Binaural Speaker Recognition for Humanoid Robots

Bastien Breteau, Sylvain Argentieri, Jean-Luc Zarader, Zefeng Wang and Karim Youssef
Institut des Systèmes Intelligents et de Robotique; Université Pierre et Marie Curie - Paris 6, CNRS UMR 7222; 4, place Jussieu; 75252 Paris Cedex 05 - France
bastien.breteau@wanadoo.fr, sylvain.argentieri@upmc.fr, jean-luc.zaraderg@upmc.fr,
zwang@isir.upmc.fr, karim.youssef@isir.upmc.fr

Abstract— **This paper deals with Automatic Speaker Recognition in a binaural context. Such a problematic, not so widely dealt with within the speech processing community, can have potential applications in humanoid robots where speech can be used as the most natural interface between humans and robots. The proposed recognition system is based on parallel Predictive Neural Networks exploiting MFCCs (Mel Frequency Cepstral Coefficients) to discriminate multiple talkers. Because of the binaural nature of the system, the sensitivity of the proposed algorithm to the speaker spatial position during the learning step is carefully studied. The influence of noise and reverberation on the recognition rate is also reviewed. Finally, preliminary experimental results based on the recorded signals from a binaural dummy head are presented.**

I. INTRODUCTION

Physicians often describe auditory perception as the most important sense in humans, playing a fundamental role in cultural learning, especially in everything related to language, and so to human communication. Significant advances in understanding the biological processes which enable the handling of acoustic data by humans have been obtained during the 80s [4], showing that our auditory system is able to turn a complex acoustic wave into a series of neuronal activity configurations transmitted to the brain. On the basis of the two perceived signals, the Robotics Community has then proposed many auditory functions, trying to mimic our amazing ability to precisely analyze an auditory scene. Many works have first focused on the sound source localization problem, which has ever been widely dealt with by the Acoustics and Signal Processing Community in the context of microphone arrays through correlation-based approaches, beamforming, or high resolution methods [9]. Then, many works have dealt with higher-level auditory functions, like speech recognition, enabling a more natural human-robot interaction.

This paper deals with Automatic Speaker Recognition (ASkR). Such a topic has been widely studied within the speech processing community [6],[10],[5]. But this problematic is most of the time apprehended in a constrained experimental framework, which leads to specialized systems devoted to very specific applications. Moreover these systems are developed nearly exclusively in a monaural context, with the signal being recorded in ideal acoustic situations. On the contrary, the robotic context arises new specific constraints, mainly related to the noisy signals perceived in a real reverberant environment, including for instance computer or air conditioning noises. In that sense, classical monaural techniques are not well suited to high-SNR (Signal to Noise Ratio) conditions, in realistic and evolutive acoustic conditions.

Surprisingly, ASkR in a binaural context is not so much convered in the litterature. Nevertheless, exploiting two coherent signals coming from a dummy head sounds like an efficient way for improving and rubustifying robots recognition capabilities. In fact, binaural audition is, from about a decade now, an increasing research area in robotics, where a lot of works focus on sound source localization, extraction and speech recognition. Most of these works are rooted in the Computational Auditory Scene Analysis (CASA) framework, which aims at providing real-time and efficient analysis of the acoustic scene surrounding a mobile robot. One can cite for example [3], where a two-channel-based system for humanoid robots is designed to reliably localize two moving sound sources without prior information. In the same vein, [7] showed that the position of a sound source in azimuth and elevation can be inferred from two artificial ears using interaural and spectral cues. Binaural audition is also exploited in [2] to perform speech detection for a humanoid robot able to separate and recognize speech signals even in noisy home environments. This paper exhibits very promising results, though it still requires manual tuning and evaluation. A completely different solution is proposed in [8], where a microphone array with 32 transducers is used to implement a speech recognition system working in a noisy housing environment. These two opposite solutions (microphone array vs. binaural audition) set out all the problems involved in sound recognition with realistic environments. It also points out the need of adaptive algorithms which are naturaly well suited to noisy and evolutive conditions.

In this paper, we champion the use of neural networks for binaural speaker recognition. Such methods, exploiting Multi-Layer Perceptrons (MLP), constitute a powerful and flexible analysis tool, as well as an efficient speaker characteristic extractor. In this work, Predictive Neural Networks (PNN) will be specifically trained to recognize one explicit talker by extracting classical speech features like MFCCs. In that sense, the purpose of this study is not to propose an all new ASkR algorithm, but rather to highlight the potentiality of binaural ASkR system with respect to estab-

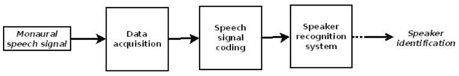

Fig. 1. The three speaker identification steps

lished monaural methods. Interestingly, the manipulation of binaural signals brings to the fore the inherent sensitivity of the speaker recognition system to the position of the talker. As a consequence, a particular attention will be paid to the constraints related to the sound sources directions, the influence of the source's position on the recognition will be carefully studied.

The paper is organized as follows. First, the ASkR system is presented. Each step, from the signal coding to the recognition process is detailed, in a monaural and binaural context. Next, simulation results follow. The recognition rates for various SNRs and a case study on the sensitivity of the method to the speakers positions are depicted. Then, preliminary experimental results are shown. Real binaural recordings from a dummy-head are exploited to assess the efficiency of the method in a real very noisy and highly reverberant environment. Finally, a conclusion ends the paper.

II. AUTOMATIC SPEAKER RECOGNITION SYSTEM

An automatic speaker recognition system is classicaly based on three successive steps. First, the (monaural) speech signal is digitally converted through an acquisition card and eventually pre-preprocessed. Then, the signal is coded by extracting multiple coefficients representing the speech information. Finally, a speaker identification algorithm exploits these coefficients to recognize one or multiple speakers (see Figure 1). Importantly, the objective of the feature extraction step is to decrease the volume of the data by deleting the redundant or useless information contained in the speech signal. So, this preliminary step condenses the initial signal into a reduced number of coefficients, which are then used by the speaker recognition system. Its role is to associate the perceived signal with one of the known speakers by using learned data.

In this section, the proposed binaural ASkR system is presented. First, the speaker database, the binaural simulations and the speech characteristics exctraction method are depicted. Then, the proposed recognition system is introduced. Finally, criteria for the evaluation of the performances are discussed.

A. Simulation and coding of the binaural signals

1) Database and simulation of the direction of sounds: The used speaker database has been generated from radiophonic sounds originating from average quality recordings. The selected sequences relate to long french monologues with a low-level ambient noise. So, this study is based on a database of $S = 9$ male speakers, with a 7-minute long signal per speaker. Next, the binaural speech signals –i.e. the perceived left and right signals– are simulated by convoluting the speaker database signals with impulse responses coming from a HRTF (Head-Related Transfer Function) database. In this paper, the KEMAR dummy-head HRTF is used, being made freely available by the CIPIC Interfaces Laboratory of the University of California [1]. 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 [1] for more detailed information). Finally, speech signals and HRTF database have been acquired with a sampling frequency $f_s = 44100\text{Hz}$.

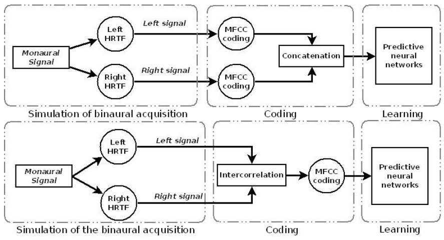

Fig. 2. Binaural methods by concatenation (BM-C) *(top)* and by intercorrelation (BM-I) *(bottom)*

2) Signal coding: Once being simulated as emitted from one or multiple directions, the resulting two signals are coded using two different strategies. In fact, binaural recognition methods will only differ from monaural ones during this coding step. More precisely, the question is: "how to combine the extracted features coming from the two signals ?" As an answer, two simple binaural methods have been tested in this paper : the binaural method by concatenation (BM-C) and the binaural method by intercorrelation (BM-I), see figure 2. It is important to notice that the second binaural method based on the intercorrelation, provides a new approach for treating binaural or two-channel signals. It also has the importance of allowing to extract the cross-spectral information and to reduce the noise effect. 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 $X[k]$ 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

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

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

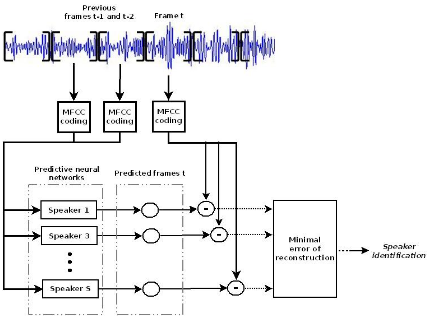

Fig. 3. Speaker recognition system using predictive neural networks

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 a high resolution for low frequencies and a weaker resolution for high frequencies.

As previoulsy mentioned, in the context of binaural audition, the two signals can now be coded by MFCC coefficients with two strategies. On the one hand, the proposed BM-C method consists in computing independently MFCCs from the left and right signals, which are then concatenated into a single vector of $2K$ elements. In all the following, we choosed $K = 16$. On the other hand, we postulate through the BM-I method the use of K MFCC coefficients of the intercorrelation $R_{xy}[m]$ of the right and left signals (see Figure 2). This intercorrelation that offers a signal reflecting the similarities between the left and right ears, and rejecting the noise, is defined by

$$R_{xy}[m] = \sum_{p=0}^{P-1} x[p]y[p-m] \tag{3}$$

where P is the length of the time frame and $x[p], y[p]$ the right and left signals respectively.

B. Recognition system

1) Predictive neural networks: Once the MFCC features are extracted, they are exploited as inputs of the recognition system. The proposed system in this paper relies on Predictive Neural Networks (PNN). It is made of $S = 9$ parallel predictive networks, each one of them being dedicated to a specific speaker of the database. The role of each PNN is to predict the MFCC coefficients of the currently analyzed time frame from the two previously coded frames, see Figure 3.

During the learning process, each PNN is trained on the basis of 3 consecutive vectors of features, each of them being extracted from a 30ms time frame with 50% ovelapping snapshots. These three vectors constitute one training example; the entire training database is computed on the basis of the speaker signal to be learned for 90%. The remaining 10% of the training data are devoted to vectors extracted from the other speakers for an unlearning process, in order to increase the specificity of each PNN for one specific speaker.

During the recognition process, a set of 3 MFCC vectors extracted from a speech signal coming from an unknown speaker is presented to the S parallel PNNs. Each network provides a prediction of the third vector on the basis of the two others. Then, the S reconstruction errors between the real and the predicted features are computed. Finally, the unknown speech signal on the current time frame is associated to the speaker linked to the PNN producing the minimal reconstruction error

2) Parallel learning: In the previous recognition process, each PNN is independently learned from the others. As a consequence, there is no control of the reconstruction performances related to the different networks during the learning process. That is to say that one PNN becomes so efficient that it is able to predict speech features for any speakers of the database.

In order to equalize the performances of each PNN –i.e. to assess that one PNN is devoted to only one speaker–, we propose in the following a parallel training method using a cross-validation technique, whose objective is to periodically control the training speed of each network. The suggested algorithm is described in Algorithm 1. This parallel training method is based upon two criteria: the recognition rate of the i^{th} network R_i, and a specificity criterion C_i^l for the i^{th} network in the l^{th} cross-validation step. The value of the recognition rate R_i is introduced to stop the training of the i^{th} network if its learning process is faster than the others, thus equalizing the recognition performances of the global system. The C_i^l criterion is also exploited to detect the increase of the specificity of the i^{th} network devoted to the i^{th} speaker. The minimization of this criterion conducts to the fall of false detections and increases the global rate of recognition R_{global}^l.

```
for the current $l^{th}$ cross validation step do
    - Computation of the confusion matrix
    - Computation of the global recognition rate $R^l_{\text{global}}$
    for $i = 1 : S$ do
        - Calculation of the recognition rate for each
        network $R_i$.
        - Calculation for each network of a criterion $C_i^l$:
            $C_i^l = \frac{\sum_{k=1;k\neq i}^{S} \text{False\_Detections}_k}{R_i}$
    end
    for $i = 1 : S$ do
        if $C_i^{l-1} > C_i^l$ then
            Memorization of the new network weights.
        end
        if $R^l_{global} < R_i$ then
            Stop learning until the next stage of
            crossvalidation.
        end
    end
end
```

Algorithm 1: Cross-validation step algorithm

C. Criteria for the evaluation of performances

Until now, all the previous descriptions were focused on a 30ms time-frame scale. As a consequence, the first immediate way to evaluate the performance of the method is to check, frame after frame, if the system correctly estimates the good speaker. However, in real robotics applications, where longer speech signals are available, different criteria can be used to produce a more reliable final decision. One possibility is to allocate the speech signal made of successive 30ms frames to the speaker whose PNN recognizes the highest number of frames. So, such a method sounds like a majority vote method. In the following, the interpretation of the results will especially focus on the recognition rate on the frames, but also on longer signals lasting 3, 5 and 15 seconds. The recognition rates obtained for the 3s-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.

III. RESULTS

We propose in this part to evaluate the performance of the proposed method in simulation on the basis of the previously described system. Because of the use of binaural signals, the position of the speaker enters as a new parameter on classical speaker recognition systems. It will be of fundamental concern. So, in the first part of this section will be studied the influence of the speaker position with respect to the recognition rate of the proposed algorithm. Next, the efficiency of the system will be assessed for different SNR values in the second subsection.

A. Influence of the speaker position on the recognition rate

In this subsection, the speech signals exploited for the learning and the cross-validation steps are simulated as coming from a single direction for each speaker. As a consequence, the ASkR system will become naturally sensitive to two kinds of competitive information. The first one is related to the specificities of speech signals which are captured into a feature extraction setup. This is actually the characteristic we are interested in. Unfortunately, the binaurality also introduces directional cues, like IPD/ITD (Interaural Phase Difference/Interaural Time Difference) or ILD (Interaural Level Difference), in the process. So, these interaural cues may conduct the ASkR system to perform a recognition of the direction of the source in spite of the speaker himself. More precisely, if the position of the speakers is not the same in the learning and test phases, the recognition rates might decrease.

To assess the sensibility of our system with respect to the speaker positions, we propose in the following to test 3 directional groups, composed of 3 different speakers each, for azimuths $\theta = \{-45°, 0°, 45°\}$ in the median plane and for elevations $\psi = \{-45°, 45°, -45°\}$ respectively (see Figure 4). Each of them is simulated thanks to the previoulsy mentioned CIPIC database, without any noise. The influence of an additionnal noise in the scene will be studied in the next subsection. The objective is thus to compare the rate of confusion within the same or different directional groups, and to highlight the influence of directional cues on the identification of the speakers for the two binaural methods BM-C and BM-I.

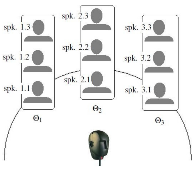

Fig. 4. Directionnal groups: spk X.Y denotes the Y^{th} speaker of the X^{th} group.

TABLE I

CONFUSION MATRIX FOR THE CONCATENATION METHOD BM-C (ELEMENT: FRAME, NUMBER OF EXAMPLES: 45000 FRAMES).

	Classification								
	Group 1			Group 2			Group 3		
	Spk 1	Spk 2	Spk 3	Spk 4	Spk 5	Spk 6	Spk 7	Spk 8	Spk 9
Spk 1	48.12	25.29	25.20	0.23	0.31	0.02	0.21	0.56	0.06
Spk 2	13.89	59.23	25.23	0.42	0.21	0.08	0.12	0.78	0.03
Spk 3	18.59	25.44	54.30	0.28	0.61	0.06	0.24	0.44	0.04
Spk 4	1.03	1.55	5.98	51.24	31.79	8.24	0.02	0.14	0
Spk 5	3.13	1.82	7.80	20.23	51.18	15.76	0.02	0.06	0
Spk 6	0.96	2.16	6.83	14.55	26.30	49.00	0.11	0.07	0.01
Spk 7	1.13	4.10	10.63	0.04	0.12	0.05	57.39	15.69	10.85
Spk 8	1.20	1.74	5.72	0.01	0.05	0.01	13.22	57.58	20.47
Spk 9	2.56	1.58	8.41	0	0	0.02	15.95	26.73	44.73

The results for the BM-C method are shown in table I. The important result to be noticed is the significant difference between the intragroup and intergroup confusion rates. More precisely, the confusion rate between two speakers from the same directional group spreads from 8.24%to 31.79%, while it is generaly less than 1% for two different directional groups. As a conclusion, this first study demonstrates a strong learning of the directional cues, supplanting the characterization of the speaker, when working with the BM-C method.

The results for the BM-I method are shown in table II. This time, the intragroup and intergroup confusion rates are

TABLE II

CONFUSION MATRIX FOR THE INTERCORRELATION METHOD BM-I (ELEMENT: FRAME, NUMBER OF EXAMPLES: 45000 FRAMES).

	Classification								
	Group 1			Group 2			Group 3		
	Spk 1	Spk 2	Spk 3	Spk 4	Spk 5	Spk 6	Spk 7	Spk 8	Spk 9
Spk 1	25.22	17.17	14.10	5.50	6.59	1.80	11.45	9.35	8.80
Spk 2	10.98	31.46	17.02	4.76	6.28	3.52	9.59	8.10	8.28
Spk 3	11.32	16.77	30.53	2.49	8.68	3.80	11.02	7.83	7.56
Spk 4	3.05	8.33	5.21	24.51	14.06	7.23	9.02	14.52	14.06
Spk 5	4.41	5.05	5.31	12.08	23.05	9.40	14.01	14.21	12.49
Spk 6	2.91	4.89	6.05	6.60	13.87	24.08	16.55	10.61	14.44
Spk 7	2.93	6.54	7.13	6.72	15.15	11.10	26.89	10.47	13.07
Spk 8	3.73	6.21	5.67	11.70	13.37	7.55	11.98	26.47	13.33
Spk 9	3.38	5.76	6.49	12.35	10.95	8.41	13.03	13.43	26.19

not so different, being equal respectively to about $10-15\%$ and 5%. These values show that even if BM-I is not totally insensitive to the directional information, the learning of directional cues is now lower. Consequently, it will give the opportunity to efficiently discriminate speakers regardless of the direction of emission. The results of the monaural method depicted on Figure 1 are not shown is this paper. Neverteless, they exhibit a mean recognition rate of 29% for a 30ms frame, while its value for the BM-I technique is 25%. So, despite the sensitivity of the BM-I to the position, this method gives quite analog performances. On the contrary, the BM-C produces an imposing 52% recognition rate. But even if this higher rate might significate that this binaural method outperforms the monaural one, one has to keep in mind that the binaural methods also learn the spatial position of each speaker, this position being identical in the learning and testing steps.

A comparison between the results of the two binaural methods exhibits their complementarity. The method of concatenation, capturing the left and right spectral densities, conserves the directional information induced by ILD or spectral notches. So, it seems that this technique could then be exploited for a localization purpose in spite of speaker recognition. On the opposite, the method of intercorrelation seems less sensitive to directional cues and thus allows efficient recognition of the speaker, independently of any position modification. This can be explained by the extraction of MFCC coefficients of the intercorrelation, which does not code the interaural level difference.

B. *Robustness to noise*

The objective of this second study is to examine the influence of the noise on the recognition rates in monaural and binaural contexts, and to highlight the eventual better robustness of binaural methods. In noisy conditions, monaural methods are known to provide poor recognition capabilities. On the contrary, the exploitation of two redundant signals may improve the robustness of the ASkR system. For this purpose, two independent white gaussian noises have been added to each left and right signal with a variable Signal-to-Noise Ratio (SNR), thus simulating a diffuse noise field. The following evaluation is focused on binaural and monaural signals affected by a SNR of $\{10\text{dB}, 0\text{dB}, -3\text{dB}\}$. Note that in ordrer to annihilate the aforementioned directional cues influence, 14 different positions for each speaker have been used during the learning step. The obtained results are depicted in Figure 5. Tests with frames and 5-seconds long signals have also been made but not presented here. tests with 3 and 15 seconds are enough to present real interaction conditions.

Generally, all the methods see a decrease of their performance with a decrease of the SNR. For example, the frame recognition rate falls from 77.78%, with SNR= 10dB when working with a 15-second long signal, while it is only 42.86% with SNR= -3dB in the monaural case. This illustrates the strong performance deterioration of monaural techniques under real conditions involving noisy and/or reverberating environments. The BM-I method is relatively robust to noise: with SNR= 10dB, it outperforms the monaural method (92.52% vs. 77.78%) for a 15-second long signal. In a very difficult situation with SNR= -3dB, it still reaches a recognition score of 68.03%, that is to say 25% higher than in the monaural context. The binaural method of concatenation does not provide such a robustness to noisy conditions. The BM-C performances remain quite identical to the monaural system, demonstrating that the binaurality itself is not a sufficient condition to improve the recognition capabilities, the key features being more in the coding step of the signals.

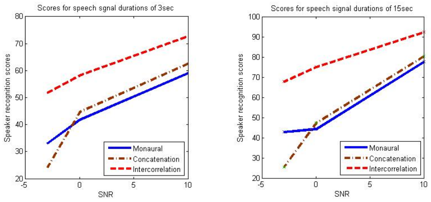

Fig. 5. Evolutions of the recognition scores for various SNR values.

TABLE III
RELATIVE SOURCE POSITIONS CAPTURED BY THE MOTION CAPTURE.

direction	azimuth	elevation	distance
left	$-52.63°$	$3.26°$	1.75m
front/left	$-29.84°$	$-0.03°$	2.40m
front	$2.15°$	$-4.03°$	1.54m
front/right	$28.33°$	$0.42°$	2.47m
right	$59.67°$	$-3.60°$	2.01m

IV. PRELIMINARY EXPERIMENTAL WORKS

In this section, we present some preliminary experimental results which have been obtained with real signals originating from a KU100 dummy head. First, the experimental setup is depicted. Next, experimental results follow.

A. *Speech recording with a dummy-head*

The first step in this experiment is to define the speaker database. In order to establish an effective comparison between the previous simulations, we exploited the same set of speaker recordings as before, but limited to only 5 effective speakers. The signals have been emitted from one high-level M-Audio AV40 monitoring loudspeaker being mounted on a 1.7meter-high tripod. Each speaker signal has been recorded from 5 different locations, each of them being very precisely measured with a motion capture process. This system consists in 3 infrared cameras capturing the position of infrared tags in real time. 4 tags have been placed on the edges of the loudspeaker to interpolate its 3D position. Next, the emitted signals are recorded through a KU100 dummy head from Neumann equipped with 2 balanced microphones.

Its outputs are then simultaneously sampled with a National Instruments acqusition card with $f_s = 100$kHz. As before, the 3D position of the head is extracted from 5 tags by the motion capture system. As a consequence, the relative position of the loudspeaker with respect to the head can be com-

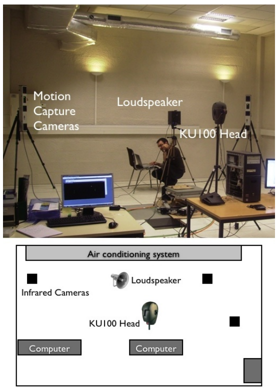

Fig. 6. Description of the experimentations. The various noise sources (computer, air conditioning) positions are approximately represented.

puted. The 5 relative positions are summarized in table III in a head-related coordinate system. This experimentation took place in a very noisy and reverberant environment. 3 computers were running and an air conditioning system was continuously humming during the experiments. The global organization of the room is shown in Figure 6.

B. Preliminary experimental results

The speaker database used for the training step of the algorithm is composed of 5 speakers, each of them being associated to one of the positions mentioned in table III. During the evaluation step, the positions of the speakers are interverted. So, this experiment makes possible the computation of the speaker rate of recognition and the confusion of the eventually learned direction of emission. These preliminary results are presented in table IV. The restricted number of speakers and positions exploited in this experiment does not strictly allow us to compare the scores obtained under real recording conditions with those obtained by pure simulation without noise. In fact, it appears that the BM-I conducts to a less effective recognition than the BM-C method, while the opposite was mentioned in Part §III-B. But one has to keep in mind that the two simulated noises were statistically independent while they are highly correlated in the real measured signals. Moreover, the acoustic environment is highly reverberant, conducting to an important degradation of the intercorrelation function. These reverberations could also explain the effective learning of the speaker by the BM-C method, the binaural cues being highly modified by the environment: apparently, sounds are not well localized in this environement. This effect is demonstrated by the increase of the speaker recognition rate for longer signals. As a conclusion, the experiments on real recorded data seem to show the potential problems of binaural methods.

V. CONCLUSIONS AND FUTURE WORKS

We have presented in this paper an Automatic Speaker Recognition system working in a binaural context. The proposed recognition method relies on parallel Predictive Neural Networks exploiting MFCC coefficients to discriminate multiple talkers. A first contribution of the paper is the definition of an original pararallel learning algorithm based on the control of the learning rate of each competitive network. On this basis, two different strategies have been evaluated, based on intercorrelation or concatenation. We have then studied the effect of the speaker spatial position during the learning and evaluation steps. Finally, we evaluated how the binaurality can be exploited to improve the recognition rates in noisy conditions. The BM-I method is the most robust to position change and whatever the SNR values. Nevertheless, the BM-C method seems to be able to estimate the talker position. Preliminary experimental results have also been presented. They exhibit the sensitivity of the BM-I method to reveberations.

TABLE IV

EXPERIMENTAL SPEAKER RECOGNITION SCORES.

	Trames	3s	5s	15s
BM-C reco. rate	36.66	77.71	83.81	88.57
BM-I reco. rate	27.60	41.71	40.00	37.14

We are now working on defining a larger experimental speaker dataset embedding at least 20 male and female speakers recorded from a high number of spatial positions with the dummy-head. Moreover, in order to robustify the proposed systems with respect to reverberations, a preprocessing step will be added before the features extraction step. This preprocessing stage will be based on equalization and dereverberation of the acquired signals.

REFERENCES

[1] V. Algazi, R. Duda, R. Morrisson, and D. Thomson. The cipic hrtf database. In *IEEE Workshop on Applications of Signal Prcessing to Audio and Acoustics*, 2001.

[2] Hyun-Don Kim, Jinsung Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2008.

[3] Hyun-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Design and evaluation of two-channel-based sound source localization over entire azimuth range for moving talkers. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2008.

[4] J.C. Middlebrooks and D.M. Green. Sound localization by human listeners. *Annual Reviews on Psychology*, 1991.

[5] I. Peer, B. Rafaely, and Y. Zigel. Room acoustics parameters affecting speaker recognition degradation under reverberation. In *Hands-Free Speech Communication and Microphone Arrays*, 2009.

[6] Douglas A. Reynolds. An overview of automatic speaker recognition technology. In *IEEE International Conference on Speech and Signal Processing (ICASSP)*, 2002.

[7] Tobias Rodemann, Gokhan Ince, Frank Joublin, and Christian Goerick. Using binaural and spectral cues for azimuth and elevation localization. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2008.

[8] Yoko Sasaki, Satoshi Kagami, Hiroshi Mizoguchi, and Tadashi Enomoto. A predefined command recognition system using a ceiling microphone array in noisy housing environments. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2008.

[9] H.L. Van Trees. *Optimum Array Processing (Detection, Estimation and Modulation Theory, Part IV)*. Wiley-Interscience, 2002.

[10] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, and Wang Lihao. Speaker recognition based on dynamic mfcc parameters. In *IEEE International Conference Image Analysis and Signal Processing (IASP)*, 2009.