From Monaural to Binaural Speaker Recognition for Humanoid Robots

Karim Youssef, Sylvain Argentieri and Jean-Luc Zarader

Université Pierre et Marie Curie
Institut des Systèmes Intelligents et de Robotique, CNRS UMR 7222
4 place Jussieu, 75005, Paris, France

***Abstract*—This paper addresses speaker recognition in a binaural context. Such an auditory sensor is naturally well suited to humanoid robotics as it only requires two microphones embedded in artificial ears. But the state of the art shows that, contrary to monaural and multi-microphone approaches, binaural systems are not so much studied in the specific task of automatic speaker recognition. Indeed, these sensors are mostly used for speech recognition, or speaker localization. This study will then focus on the benefits of the binaural context in comparison with monaural techniques. The proposed approach is first evaluated in simulation through a HRTF database reproducing the head shadowing effect and with a 10-speakers database. Next, the method is assessed with an experimental binaural 15-speakers database recorded in our own almost-anechoic room for various SNR conditions. Results show that the speaker positions during the learning step of the proposed approach strongly influence the recognition rates.**

***Index Terms*—Speech processing, speaker identification, binaural hearing, humanoid robot, GMM, MFCC.**

I. Introduction

Thanks to the growing interest in robotics during the last decade, many kinds of robots have been designed and developed for interaction among humans. In this topic, humanoid robots are probably the most appropriate, and a lot of works focus now on trying to make them sense and look like Humans. For that purpose, auditory perception is a must-have capability. Indeed, it is a very important sense for humans and other living creatures, helping them to communicate in their surrounding environment. So, giving robots such capabilities is clearly of interest, thus making us able to use our best means of communication and interaction: our voice.

Robot Audition is a growing field of research, with an increasing Community interested in trying to reproduce the amazing auditive human capabilities. This includes sound source localization, but also sound extraction, sound/speaker recognition, speech recognition, etc. Each of these topics has been already deeply dealt with, but not necessarily in a robotic context, which imposes specific and original constraints (embeddability, real-time, etc.). Numerous recent works in the Robotics Community have integrated these limitations and proposed very interesting solutions, but mainly for localization [1] and/or speech recognition purposes [2], [3]. So this paper mainly focuses on Automatic Speaker Recognition (ASkR), for humanoid robots equipped with two ears. Surprisingly, such a binaural framework has not been so much studied in the specific task of automatic speaker recognition.

Speaker identification has already been widely studied in the single microphone case. A variety of operations can be performed, and very good results can be achieved in adequate environments. For instance, [4] proposes a method using the Mel Frequency Cepstral Coefficients (MFCCs) together with Support Vector Machine (SVM) classifiers to perform the recognition. In the same vein, [5] and [6] exploit spectral subtraction in order to reduce noise influence. Nevertheless, these approaches are not so robust against high noise level or reverberations, and present a loss of performance when compared to systems working with more than one microphone. But two different approaches to the identification problem can be exhibited in this multiple signals case. On the one hand, a lot of works deal with an appropriate combination of multiple signals into a single one being generally less corrupted by noise. Classical monaural methods can then be exploited to perform the recognition. One can cite beamforming approaches exploiting the microphone array directivity [5], [6], or adaptive noise cancellation [7]. Identically, matched filter arrays are used in [8] where a parameterization analysis of an ASkR system is presented. On the other hand, other works propose to extract features from each available signal before the recognition algorithm. In this vein, [9] proposes to combine multiple GMMs classification results on the basis of a 8 microphones array. In the binaural context, [10] developed a feature vector combination method optimizing a mixture weight value.

This paper is more concerned by this second approach, envisioned in a binaural context. But existing binaural studies specifically focused on noise reduction and simulation of the human auditory system for speech recognition and localization, and not so much on speaker identification. For instance, [11] developed a binaural model for speech recognition, simulating the functioning of the cochlea. The design of an artificial ear is presented in [12], by taking into account the spectral changes induced by the pinna and the concha in the speech signal. The resulting system is then exploited for localization. The binaural case has also

been used in [13] to reduce noise and reverberations effects through blind source separation. One can also cite [14], where adaptive noise reduction permits voice activity detection through neural networks, but also speech localization and recognition with a binaural sensor. Similarly, noise estimation techniques applied to one of the two available signals allow the cancellation through adaptive filtering of the noise in the second signal [5], [6], [15].

The paper is organized as follows. The proposed monaural and binaural speaker recognition systems are described in section II. They are next both compared in simulation in Section III. The influence of the noise and of the speaker position is carefully addressed. Then, an experimental evaluation of the approach is presented in section IV. For that purpose, a 15-speakers database has been recorded in an almost-anechoic room with a binaural dummy head. Finally, a conclusion ends the paper.

II. Monaural and Binaural Recognition Systems

The proposed ASkR system is presented in this section. It is text-independent, and mainly relies for the moment on MFCC features combined with GMM classification, both being evaluated in a one channel (monaural) or two channels (binaural) configuration. The later is addressed as a bioinspired system, simulating the auditory human perception. Consequently, such a binaural system is naturally well suited to humanoid robotics. For each case, the influence of noise, speech duration and location will then be investigated in §III and §IV.

The overall evaluation of the approach is based on two successive studies. First, simulations are used to assess the performance of the approach. It relies on a high quality audio database, acquired from long French monologues in identical and good conditions. Second, experimental measurements are exploited to stress the method with real binaural signals acquired from a dummy head in an acoustically-prepared room. In these two cases, the following monaural and binaural ASkR systems are exploited.

A. Monaural speaker identification system

The proposed monaural system is based on the following successive computation steps, see Figure 1. The major steps of this conditioning are described hereafter.

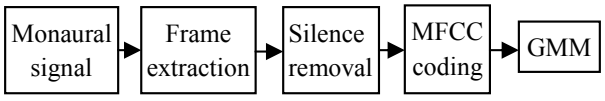

Fig. 1. Major steps of the monaural system.

1) Frame extraction: First of all, 512 successive points, corresponding to about 22ms-length frames, are extracted from the signal. The energy E_i of each i^{th} frame is computed and compared with a threshold T to eliminate non-speech portions, T being defined as

$$T = E_{\text{min}} + K(E_{\text{max}} - E_{\text{min}}), \quad (1)$$

where $E_{\text{min}} = \min_i(E_i)$, $E_{\text{max}} = \max_i E_i$ and K is the threshold parameterization in percent. In all the following, K is set to 1%, resulting in the classification of about 65% of all the frames as being speech. Next, pre-accentuation filters and Hamming windows are exploited to obtain useful speech frames. Finally, 16 MFCC and 16 Δ-MFCC coefficients are extracted from these frames, with an overlapping factor set to 0.5. These features are then used to train and test the recognition algorithm.

2) MFCC coding: MFCCs are commonly used as features in speech and speaker recognition systems. They can be interpreted as a representation of the short-term power density of a sound. These coefficients are commonly derived as follows:

- Compute the Fourier Transform (FFT) $X[k]$ of the considered time frame.
- Apply to $|X[k]|^2$ a set of $N = 24$ triangular filters regularly spaced on the mel scale defined by

$$\text{mel}(f) = 2595 \log_{10}\Big(1 + \frac{f}{700}\Big), f \in [0, f_s/2], \quad (2)$$

- Compute the N output energies $S[n]$ of each filter.
- Compute the k^{th} MFCC coefficient MFCC_k value with

$$\text{MFCC}_k = \sum_{n=1}^{N} \log_{10}(S[n]) \cos\Big(\frac{k\pi(2n-1)}{N}\Big). \quad (3)$$

The objective of the mel-scale introduced in the MFCC computation is to approximate the human auditory system response more closely than the classical linearly-spaced frequency bands. More precisely, the mel scale is shown to be a perceptual scale of pitches judged by listeners to be equal in distance from one to another. As a consequence of this decomposition, the representation of the speech signal information is close to the human perception of sounds, providing high resolution for the low frequencies and a weaker resolution for high frequencies.

Additionally, 16 Δ-MFCC coefficients are also computed. They represent the variations of the original MFCC features as a function of time and are simply obtained from a 9^{th}-order FIR filter applied on the MFCC vectors along time.

3) GMM: In statistics, a mixture model (MM) is a probabilistic model for density estimation using a mixture distribution. In the Gaussian case, a Gaussian MM (GMM) is a simple linear superposition of Gaussian components, which aims at providing a richer class of density models than a single Gaussian [16]. For a model of M Gaussian states, a GMM density function of a variable x_n can be defined as

$$p(x_n|\lambda) = \sum_{i=1}^{M} p_i b_i(x_n), \quad (4)$$

where p_i is the probability of being in the state i and b_i the Gaussian density function of mean μ_i and covariance Σ_i. λ writes as

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, i = \{1, \ldots, M\}, \quad (5)$$

and represents the set of weights p_i, mean vectors μ_i and covariance matrices Σ_i of the GMM.

In a speaker identification task, an M state GMM is associated with each of the S speakers to be discriminated. On this basis, the aim is to determine which model number $\hat{S}$ has the biggest *a posteriori* probability over a set $X = \{x_1, x_2, \ldots, x_N\}$ of measured MFCC and ΔMFCC features , that is, according to Bayes rules,

$$\hat{S} = \text{Arg} \max_{1 \leq k \leq S} p(\lambda_k|X) = \text{Arg} \max_{1 \leq k \leq S} \frac{p(X|\lambda_k)p(\lambda_k)}{p(X)}. \quad (6)$$

In this case, $\lambda_k = \{p_i^{(k)}, \mu_i^{(k)}, \Sigma_i^{(k)}\}, i = \{1, \ldots, M\}$, represents the mixture parameterization of the M-state GMM associated to the k^{th} speaker. Assuming that the *a priori* probability $p(\lambda_k)$ is the same for all speakers, and for one set of measured data X, equation (6) can then be simplified as

$$\hat{S} = \text{Arg} \max_{1 \leq k \leq S} p(X|\lambda_k). \quad (7)$$

All the problem now is to determine the $3 \times M$ parameters included in λ_k describing the GMM related to the k^{th} speaker. This is achieved through the classical iterative Expectation - Maximization (EM) algorithm [17]. Such a method exhibits a fast convergence of the parameters and is based on two successive steps: expectation (E) and maximization (M). These two steps are iterated until convergence of the set λ_k; the convergence of the algorithm is evaluated through the log-likelihood $\log(p_l(X|\lambda_k))$, with l denoting the l^{th} iteration of the algorithm. The learning is initialized with a first clustering of the data obtained with a K-means algorithm. Note that during this learning step, no interaction occurs between the GMMs of different speakers.

Once the $3 \times M \times S$ GMM parameters of the S speakers are known, these Gaussian models are exploited to perform the recognition as follows. As soon as a set of new features X is available, the predicted speaker is selected as being the speaker having the GMM with the highest *a posteriori* probability $p(\lambda_k|X)$, see Equation (7).

B. Binaural speaker identification system

The overall functioning of the monaural system has just been described. In the binaural context, the proposed method only differs from the previous one in the frame and feature extraction steps. Indeed, there are now two signals corresponding to the left and right perceived auditory signals.

1) Frame extraction: The same strategy in the monaural case, relying on 512-points frames, is exploited. The speech detection is still based on the simple energy criterion (1), but this process must be coherently performed between the left and right signals. Indeed, some frames in one channel can be classified as being speech, while being categorized as silence in the other one. This fact is a direct consequence of the shadowing induced by the head, represented in Figure 2 by the two HRTF blocks. As a solution, the left and right signals are normalized so that they have the same energy. Each of them is then respectively compared with a threshold T_{left} and T_{right} computed with (1). Finally, only the frames being classified as speech in the left and right signals simultaneously are gathered and exploited in the following. This results in the classification of about 50% of all the frames as being speech.

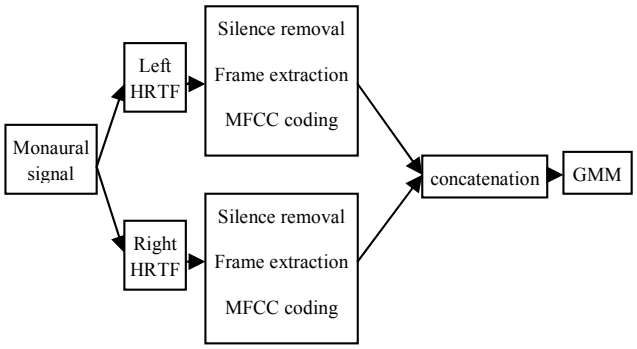

Fig. 2. Major steps of the proposed binaural system.

2) MFCC coding: Concerning the features extracted from the previously collected frames, the question is now: how to combine the available auditory features? In this paper, we only focus on a simple concatenation of the two feature vectors originating from the left and right signals, see Figure 2. Other strategies are currently in investigation and will be presented in future works.

III. Evaluation of the method in simulation

In this section, monaural and binaural speaker recognitions are compared in simulation. First, the simulation setup is presented. Next, classical monaural recognition rates are obtained in the second subsection. These results are then exploited to show the benefits of the binaural case in a third subsection. The effectiveness of the recognition with respect to noise level and speaker position is also tested.

A. Simulation setup

As was mentioned in §II, the used speaker database comes from long radiophonic French monologues recorded in identical and good conditions. It is made of $S = 10$ speakers, with 28 tracks per speaker, each track lasting 15 seconds. So, 7 minutes per speaker are available, for a total of 70 minutes-length audio signals. The original sampling rate is $f_s = 44100$Hz, but all tracks have been downsampled to $f_s = 22050$Hz, and so treated by a Chebychev anti-aliasing filter.

Then, the binaural speech signals are simulated by convolving the monaural speaker database signals with impulse responses coming from a HRTF database. The Head Related Transfer Function (HRTF) describes how a sound signal is altered by the acoustical properties of diffraction and/or reflection of our head, outer ear and torso, before reaching the transduction stages of the inner ear. This effect is traditionally modeled as a filter whose impulse response is a function of the sound source's position with respect to the head. In this paper, the KEMAR dummy-head HRTF is used, being made freely available

by the CIPIC Interfaces Laboratory of the University of California [18]. This HRTF Database is public, and made of high spatial resolution HRTF measurements for 45 different subjects. The database includes 1250 HRTF-identifications for each subject, recorded at 25 interaural-polar azimuths and 50 interaural-polar elevations (see [18] for more detailed information). Finally, the speech signals and HRTF database have been acquired with a sampling frequency $fs = 44100$Hz, and then downsampled to $f_s = 22050$Hz as in the monaural case.

Finally, the speaker database is divided into two distinct parts. The first one, representing about 66% of the entire database, is employed for the learning of the GMMs (see §II-A3). The remaining database part (33%) is devoted to the evaluation of the recognition capabilities of the proposed system. We recall that the threshold parameterization K is set to 1%, and the number M of GMM states is $M = 16$[1]. For such a value, 40 iterations are sufficient for the convergence of the GMM parameters, like in [17].

B. Monaural case

In this subsection, the influence of the Signal to Noise Ratio (SNR) and of the duration testing sets is assessed.

1) Influence of noise: In order to test the robustness of the monaural approach to noise, a white Gaussian noise is added to the speech signal to produce various SNR conditions. Next, the silence removal process is applied on the resulting noisy signal. The recognition is then performed on the basis of the extracted features, and the recognition ratio is obtained by dividing the number of well recognized frames by the total frame number of the considered testing set. The recognition results are reported in Table I (Monaural column). Logically, the recognition performance increases when the signal to noise ratio also raises.

2) Influence of the testing duration: The previous study has been performed on the basis of about 22ms-length frames. But considering real-life applications, recognition rates for longer durations are clearly more realistic and meaningful. Interestingly, this might also produce higher performance, as the recognition can now be consolidated along time. This integration is achieved by a majority vote algorithm performed over consecutive frames. In the following, the interpretation of the results will especially focus on the recognition rate on the frames, but also on longer signals lasting 1, 3 and 5 seconds. The recognition rates obtained for the 1s-long signals are of particular interest when trying to recognize the speaker on the basis of only one pronounced word. In the same way, 3s-long signals may provide a more efficient speaker recognition of an entire phrase. These two specific scenarios respectively correspond to 2 different interaction conditions: on the one hand, the recognition capabilities of the robot must be good enough to guarantee its reactivity in emergency situations where short words are likely to be used. On the other hand, longer speech signals relate to more classical situations during the interaction. The obtained recognition ratios are reported in Table I (Monaural column). As expected, the recognition rates increase for longer durations, and reach up to almost 100% for a 3s-long signal for high SNR values. This table will now serve as a reference for comparison with binaural methods.

[1]While it is not presented here, various M values have been tested, resulting in this optimal choice between good speaker modeling and computing cost.

TABLE I
BEST MONAURAL VS. BINAURAL RECOGNITION RATES, FOR VARIOUS INTEGRATION TIMES AND SNR CONDITIONS.

Frame length	SNR	Monaural	Binaural
23 ms	-3	19.4	29.28
	0	24.4	34.5
	10	39.9	51.6
1s	-3	55.8	73.9
	0	65.2	85.6
	10	94.9	98.9
3s	-3	76.4	85
	0	80.2	92.7
	10	98	100

C. Binaural case

We propose in this part to evaluate the performance of the proposed method in simulation on the basis of the previously described binaural system (see §II-B). Because of the use of binaural signals together with a learning algorithm, the position of the simulated speaker will be of fundamental concern. Actually, the questions are: "will the system learn the speaker position instead of the speaker himself"? And in the case of a good speaker recognition, "can the sensitivity of the approach to the position be evaluated?" This inherent position dependence is carefully addressed in the following paragraphs. In all the following, -3, 0 and 10 dB SNR values are considered. Sources positions are given in the form (θ, ϕ), with θ being the azimuth measured in the horizontal plane, and ϕ the elevation in the vertical plane. $\theta = 0°$ and $\phi = 0°$ both corresponds to a sound source in front of the head.

1) One direction for all speakers: In this first scenario, the 10 speakers are all regrouped as emitting from the same spatial direction. A first evaluation consists then in learning the GMMs parameters and testing them while this position remains the same. The resulting recognition rates are reported in Figure 3 (left), and are quite similar to the previous monaural case. Indeed, as the speakers position remains the same during the learning and evaluation steps, no effect of the position can be brought to the fore. But if the 10 speakers orientation is changed between the learning and test phases, one can show that the best performances are obtained only in the training direction, see Figure 3 (right) for SNR $= 10$ dB. Such a phenomenon remains valid for other SNR values. This clearly shows that GMMs model both the speaker and the direction.

2) Same direction for a group of speakers: In order to capture how the position influences the algorithm's performances, a second scenario has been tested. It consists in forming 3 speakers groups respectively emitting from the 3

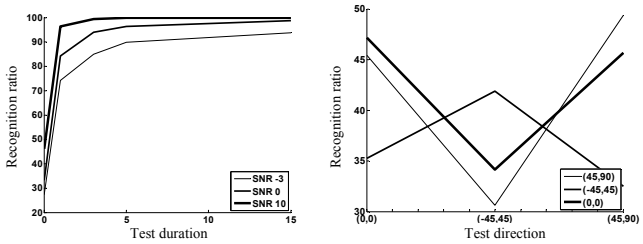

Fig. 3. Study for the same direction for all the speakers. (Left) Mean binaural recognition ratio with GMMs trained and tested in the same direction. (Right) Binaural frame recognition ratio as a function of the test direction, for SNR = 10dB.

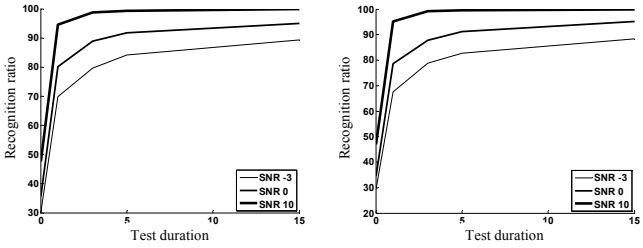

Fig. 5. Study for multiple learning directions. (Left) Binaural recognition ratio with GMMs trained and tested in the same multiple directions. (Right) Binaural recognition ratio with testing on 10 unlearned directions for all speakers.

angular positions $(\theta, \phi) = \{(0°, 0°); (0°, 45°); (0°, -45°)\}$ during the learning step. Maintaining these same positions during the evaluations leads to the recognition rates reported in Figure 4 (left). While the method shows good

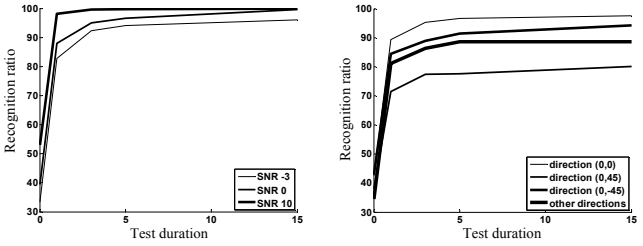

Fig. 4. Study for a group of speakers. (Left) Binaural recognition ratio with GMMs trained and tested in the same direction. (Right) Binaural recognition ratio with GMMs tested when all the speakers are simulated from the direction of training of one group.

performances, it also demonstrates the effectiveness of the binaural recognition to speaker situation. Indeed, one can see that better rates are obtained in Figure 4 (left) than in Figure 3 (left): this can be explained by the lower number of speakers per direction, thus reducing the intra-group confusion.

The second experiment consists in regrouping all the 10 speakers into the same position during the testing phase. Note that this position is chosen as being one of the 3 previously mentioned or a new one. In this case, the best performances are obtained in the position $(0°, 0°)$, see Figure 4 (right). In fact, this specific position is *central*, being the closest place to the other learned positions. In that sense, it represents the orientation minimizing the position influence, and thus also minimizing the speaker confusion.

3) Multiple directions for each speaker: In order to minimize the position influence, the GMM's learning is performed with 10 different directions per talker, covering a large part of the surrounding space of the binaural head. The resulting recognition ratios are shown in Figure 5 (left and right). As before, left Figure is obtained when considering the same positions during the learning and testing steps. It appears that the algorithm's performances are more sensitive to the SNR value, and this effect is clearly more obvious in this last scenario. The same holds when considering the recognition performed from unknown positions, see Figure 5 (right). But it now appears that the system is robust to changes in speaker positions, which is a fundamental property for real life applications. This seems to indicate that the learning has to be conducted from a lot of potential positions in order to achieve acceptable performances. This is a major issue intrinsically linked to the binaural nature of the exploited sensor. From an experimental point of view, it will make necessary to perform the learning step on a sufficient position set to obtain valuable and more realistic performances. This intuitive fact, actually demonstrated here in simulation, will now be assessed with real binaural signals in the following section.

IV. Experimental results

In this section, real binaural signals coming from a dummy head are exploited within the preceding binaural framework. The experimental setup and the binaural speaker database creation is outlined in the first subsection. The resulting two signals are then used to perform the speaker recognition. The experimental recognition rates and the sensitivity of the approach to directions and noises is then investigated in the second subsection.

A. Experimental setup

In order to assess the proposed approach with real signals, a binaural speaker database has been recorded. To our knowledge, such a database does not exist in the literature, and so we plan to make it public in a close future for ongoing works in the field. For now, it is made of only 15 different speakers, each of them being recorded during 50 minutes from 7 distinct positions. The people are asked to utter with their classical way of speaking, while reading a newspaper or freely talking.

The experiment takes place in an acoustically prepared room, equipped with 3D-pyramid pattern studio foams placed on the roof and on the walls (see the two pictures in Figure 6). A binaural KU100 dummy head from Neumann, equipped with two high-quality balanced microphones embedded inside two ears imitating the human pinnae, provides the binaural signals. An additional wireless microphone, attached to each speaker, provides a third clean speech signal. Importantly, this signal is not a function of the position and can be used to perform monaural recognition if necessary. These three signals are then simultaneously sampled and acquired with a National

Instruments PCI acquisition card through 24 bits delta-sigma converters and with a sampling frequency f_s set to 48kHz. All speakers are recorded from a constant distance

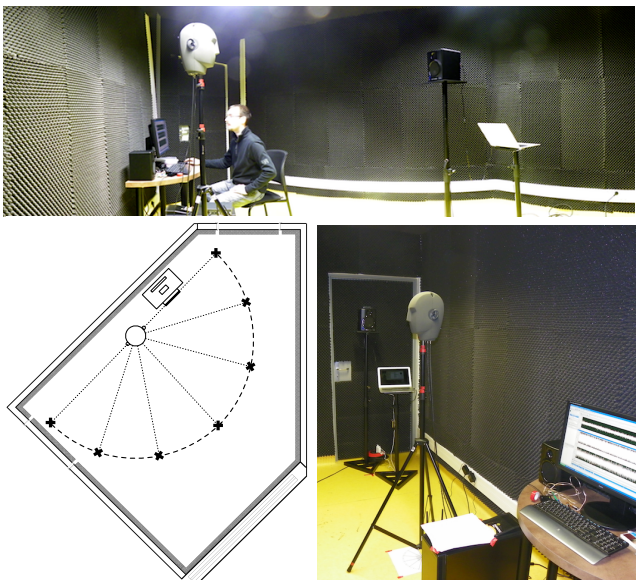

Fig. 6. Experimental setup. (Top) Overview of the acoustically-prepared room during the white noise recording. (Left) Representation of the room and of the 7 positions from where the speakers were recorded. (Right) Focus on the binaural dummy head and on the acquisition computer.

$d = 1.7$m to the head center and from the 7 azimuth angles $\theta = \{-90, -60, -30, 0, 30, 60, 90\}$ degrees, $\theta = 0°$ being in front of the head. The elevation is specific to each speaker, and so entirely determined by their size and the stand height supporting the dummy head. In all the following, the elevation value –ranging from $-1°$ to $4°$– will not be taken into account as only azimuth θ will have significant influence on the recognition task in this experiment.

In the same way, a binaural noise database is recorded by emitting a white noise through a loudspeaker for each of the 7 aforementioned positions. This database will then be exploited to test the robustness of the approach to directional noises.

B. Recognition results

1) One direction for all speakers: In this first scenario, the 15 speakers of the database are all grouped as emitting from the same azimuth. As in § III-C1, the GMM's learning and testing steps are then performed from the same positions, resulting in the recognition rates reported in Figure 7 (left). Interestingly, simulation and experimentation exhibit similar results, showing in this specific case very good recognition ratios. But if the recognition is now performed from a distinct position during the testing phase, then the recognition performances drastically fall, see Figure 7 (right). So, from an experimental point of view, it is obvious that the recognition is very sensitive to the learning position. As already stated during the simulation subsection §III-C, the GMM's learning step has to be performed from multiple positions.

2) Multiple directions for each speaker: In order to minimize the position influence, the learning step is now performed by presenting each speaker as uttering from the 7 database azimuths. The resulting recognition rates are reported in Figure 8 (left), with GMMs being trained and tested in the same multiple azimuths. As already explained in previous simulations (see III-C3), the raw recognition ratios are now smaller than in Figure 7, while the method is less sensitive to the speaker position.

The proposed database also includes one specific 1 min-length record per talker, during which the people were asked to continuously move around the head. Performing the recognition task with these distinct data leads to the results reported in Figure 8 (right). One can see that the recognition performances decrease, but still reach up about 70% for a 10dB Signal-to-Noise Ratio, and for an integration time set to 1s. This loss of performance can be partially explained by the small duration of the signal related to each moving speaker, but also by the footstep noise generated during the walk, which jams the silence removal algorithm. The first problem will be assessed in the future by recording a longer moving sequence. The second one brings to the fore the need of a more sophisticated voice activity detection (VAD) technique, relying on the two perceived signals, that is, a *binaural* VAD.

3) Directional noise influence: A final evaluation has been performed in relation with the noise and the position sensitivity of the approach. It consists in blurring the 2 binaural signals by a directional white noise which has been recorded during the database creation. In this scenario, a noise is continuously emitted from the azimuth $30°$ while the speakers are uttering. Note that the noise superimposition is made offline, by simply adding the recorded noise to the left and right speaker signals. This noise level is then adjusted in order to simulate various SNR conditions with respect to the left ear only, in order to preserve the inherent interaural level difference between the left and right signals. The resulting recognition ratios are reported in Table II, with GMMs being learned and tested in the same multiple azimuths. Comparing these results with those in Figure 8 (left) exhibits very interesting outcomes. Indeed, it appears that the approach shows a smaller sensitivity to noise, especially when working on

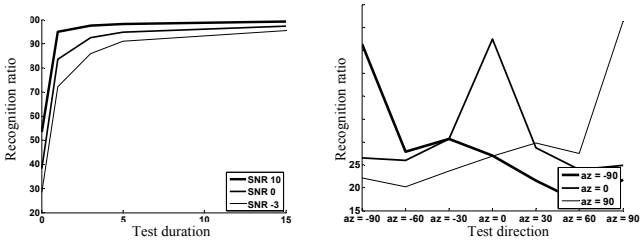

Fig. 7. Experimental study for the same direction for all the speakers. (Left) Mean binaural recognition ratio with GMMs trained and tested in the same direction. (Right) Mean frame binaural recognition ratio as a function of the test direction, for SNR = 10dB and 3 different learning directions

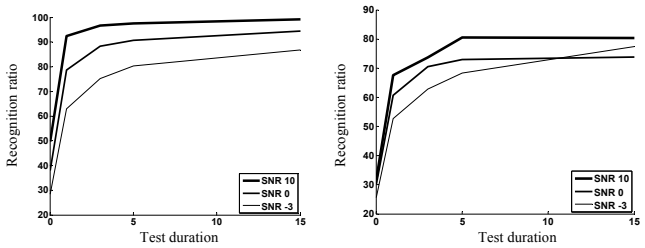

Fig. 8. Experimental study for multiple learning directions. (Left) Mean binaural recognition ratio with GMMs trained and tested in the same multiple directions. (Right) Mean binaural recognition ratio with testing performed on moving speakers.

TABLE II
BINAURAL RECOGNITION RATES, FOR VARIOUS INTEGRATION TIMES AND SNR CONDITIONS IN THE PRESENCE OF A DIRECTIONAL WHITE NOISE.

SNR/Length	23ms	1s	3s	5s
-3 dB	48.9	84.3	93.4	95.3
0 dB	50	92.4	98	98.7
10 dB	56.5	94.2	98.4	99.2

frame lengths, together with better recognition ratios. It might indicate that all the previous studies in §III and §IV are quite pessimistic, as they all consider independent white noise between the left and right signals. In fact, our first acquisitions in a classical acoustic environment indicates that the left and right noises can be highly correlated, the noise origin being generally well localized (air-conditioning systems, open windows, etc.) More precisely, the additive noise in real environment can be seen as a mixture of high-level directional noises (generally originating from known interfering sound sources) and low-level independent noises (like measuring noises).

V. CONCLUSION

A binaural speaker recognition system has been presented in this paper. It relies on MFCC features and GMM to perform the identification in noisy conditions. It has been shown, in simulation and in experimental conditions, that the speaker positions during the testing step affect the recognition depending on their gap with the training directions. More generally, it appears that better performances are produced when increasing the number of learning directions. We also showed the advantage of the binaural hearing and its benefits, being in a world where the humanoid robots become a need and a highly performing machine. Future works will have other theoretical and practical aspects. First, we will focus on the features themselves, and on their combination. Indeed, MFCCs are very classically used in ASkR, but other features might induce a smaller dependence on the speaker positions. For instance, spectral methods based on the correlation of the two signals are good candidates. Next, the combination of the left and right features is also of particular interest. A simple concatenation, while still providing better recognition ratios than in the monaural case, is a very naive approach which might be bettered through adaptive approaches. Finally, the proposed binaural database will include in a close future a larger set of speakers recorded from multiple directions and for various scenarios, in controlled as well as in daily environment. This database will then be accessible for other works in the field of Robot Audition.

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. Bonnal, S. Argentieri, P. Danès, and J. Manhès, "Speaker localization and speech extraction with the ear sensor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[2] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego noise suppression of a robot using template subtraction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[3] J. Even, H. Sawada, H. Saruwatari, K. Shikano, and T. Takatani, "Semi-blind suppression of internal noise for hands-free robot spoken dialog system," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[4] S. S. Kajarekar, "Four weightings and a fusion / a cepstral-svm system for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.

[5] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *ISCLP proceedings, fourth International Conference on Spoken Language*, 1996.

[6] ——, "Providing single and multi-channel acoustical robustness to speaker identification systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997.

[7] J. Ortega-Garcia, J. Gonzalez-Rodriguez, C. Martin, and L. Hernandez, "Increasing robustness in gmm speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proceedings of ICSLP*, 1996.

[8] Q. Lin, E.-E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," in *IEEE Transactions on Speech and Audio Processing*, vol. 2, 1994.

[9] M. Ji, S. Kim, H. Kim, K. Kwak, and Y. Cho, "Reliable speaker identification using multiple microphones in ubiquitous robot companion environment," in *16th IEEE International Conference on Robot & Human Interactive Communication, Jeju, Korea*, 2007.

[10] Y. Obuchi, "Mixture weight optimization for dual-microphone mfcc combination," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.

[11] T. Usagawa, M. Bodden, and K. Rateitscheck, "A binaural model as a front-end for isolated word recognition," in *Fourth International Conference on Spoken Language, ICSLP Proceedings*, 1996.

[12] S. Hwang, K.-H. Shin, and Y. Park, "Artificial ear for robots," in *IEEE Sensors*, 2006.

[13] F. Keyrouz, W. Maier, and K. Diepold, "A novel humanoid binaural 3d sound localization and separation algorithm," in *IEEE-RAS International Conference on Humanoid Robots*, 2006.

[14] R. Brueckmann, A. Scheidig, and H.-M. Gross, "Adaptive noise reduction and voice activity detection for improved verbal human-robot interaction using binaural data," in *IEEE International Conference on Robotics and Automation*, 2007.

[15] P. Brayer and S. Sridhatan, "Robust speaker identification using multi-microphone systems," in *Speech and Image Technologies for Computing and Telecommunications IEEE Region 10th Annual Conference*, 1997.

[16] C. M. Bishop, "Mixtures of gaussians," in *Pattern Recognition and Machine Learning*, 2006.

[17] K. Kroschel and D. Bechler, "Demonstrator for automatic text-intependent speaker identification," in *Revue Fortschritte der Akustik*, 2006.

[18] V. Algazi, R. Duda, R. Morrisson, and D. Thompson, "The cipic hrtf database," *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to audio and Acoustics*, pp. pp. 99–102, 2001.