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*—In this paper, an original study of a binaural speaker identification system is presented. The state of the art shows that, contrarily to monaural and multi-microphone approaches, binaural systems are not so much studied in the specific task of automatic speaker recognition. Indeed, these systems are mostly used for speech recognition, or speaker localization. This study will focus on the benefits of the binaural context in comparison with monaural techniques. It demonstrates the interest of the binaural systems typically used in humanoid robotics. The system is first tested with monaural signals, and then with a binaural sensor, in many signal to noise ratios, speech durations and speaker directions. Up to 11 percent of improvement in recognition ratios of 23 ms frames can be obtained. The used database is a set of audio tracks recorded for 10 speakers, and filtered by HRTFs to obtain binaural signals in the directions of interest, for the binaural training and testing steps. This way, we study the sensitivity of the system to the speaker's location in an environment where a maximum of 10 speakers is present.**

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

I. Introduction

The auditory perception is a very important sense for humans and other living creatures, helping them to communicate in their surrounding environment. Indeed, humans can understand speech and recognize speakers and other sound sources. So, giving robots such capabilities is clearly of interest, thus making us able to use our best means of communication: our voice. Robot audition is a growing field of research, and a lot of recent works have tried to reproduce the amazing auditive human capabilities, including sound localization, noise filtering, sound extraction and recognition, etc. This paper focuses on Automatic Speaker Recognition (ASkR), for humanoid robots equipped with two ears. More precisely, ASkR is the process of knowing who is speaking to a machine among a number of persons, based on their vocal characteristics. This identification can be done with a closed set or an open set of persons (identifying a known or an unknown speaker, an impostor), and can be text-dependent or independent. The first studies in this field took place fifty years ago and their progress continues until nowadays [1]. ASkR interest is actually growing thanks to the numerous various fields of applications it covers. For instance, it can be used for audio surveillance, with aged and sick persons at home. It still faces the effects of noise and reverberations, and the mismatch between the learning and testing phases of the classifiers. Other problems exist, such as the insufficient learning data, and the intra-speaker variability of speech.

Speaker identification has already been widely studied in the single microphone case, where only one signal is present. A variety of operations can be performed, and very good results can be achieved in adequate environments. For instance, [2] 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, [3] and [4] 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. Indeed, the redundancy brought by microphones array could be exploited to better the recognition performances. But two different approaches to the identification problem can be exhibited in this multiple signal case:

- on the one hand, a lot of works deal with the intelligent 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, whose goal is to focus a microphones array in a specific direction, thus improving the speech signal [3], [4]. Gaussian Mixture Model (GMM) robustness to noise in a speech/pause system has been evaluated in [5] through adaptive noise cancellation methods based on beamforming. Identically, matched filter arrays are used in [6] 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. As an example, one can cite [7], where the identification results reached by GMMs are combined on the basis of a 8 microphones array. In the binaural context, [8] developed a feature vector combination method optimizing the mixture weight value.

This paper is more concerned by this second approach, envisioned in a binaural context. But binaural ASkR, exploiting only the two auditory signals perceived by our two ears, has not been so covered by the literature. Actually, existing 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, [9] developed a binaural model for speech recognition, simulating the functioning of the the cochlea. The design of an artificial ear is presented in [10], 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 [11] to reduce noise and reverberations effects through blind source separation. One can also cite [12], 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 [3], [4], [13]. Finally, not so much works deal with speaker recognition in the binaural context.

The paper is organized as follows. The proposed monaural and binaural speaker recognition systems are depicted in section II. They are next both compared in Section III. The influence of the noise and of the speaker position is also carefully addressed. Finally, a conclusion ends the paper.

II. Monaural and Binaural Recognition Systems

The proposed ASkR system is presented in this section. It is text-independant, and mainly relies 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 and speech duration will then be investigated in §III.

The evaluation of the approach is based on a high quality audio database, acquired from long French monologues in identical and good conditions. It is made of 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 the tracks have been downsampled to $f_s = 22050$Hz, and so treated by Chebychev anti-aliasing filters.

A. Monaural speaker identification system

The proposed monaural system is based on the following successive computation steps, see Figure 1. First of all, 23ms-length frames are extracted from the acquired signal. The energy of each frame is computed and compared with a threshold T to eliminate non-speech portions. Next, pre-accentuation and Hamming filters 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. The major steps of this conditioning are described hereafter.

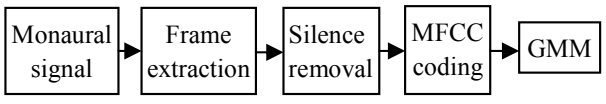

Fig. 1. Major steps of the monaural system.

1) 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 follow (see Figure 2):

- Compute the Fourier Transform (FFT) $X[k]$ of the considered time frame.
- Apply to $X[k]$ a set of $N = 25$ triangular filters regularly spaced on the mel scale defined by

$$\text{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

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

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

Note that in order to increase the robustness of the method in the presence of noise, the 16 MFCC coefficients are normalized. 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.

Fig. 2. MFCC coding

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 8^{th}-order FIR filter applied on the MFCC vectors.

2) 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 [14]. For a model of M Gaussian states, a GMM density function function of a variable x_n can be defined as

$$p(x_n|\lambda) = \sum_{i=1}^{M} p_i b_i(x_n), \quad (3)$$

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

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, i = \{1, \ldots, M\}, \quad (4)$$

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

In a speaker identification task, a M state GMM is associated to 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,

$$\hat{S} = \text{Arg} \max_{1 \leq k \leq S} p(\lambda_k|X) = \text{Arg} \max_{1 \leq k \leq S} \frac{p(X|\lambda_k)p(\lambda_k)}{p(X)}. \quad (5)$$

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 (5) can be simplified as

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

3) Expectation - Maximization: The objective is now to learn 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 [15]. Such a method exhibits a fast convergence of the parameters and is based on two successive steps: expectation (E) and maximization (M).

In the E step, responsibility functions $f_k(i|x_n, \lambda_k)$ are estimated, with

$$f_k(i|x_n, \lambda_k) = \frac{p_i(k)b_i(x_n)}{p(x_n|\lambda_k)}, \quad (7)$$

where i represents i^{th} state among the M states of the k^{th} speaker GMM. In the M step, the GMM parameters are updated on the basis of the previous function computed during the E step, that is

$$\begin{aligned} p_i^{(k)} &= \frac{1}{N} \sum_{n=1}^{N} f(i|x_n, \lambda_k), \\ \mu_i^{(k)} &= \frac{\sum_{n=1}^{N} x_n f(i|x_n, \lambda_k)}{\sum_{n=1}^{N} f(i|x_n, \lambda_k)}, \\ \Sigma_i^{(k)} &= \frac{\sum_{n=1}^{N} (x_n - \mu_i^{(k)})(x_n - \mu_i^{(k)})^T f(i|x_n, \lambda_k)}{\sum_{i=1}^{N} f(i|x_n, \lambda_k)}, \end{aligned} \quad (8)$$

with $i = \{1, \ldots, M\}$. These two steps are then 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 (6). Interestingly, such easy computations are not time consuming, thus allowing a real time implementation of the method.

B. Binaural speaker identification system

The overall functioning of the monaural system has been just described. In the binaural context, the proposed method only differs from the previous one in the feature extraction step. Indeed, there is now two signals corresponding to the left and right perceived auditive signals. 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 3. Other strategies are currently in investigation and will be presented in future works.

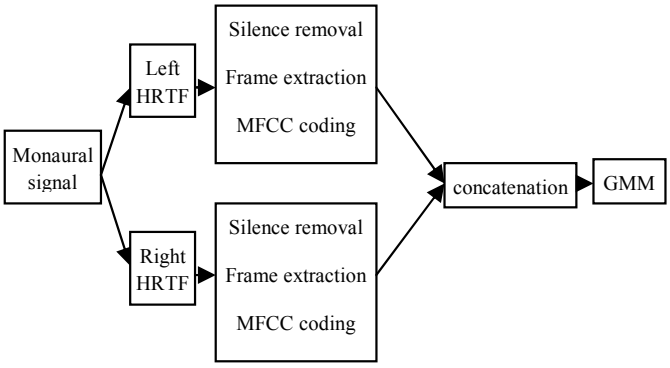

Fig. 3. Major steps of the proposed binaural system.

In the following, the binaural speech signals are simulated by convoluting 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 of the sound sources position with respect to the head. Biologically, this specific position-related filtering helps the determination of the source's position. For instance, it has been shown that two binaural cues named Interaural Time Difference (ITD) and Interaural Level Difference (ILD) are responsible for our horizontal sound localization. These cues can be directly extracted from the aforementioned HRTF filters. Practically, the frequency responses of these filters are identified through the Fourier Transform of the HRIR (Head Related Impulse Response). The HRTFs are typically measured in an anechoic room, in order to minimize the influence of spontaneous reflections and reverberations on the measured response. In this paper, the KEMAR dummy-head HRTF is used, being made freely available by the CIPIC Interfaces Laboratory of the University of California [16]. 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 [16] for more detailed information). Finally, 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.

III. Evaluation of the method

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

In the following, the speaker database is divided into two distinct parts. The first one, representing about 66% of the database, is employed for the learning of the GMMs (see §II-A2). We recall that this learning is achieved when all the GMM parameters have converged, which is indicated by the limitation of the recognition ratio's log-likelihood growth. The remaining database part (33%) is devoted to the evaluation of the recognition capabilities of the proposed system.

A. Monaural case

In this subsection, the influence of the Signal to Noise Ratio (SNR), the silence threshold T, the ΔMFCC coefficients on the frame recognition rate is assessed. Next, an evaluation of the method with longer duration testing sets is proposed.

1) Influence of noise, silence threshold and features: Here, the learning and testing steps are performed on 23ms-frames. The recognition ratio is then obtained by dividing the number of well recognized frames by the total frame number of the considered set. Next, additional white Gaussian noise is added to the speech signal to produce various SNR conditions. Finally, the silence removal process is applied on the resulting noisy signal. The subsequent recognition ratios are depicted in Figure 4 (left). Logically, the recognition performance

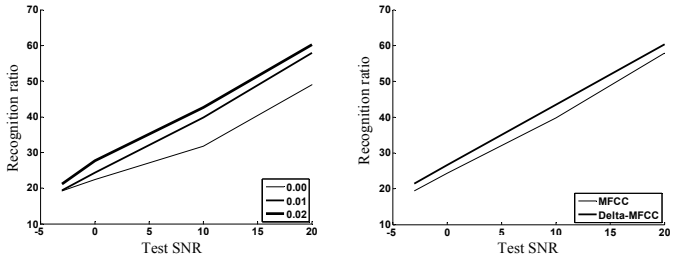

Fig. 4. (Left) Monaural recognition ratio as a function of the SNR for distinct silence threshold T (set to 0, 1 or 2%, 0% meaning no silence removal). (Right) Recognition ratio with and without ΔMFCC.

increases when the signal to noise ratio also raises. In the same vein, the highest performances are obtained with a high-value silence threshold $T = 0.02$. But note that with this value, the speech signal is highly degraded as a lot of frames are classified as being silence. This results in a low frame number available for the recognition process. Consequently, a value of $T = 0.01$ is used in all the following.

While it is not presented here, the influence of the GMM states number M has also been evaluated, for $M = 8$ to 32. As the database is only made of 10 different speakers, the M value does not have any significant influence on the performances. So, in the following $M = 16$ has been selected. For such a value, 40 iterations are sufficient for the convergence of the GMM parameters, like in [15].

Duration	No Noise	10 dB	0 dB	-3 dB
1 s	99.32	95.69	74.14	63.85
3 s	100	99.53	90.38	84.88
5 s	100	99.84	95.89	83.72
15 s	100	100	100	100

Fig. 5. Monaural recognition rates, for various time integration and SNR conditions.

Previously, MFCC coefficients only have been used during the learning and testing steps. 16 ΔMFCC coefficients are now also considered during these two steps, resulting in a features vector of size 32. The subsequent recognition rates are exhibited in Figure 4 (right). Clearly, considering ΔMFCC coefficients can improve the recognition rate up to 8.5%. So, in all the following, the features vector will always be composed of 16 MFCC and 16 ΔMFCC features.

2) Influence of the testing duration: The previous study has been performed on the basis of 23ms-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 overs previous 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, 5 and 15 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, 15s-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. As expected, the recognition rates increase for longer durations, and reach up to 100% for a 15s-long signal even for low SNR values. This table will now serve as a reference for comparison with binaural methods.

B. 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 value are considered, together with ΔMFCC coefficients. 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 place. $\theta = 0^\circ$ and $\phi = 0^\circ$ 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 6 (left), and are quite similar to the previous monaural case. Indeed, as the speakers position remains the same during

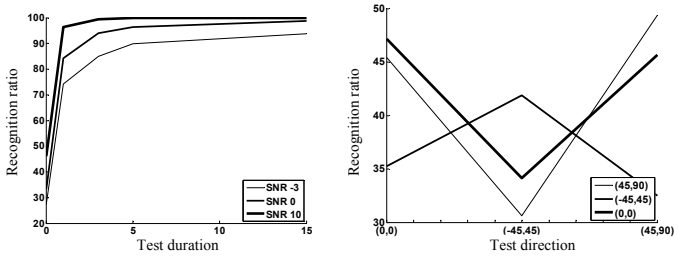

Fig. 6. Study for the same direction for all the speakers. (Left) Mean binaural recognition ratio with GMMs trained and tested in the same direction as a function of test duration in second. (Right) Binaural recognition ratio as a function of the test direction, for SNR = 10dB.

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 6 (right) for SNR $= 10$ dB. Such a phenomenon remains valid for other SNR values. This clearly shows that the 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 performances, a second scenario has been tested. It consists in forming 3 speakers groups respectively emitting from the 3 angular positions $(\mathrm{Az}, \mathrm{El}) = \{(0^\circ, 0^\circ); (0^\circ, 45^\circ); (0^\circ, -45^\circ)\}$ during the learning step. Maintaining these same positions during the evaluations leads to the recognition rates reported in Figure 7 (left). While the method shows good performances, it also demonstrates the sensibility of the binaural recognition to speaker situation. Indeed, one can see that better rates are obtained in Figure 7 (left) than in Figure 6 (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 $(0^\circ, 0^\circ)$, see Figure 7 (right). In fact, this specific position is *central*, being the closest place to the other learned positions. In that sense, it represents the

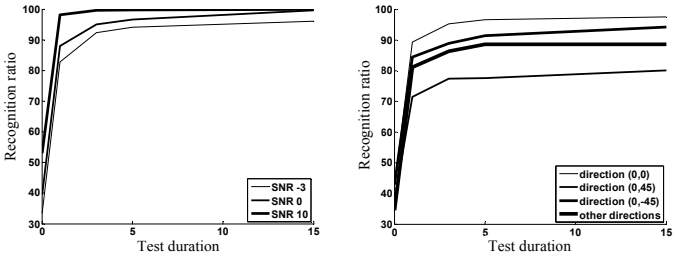

Fig. 7. Study for a group of speakers. (Left) Binaural recognition ratio with GMMs trained and tested in the same direction. (Right) Binaural recognition ratio with GMMs tested when all the speakers are simulated from the direction of training of one group. Test duration is indicated in second.

orientation minimizing the position influence, and thus also minimizing the speaker confusion.

3) One direction for each speaker: This time, one position is linked to one specific speaker during the learning step. As ever mentioned, if these positions are the same during the testing phase, then better results can be obtained if a smaller number of speaker is associated to one direction. So, results in Figure 8 (left) could be considered as the best reachable ratios in this condition (one direction per talker), minimizing the position influence. But if the speakers position is changed

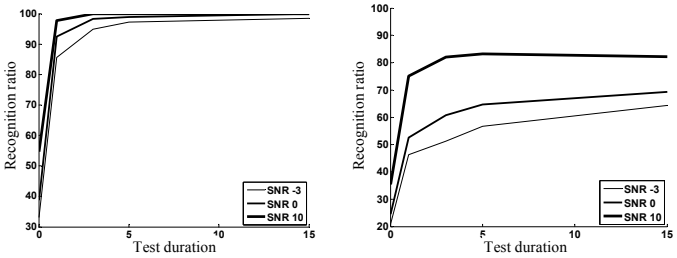

Fig. 8. Study for one direction per speaker. (Left) Binaural recognition ratio with GMMs trained and tested in the same direction. (Right) Binaural recognition ratio with testing on 3 unlearned directions for all speakers. Test duration is indicated in second.

during the evaluation step, the algorithm performances drastically decrease (see Figure 8 (right)): this clearly shows that one has to perform the learning with multiple positions per talker.

4) Multiple directions for each speaker: In order to minimize the position influence, the GMMs 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 9 (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 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 9 (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.

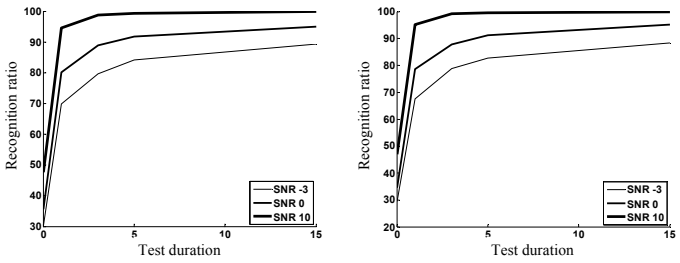

Fig. 9. Study for multiple learning directions. (Left) Binaural recognition ratio with GMMs trained and tested in the same multiple directions. (Right) Binaural recognition ratio with testing on 10 unlearned directions for all speakers. Test duration is indicated in second.

This is a major 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.

IV. 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 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: we will use spectral methods based on the correlation of left and right signals, and will use a real recorded database for the speaker's directions, without passing through simulated HRTFs from monaural signals. We have conducted such preliminary experiments on real data in [17], demonstrating the effectiveness of the approach in a controlled acoustic environment.

Acknowledgment

This work is 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] S. Furui, "40 years of progress in automatic speaker recognition," in *Lecture Notes in Computer Science, volume 5558*, 2009.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] K. Youssef, S. Argentieri, and J.-L. Zarader, "From monaural to binaural speaker recognition for humanoid robots," in *IEEE-RAS International Conference on Humanoid Robots*, 2010.