ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Approaches for Automatic Speaker Recognition in a Binaural Humanoid Context

Karim Youssef, Bastien Breteau, Sylvain Argentieri, Jean-Luc Zarader and Zefeng Wang

Institut des Systèmes Intelligents et de Robotique – Université Pierre et Marie Curie, Paris 6
4, Place Jussieu, 75252 Paris Cedex 05 - France

Abstract. This paper presents two methods of Automatic Speaker Recognition (ASkR). ASkR has been largely studied in the last decades, but in most cases in mono-microphone or microphone array contexts. Our systems are placed in a binaural humanoid context where the signals captured by both ears of a humanoid robot will be exploited to perform the ASkR. Both methods use Mel-Frequency Cepstral Coding (MFCC), but one performs the classification with Predictive Neural Networks (PNN) and the other performs it with Gaussian Mixture Models (GMM). Tests are made on a database simulating the functioning of the human ears. They study the influence of noise, reverberations and speaker spatial position on the recognition rate.

1 Introduction

Audition is a very important sense for humans. Communication with others is a need and speech holds much information. Based on the speech signals that they perceive, humans are able to recognize the words that they hear, the person saying them and the position of this person. The growth of robotic technologies has made it possible to create robots that are able to use auditory capabilities that resemble to those of the humans. And especially humanoid robots are likely to be used in scenarios where interactions with humans are needed.

Our paper deals with automatic speaker recognition. This task has been widely studied in the last decades [1] [2]. But speaker recognition systems often rely on a single-microphone or a microphone array data acquisition. In the second case, monaural recognition approaches can be operated on a single signal that is obtained from the array using methods like beamforming [4]. Here, classical recognition systems require a speech-specific coding combined with a pattern recognition method. For instance, [5] used MFCCs and SVMs.

Binaural or two-channel based methods are almost not studied in the exact object of speaker recognition. Two channels can be used in signal enhancement approaches like adaptive noise cancellation [6]. Localization has also been studied in two-channel contexts [7]. Our systems address the binaural humanoid speaker recognition: the signals of two ears are exploited in the same time to efficiently recognize speakers.

This paper is organized as follows. First, the two speaker recognition systems are presented. Then a simulated database used for their evaluation is depicted. Test results are shown; they study the effects of noise and speaker position on the recognition rates. Both systems' results are then compared and analyzed. Finally, a conclusion ends the paper.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

2 Recognition Systems

In this section, we present the recognition systems we designed and we detail their main steps. The approach is text-independent and applied on a closed set of persons. Mel Frequency Cepstral Coding is adopted and classification is made once with predictive neural networks and once with Gaussian mixture models.

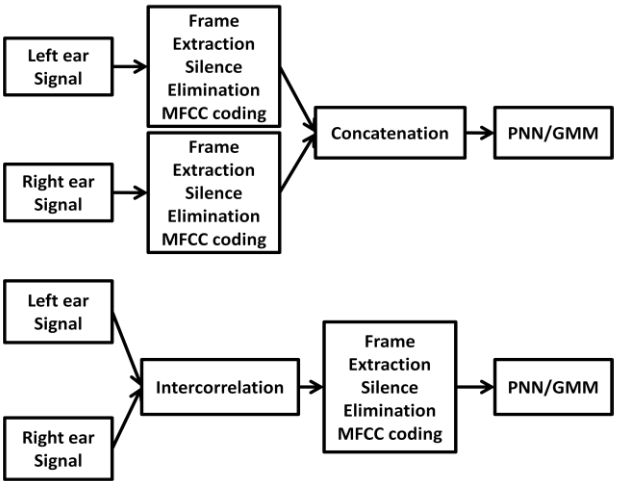
Left ear Signal
Frame Extraction Silence Elimination MFCC coding
Right ear Signal
Frame Extraction Silence Elimination MFCC coding
Concatenation
PNN/GMM
Left ear Signal
Right ear Signal
Intercorrelation
Frame Extraction Silence Elimination MFCC coding
PNN/GMM

Fig. 1: speaker recognition systems. Upper part: binaural concatenation system, lower part: binaural intercorrelation system

As we can see, the systems consist mainly in: extracting frames from the signals, eliminating the silent frames, coding the speech frames with MFCCs and then learning and testing the PNNs and the GMMs. The binaural concatenation system consists in concatenating the MFCC codevectors obtained with the two ears. And the binaural intercorrelation system consists in performing the previously mentioned steps on a signal obtained by intercorrelation of the signals in both ears.

The frames have the length of 22ms. In this duration, the configuration of the vocal tract is constant so the content of speech can be studied and speaker-related info can be extracted. The silence elimination method is based on energy calculations. Fames of energy above a threshold are kept while frames below it are discarded. The threshold is given by:

$$E_{th} = E_{min} + k(E_{max} - E_{min}) \qquad (1)$$

where E_{min} and E_{max} are respectively the minimum and the maximum frame energies in the studied speech segment. K is a parameter that controls the silence elimination. This step is followed by pre-treating the resulting frames with a Hamming window and a pre-accentuation filter. MFCC codevectors are then extracted and used as features for the PNN and the GMM classification. MFCC, PNN and GMM are detailed in the following sub-sections:

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

2.1 Mel-Frequency Cepstral Coefficients

MFCC features are well known for their utility in sound coding for speech and speaker recognition. The main steps of this coding are as follows: the Fourier transform of the considered time frame is weighted with a set of 24 triangular filters regularly spaced on the Mel scale. The logarithms of the filters' output energies are submitted to a discrete cosine transform that gives the MFCC coefficients. We use the first 16 coefficients to form a feature vector. In the GMM system, delta-MFCCs are used with MFCCs. They are calculated based on a finite impulse response filter that takes in consideration the 9 frames surrounding the current frame.

2.2 Predictive Neural Networks

Our first system classifies with predictive neural networks. We associate a PNN to each speaker. The training consists in learning to output a codevector corresponding to a time frame based on the two input codevectors of the frames that precede it. So each network has 64 inputs (16*2: left+right*2: 2 frames) and 32 outputs (16*2:left+right*1:current frame). 90% of the training data of each PNN comes from the corresponding speaker, and the rest comes from all the other speakers to apply an unlearning process and improve the specificity of each network to his speaker. The PNNs are independently trained from each other, but this way their performances may vary and some networks may become more efficient than others. So we try to control the training speeds of the networks. We use cross-validation steps to measure their recognition performances and their learning speeds, and we hold the learning of fast networks until the others reach their levels of performance.

When testing the PNNs, a set of three consecutive MFCC codevectors belonging to an unknown speaker is presented to all the networks. They will use the first two in order to make a prediction of the third and then the reconstruction errors, i.e. the errors between the real third and the predicted third are calculated. The network with the minimal reconstruction error belongs to the classified speaker.

2.3 Gaussian Mixture Models

GMMs form a strong statistical pattern recognition method. A GMM is a set of Gaussians, also called states in this context. Each state has its own parameters: weight, mean vector and covariance matrix. The probability for a vector to belong to a GMM is then the weighted sum of its probabilities according to all the Gaussians in this mixture model:

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(x) \quad (2)$$

Where p_i is the weight of the i-th state, M is the total number of states, and $\lambda = \{p_i, \mu_i, \Sigma_i\}$, i={1, …, M} is the set of characteristics (respectively: weights, mean vectors and covariance matrices) of all the states in the GMM.

We associate a 16-state GMM to each speaker. To reach the optimal parameters of the GMMs states, each GMM is trained independently from the others with the iterative Expectation-Maximization algorithm that runs until the convergence of the parameters. Once the models are trained, we can use them for recognition: this consists in presenting each vector whose speaker is unknown to all the models.

Classification is based on *a posteriori* probability: the model that gives the biggest probability corresponds to the classified speaker.

$$S_{pr} = Arg\ max_{1 \le k \le S} p(\lambda_k|x) = Arg\ max_{1 \le k \le S} \frac{p(x|\lambda_k)p(\lambda_k)}{p(x)} \qquad (3)$$

where S is the total number of speakers. Considering that all the speakers have the same probability of appearance before the robot, and having the same test vector for all the GMMs, we obtain that the maximum of the *a posteriori* probability reflects the maximum of the probability $p(x|\lambda_k)$.

3 Training and testing strategies

Our main object is to perform the binaural speaker recognition. We study the effects of noise on the recognition ratio, and those of the speaker's positions during training and testing, since binaural signals offer information that reflect the speaker's position. Previously made tests showed that when training taking speech from only one direction for a speaker, good testing performances are obtained when the speaker is in this same direction. But when the testing direction is different, the recognition ratios decrease and the fall becomes bigger as the test direction gets further from the training direction. In a robotic context, the robot is constantly moving and the hypothesis of a speaker talking to it from the same direction where he trained it is almost impossible to achieve. That is why, in this paper, we present the results when training the robot from multiple directions. This way, the robot knows positions that cover a big part of its surrounding space and is capable of correctly identifying speakers located anywhere. The results obtained with tests on the training directions and on different directions give almost the same result. Very minimal differences are present so we only show the results with tests on the training directions. The tests are made with multiple SNRs (Signal to Noise Ratios). They also study the effects of speech duration on the recognition process. Speech durations of frames, 3, 5 and 15 seconds were tested. The last three are long enough to pronounce small or long sentences (or commands). When classifying in these durations, the results obtained with frame sequences that constitute them are taken in consideration, in a majority vote technique.

4 Database

This database originates from long radiophonic French monologues recorded in identical and good conditions, with 10 speakers and seven minutes per speaker. Since these recordings are obtained with a single microphone, they can be used in a monaural system using the same main steps of our systems. But to test our binaural system, we need to obtain left and right ear signals that correspond to multiple directions. For this purpose, we convolute our recorded signals with Head-Related Impulse Responses corresponding to Head-Related Transfer Functions (HRTFs). A HRTF describes how a signal is altered by the acoustical properties of the head, the outer ear and the torso, before reaching the final transduction stages of the inner ear. This effect is modeled by a filtering process whose impulse response is specific to each ear, and to each sound source position. So to measure a HRTF, we create an impulse in the wanted position, and we measure the responses in two microphones placed in the tympanums of both ears of a dummy head. Such measurements are done in anechoic chambers to minimize the effects of reverberations and noises. The HRTF database we used belongs to the KEMAR dummy-head [3].

5 Results

PNNs used a cross-validation direction that measured their performances and ruled the operation of parallel learning that sets them to close recognition levels (cf. parag. 2.2). The testing results are shown in Figure2.

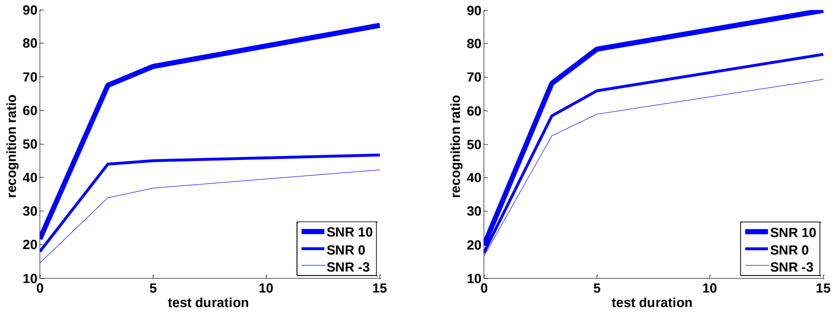

Fig. 2: speaker recognition ratios with Predictive Neural Networks.
Left: concatenation method. Right: intercorrelation method.

GMMs used delta-MFCC vectors that were concatenated with corresponding MFCC vectors to show their dynamic temporal variation, the results are shown in Figure 3.

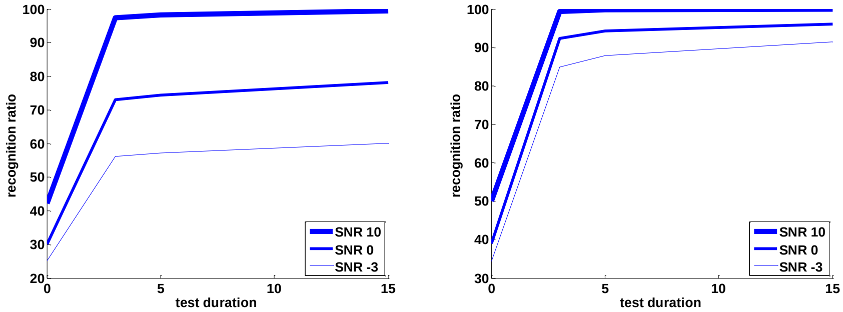

Fig. 3: speaker recognition ratios with Gaussian Mixture Models.
Left: concatenation method. Right: intercorrelation method.

From the shown results, we can reach the following statements:

- The recognition ratio improves when the signal-to-noise ratio increases and when the testing duration increases, which is a logical conclusion since the signal becomes cleaner and contains more information from the speaker.
- In both cases, the intercorrelation method gives better performances and is more robust to noise. This is explained by the fact that the intercorrelation of two ear

signals representing the same speech rejects the noise components that are not intercorrelated between the two signals. On the contrary, it highlights the signal information that reflects the identity of the speaker, well detected by MFCCs.

- When comparing the PNN results to the GMM results, we can state that GMMs give better performances. This can be explained by the fact that it is easier to group MFCC codes of a single speaker in well organized clusters and then determine the belonging of a new code to one of the groups than to predict a code based on only two others.

6 Conclusions and perspectives

In this paper, two systems of automatic speaker recognition have been presented. They place the approach in a robotic humanoid context. They are based on MFCC coding combined with PNNs once, and GMMs once. Two binaural treatments are considered: concatenation of both ears' MFCC vectors and intercorrelation of both ears' signals then MFCC coding. The results show that the intercorrelation method is more robust to noise and that GMMs work better than PNNs. A comparison with monaural methods that are classically used shows that binaural methods offer better performances. (A binaural GMM method gives at least 10 percent of improvement in the recognition ratios compared to a monaural GMM method).

Current works are based on using these methods with a database built in our laboratory. Recordings are made in an anechoic chamber with multiple speakers. The recordings are made with the Neumann KU100 dummy head. Future works will build a Voice Activity Detection system that overperforms the limitations of this energy-based silence elimination system. Other perspectives take in consideration the effect of the robot's and/or the speaker's movement, they combine VAD with speaker localization and recognition in a global system.

References

[1] S. Furui, 40 years of progress in automatic speaker recognition, *Lecture notes in Computer Science. Volume* 5558, 2009.

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

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

[4] 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, *Proceedings of ICSLP*, 1994.

[5] S. S. Karajekar, Four weightings and a fusion / a cepstral-SVM system for speaker recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.

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

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