Security Monitoring Based on Joint Automatic Speaker Recognition and Blind Source Separation

Michele Scarpiniti and Fabio Garzia
Department of Information Engineering, Electronics and Telecommunications (DIET)
"Sapienza" University of Rome
via Eudossiana 18, 00184 Rome, Italy
Email: {michele.scarpiniti, fabio.garzia}@uniroma1.it

***Abstract*—The aim of this paper is to introduce an enhanced approach for standard Automatic Speaker Recognition (ASR) systems in noisy environment in conjunction with a Blind Source Separation (BSS) algorithm. This latter is able to discern between interfering noise signals and the reference speech signal, hence it can be consider as a necessary preprocessing step. The main problem of the proposed approach lies in the not removable ambiguities typically of the BSS algorithms. In order to overcome to this drawback, a geometrical constraint is also added to the learning algorithm. A practical example shows the effectiveness of the proposed approach in terms of recognition accuracy.**

***Index Terms*—Automatic speaker recognition, Blind source separation, Security monitoring, Cepstral coefficients.**

I. Introduction

In last decades an increasing interest in security systems, in particular of access control systems, has arisen. In specific, a huge number of research efforts was directed to speaker recognition problem. To this aim a large number of Automatic Speaker Recognition (ASR) systems, with different accuracy, are available in literature [1]. ASR architectures are very simple to use, in fact there is need only that the user simply pronounce a word or a sentence and the system can be able to accept or discard him [2].

However, one of the main issue in ASR systems is the presence of different sound sources in the controlled environment. It is in fact very improbable that the user is the sole active source, but, on the contrary, it is plausible that in the monitored environment there are some speaking people, some noisy electronic devices, outdoor noises and other interfering environmental noises. The discrimination from all these active noises could be a very complicated task for the ASR system [3].

Fortunately, in literature there exist several techniques able to separate or extract sources of interest from a mixture of different sound sources [4]–[6]. This problem is known by scientific community as the Blind Source Separation (BSS) problem. The term blind is meaning that no *a priori* information is known neither on the mixing environment nor on the mixed sources. The sole information is about the statistical independence between sources. Since independence is a key concept, the BSS problem is usually solved by applying the Independent Component Analysis (ICA) approach [7], [8]. While proposed algorithms work very well when the mixing environment is an unrealistic instantaneous one, several solutions were proposed to solve BSS in a convolutive environment too [9], but results seem not to be so convincing. Some of these solutions work in time domain, others in frequency domain. Each of them have some advantages and disadvantages, but there is not a unique winning approach [10].

In order to solve the BSS problem in a reasonable amount of time the problem is transformed into the frequency domain: the algorithm solves an instantaneous BSS problem for every frequency simultaneously [11], [12]. Unfortunately in frequency domain two trivial ambiguities occur that could be particular troublesome [8]. The permutation ambiguity is particularly tiresome: when converting a signal back to time domain, contributions from different sources will appear into a single channel, thus destroying the separation achieved in the frequency domain; in addition the scaling indeterminacy at each frequency bin will result in an overall filtering effect of the sources. Different solutions to these problems can be found in literature [13], [14]. In order to solve these indeterminacies a geometrical constraint has been introduced [15] providing good results.

In this paper we propose a preprocessing algorithm using BSS algorithms in reverberant environment, based on the geometric constraint [15], in order to extract the useful speech and then recognize the user. The entire system is shown in Figure 1, where the estimated signal $\hat{s}$ of the source of interest, extracted from noisy mixtures x_1 and x_2, is passed to the ASR module that returns the user s.

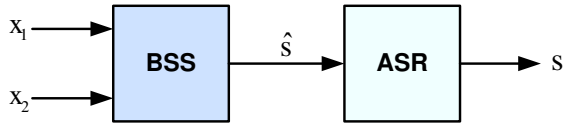

Fig. 1. Proposed system composed of the cascade of a BSS preprocessing step and an ASR step.

The rest of this paper is organized as follows. Section II introduces the basic operation of an ASR system while Section III explains as the BSS algorithm is implemented. Then Section IV shows some experimental results and finally Section V concludes the paper.

978-1-4799-3532-1/14/$31.00 ©2014 IEEE

II. Automatic Speaker Recognition

The ASR system must be able to recognize each person by anatomical differences, the type of vocal tract and habits acquired in the way of speaking. To create a prototype of the voice, the system has to extrapolate the tone of voice and frequency characteristics. In order to obtain these characteristics, each person in the training phase pronounces a particular word or phrase that is saved in a database archive. Later, in the recognition phase, the user is invited to say a word or a phrase. There are two types of ASR systems: if both in the Training phase and the recognition phase, the user must say the same word or phrase, the system is called *Text Dependent* (TD); if there is no need to pronounce the same word or sentence, the system will be called *Text Independent* (TI) [16].

There are many problems during the use of a particular method of voice recognition. Just think, in fact, to some possible background noises, an incorrectly pronounce of the word or phrase (in the case of a TD system), the change in the pitch over the years, and, in particular the health states that may in some way alter the voice. Among the advantages, however, there is above all the convenience of the adoption of a such method of user verification, since the voice recognition is a not invasive method; also both hands and eyes are free to make any other thing during authentication.

A recognition system should be able to manage invariants: the characteristics of the word, the characteristic features of the speaker that must be extracted from speech, while insignificant elements, that do not lead to useful information, could be ignored. In order to achieve a good recognition system, it is necessary to obtain, from the waveform of the word or speech, the peculiar elements that can uniquely distinguish the speaker. Such elements are simply called *features*.

Both in the verification and the identification of a user, the ASR system must be able to accomplish:

1) The feature extraction, that is the process of extracting peculiar characteristics from the signal which can be used for represent each speaker in the database;
2) The correspondence of features, that implies an effective procedure to verify or identify the speaker by comparing the extracted features from the speaker with those of a known set of speakers.

The ASR systems have a first phase of *training* (or learning), where each speaker provides samples of speech in such a way that the system can build or train a model for each user. In the successive phase of *testing* (recognition) a user utters a word, or a phrase, that is compared to all reference models created during the training phase. The system must be able to establish a correspondence between the vocal characteristics, in order to authenticate or identify the user. If a user is new, the system must provide to allocate in memory all the features extracted by the new user and create a new model for the speaker in order to make possible a future test phase.

All the salient passages of the speaker recognition procedure are summarized in the block diagram shown in Figure 2. Generally, an ASR system consists of:

- a set of sensors (microphones) which make the acquisition of data and its subsequent sampling;
- the extraction of the peculiar characteristics (feature extraction), possibly after simple preprocessing techniques;
- the generation of a specific template for each speaker;
- a data-base where all user templates are loaded;
- a comparison (verification) procedure, that determines which profile matches the generated template of the test speech.

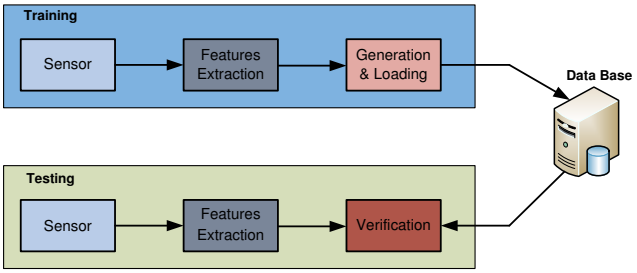

Fig. 2. A typical ASR system.

A. Feature Extraction

Several sets of features are demonstrated to be valid in order to well recognize a speaker, but the most robust set, considered in this paper, is represented by the Mel-Frequency Cepstrum Coefficients (MFCC) [17], [18]. The whole feature extraction procedure for the MFCC coefficients, is graphically shown in Figure 3 and deeply described in the following.

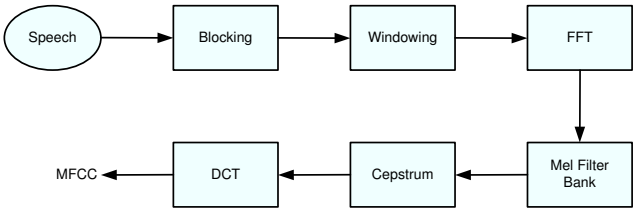

Fig. 3. A typical feature extraction procedure.

In order to perform the Fourier Transform, via the FFT algorithm [19], the speech signal recorded by the microphone arrays, is blocked in frames of 256 samples with an overlap of 128 samples. Then a windowing operation, using a Hamming window, is performed to reduce the Gibss phenomenon.

Psychoacoustic studies [20]–[22] have shown that the mind of perception of the frequency content, over 1 kHz, follows a nearly logarithmic scale. In this way it was introduced the Mel scale, as

$$mel(f) = \begin{cases} f, & \text{if } f \leq 1 \text{ kHz} \\ 2595 \log\left(1 + \frac{f}{700}\right), & \text{if } f > 1 \text{ kHz} \end{cases} \quad (1)$$

and shown in Figure 4. Then the signal is filtered in the frequency domain with a filter bank constructed in the Mel scale, for which the pitches are judged to be equally distant from one other. This filter bank is called Mel Filter Bank. The

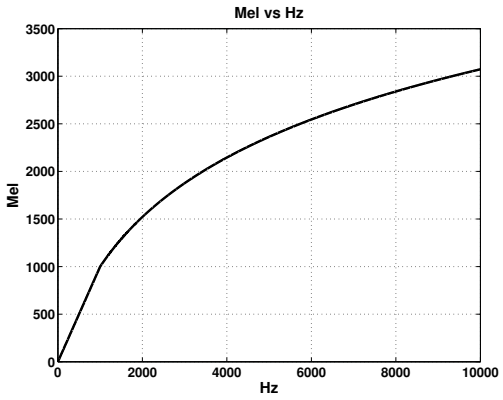

Fig. 4. The Mel scale of frequency.

cepstrum $x_c[n]$ is defined as the inverse Fourier transform of the logarithm of the spectrum of a signal $x[n]$ [19]

$$x_c[n] = \mathrm{FFT}^{-1}\left\{\log\left|\mathrm{FFT}\left\{x[n]\right\}\right|\right\}. \tag{2}$$

The cepstrum transforms the signal from the frequency domain to the *quefrency* domain.

After evaluating the energy X_n of the cepstrum of a speech signal in the mel domain, the MFCC coefficients are estimated through the P-point Discrete Cosine Transform (DCT), by

$$c_k = \sum_{n=0}^{P-1} X_n \cos\left[k\left(n+\frac{1}{2}\right)\frac{\pi}{P}\right],\ k = 0, 1, \ldots, K-1 \tag{3}$$

The spectral envelope is then rebuilt with the first $K_m < K$:

$$C(mel) = \sum_{k=0}^{K_m} c_k \cos\left(2\pi k \frac{mel}{B_m}\right), \tag{4}$$

where B_m is the bandwidth analyzed in Mel domain and $K_m = 20$ is a typical value.

B. Classification

In general terms, the problem of ASR belongs to the branch of pattern recognition [23]. The purpose of pattern recognition is to classify given objects into a number of classes. In the case of an ASR, patterns are the vectors c_k of MFCC coefficients, while the classes are each single speaker. The classification is performed by the features extracted from each speech signal.

It is necessary to convert data from the high-dimensional input space, into a new space of smaller dimension, consisting of small discrete points. The vector quantization (VQ) is the problem of the discretization of a vector space [24]. The quantization of a vector space allows to treat a limited number of data, and associating the points of a region of the input space to a single reference vector. This certainly introduces an error that can be minimized with an appropriate arrangement of all reference vectors. Therefore, a VQ algorithm has the task of determining the best arrangement of the reference vectors with respect to the optimization of a criterion that can be, for example, the reduction of the mean square distance of the points of these vectors. Therefore the objective of VQ is to reduce the size of a database in an efficient manner, i.e. such as to limit the loss of information [25]. To accomplish this task, the VQ algorithm divides the space into continuous regions, called Voronoi regions, and all observations belonging to a region (cluster) are made to coincide in a single point said *centroid*. The goal is to identify a set of centroids with appropriate characteristics, in reduced number compared to the cardinality of the input space and with reduced loss of information. Each centroid is called *codeword*, and in the case of speaker recognition, each speaker can be recognized based on the location of its codeword. The set of codewords forms a *codebook* that characterizes each speaker. An example of Voronoi regions and codewords is shown in Figure 5.

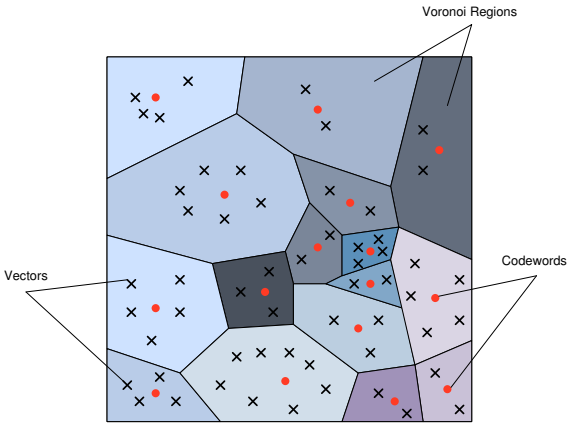

Fig. 5. The Voronoi regions.

The vector quantization associates to each vector $\mathbf{x}$ in the K-dimensional input space, one of the Q (with $Q << K$) vectors $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_Q]$ that forms the codebook. Conversely, to each codeword $\mathbf{c}_i$ it is associated the i-th Voronoi region S_i, whose points satisfy the condition

$$S_i = \left\{\mathbf{x} \in \mathbb{R}^K : \|\mathbf{x} - \mathbf{c}_i\| \leq \|\mathbf{x} - \mathbf{c}_j\|,\ \forall j \neq i\right\}. \tag{5}$$

The training of the codebook is performed by the LBG algorithm proposed in [26].

III. Blind Source Separation

Let us consider a set of N unknown and independent sources $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]^T$, such that the components $s_i(n)$ are zero-mean and mutually independent. Signals received by an array of M sensors are denoted by $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ and are called mixtures. For simplicity we consider the case of $N = M$.

The convolutive model introduces the following relation between the i-th mixed signal and the original source signals

$$x_i(n) = \sum_{j=1}^{N}\sum_{k=0}^{K-1} a_{ij}(k) s_j(n-k), \qquad i = 1, \ldots, M \tag{6}$$

The mixed signal is a linear mixture of filtered versions of the source signals, $a_{ij}(k)$ represents the k-th mixing filter coefficient and K is the number of filter taps. The task is to

estimate the independent components from the observations without resort to a priori knowledge about the mixing system and obtaining an estimate $\mathbf{u}(n)$ of the original source vector $\mathbf{s}(n)$:

$$u_i(n) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} w_{ij}(l) x_j(n-l), \qquad i = 1, \ldots, N \quad (7)$$

where $w_{ij}(l)$ denotes the l-th mixing filter coefficient and L is the number of filter taps.

When a mixing environment is quite complex, filters of the ICA network may require thousands of taps to appropriately invert the mixing. In such cases, the time domain methods have a large computational load to compute convolution of long filters and are expensive for updating filter coefficients. The methods can be implemented in the frequency domain using the Fast Fourier Transform (FFT) [19] in order to decrease the computational load because the convolution operation in the time domain can be performed by element-wise multiplication in the frequency domain. Note that the convolutive mixtures can be expressed as

$$\mathbf{x}(f,k) = \mathbf{A}(f)\,\mathbf{s}(f,k), \quad \forall f \quad (8)$$

where $\mathbf{x}(f,k)$ and $\mathbf{s}(f,k)$ are the frequency components of mixtures and the independent sources at frequency f, respectively. $\mathbf{A}(f)$ denotes a matrix containing elements of the frequency transforms of mixing filters at frequency f. From (8), it is clear that convolutive mixtures can be represented by a set of instantaneous mixtures in the frequency domain. Thus, the independent components can be recovered by applying ICA for instantaneous mixtures at each frequency bin and then transforming the results in the time domain:

$$\mathbf{u}(f,k) = \mathbf{W}(f)\,\mathbf{x}(f,k), \quad \forall f \quad (9)$$

where $\mathbf{W}(f)$ denotes the demixing matrix in the frequency domain. Note that $\mathbf{s}(f,k)$, $\mathbf{x}(f,k)$ and $\mathbf{u}(f,k)$ are vectors of complex elements.

Usually some preliminary preprocessing steps are required in order to simplify the identification of the demixing matrix. The first one preprocessing step is simply the mean value removing or centering. The second step is a canonical whitening preprocessing $\tilde{\mathbf{x}} = \mathbf{Q}\mathbf{x}$, where $\mathbf{Q}$ is an orthogonal matrix, in order to obtain an identity correlation matrix $E\left\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H\right\} = \mathbf{I}$. In the following we suppose that centering and whitening preprocessing are always performed, so we simply denote the whitened mixtures $\tilde{\mathbf{x}}$ with $\mathbf{x}$.

In order to solve the BSS in the convolutive environment Bingham & Hyvarinen in [27] have proposed a complex-valued version of the well-know and best performing FastICA algorithm introduced in [28], [29]. The k-th column $\mathbf{w}_k$ of the $\mathbf{W}(f)$ matrix is obtained by maximizing an approximation $G\left(\left|\mathbf{w}_k^H\mathbf{x}\right|\right)$ of the negentropy function [7], where $G(\cdot)$ is a suitable non-quadratic function, as in the following optimization problem

$$\begin{aligned} &\arg\max_{\mathbf{w}_k} G\left(\left|\mathbf{w}_k^H\mathbf{x}\right|\right) \\ &s.t. \quad \left\|\mathbf{w}_k\right\|^2 = 1 \end{aligned} \quad (10)$$

The constraint in the previous problem is needed in order to avoid the trivial null solution. The solution of (10), can be obtained by the following fixed-point iterations [27]

$$\begin{aligned} \mathbf{w}_k^+ &= E\left\{\mathbf{x}(\mathbf{w}_k^H\mathbf{x})^* g\left(\left|\mathbf{w}_k^H\mathbf{x}\right|^2\right)\right\} \\ &- E\left\{g\left(\left|\mathbf{w}_k^H\mathbf{x}\right|^2\right) + \left|\mathbf{w}_k^H\mathbf{x}\right|^2 g'\left(\left|\mathbf{w}_k^H\mathbf{x}\right|^2\right)\right\}\mathbf{w}_k, \quad (11) \\ \mathbf{w}_k &= \frac{\mathbf{w}_k^+}{\left\|\mathbf{w}_k^+\right\|}, \quad (12) \end{aligned}$$

where $g(\cdot)$ is the derivative of $G(\cdot)$ and $g'(\cdot)$ its second derivative. We adopt the following function $g(y) = 1/(y+a)$, with a a constant value, usually set to $a = 0.1$. Note that it is necessary to choose no learning rates in this algorithm.

Unfortunately, the two scaling and permutation ambiguities of the ICA algorithms results in a undesired and fastidious distorted reconstructed signals, after applying eqs. (11) and (12). This kind of distortions can be hugely reduced using some particular constraints on the learning algorithm.

A. Geometrical Constraints

In order to overcome the ICA ambiguities and reduce the undesired distortions, [15] proposed a geometrical approach to ICA, using the hint that frequency-domain blind source separation is equivalent to a set of frequency-domain adaptive beamformers (ABFs) under certain conditions.

Since the equivalence of the FD-ICA and FD-ABF described in [30], the k-th demixing vector $\mathbf{w}_k$ has to be constrained as follows

$$\mathbf{w}_k^H \hat{\mathbf{h}}_k(f) = c. \quad (13)$$

The estimated steering vector $\hat{\mathbf{h}}_k(f)$, since the true impulse response is not available, is evaluated solely from the time delays of the direct sound, hence, for each bin f, has the form

$$\hat{\mathbf{h}}_k(f) = \begin{bmatrix} 1 \\ e^{-j2\pi\frac{d}{c}f\cos\hat{\theta}_k} \\ \vdots \\ e^{-j2\pi(N-1)\frac{d}{c}f\cos\hat{\theta}_k} \end{bmatrix}, \quad (14)$$

where d is the distance between microphones, c the sound speed and $\hat{\theta}_k$ is the direction of arrival (DOA) estimated by a beamformer. Figure 6 describes the geometry of the sources and the microphone array. Because we are interested only in the direction of the demixing vector $\mathbf{w}_k$ and not to its norm, we project this solution to the constraint in (13), thus the resulting new optimization problem is solved by substituting the normalization in (12) by the following one:

$$\mathbf{w}_k = \frac{\mathbf{w}_k^+}{\left\|\mathbf{w}_k^+\hat{\mathbf{h}}_k\right\|}. \quad (15)$$

Unfortunately, at low frequencies a certain number of permutations still occur. In order to overcome the previous problem at low frequencies, a DOA estimation is performed. Such estimation is done analyzing the directional patterns which

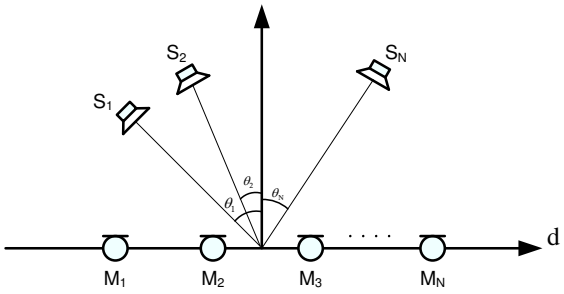

Fig. 6. Assumptions on microphones and sources geometry: M_j is the j-th microphone while S_k is the k-th source.

allow us to associate a single source to a local minimum, as described in [31].

The k-th directional pattern can be expressed, for the non-restrictive case of $N = 2$ sources and $M = 2$ microphones, as follows

$$F_k(f,\theta) = \sum_{l=1}^{2} W_{kl}(f) \exp\left[\frac{j2\pi f d \sin\theta}{c}\right], \qquad (16)$$

where c is the sound speed, d is the distance between microphones and $W_{kl}(f)$ is the l-th entry of the $\mathbf{w}_k$ column vector. These directional patterns have a minimum value in correspondence of an estimated disturbing source [31]. In particular the two source directions are evaluated as

$$\begin{aligned}\theta_1(f) &= \min\left[\arg\min_\theta |F_1(f,\theta)|, \arg\min_\theta |F_2(f,\theta)|\right], \\ \theta_2(f) &= \max\left[\arg\min_\theta |F_1(f,\theta)|, \arg\min_\theta |F_2(f,\theta)|\right].\end{aligned} \qquad (17)$$

Assuming that the minimum value of the pattern $F_1(f_k,\theta)$ is θ_2 for the k-th frequency bin f_k, if for another frequency bin f_j the minimum value is θ_1, then a permutations occurred and the filter coefficients must be swapped. In addition the value of the directional pattern in correspondence of the source, can be used as scaling factor in order to solve the scaling ambiguity. An optional beamformer can be used for the evaluation of the directions $\hat{\theta}_1$ and $\hat{\theta}_2$. The use of the directional patterns allows us to solve both the permutation and scaling ambiguity: the patterns decide to choose the ICA solution or a swapped one, if a permutation occurs. Then a generalization for $N > 2$ sources can be easily derived.

The whole BSS system implementing the geometrically constraint is shown in Figure 7.

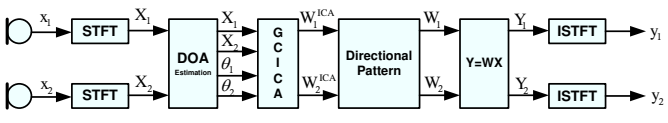

Fig. 7. BSS system using geometrical constraint and directional pattern in the case of $N = M = 2$.

IV. Experimental results

In order to verify the effectiveness of the proposed approach, we have performed several experimental test in different noisy conditions. We have imaged that the proposed system is posed in a standard room of dimensions $5 \times 4 \times 3$ m, and the speaker to be recognized is located in front of the ASR system and very close to it. The noisy source is randomly posed in the range $1 \div 3$ m from the microphone array and an angle $\theta_1 \in [10 \div 170]^\circ$. Then a mean over 100 separate tests are performed. The experimental set-up is shown in Figure 8.

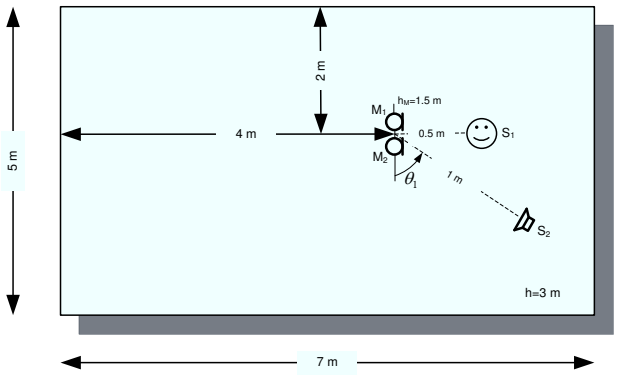

Fig. 8. Experimental set-up used in this paper.

The performances of the proposed architecture are evaluated in terms of Accuracy (Acc), False Rejection Rate (FRR) and False Identification Rate (FIR), defined by:

$$Acc = \frac{N_{TR}}{N_T}, \quad FRR = \frac{N_{FR}}{N_T}, \quad FIR = \frac{N_{FI}}{N_T}, \qquad (18)$$

where N_T is the total number of performed tests, N_{TR} is the number of true recognized speakers, N_{FR} is the number of speakers not recognized, and N_{FI} is the number of speakers recognized but associated to another identity.

To validate the proposed approach a background noise with different *Signal to Noise Ratio* (SNR) in the range $[\infty, -10]$ dB, is added to the present speaker signal. This background noise compromises the performance of the traditional ASR system. When the SNR is high, the accuracies of the traditional and proposed systems are comparable, but when SNR is very low, the accuracy of the standard ASR system is poor, while FRR and FIR is increasing. Using instead the BSS preprocessing step, the proposed system is able to guarantee an acceptable accuracy, even in the presence of a very strong background noise, as shown by results, whose summary can be found in Table I and graphically shown in Figure 9.

V. Conclusions

In this paper we have introduced a preprocessing procedure to enhance the accuracy of a standard ASR system in noisy environment. This step is consisting in a frequency-domain ICA algorithm, performing source separation at microphone array. In order to avoid ambiguities of standard ICA algorithms, a geometric constraint is added. Hence a conventional

SNR	With BSS			Without BSS		
[dB]	Acc [%]	FRR [%]	FIR [%]	Acc [%]	FRR [%]	FIR [%]
∞	96.2	0.3	3.5	96.2	0.3	3.5
30	91.2	0.6	8.2	62.6	1.9	35.5
20	90.6	0.8	8.6	54.5	3.4	42.1
15	90.0	1.0	9.0	42.4	6.3	51.3
10	89.8	1.1	9.1	31.2	10.1	58.7
5	87.8	2.0	10.2	24.0	15.5	60.5
0	85.3	2.9	11.8	12.8	58.2	29.0
-3	84.5	3.3	12.2	9.3	55.5	35.2
-10	82.9	7.0	10.1	2.3	96.0	1.7

TABLE I
SUMMARY OF THE ACCURACY, FRR AND FIR, FOR THE PROPOSED APPROACH WITH AND WITHOUT THE PREPROCESSING BSS ALGORITHM.

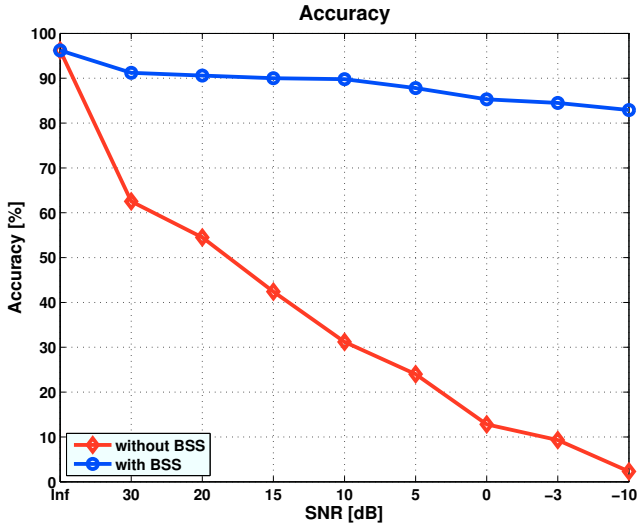

Fig. 9. Accuracy of the proposed approach with and without the preprocessing BSS algorithm at different SNR values.

ASR algorithm based on MFCC coefficients classification is adopted.

Some experimental results show the effectiveness of the proposed approach in terms of accuracy reached by the whole system, and its robustness with respect additive background noise.

REFERENCES

[1] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002)*, vol. 4, Orlando, FL, USA, 13-17 May 2002, pp. 4072–4075.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[3] A. Maesa, F. Garzia, M. Scarpiniti, and R. Cusani, "Text independent automatic speaker recognition system using mel-frequency cepstrum coefficients and gaussian mixture models," *Journal of Information Security*, vol. 3, no. 4, pp. 335–340, October 2012.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. John Wiley, 2002.

[5] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation*. Academic Press, 2010.

[6] S. Haykin, Ed., *Unsupervised Adaptive Filtering, Volume1: Blind Source Separation*. John Wiley & Sons, Inc, 2000.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.

[8] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[9] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Springer, 2007.

[10] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, March 2003.

[11] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, 1999, pp. 365–370.

[12] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Proc. International workshop on Indipendence and Artificial Neural Networks*, Tenerife, Spain, February, 9-10 1998.

[13] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Transaction on Speech and Audio Processing*, vol. 13, no. 1, pp. 1–13, January 2005.

[14] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Transaction on Speech and Audio Processing*, vol. 8, pp. 320–327, 2000.

[15] M. Scarpiniti, F. Di Palma, R. Parisi, and A. Uncini, "A geometrically constrained ICA algorithm for blind separation in convolutive environments," in *Neural Nets WIRN10*, ser. Frontiers in Artificial Intelligence and Applications, B. Apolloni, S. Bassis, A. Esposito, and C. F. Morabito, Eds. Amsterdam: IOS Press, February 2011, vol. 226, pp. 79–88.

[16] H. Beigi, *Fundamentals of Speaker Recognition*. New York: Springer, 2011.

[17] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, April 2012.

[18] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science /& Technology*, vol. 16, no. 6, pp. 582–589, June 2001.

[19] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice-Hall, 1999.

[20] M. Gold, *Speech an Audio Signal Processing*. John Wiley & Sons, 2002.

[21] O'Shaughnessy, *Speech Communication: Human and Machine*. Boston: Addison-Wesley, 1987.

[22] S. Stevens, J. Stanley, and E. B. Volkman, J. andNewman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, March 1937.

[23] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[24] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, 1985, pp. 387–390.

[25] S. Singh and E. G. Rajan, "Vector quantization approch for speaker recognition using MFCC and inverted MFCC," *International Journal of Computer Application*, vol. 17, no. 1, pp. 1–7, March 2011.

[26] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[27] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex-valued signals," *International Journal of Neural Systems*, vol. 10, no. 1, pp. 1–8, 2000.

[28] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[29] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[30] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *Journal on Applied Signal Processing*, vol. 11, no. 1, pp. 1157–1166, 2003.

[31] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 666–678, March 2006.