2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

A Learning-Based Approach to Robust Binaural Sound Localization

Karim Youssef, Sylvain Argentieri, and Jean-Luc Zarader

Abstract— **Sound source localization is an important feature designed and implemented on robots and intelligent systems. Like other artificial audition tasks, it is constrained to multiple problems, notably sound reflections and noises. This paper presents a sound source azimuth estimation approach in reverberant environments. It exploits binaural signals in a humanoid robotic context. Interaural Time and Level Differences (ITD and ILD) are extracted on multiple frequency bands and combined with a neural network-based learning scheme. A cue filtering process is used to reduce the reverberations effects. The system has been evaluated with simulation and real data, in multiple aspects covering realistic robot operating conditions, and was proven satisfying and effective as will be shown and discussed in the paper.**

Keywords — Robot audition, binaural cues, azimuth estimation, sound localization, sound processing.

I. INTRODUCTION

With the growth of sound processing and intelligent systems technologies, it has become possible to endow robots and machines with artificial sound source localization capabilities. Indeed, many applications require the sound source(s) position(s), like human-robot interaction, surveillance, hearing aids... Sound-based localization has been addressed in multiple aspects: estimating the azimuth, elevation or distance and using hardware ranging from microphone arrays [14], [15] to the biologically-inspired binaural hearing.

Binaural artificial audition uses two microphones placed inside two ears separated by a head. This allows to take benefit of signals similar to those exploited by the impressive human auditive system, imposing light hardware constraints. Robotic binaural audition systems are still far from reproducing human auditive capabilities, but satisfying results can be obtained using binaural inputs for specific tasks, like localization. Most artificial binaural localization systems mainly focus on estimating the sound source-receiver relative azimuth angle. This can be inferred by computing differences between the signals reaching the two ears, mainly in arrival times and levels. These differences are then exploited to provide the azimuth, through geometrical cue-azimuth mapping [9] or learning approaches [8], [16] for example. Sound-based localization systems are also constrained to multiple types of problems, notably limited computational capabilities, interfering noises and room reflections. And whereas interfering sound sources are restrained in time and/or frequency, room reflections constantly affect the perception of signals of interest. While they can be useful for distance estimation [19], they negatively affect azimuth estimation cues. Their drawbacks are easily avoided by the human auditive system, but they are not clearly treated in artificial audition yet. Most systems perform a frame selection where only frames judged to be not much affected are used [3], [4], [6]. In addition, a robot operating in real environments witnesses multiple changes affecting the signals it receives, and thus its performances. In this context, one can mention changing the robot placement inside the room or changing rooms and the source moving or getting closer or further from the robot. Very few studies evaluated their own capabilities to cope with problems possibly caused by these changes. But before putting a system into real-world operation, its generalization capabilities are to be evaluated.

This paper addresses sound source azimuth estimation in the binaural context. In the proposed system, binaural cues are computed on multiple frequency bands, their fluctuations induced by the presence of sound reflections are smoothed and a neural network exploits them to estimate the azimuth. Simulated and real data are used, providing realistic and challenging robot operation scenarios. Evaluations are made in identical and mismatched training and testing conditions, in order to judge about the generalization performances. The paper is organized as follows; the next section details the approach . Section III presents simulated and real databases, as well as the obtained corresponding results. Finally, a conclusion ends the paper, with prospects for future works.

II. LOCALIZATION APPROACH

A binaural sound source azimuth estimation system tyically relies on the consecutive steps shown in Fig. 1. The left and right ears signals pass through models taking mainly the inner ears effects into consideration. Features can then be extracted on the resulting signals, and the azimuth can be obtained through a cue exploitation algorithm. In our system, ITDs and ILDs are extracted on multiple frequency bands and provided as inputs to a Neural Network (NN) that estimates the source azimuth. We also implement a process

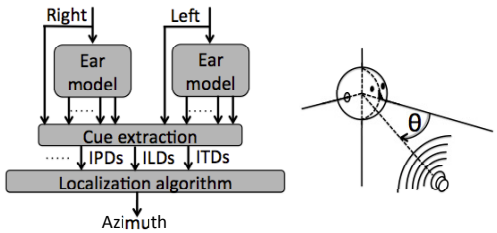

Fig. 1. Left: a typical binaural sound source azimuth estimation system. Right: a sound source located in the horizontal plane of a binaural sensor, with azimuth θ.

K. Youssef, S. Argentieri and J.-L. Zarader are with UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005, Paris, France and CNRS, UMR 7222, ISIR, F-75005, Paris, France. E-mail: lastname@isir.upmc.fr

that reduces the effects of reverberations without reducing the number of used time frames. These steps are detailed in the following subsections. §II-A exhibits the ITD and ILD extraction and selection strategies. §II-B shows the method used to reduce the reverberations bad effects. And §II-C presents the used NN structure.

A. Cue extraction

Most state-of-the-art binaural azimuth estimation approaches use interaural differences. Whereas these approaches seek to accomplish the same task with the same cues, they present substantial differences. From these differences, one can mention time frame duration, sampling frequency, used frequency range and mainly, cue computation techniques. Indeed, original temporal signals, their Fourier transforms or filterbank outputs corresponding to them can be used for this computation. Reviewing these strategies was the object of a previous paper where we presented a systematic study comparing them in terms of the resulting cues appropriateness for azimuth discrimination [18]. It showed that the most effective cues result from a cochlear filtering of the signals, performed with a gammatone filterbank and not followed by any inner hair cells transduction reproduction. Thus, this study uses a gammatone filterbank of $N_{filters} = 30$ filters with frequencies reaching up to half of the sampling frequency denoted by F_s. ITDs and ILDs are extracted on approximately 23ms frames taken with no overlap. The next two paragraphs show their respective computation techniques, and the method used to select the used cues among all the computed ones is detailed in §II-A.3.

1) ITD: the differences in the trajectories crossed by the sound to reach the two ears, and thus in times of arrival to the ears provide information about the source azimuth. This information is exploited across frequency, at the outputs of left and right gammatone filters. In the following, a gammatone filter is indexed by i, $i = 1 \dots N_{filters}$, the left and right i^{th} channel signals being denoted respectively $l^{(i)}[n]$ and $r^{(i)}[n]$, with n the time index. This way, the ITD on channel i is initially equal to $T_s m_1^{(i)}$ where $T_s = \frac{1}{F_s}$,

$$m_1^{(i)} = \arg\max_m C_{lr}^{(i)}[m], \tag{1}$$

and

$$C_{lr}^{(i)}[m] = \sum_{n=0}^{N-m-1} l^{(i)}[n+m] r^{(i)}[n], \tag{2}$$

where N is the total number of samples in a single frame. This operation provides ITDs as discrete multiples of T_s. An exponential interpolation of the cross-correlation function around $m_1^{(i)}$ is used [8], [20]. It provides the fractional part $m_2^{(i)}$ that helps to estimate the real position of the cross-correlation maximum, not restrained to a step of T_s, with

$$m_2^{(i)} = \frac{logC_{lr}^{(i)}[m_1^{(i)}+1] - logC_{lr}^{(i)}[m_1^{(i)}-1]}{4logC_{lr}^{(i)}[m_1^{(i)}] - 2logC_{lr}^{(i)}[m_1^{(i)}+1] - 2logC_{lr}^{(i)}[m_1^{(i)}-1]}. \tag{3}$$

Thus, finally, the i^{th} channel ITD for each time frame, denoted $ITD^{(i)}$ is, given by

$$ITD^{(i)} = T_s(m_1^{(i)} + m_2^{(i)}). \tag{4}$$

2) ILD: the sound absorption and diffraction effects of the head imply frequency and azimuth-dependent energy differences between both ears signals. For each time frame, the level difference is computed on each frequency band i as:

$$ILD^{(i)} = 20 \log_{10} \left(\frac{\sum_{n=0}^{N-1} l^{(i)}[n]^2}{\sum_{n=0}^{N-1} r^{(i)}[n]^2} \right). \tag{5}$$

3) Cue selection: as shown, frequency-dependent ITDs and ILDs are computed on multiple frequency bands. According to Lord Rayleigh's Duplex theory [11], humans use ITDs in the low frequencies and ILDs in the high frequencies to localize sound sources. In our case, for each frame j, the first 15 ITDs and the last 15 ILDs (corresponding to channels with center frequencies of up to and as of 2kHz respectively) are concatenated into a single vector V_j provided as input to the azimuth-estimating NN that will be presented in §II-C.

$$V_j = [ITD^{(1)}, \dots, ITD^{(15)}, ILD^{(16)}, \dots, ILD^{(30)}]. \tag{6}$$

B. Reflections effects reduction

In a closed space, sound reflections reach the receiver from all sides as delayed, attenuated and spectrally modified versions of the emitted signals. A closed space with reflective surfaces can be acoustically characterized by its reverberation time RT60. RT60 is frequency-dependent, and specifies the time taken by the present sound to decrease by 60dB after the source stops emitting. Higher RT60 means harder acoustic conditions for speech intelligibility and speech or speaker recognition and localization for example. The human auditive system has the fascinating ability to neglect the reflections using only the main signal of interest. This is described by a model called the precedence effect [3]. It states that the auditory system can localize/perceive only the main source signal, unless the reflection is powerful or delayed enough not to be neglected. Most of the artificial audition systems try to eliminate frames affected by reflections using energy or coherence criteria [4], [6], [3]. Indeed, they state that signal parts that are not energetic enough, or that do not present enough coherence between the two ears signals are not reliable. This leads to reduce the amounts of used data. We propose a simple approach that keeps all the frames while reducing the reflections bad effects. Reflections cause interaural cues to fluctuate around the values corresponding to the sound source position, increasing their variance [18]. Smoothing the cue values reduces the fluctuations, and equivalent smoothing between the training and testing phases of the system leads to satisfying localization results. For each frame indexed by j, a new vector SV_j is computed as the weighted sum of the vectors belonging to a surrounding ranging between the frames $j - N_{smooth}$ and $j + N_{smooth}$, with $N_{smooth} = 10$. The highest weight is attributed to the current vector V_j and weights decrease linearly as corresponding vectors get further from it:

$$SV_j = \frac{1}{(N_{smooth}+1)^2} \sum_{l=j-N_{smooth}}^{l=j+N_{smooth}} (- \mid l-j \mid + N_{smooth}+1) V_l. \tag{7}$$

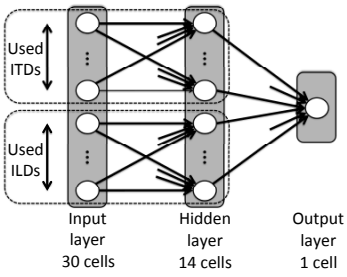

Fig. 2. Used neural network architecture.

Note that the value of N_{smooth} should be set in relation with RT60 as the reverberation effects spread on larger or smaller numbers of frames with higher or lower RT60. But the robot operates in some cases in environments with no *a priori* information about the acoustic conditions, thus not being able to estimate an adequate value for N_{smooth}. 10 is used after a series of evaluations as a value of N_{smooth} that is capable of providing satisfying results with multiple reverberation times.

C. Cue exploitation

Smoothed interaural cue vectors are provided as inputs to a NN trained to output the corresponding source azimuth. The NN architecture is shown in Fig. 2. It has one hidden layer of 14 cells and the connections between the input and hidden layers are not regular (every hidden cell is not connected to all the input cells). Indeed, since the input data contains time and level differences, it is physically more plausible to dedicate specific hidden cells to each of the ITDs and ILDs. Moreover, this partial connectivity architecture is computationally more efficient than the regular one, providing better results with less training time [10]. In the training phase, data is divided into two parts: three quarters for weights optimization and one quarter for cross-validation. The NN receives a matrix regrouping the codevectors of the first part at the input level, and the corresponding azimuths at the output level. It updates its weights iteratively with the back-propagation algorithm for N_{iter} iterations. During this process, cross-validation is applied regularly in order to keep the best validation NN weights. In the testing phase, the NN is provided with input vectors that it exploits to estimate the corresponding azimuths through its optimized weights.

Learning-based approaches to sound localization usually rely on large datasets of recordings. But if the proposed algorithm was only able to estimate the source position after hours of audio recordings and learning, it would not be applicable for robotics applications in dynamic and changing environments. We chose to put the system in very constraining conditions where the numbers of training iterations and examples are limited. Approximately 5000 iterations and 4000 examples are used for the NN training at each of the presented evaluations. This corresponds to approximately 90 seconds gathered from all the source positions after voice activity detection based on a frame energy thresholding.

III. Simulations and Experiments

This section will present the evaluation data and results, on the basis of simulations and real recordings. They provide white noise and speech signals reaching the insides of two ears of a humanoid robot in a reverberant environment. Multiple source-receiver relative azimuths and distances, receiver positions and room conditions are considered as will be detailed in the following. This provides a basis for challenging tests. Simulation data and results are first shown in III-A, and the same is done for the real data in III-B.

A. Simulations

This subsection first presents the database used to evaluate the system in a virtual but realistic environment simulating a shoebox room with a sound source and receiver inside. Second, tests are performed and results are discussed.

1) Database: binaural signals are obtained using Roomsim [2], proven to be an efficient shoebox room acoustics simulator [8], [5]. It runs on Matlab and uses Head-Related Impulse Responses (HRIRs) to compute Binaural Room Impulse Responses (BRIRs) through the Images method [1]. Acoustic parameters (surface types and absorption patterns, humidity, air absorption, temperature, distance attenuation...) are taken into account. This allows to simulate a realistic shoebox room containing the receiver and the source. The simulated room in our case has respectively a length, a width and a height of 5m, 4m and 2.75m. The surfaces are acoustic plaster walls, wooden floor and painted concrete roof and humidity is of 50%. In these conditions, the reverberation time at 1kHz is of 200ms. Additionally, the walls absorption coefficients were scaled appropriately to obtain other datasets with reverberation times of 450ms and 700ms at 1kHz. An anechoic room of the same size is also simulated. Four receiver positions are used at different places of the room, and for each receiver position, the source occupies azimuths between $-45°$ and $45°$ and distances between 1m and 2.8m with steps of $5°$ and 45cm respectively (See Fig. 3). Both ears and source are maintained at the same distance of 1.5m to the ground, keeping a null source-receiver relative elevation.

2) Evaluations and Results: using the pre-described simulated database, the system has been evaluated in multiple aspects. First, its ability to learn and estimate the azimuth in known receiver position and acoustic conditions. Then, its abilities to estimate the azimuth with unknown source distances, receiver positions and acoustic conditions are consecutively evaluated. The term "known" is used here to refer to data provided to the NN during the training. Details and results are shown in the following paragraphs.

a) Identical training and testing conditions: in this paragraph, the system is evaluated by training and testing done with data from the same receiver position, reverberation time and source distance of 1m. The upper part of Fig. 4 shows the estimation errors obtained with all available reverberation times at Position1 (near the room center). Noise and speech signals are used separately. A study with no cue smoothing is also performed, to show the benefits

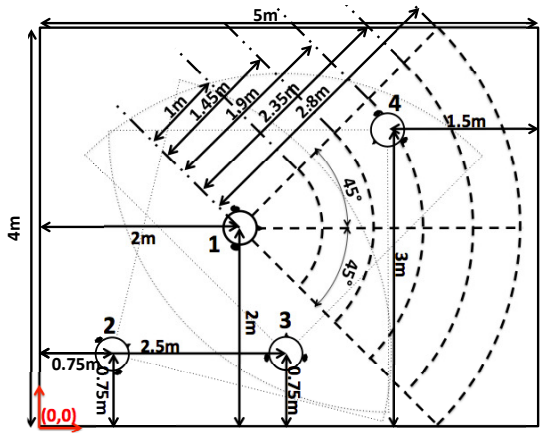

Fig. 3. Simulated echoic room, all available receiver and source positions.

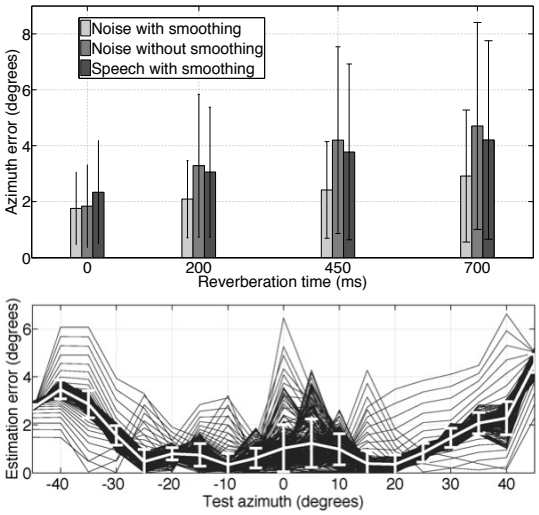

Fig. 4. Up: azimuth estimation errors for speech and white noise, with and without smoothing, in Position1 and distance of 1m and all available reverberation times. Down: estimation error in function of the test azimuth, white noise signals with smoothing, RT60=0ms. Each black curve corresponds to a test covering all the azimuths. The white curve represents the mean and standard deviation of the black plots.

of the smoothing operation. The plotted errors show the high accuracy of the method. Indeed, mean errors are of approximately $2°$ in the anechoic room and remain near $5°$ for the highest reverberation time. White noise signals provide slightly better results than speech, which is expected from their spectral nature and the extracted cues. And a comparison of the white noise results with and without smoothing shows the advantage of this method especially as the reverberation time grows. Results on other receiver positions exhibit fairly similar results.

The lower part of Fig 4 shows the estimation errors of a test using white noise signals, with the receiver at Position1 in the anechoic room. Each curve corresponds to the same test time-frame simulated as emitting from all the azimuths. These results show the variance of performances obtained with 250 test frames, together with the mean error curve. In

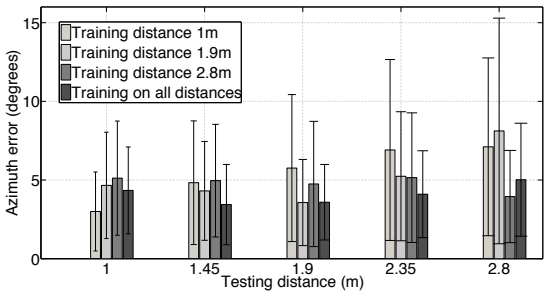

Fig. 5. Azimuth estimation errors with training on a single source distance or all the distances and testing on all the available distances separately. Speech signals, RT60=200ms.

the illustrated case, some azimuths are better estimated than others, especially in the frontal positions. Consequently, this loss of performance on the two sides of the head increases the overall variance of the errors, which is evaluated for the 250 tests on the 19 azimuth angles in all the following.

These results are obtained in controlled situations where only the source azimuth changes, but more realistic tests should take into consideration modifications in the operation conditions. Indeed, the robot is expected to operate inside an environment where, for example, it can move in and between rooms. In the following, speech signals are used, although not providing the best possible performances of the system, but being a realistic case of robot operation.

b) Generalization capabilities, source distance: the source distance has been reported to affect the values of the interaural cues. Some approaches used these cues to estimate the source distance [12], [13]. In the last evaluation, training and testing have been made with data extracted from all the available source azimuths and a fixed distance of 1m. The current evaluation performs the training with data from all the azimuths but on specific distances and the tests on data from all the distances separately. Speech signals are used, the receiver is in Position1 and the reverberation time is of 200ms. Fig. 5 exhibits the corresponding results, they show a performance variability in function of the distance. Errors are the smallest when tests are made on the training distance and they increase when the distance changes. This generalization constraint is avoided with a multi-distance training, providing errors smaller than the mono-distance generalization errors.

c) Generalization capabilities, receiver position: another test of the change in the system's operating conditions can be in the mismatch of robot positions between training and testing, while staying inside the same room. Thus, an evaluation is made is such a way that the training is performed with data extracted from a given receiver position and testing is performed with data from each of the available positions at a time. Two trainings have been first made on Position1 and Position2. These two positions represent two different aspects of the robot placement inside the room: near the center and near a corner with close reflective walls. Then, a training is performed on all the positions. Fig. 6 plots the estimation errors obtained in this study. Speech signals are

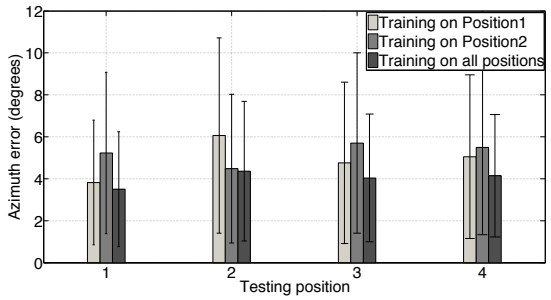

Fig. 6. Azimuth estimation errors with training on a single position or all the positions and testing on all the available positions separately. Speech signals, RT60=200ms.

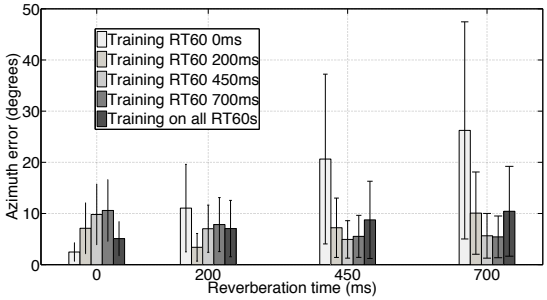

Fig. 7. Azimuth estimation errors with training on a single RT60 or all the RT60's and testing on all the available RT60's separately. Speech signals, receiver Position1.

used, with a reverberation time of 200ms. Note that in this test, training and testing data for each of the taken positions are extracted from both all available azimuths and distances. Extracting data from all the source azimuths but only one distance leads to smaller error variances. It can be seen that the smallest errors are obtained when testing is made on the training position itself. Errors slightly increase with the position change but keep acceptable levels. Notably, testing in Position2 leads to the highest errors when the training is made in Position1. Moreover, the training in Position2 leads to higher generalization errors than the training in Position1. This is due to the fact that Position2 can be considered as the most acoustically constrained position, being the closest to the room reflective walls, which increases the estimation errors. Finally, the multi-position training provides more stable results, with higher errors at Position2, showing once more that the neural network is harder to train on this position than it is on the others.

d) Generalization capabilities, acoustic conditions: the last two paragraphs evaluated generalization in terms of source distance and receiver position with fixed room reverberation time. Another generalization evaluation lies in the room conditions. Thus, training is made on a reverberation time, or on all reverberation times at once, followed by testing on all the reverberation times separately. The receiver is placed in Position1 for both training and testing and speech signals are used. The results are shown in Fig. 7. It is seen that the errors increase when there's a mismatch between the training and testing RT60's. The anechoic training is the hardest to generalize and the multi-RT60 training provides better results with lower RT60's. And as in the previous generalization evaluation, training and testing are made using examples from all azimuths and distances. Performing the training and the testing on the same distance provides smaller error variances.

The results obtained in the last two paragraphs show that the acoustical constraints are more dependent on the room itself than they are on the receiver position. Indeed, it is harder to generalize while changing RT60 than it is while changing the receiver position, maintaining the same RT60. Such a conclusion is also stated in other studies, like in [7].

B. Experiments

Simulation evaluations have been exhibited. It is also important to evaluate the system in real environments. This subsection first presents the established recorded database, and later the obtained corresponding localization results.

1) Database: recordings have been made inside a meetings room of $10 \times 7.5 \times 2.8$m. The walls and the roof are of painted concrete and the floor is resilient. One of the walls has glass windows covered by curtains, thus with different absorption patterns than those of the other walls which increases the environment's acoustic non-symmetry. The receiver has been placed in three positions as shown in Fig. 8. For each receiver position, the source occupies azimuths between $-45°$ and $45°$ with a $5°$ step and distances between 1 and 3m with a 50cm step. The elevation is kept null as both the receiver and the source are placed at the same height of 1.5m. The recorded signal at all the positions is a 3.5min sequence of sentences pronounced by 10 french male speakers. The recordings are made using the Neumann KU100 dummy head, with its human-like outer ears and microphones placed inside them. Signals are transferred to a computer via a NI portable sound card with a sampling frequency of 25600Hz. To verify the correct placement of the source relatively to the receiver, a Codamotion 3-

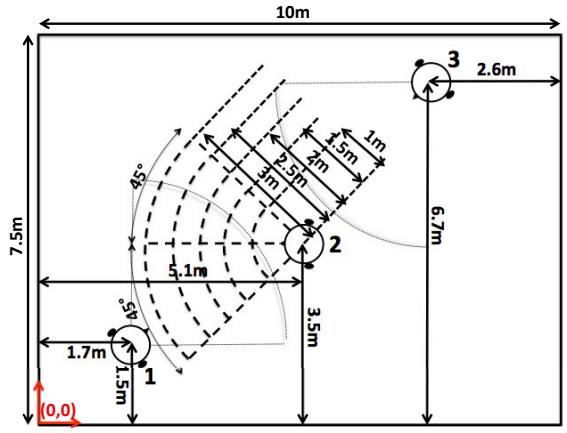

Fig. 8. Recordings meetings room, taken receiver and source positions.

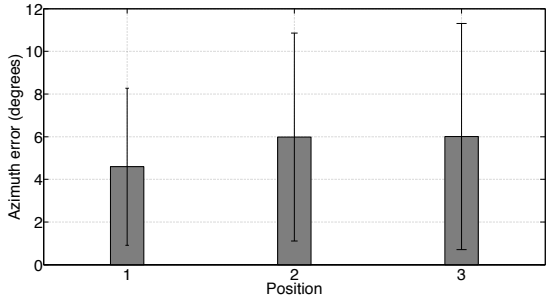

Fig. 9. Azimuth estimation errors in the real database case.

D motion capture system is used, providing positions of multiple infrared markers carried by the receiver and the source in a common landscape. Thus, at each recording, the actual source-receiver position is set in order to match the theoretical conceived position. This allowed to place the sensor at real positions that are very close to the theoretical ones with an azimuth mean error of 0.18° and a distance mean error of 1.46cm.

2) Evaluations and results: with these real recordings, the NN is set in the same way as in the simulation case. The results with training and testing made on the same receiver position with all azimuths and distances, for the three positions separately, are plotted in Fig. 9. Relatively hard conditions are imposed on the system in this case. Indeed, recording noises and the acoustic non-symmetry of the environment are additional constraints that the system training encounters. Smaller errors at Position1 can be justified by the fact that it is close to the most sound absorbing and thus least reflecting wall with curtains. This reduces the reflections constraint in this position, contrarily to the corner Position2 in the simulation case that was the most constrained and showed the higher errors.Note that, in both this case and the simulation case, estimation errors can be further reduced by pushing the NN training to higher numbers of iterations or examples. But as previously said, the goal is to evaluate the system with a fast and light training requiring relatively small times for training data acquisition and exploitation.

IV. Conclusion

This paper presented a robust system using binaural inputs in a humanoid robotic context to estimate the azimuth of a sound source in reverberant environments. The method computes interaural cues on the outputs of gammatone filterbanks and provides them to a neural network that estimates the corresponding azimuth. The system is conceived to be fast and able to adapt to relatively small datasets. It is evaluated in constraining conditions and proved robust. Current works are focused on distance estimation. Both systems will be combined in the future, with a multi-source localization approach based on a visio-auditive learning as we proposed in [17] to provide a final robust robotic binaural sound source localization system.

Acknowledgment

This work was conducted within the French/Japan BINAAHR (BINaural Active Audition for Humanoid Robots) project under Contract n°ANR-09-BLAN-0370-02 funded by the French National Research Agency.

References

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4), 1979.

[2] D. R. Campbell, K. Palomäki, and G. Brown. A matlab simulation of "shoebox" room acoustics for use in research and teaching. *Computer Information Systems*, 9(3), 2005.

[3] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustical Society of America*, 116(5), November 2004.

[4] M. Heckmann, T. Rodemann, F. Joublin, C. Goerick, and B. Schölling. Auditory inspired binaural robut sound source localization in echoic and noisy environments. *International Conference on Intelligent Robots and Systems*, 2006.

[5] B.-g. Lee, J. Choi, D. Kim, and M. Kim. Sound source localization in reverberant environment using visual information. In *IEEE International Conference on Intellligent Robots and Systems*, pages 3542 – 3547, October 2010.

[6] R. Liu and Y. Wang. Azimuthal source localization using interaural cpherence in a robotic dog: Modeling and application. *Robotica, Cambridge University Press*, 28:1013–1020, 2010.

[7] S. J. Loutridis. Quantifying sound-field diffuseness in small rooms using multifractals. *Journal of the Acoustical Society of America*, 125(3), 2009.

[8] T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech and Language Processing*, 19(1), 2011.

[9] M. Raspaud, H. Viste, and G. Evangelista. Binaural source localization by joint estimation of ild and itd. *IEEE Transactions on Audio, Speech and Language Processing*, 18(1), 2010.

[10] A. Rast, S. Welbourne, X. Jin, and S. Furber. Optimal connectivity in hardware-targetted mlp networks. *International Joint Conference on Neural Networks*, 2009.

[11] L. Rayleigh. On our perception of sound direction. *Philosophical magazine*, 13(74):214–232, 1907.

[12] T. Rodemann. A study on distance estimation in binaural sound localization. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[13] P. Smaragdis and P. Boufounos. Position and trajectory learning for microphone arrays. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1), 2007.

[14] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, and T. Takano. Real-time 2 dimensional sound source localization by 128-channel huge microphone array. *IEEE International Workshop on Robot and Human Interactive Communication*, 2004.

[15] J.-M. Valin, F. Michaud, and J. Rouat. Robust 3d localization and tracking of sound sources using beamforming and particle filtering. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2006.

[16] J. Woodruff and D. Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[17] K. Youssef, S. Argentieri, and J.-L. Zarader. A binaural sound source localization method using auditive cues and vision. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2012.

[18] K. Youssef, S. Argentieri, and J.-L. Zarader. Towards a systematic study of binaural cues. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[19] P. Zahorik. Assessing auditory distance perception using virtual acoustics. *Journal of the Acoustical Society of America*, 111(4), April 2002.

[20] L. Zhang and X. Wu. On cross correlation based discrete time delay estimation. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2005.