K(\mathbf{x}) = \frac{g(\mathbf{x})}{g(\mathbf{x}) + \varepsilon_g(\mathbf{x})} \tag{4}$$

$\varepsilon_f(\mathbf{x})$ and $\varepsilon_g(\mathbf{x})$ can have any value and, supposing any volume is possible, so do $f(\mathbf{x})$ and $g(\mathbf{x})$. Thus there exists a possibility that the error be "amplified" after normalization. From (3) it can be seen that this possibility is smaller as $\varepsilon_f(\mathbf{x})$ and $\varepsilon_g(\mathbf{x})$ tend to zero. Also note that part of the error is caused by the use of signal segments in which the sound is beginning or ending.

With regard to spectral cues, the frequency banks used in [7] can be useful for modelling the frequency dependence of ILD cues. In any case, they should also be normalized. On the other hand, there exists a strong dependence between the reliability of ITD cues and the frequency of the sound signal, which can produce bad estimates. As explained in Section 2, an additional test was used in [10] in order to reduce such (and other) errors in the ITD estimate. If the ratio between peak and average values of the correlation result was lower than a threshold, the sample was rejected. In our system that test is used too, though the sample is never rejected. If the obtained ratio is lower than the threshold, the value of the ITD cue for the sample is substituted by the last higher-than-the-ratio value obtained. As the sample is not discarded, this allow us to take advantage of the useful ILD information in the sample.

The same mechanism was used for the second cue (difference in the positions of the maxima): if the correlation ratio is lower than the threshold, the value of the second cue is substituted by the last second cue value obtained in which the correlation ratio was higher than the threshold.

4 Experimental Results

In order to carry out experiments, sounds were recorded using two *Philips Lavalier* omnidirectional microphones, pre-amplifier circuits and a professional sound card (*Terratec EWS88*). A *DirectX* application was developed to integrate all the processing stages: low-pass filtering, sound source detection, feature extraction, data recording and playing (for off-line analysis), and classifying (see Figure 4).

This application was developed for a robot head currently under construction. For these experiments, the two microphones were placed 28 cm apart on both sides of a custom-made plastic head (see Figure 5).

Four different sounds were used in the experiments: hand claps, a call tone from a mobile phone, a maraca and a whistle, see Figure 6. The objective was to detect if the sound was coming from the left, right or frontal side. Sounds were recorded in front of the head and at between 35 and 45 degrees on the left and

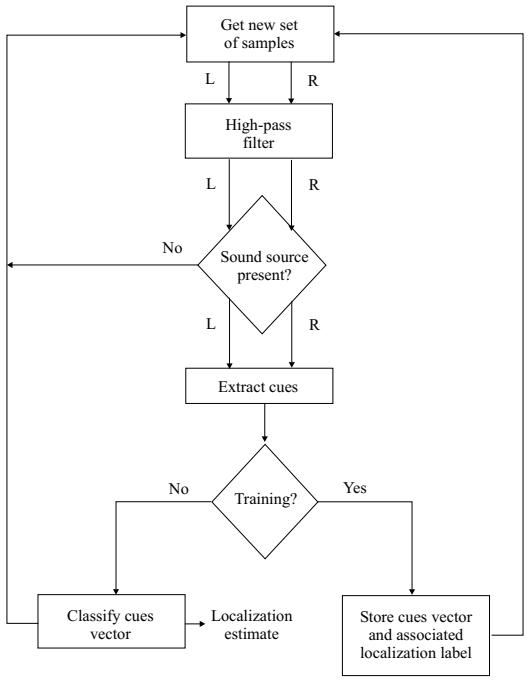

Figure 4: Steps performed by the developed sound localization module. The work described in this paper focuses on the cue extraction stage.

right sides, at a distance of at least one meter to the head. As indicated before, we have compared our feature extraction method with that used in [7], which will be referred to as 'Cog'.

In order to study the reliability of the extracted cues, the ratio between inter-class to intra-class variances will be used as a measure of overlap between samples:

$$r = \frac{\sum_{i=1}^{z}(\mu_i - \mu)^2/z}{\sum_{i=1}^{z}\sum_{j=1}^{n_i}(x_j - \mu_i)^2/n_i} \qquad (5)$$

This is actually the Fisher criterion for feature evaluation. A classifier was not used because our interest is only in the error present in the individual extracted features. The larger the ratio the better the separation of the samples for a given feature. On the other hand, a number F of consecutive cue vectors was extracted and the mean of them was given as features. The results obtained for F=250 are shown in Table 1.

The results obtained with the proposed method achieve in general a higher separation ratio for

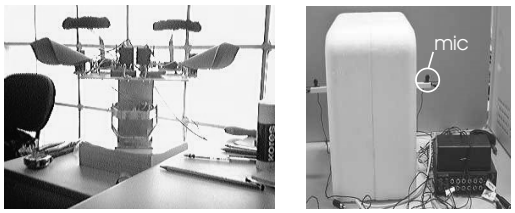

Figure 5: Left: Robot head the sound localization system is being developed for, currently under construction. Right: Plastic head used in the experiments, next to the sound card external rack and preamplifiers.

Sound	Cue1	Cue2	Cue3	Cue4
phone	6.986	0.561	0.418	4.016
claps	2.286	0.195	1.034	0.396
maraca	6.297	0.163	0.076	3.689
whistling	5.820	2.000	1.368	5.812
phone	**8.187**	**1.374**	0.100	**5.920**
claps	**3.546**	**0.687**	**1.175**	**0.409**
maraca	5.074	**0.890**	**1.884**	**4.093**
whistling	**16.71**	1.762	**2.347**	**17.09**

Table 1: Results obtained for F=250. The top half of the table shows the results obtained using *Cog's* system, while the bottom half shows the results obtained with the proposed method. Improved ratios appear in bold.

the four features used. Note that this results are achieved with a high value for F. In Table 2 the same values are shown, for F=0 (cue values are not averaged). In this case, the ratio values for the normalized cues (1 and 4) are worse in the first and second sounds. As F is low, the error is higher and, as indicated in Section 3, it could be amplified. This reflects negatively in the two first sounds because these sounds contain no significative changes in volume. The other two sounds still give a better ratio because they contain significative changes in volume, as can be seen in Figure 6.

The results using the four sounds together appear in Table 3, for both F=0 and F=250. Again, there is a significant improvement with the proposed method.

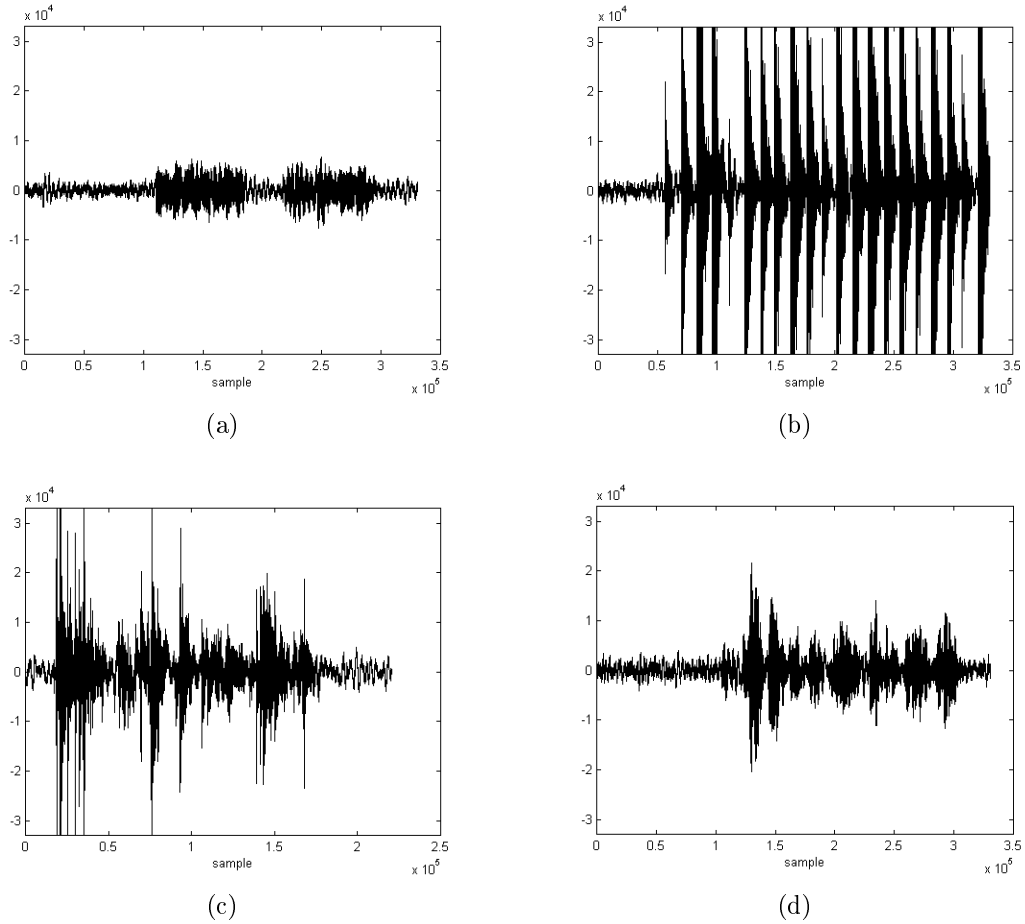

Figure 6: Sounds used in the experiments: a) mobile phone, b) hand claps, c) maraca, d) whistling.

Sound	Cue1	Cue2	Cue3	Cue4
phone	0.431	0.001	0.001	0.359
claps	0.015	7e-4	1e-4	0.004
maraca	0.078	0.002	0.004	0.116
whistling	0.141	7e-4	0.006	0.137
phone	0.300	**0.004**	**0.002**	0.296
claps	0.002	**0.004**	**0.001**	2e-4
maraca	**0.119**	**0.015**	**0.014**	**0.239**
whistling	**0.251**	**0.007**	**0.030**	**0.247**

Table 2: Results obtained for F=0.

Sound	Cue1	Cue2	Cue3	Cue4
F=0	0.053	2e-4	4e-4	0.071
F=250	0.458	0.097	0.059	0.429
F=0	**0.119**	**0.001**	**0.002**	**0.129**
F=250	**1.059**	**0.099**	**0.166**	**0.607**

Table 3: Results obtained considering the four sounds together.

5 Conclusions and Future Work

This paper describes a new method for feature extraction in the context of sound localization. The method has been implemented and will be used in a robot head currently under construction. Using the proposed procedure, extracted cues are more reliable, though a reject possibility is introduced. Our robot head is expected to work on a highly dynamic environment, where changes in the volume of the sound signals are commonplace. Such changes are due to variations in the volume of the signal itself and changes in the distance to the sound source. In the proposed procedure ILD cues are normalized, which allows for changes in the intensity of the sound signals. As sound intensity decreases with the square of the distance, moving sources are also addressed. Experiments con-

firm that the method is specially useful when the error in the signals is not too high. For all the sounds used in the experiments the extracted cues show a separation higher than the system used for comparison.

Future work will include the use of a pan/tilt (neck) unit to complete an active sound localization system. A higher localization precision would then be achievable through successive neck movements, though motor sounds would have to be considered and somehow eliminated from the sound signals.

References

[1] Jennifer Alexander. Sound localization in a meeting room, 1995. Available at `http://citeseer.nj.nec.com/136635.html`.

[2] Jens Blauert. *Spatial hearing*. MIT press, Cambridge, MA, 1983.

[3] GCAT. Perception of direction, 1999. Available at `http://www.gcat.clara.net/Hearing/perception_of_direction.htm`.

[4] A. Härmä and K. Palomäki. HUTear - a free matlab toolbox for modeling of human auditory system. In *Procs. Matlab DSP conference*, pages 96–99, Espoo, Finland, November 1999.

[5] William M. Hartmann. How we localize sound. *Physics Today*, 52(11):24–29, 1999.

[6] Seeing, hearing and smelling the world. Technical report, Howard Hughes Medical Institute, 1997. Available at `http://www.hhmi.org/senses/c210.html`.

[7] Robert R. Irie. Robust sound localization: an application of an auditory perception system for a humanoid robot. Master's thesis, Massachusetts Institute of Technology, June 1995.

[8] J.Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie. A model based sound localization system and its application to robot navigation. *Robotics and Autonomous Systems (Elsevier Science)*, 27(4):199–209, 1999.

[9] D.V. Rabinkin, R.J. Renomeron, A. Dahl, J.C. French, J.L. Flanagan, and M.H. Bianchi. A DSP implementation of source location using microphone arrays. In *Procs. of the SPIE*, volume 2846, pages 88–99, Denver, Colorado, August 1996.

[10] Greg L. Reid and E. Milios. Active stereo sound localization. Technical Report CS-1999-09, York University, 1999.

[11] Kyung B. Ryu. Application of the sound localization algorithm to the global localization of a robot. Technical Report UG 2001-5, Institute for Systems Research, University of Maryland, 2001.

[12] W. A. Yost and G. Gourevitch. *Directional hearing*. Springer-Verlag, New York, 1987.