International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:4, 2017

Forensic Speaker Verification in Noisy Environmental by Enhancing the Speech Signal Using ICA Approach

Ahmed Kamil Hasan Al-Ali, Bouchra Senadji, Ganesh Naik

***Abstract*—**We propose a system to real environmental noise and channel mismatch for forensic speaker verification systems. This method is based on suppressing various types of real environmental noise by using independent component analysis (ICA) algorithm. The enhanced speech signal is applied to mel frequency cepstral coefficients (MFCC) or MFCC feature warping to extract the essential characteristics of the speech signal. Channel effects are reduced using an intermediate vector (i-vector) and probabilistic linear discriminant analysis (PLDA) approach for classification. The proposed algorithm is evaluated by using an Australian forensic voice comparison database, combined with car, street and home noises from QUT-NOISE at a signal to noise ratio (SNR) ranging from -10 dB to 10 dB. Experimental results indicate that the MFCC feature warping-ICA achieves a reduction in equal error rate about (48.22%, 44.66%, and 50.07%) over using MFCC feature warping when the test speech signals are corrupted with random sessions of street, car, and home noises at -10 dB SNR.

***Keywords*—**Noisy forensic speaker verification, ICA algorithm, MFCC, MFCC feature warping.

I. Introduction

AUTOMATIC speaker recognition systems face numerous challenges in real forensic situations. Firstly, covert audio recordings are often corrupted with various types of real environmental noise [1]. The police agencies often use hidden microphones to record the speech from the criminals in public places. Such forensic audio recordings may be far away from the hidden microphones and these recordings are often corrupted with car, and street noises. Secondly, channel mismatch often occurs in forensic situations. For example, enrolment speech may be recorded by using a microphone, while the test speech signals are recorded by using telephone intercept from a mobile phone [2]. Finally, the speech samples from the suspect may be very short duration and they may not have enough information to verify the identity of the speaker [3]. These factors lead to decreased the performance of speaker verification for forensic applications. Choosing the most reliable methods for speech enhancement play an important role in noisy forensic speaker recognition systems.

Speech enhancement algorithms can be classified into single and multiple channel algorithms based on the number of the microphones are used for collecting the noisy speech signals. Various algorithms for single channel speech enhancement have been proposed in the previous studies [4], [5], but these techniques achieve less improvement in speech quality compared with multiple channel speech enhancement algorithms [6].

Independent component analysis (ICA) was used in multi-channel speech enhancement algorithm to remove the noise from the noisy speech signals [7], [8]. It is based on transforming the mixed signals into components which are statistically independent [9]. The principle of estimating independent component is based on maximizing the non-Gaussian distribution of one of the source signals. The difference between the Gaussian distribution and the distribution of the independent component can be estimated by using various non-Gaussian measurements such as Kurtosis, negentropy, and approximation of the negentropy [9].

Although independent component analysis was used to suppress the noise from the noisy speech signal in previous studies [7], [8], [10], The effectiveness of this algorithm has not investigated yet to improve the performance of noisy forensic speaker verification system in state-of-the-art i-vector and PLDA for short noisy test durations (10 seconds). This is the main original contribution of this research.

The second contribution of this research is to design a benchmark for noisy forensic speaker verification systems by using forensic and QUT-NOISE databases [11]. In real forensic situations, the speech segments from different suspects are often corrupted with different sessions of real environmental noise. Previous studies [1], [12] have used a limited duration of noise corpus (NOISEX-92 database) [13] to mix with existing clean speech databases for evaluation of noisy forensic speaker recognition. However, while the large number of speakers in speech corpus available to the researchers for forensic applications allow a wide variety of speakers to be evaluated for speaker verification systems, limited duration of noise corpus lacks the ability to evaluate forensic speaker verification under real environmental noise. Therefore, in this research, we mix a random session of noise from QUT-NOISE database with clean forensic audio recordings to more closely approach in forensic situations. It is important to use realistic forensic and noise databases to evaluate noisy forensic speaker recognition because law enforcement agencies and police investigations can use this algorithm to improve the robustness of speaker verification systems in the casework conditions.

In this work, we propose a new approach for noisy forensic

Ahmed Kamil Hasan Al-Ali and Bouchra Senadji are with the School of Electrical and Computer Engineering, Queensland University of Technology, Brisbane, Australia (e-mail: ahmedkamilhasan.alali@hdr.qut.edu.au, b.senadji@qut.edu.au).
Ganesh Naik is with the University of Technology, Sydney, Australia (e-mail: Ganesh.Naik@uts.edu.au).

speaker recognition systems. This approach is based on using independent component analysis algorithm to reduce the effect of real environmental noise. The enhanced speech signals from ICA algorithm are applied to MFCC or MFCC feature warping to capture the essential characteristics of the speech signal. The features extracted from MFCC or MFCC feature warping can be used to train the state of the art i-vector and PLDA speaker verification systems. The performance of this approach is compared with traditional feature extraction technique (MFCC and MFCC feature warping).

II. Model of ICA

Let the speech and noise signals emitted from N sources be represented as $s(t) = \{s_1(t), s_2(t), ..., s_N(t)\}$. The noisy speech signals can be recorded instantaneously by using M microphones in a street for forensic applications and be expressed as $x(t) = \{x_1(t), x_2(t), \cdots, x_M(t)\}$. Instantaneous ICA can be defined as a linear transformation of noisy speech signals into components which are statistically independent, and can be represented as [9]

$$x = As \tag{1}$$

where A is an unknown mixing matrix.

The goal of ICA is to estimate the original sources from the mixed signals. The estimates of speech and noise signals $(\hat{s})$ can be represented by the following equation [9]

$$\hat{s} = Wx \tag{2}$$

where W is the unmixing matrix which equals the inverse of the mixing matrix A when the matrix is square.

In this paper, we use two sources (speech and noise) and two microphones to record the noisy speech signals$(M = N = 2)$. Therefore, the mixing and unmixing matrices are square and they have a size of 2×2.

A. Fast ICA Algorithm

The procedure for a fast ICA algorithm for one unit can be illustrated by the following steps [9]:

1) Remove the mean value from the noisy signal and center its distribution.
2) Whiten the noisy speech signal (x) to get (x_w) by using eigenvalue decomposition of the covariance of the noisy speech signal.

$$x_w = VD^{-1/2}V^T x \tag{3}$$

where V is the eigenvector matrix of the covariance of the noisy speech signal, and $D^{-1/2}$ is the inverse square root diagonal of the eigenvalue matrix.

3) Choose an initial vector of unmixing matrix W.
4) Estimate a row vector of unmixing matrix

$$w^+ = E\{x_w\, g(w^T\, x_w)\} - E\{g'(w^T\, x_w)\}w \tag{4}$$

where w^+ is the new value of the row vector of the unmixing matrix, E is the sample mean, g and g' are the first and the second derivatives of the contrast function respectively.

5) Normalize the row vector of w^+

$$w^* = \frac{w^+}{\|w^+\|} \tag{5}$$

where w^* is the normalization of the new row vector of the unmixing matrix.

6) Insert $w = w^*$ in step 4 and repeat the procedure until there is convergence.

The criterion of convergence is that the direction of previous and new values of w must be in the same direction, i.e. the dot product of these w points is almost equal to one.

This algorithm is based on separating one non-Gaussian component each time under the assumption that the sum of the others has a Gaussian distribution. It is necessary to prevent different row vectors of w from converging to the same maxima and this can be performed by using a deflation decorrelation of the output $w_1^T x, w_2^T x, \cdots, w_p^T x$ after every iteration.

III. Proposed Speaker Verification System

Forensic audio recordings are often corrupted by real environmental noises. These noises decrease the performance of the speaker verification systems. It is important to reduce the effect of real environmental noise from the noisy speech signals before using these speech signals in speaker verification systems. Therefore, using speech enhancement algorithm (independent component analysis) in our algorithm may improve the performance of noisy speaker verification systems under high noise conditions.

The proposed algorithm can be described by the following steps. Firstly, the noisy speech signals are applied to independent component analysis algorithm to suppress the noise and improve the performance of the noisy speaker verification systems. Mel frequency cepstral coefficient or MFCC feature warping is used to extract the essential features of the enhanced speech signals. These features can be used to train i-vector PLDA speaker verification framework. The flowchart of the proposed speaker verification system in the presence of noise can be illustrated in Fig. 1.

IV. Speaker Verification System

The speaker verification system used state of the art i-vector PLDA speaker verification systems. Mel frequency cepstral coefficient (MFCC) or MFCC feature warping was used as the front end. The MFCCs were computed over Hamming window frame with 30 msec and 10 msec shift. The MFCCs were obtained using mel filterbank of 32 channels, followed by transformation to the cepstral domain, keeping 13 coefficients. The first and second derivatives of the cepstral coefficients were appended to the MFCCs. Feature warping was used to transform the distribution of the cepstral feature into a standard normal distribution over sliding window that typically spans 301 frames.

The Universal background model (UBM) with 256 components was used in our experimental results. The UBM was trained on telephone and microphone speech using 348 speakers from the forensic voice comparison database and used

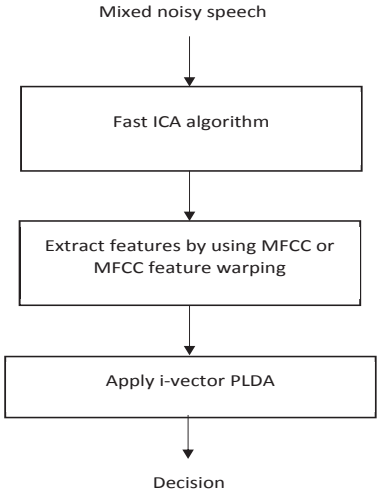

Fig. 1 Proposed speaker verification in the presence of noise

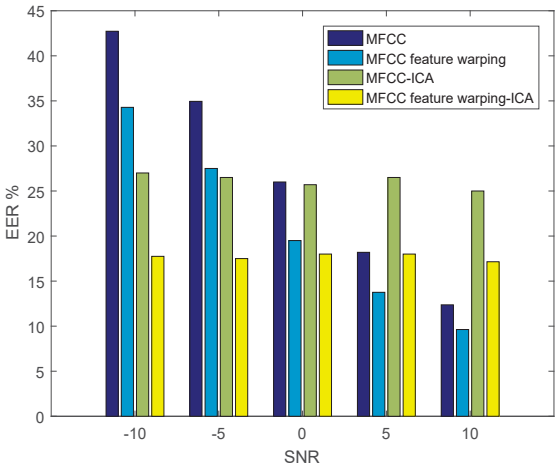

Fig. 2 Comparison of the proposed speaker verification systems with other techniques when the test speech signals were corrupted with street noise

to compute the Baum-Welch statistic for calculation of total variability space of dimension 400. These total variabilities were used to calculate the i-vector speaker representation. The i-vector dimension was reduced to 200 eigenvectors by using linear discriminant analysis (LDA). I-vectors were length normalized before Gaussian probabilistic linear discriminant (GPLDA) training using i-vector centering followed by the whitening approach. Scoring was performed using the batch likelihood ratio. The performance of i-vector PLDA speaker verification system was evaluated using MSR identity toolbox by Microsoft [14].

V. Experimental Configuration

The speaker verification system was evaluated using the forensic voice comparison database of 500+ Australian English speakers [15]. This database consists of 552 speakers. Each speaker was recorded in three speaking styles: informal telephone conversations, information exchange over the telephone and pseudo-police styles [2]. Informal telephone conversations and information exchange over the telephone were recorded between two speakers by using a telephone channel. For the pseudo-police style interview, each speaker was interviewed by an interviewer sitting in the same room as the speaker. In this task, microphones were used instead of the telephone used in informal telephone conversations and information exchange over the telephone styles. The clean speech signal was sampled at 44.1 kHz and 16 bit/sample resolution [2].

The enrolment of the speech signals was obtained by using full duration utterances from 200 speakers (100 males and 100 females) from the pseudo-police style. The test speech signals were obtained by using 10 seconds duration from 200 speakers in informal telephone conversation style after removing the silence region by using voice activity detection (VAD). Voice activity detection based on the statistical model [16] was used in this research to remove the silence frames from the enrolment and test speech signals. Each of the test speech signal was mixed with one random session of the real environmental noises (car, street, and home noise) from QUT-NOISE database [11], resulting in a two channel noisy speech signal. These noises were sampled at 48 kHz and 16 bit sample resolution and they were down sampled from 48 kHz to 44.1 kHz before mixing with the clean speech signal. The down sampled is necessary to match the sampling frequencies of the clean speech and noise signals. The noisy speech signals in an ICA algorithm can be represented as

$$x_1 = s(n) + e(n) \tag{6}$$

$$x_2 = s(n) + 0.6 * e(n) \tag{7}$$

The noisy speech signals were separated using fast ICA algorithm and the contrast function which was used in ICA algorithm has Gaussian function and it can be represented as [9]

$$G(u) = -\exp\left(\frac{-u^2}{2}\right) \tag{8}$$

The enhanced speech signals from ICA algorithm was applied to MFCC or MFCC feature warping techniques to extract the features extraction. Theses features were used to train the state of the art i-vector PLDA speaker verification systems.

VI. Results

In this section, we present the simulation results of the speaker recognition systems when the enhanced speech signals from ICA algorithm were applied to MFCC or MFCC feature warping technique. This system was compared with feature extraction techniques (MFCC, and MFCC feature warping). The equal error rate (EER) was used to evaluate the performance of noisy speaker verification systems. Figs. 2-4 show the comparison of the proposed speaker verification systems with other techniques when the test speech signals were corrupted with street, car, and home noise respectively. From Figs. 2-4 we conclude the following points:

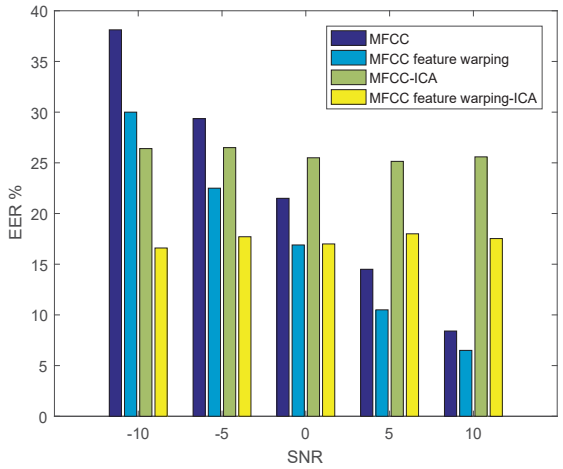

Fig. 3 Comparison of the proposed speaker verification systems with other techniques when the test speech signals were corrupted with car noise

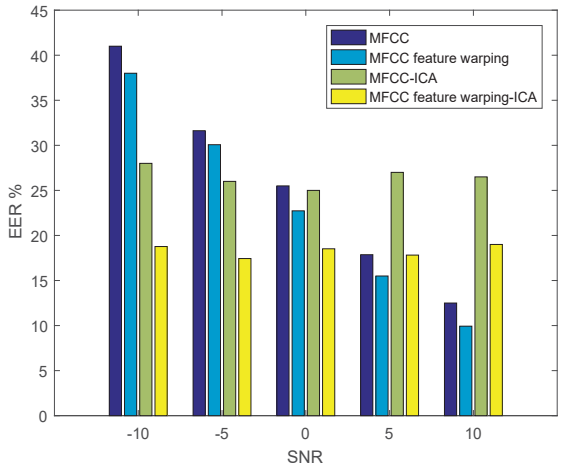

Fig. 4 Comparison of the proposed speaker verification systems with other techniques when the test speech signals were corrupted with home noise

- Mel frequency cepstral coefficient feature warping achieves improvement in EER, compared with MFCC only, when the test speech signals were corrupted with street, car, and home noise for input SNR ranging from -10 to 10 dB. At -10 dB, the reduction in EER of MFCC feature warping over MFCC is about (19.76%, 21.30%, 7.31%) when the test speech signals were corrupted with street, car, and home noise respectively.
- Independent component analysis achieves significant improvement in EER compared with using traditional feature extraction techniques(MFCC and MFCC feature warping) at low SNR. At -10 dB, the reduction in EER of the MFCC feature warping-ICA over MFCC feature warping is about(48.22%, 44.66%, and 50.07%) when the test speech signals were mixed with street, car, and home noise respectively. The performance of the proposed algorithm degrades compared with conventional feature extraction at high SNR.

VII. Conclusion

This paper has presented the effectiveness of ICA algorithm to reduce the real environmental noise and improve the performance of speaker verification systems. This method is based on suppressing the effect of the noise by using ICA algorithm. The enhanced speech signals were applied to MFCC or MFCC feature warping to extract the essential characteristics of the speech signals. A state of the art i-vector PLDA speaker framework was used in this research to reduce the effect of channel mismatch. Experimental results demonstrate that independent component analysis is robust to high levels of noise compared with traditional cepstral feature techniques. The proposed speaker verification systems achieve a significant reduction in EER over using MFCC and MFCC feature warping when the test speech signals were corrupted with various types of real environmental noise at low SNR (-10 dB to -5 dB).

References

[1] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4249-4252.
[2] G. S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice,"*Australian J. Forensic Sci.*, vol. 44, pp. 155-167, 2012.
[3] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Process. Mag.*, pp. 95-103, 2009.
[4] Berouti , M., Schwartz, R. and Makhoul, J., "Enhancement of speech corrupted by acoustic noise", *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 1979, pp. 208-211.
[5] Donho, D.L and Johnston, I.M., "Ideal spatial adapation by wavelet shrinkage", *Biometrika J.*, vol. 81, pp. 425-455,1994.
[6] A. K. H. AL-ALI, D. Dean, B. Senadji, and V. Chandran,"Comparison of speech enhancement algorithms for forensic applications,"in *16th Speech science and technology conference*, Sydney, 2016.
[7] H. Liang, J. Rosca, and R. Balan, "Independent component analysis based single channel speech enhancement," in *3rd IEEE Int. Symp. Signal Process. Inform. Technology*, 2003, pp. 522-525.
[8] H. Li, H. Wang, and B. Xiao, "Blind separation of noisy mixed speech signals based on wavelet transform and Independent Component Analysis," in *8th Int. Conf. Signal Process.*, 2006.
[9] Hyvarinen, A. and Oja, E., "Independent component analysis: algorithms and applications", *Neural Netw.*, vol. 13, no. 4, pp. 411-430, 2000.
[10] H.-y. Li, Q.-h. Zhao, G.-l. Ren, and B.-j. Xiao, "Speech Enhancement Algorithm Based on Independent Component Analysis," in *5th Int. Conf. Natural Computation*, 2009, pp. 598-602.
[11] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 26-30.
[12] R. S. Holambe and M. S. Deshpande, "Noise Robust Speaker Identification: Using Nonlinear Modeling Techniques," in *Forensic Speaker Recognition*, Ed: Springer, 2012, pp. 153-182.
[13] A. Varga and H. J. M. Steeneken,"Assessment for automatic speech recognition:II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247-251, 1993.
[14] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox - A matlab toolbox for speaker recognition research", Microsoft Research,Conversational Systems Research Center (CSRC), 2013.
[15] G. S. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. K. Folky, S. Desouza, N. Cumminus, D. Chow. (2015). Forensic database of voice recordings of 500+ Australian English speakers. (Available: http//databases.forensic-voice-comparison.net/).
[16] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Pocess. Lett.*, vol. 6, no.1, pp. 1-3, Jan. 1999.