Evolutionary Intelligence (2021) 14:1475–1494
https://doi.org/10.1007/s12065-020-00406-8

RESEARCH PAPER

Enhanced forensic speaker verification performance using the ICA-EBM algorithm under noisy and reverberant environments

Ahmed Kamil Hasan Al-Ali[1] · Vinod Chandran[2] · Ganesh R. Naik[3]

Received: 27 May 2019 / Revised: 5 March 2020 / Accepted: 7 April 2020 / Published online: 7 May 2020

Abstract

Forensic speaker verification performance reduces significantly under high levels of noise and reverberation. Multiple channel speech enhancement algorithms, such as independent component analysis by entropy bound minimization (ICA-EBM), can be used to improve noisy forensic speaker verification performance. Although the ICA-EBM was used in previous studies to separate mixed speech signals under clean conditions, the effectiveness of using the ICA-EBM for improving forensic speaker verification performance under noisy and reverberant conditions has not been investigated yet. In this paper, the ICA-EBM algorithm is used to separate the clean speech from noisy speech signals. Features from the enhanced speech are obtained by combining the feature-warped mel frequency cepstral coefficients with similar features extracted from the discrete wavelet transform. The identity vector (i-vector) length normalized Gaussian probabilistic linear discriminant analysis is used as a classifier. The Australian Forensic Voice Comparison and QUT-NOISE corpora were used to evaluate forensic speaker verification performance under noisy and reverberant conditions. Simulation results demonstrate that forensic speaker verification performance based on ICA-EBM improves compared with that of the traditional independent component analysis under different types of noise and reverberation environments. For surveillance recordings corrupted with different types of noise (CAR, STREET and HOME) at − 10 dB signal to noise ratio, the average equal error rate of the proposed method based on ICA-EBM is better than that of the traditional ICA by 12.68% when the interview recordings are kept clean, and 7.25% when the interview recordings have simulated room reverberations.

Keywords Independent component analysis · ICA-EBM · Noisy and reverberant environments · Forensic speaker verification systems

1 Introduction

The purpose of forensic speaker verification is to investigate a suspect or confirm a judgment of guilt or innocence by analyzing their speech signals [1]. The speech signal from a criminal is compared with a corpus of speech signals of known suspects in forensic speaker verification to prepare legal evidence for the court [2].

Speaker verification systems face many challenges in real forensic situations. Reverberation often occurs when interview speech signals from the suspect are recorded in a police interview office. In reverberation environments, the interview speech signals are often combined with a multiple reflection version of the speech due to the reflection of the interview speech signals from the surrounding room. The reverberated interview speech can be modeled by the convolution of the impulse response of the room with the interview speech signal [3]. The amount of reverberation can be characterized by (T_{20} or T_{60}) which describes the amount of time for the direct sound to decay by 20 dB or 60 dB, respectively [4]. The surveillance speech signals from criminals are usually recorded using hidden microphones in public places. Such forensic surveillance data are usually mixed with different types of environmental noise [5].

✉ Ahmed Kamil Hasan Al-Ali
ahmedkamilhasan.alali@hdr.qut.edu.au

[1] Department of Electromechanical Engineering, University of Technology, Baghdad, Iraq

[2] Queensland University of Technology, Brisbane, QLD 4001, Australia

[3] MARCS Institute, Western Sydney University, Sydney, NSW 2747, Australia

The distortion of speech by environmental noise and reverberation conditions severely degrades speaker verification performance [6].

Speech enhancement algorithms can be divided into single channel and multiple channel algorithms based on the number of microphones used for recording the noisy speech signal [5]. Multiple channel speech enhancement algorithms improve the quality of noisy speech signals compared with single channel speech enhancement algorithms [7]. Multiple channel speech enhancement algorithms can be used to improve speaker recognition performance under noisy and reverberant environments [6, 8].

Beamforming techniques were used widely as multiple channel speech enhancement algorithms, and they are based on using microphone arrays for directional speech signal transmission only. The generalized sidelobe canceller techniques were proposed in [9, 10] as effective adaptive beamformer approaches which improve the gain of the desired speech signals and suppresses the interference signals by forming a main lobe toward the direction of arrival of the desired speech signals. The performance of the generalized sidelobe canceller is based on the steering vector and block matrix estimate. These parameters require some knowledge about the speech and noise direction which are unknown in real-world applications. The enhanced speech signals from the generalized sidelobe canceller could be affected by speech leakage if the steering vector is estimated inaccurately [11]. Spectro-temporal filtering technique was used as a multiple channel speech enhancement algorithm in [12]. This technique is based on estimating the power spectral density (PSD) of the speech and noise signals. The enhanced speech signal in spectro-temporal filtering can be obtained by applying the parametrized multichannel non-causal Wiener filter to the input microphone array. The speech presence probability approach was used to estimate the PSD of noise and this approach assumed the speech and noise components as multivariate Gaussian distribution [12]. In fact, the speech signals are usually either super-Gaussian or skewed distribution in nature [13]. Thus, modeling speech signals with multivariate Gaussian distribution in spectro-temporal filtering could estimate the PSD of the noise inaccurately and lead to distortion in enhanced speech signals.

Independent component analysis (ICA) is used widely as a multiple channel speech enhancement algorithm [14–16]. ICA is used to separate the source signals (speech and noise) by transforming the noisy signals into components which are statistically independent. The principle of estimating independent components is based on maximizing the contrast function of one independent component. Various contrast functions such as negentropy, Kurtosis and approximation of negentropy have been used to separate the mixed signals by estimating the difference between the distribution of the independent component and the Gaussian distribution [17].

Various algorithms of ICA have been proposed in previous studies such as Fast ICA [18], information maximization [19] and efficient fast ICA (EFICA) [20]. These algorithms used a fixed nonlinear contrast function or model which makes them computationally attractive for estimating source signals. However, the quality of separation of the source signals in the ICA algorithm degrades when the density of the source signals deviates from the assumed underlying model. Recently, the ICA-EBM algorithm has been used as an effective technique for source separation [13]. It is based on calculating the entropy bounds from four contrast measuring functions (two odd and two even functions) and choosing the tightest maximum entropy bound. The tightest entropy bound is the one closest to the entropy of the true source, and it can be used to estimate the entropy of the source signals. The ICA-EBM algorithm can be used to separate the source signals that come from different distributions and achieve superior separation performance to other ICA algorithms [13]. Thus, we use the ICA-EBM as a speech enhancement algorithm for separation of the speech from the noisy speech signals.

In this paper, we propose a forensic speaker verification system based on the ICA-EBM algorithm to improve forensic speaker verification performance in the presence of different types of environmental noise and reverberation conditions. The ICA-EBM algorithm is used to separate the clean speech from noisy speech signals. A fusion of feature warping with MFCC and DWT-MFCC is used to extract the features from the enhanced speech signals. These features are used to train a modern i-vector based speaker verification system.

Although the ICA-EBM algorithm was used to separate mixed speech signals under clean conditions [13], the effectiveness of using the ICA-EBM to separate speech from noise has not been investigated yet for improving state-of-the-art i-vector forensic speaker verification performance in the presence of various types of environmental noise and reverberation conditions. This is the original contribution of this research.

The structure of the paper is organized as follows: Sect. 2 presents a model of independent component analysis. Speech and noise data set describe in Sect. 3. Section 4 presents the construction of noisy and reverberant data. The proposed approach of speaker verification based on the ICA-EBM is described in Sect. 5. Section 6 presents the simulation results, and Sect. 7 concludes the paper.

2 Model of independent component analysis

In real forensic situations, police agencies often record the interview speech from a suspect in a police interview room where reverberation is usually present. However, the surveillance speech signals from the criminal are usually recorded using hidden microphones in open areas. These surveillance speech signals are often mixed with different types of environmental noise [21]. The effect of reverberation is not as much of a problem as environmental noise when the surveillance forensic recordings are made in open areas. Thus, we assume that using instantaneous ICA in forensic speaker verification could separate the environmental noise from the noisy surveillance recordings and improve forensic speaker verification performance under noisy environments. Let the source signals (speech and noise) emitted from n sources be represented as $s(t) = \{s_1(t), s_2(t), \dots, s_n(t)\}$. These signals can be recorded using m microphones and the observed noisy signals represented as $x(t) = \{x_1(t), x_2(t), \dots, x_m(t)\}$. Instantaneous ICA assumes the mixing process is linear so that the observed noisy surveillance speech signals can be represented as:

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

where A is mixing matrix.

The purpose of ICA is to estimate the source signals from the observed noisy surveillance signals when both sources and the parameters of the mixing matrix are unknown. The estimate of the source signals, $\hat{s}$, can be represented by:

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

where W is the unmixing matrix. The unmixing matrix can be defined as:

$$W = A^{-1} \tag{3}$$

where A^{-1} is the inverse of the mixing matrix and the ICA assumes the number of observation and source signals is equal [22].

To simplify the ICA algorithm, preprocessing must be applied before using the fast ICA algorithm and can be divided into two stages: centering and whitening. Centering can be performed by removing the mean from the noisy surveillance recordings so that the noisy recordings (x) have a zero mean. Whitening is another preprocessing stage in the ICA algorithm. It involves linearly transforming the noisy speech signal such that its components are uncorrelated and their variances equal unity [17]. The noisy surveillance recordings can be whitened by

$$x_w = \mathrm{V}D^{-1/2}\mathrm{V}^T x \tag{4}$$

where V is the matrix of eigenvectors of the covariance of the noisy surveillance recordings x, and $D^{-1/2}$ is the inverse square root diagonal matrix of eigenvalues.

2.1 Fast ICA

Hyvärinen proposed the fast ICA algorithm [18]. Fast ICA for one unit is a simple algorithm to estimate one row vector of the unmixing matrix by finding the maximum non-Gaussian value of one independent component. There are four steps to estimate one unit for the fast ICA algorithm:

1. Select an initial guess for w.
2. Estimate $w^+ = E\{x_w\, g(w^T x_w)\} - E\{g'(w^T x_w)\}w$

   where w^+ is the new row vector of the unmixing matrix, E is the sample mean, and the first and the second derivatives of the contrast function can be represented by g and g' respectively. The Gaussian contrast function is used in this work and is defined as:

   $$G(u) = -exp(-au^2/2) \tag{5}$$

   where a is a constant equal to one.
3. Let $w^* = \frac{w^+}{\|w^+\|}$
4. If not converged, return to step 2.

The criterion of convergence is that the dot product between previous and new row vectors of the unmixing matrix is almost equal to one. To estimate all source signals, the one unit fast ICA algorithm runs for n times and the deflation decorrelation was used in this work to prevent different row vectors of the unmixing matrix from converging to the same maxima.

3 Speech and noise data set

This section describes the Australian Forensic Voice Comparison (AFVC) and QUT-NOISE databases which will be used to construct the noisy and reverberant data described in the next section.

3.1 AFVC database

The AFVC database [23] contains 552 speakers recorded in three speaking styles. The speaking styles for each speaker include informal telephone conversation, pseudo-police interview and information exchange over the telephone. A telephone was used to record the information exchange over the telephone and informal telephone conversation styles, while a microphone was used to record the pseudo-police interview style. The sampling frequency of the speech data for all speaking styles was 44.1 kHz with 16 bit/sample resolution [24]. The AFVC database is used in the proposed system because this database contains different speaking styles for each speaker and these speaking styles are often found in most real forensic situations. We used informal telephone conversation and pseudo-police interview styles in this paper because in most real forensic scenarios, the enrolment forensic audio recordings are often recorded in a police interview room and the criminal may use a mobile phone in public places to commit a criminal offence [25]. A brief description of speaking styles used in this paper is given below:

1. Informal telephone conversation style

   Informal telephone conversation style was recorded when two speakers (friends or colleagues) talked through telephone for 10 min about whatever they want. Thus, it is easy to record informal conversation with minimal instruction.
2. pseudo-police interview style

   This style is a live interview and it needs three researchers for recording pseudo-police style, one to interview one speaker, a second to simultaneously interview the other speaker and third to monitor the recording quality.

   The success of the interview task depends on the skill of the interviewers. Questions asked are based on the response to the previous questions. For real modern police interview style, the police asked the suspect to describe events, initially eliciting mostly uninterrupted monologues. Questions asked in the interview are to help the suspect keep speaking and give more information. Narration of the past events and opening questions are most commonly questions in pseudo-police interview style.

3.2 QUT-NOISE database

The QUT-NOISE database [26] consists of 20 noise sessions. The duration of each session is approximately 30 min. QUT-NOISE was recorded in five common noise (CAR, STREET, HOME, CAFE, and REVERB). The noise was sampled at 48 kHz with 16 bit/ sample resolution. We used three types of noise (CAR, STREET, and HOME) from the QUT-NOISE database because these types of noise are more likely to occur in real forensic scenarios [21]. A brief description of the noise scenario used in this paper is given below:

1. STREET

   The STREET noise was recorded in two locations: an inner-city and outer-city. The inner-city recordings consist of pedestrian traffic and bird noise. The outer-city recordings consist mainly of cycles of traffic noise and traffic light changes.
2. CAR

   The CAR noise was recorded in driving window-down and window-up conditions. These recordings consist of car-interior noise such as bag movement, keys, and indicators, as well as characteristics of the wind for car window-down.
3. HOME

   This noise was recorded in two locations for home: kitchen and living room. The kitchen noise consists of silence interrupted by kitchen noise. The living room was recorded in the presence of children singing and playing alongside a television.

For most speaker recognition algorithms [27–29], the clean speech signals from existing speech corpora are mixed with short periods of environmental noise collected separately and the SNR is varied as desired using the mixing coefficients. However, while a large number of speakers in the speech corpora available to researchers through these algorithms allow a wide variety of speakers to be evaluated for speaker verification systems, most existing noise corpora such as NOISEX-92 [30], freesound.org [31] and AURORA-2 [32] have limited duration (less than 5 min). The limited duration of noise corpus has hindered the ability to evaluate speaker verification systems in a wide range of environmental noise in real forensic scenarios. Therefore, we mixed a random session of environmental noise from the QUT-NOISE database with clean speech signals from the AFVC database resulting in two-channel noisy speech signals to achieve a closer approximation to real forensic scenarios.

4 Construction of noisy and reverberant data

The robustness of i-vector speaker verification based on the ICA-EBM algorithm under conditions of noisy surveillance and/or reverberant interview data cannot be evaluated by using the clean speech signals from the AFVC only. Thus, we designed new data to evaluate forensic speaker verification performance based on the ICA-EBM algorithm when surveillance data are mixed with different types of environmental noise and/or interview data reverberated. Firstly, we designed a multi-channel noisy surveillance speech data by mixing clean telephone speech from the AFVC corpus for surveillance with noise from the QUT-NOISE corpus [26]. Secondly, the reverberant interview data was generated by convolving impulse response of the room with the microphone data from the AFVC corpus for interview.

4.1 Simulating multi-channel noisy surveillance data

The aim of designing the multi-channel noisy surveillance speech signals was to evaluate the effectiveness of i-vector speaker verification based on the ICA-EBM algorithm when surveillance recordings are mixed with different types of noise only. We extracted the surveillance recordings from short segments (10–40 s) of 200 speakers using the informal telephone conversation style after applying the VAD algorithm [33] to remove silent regions. We applied the VAD algorithm to clean speech rather than noisy surveillance recordings in this work because noisy surveillance recordings can be manually segmented in forensic applications when encountering noisy speech [2].

In most real forensic situations, the criminal may use mobile to commit criminal offenses. The surveillance recordings from the criminal are usually corrupted by different types of environmental noise in open areas. Thus, the effect of reverberation on noisy surveillance recordings will not be investigated in this work [34]. Each segment of the surveillance data was corrupted by one random segments of the environmental noises (HOME, CAR, and STREET) from the QUT-NOISE corpus [26], resulting in a two-channel noisy surveillance recordings at SNRs ranging from − 10 to 10 dB. The sampling frequency of the noise was 48 kHz and down-sampled to 44.1 kHz before mixing with the clean speech signal to match the sampling frequency with the surveillance speech signals.

Figure 1 shows the configuration of sources (z(n) and e(n)) and microphones (x_1 and x_2) in instantaneous ICA algorithm. The observed noisy surveillance data recorded by the microphones, x, can be modeled as follows:

$$x = As(n) \tag{6}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \mathrm{z(n)} \\ \mathrm{e(n)} \end{bmatrix} \tag{7}$$

$$A = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 0.6 \end{bmatrix} \tag{8}$$

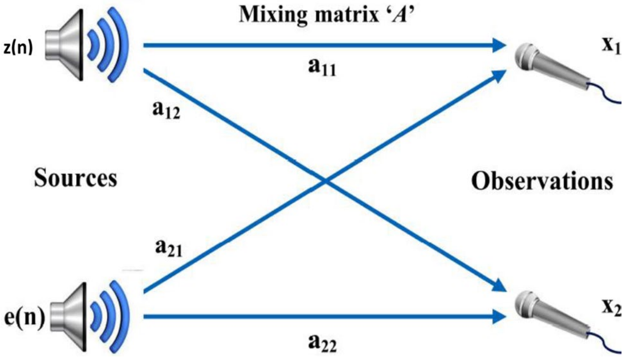

Fig. 1 Configuration of sources and microphones in instantaneous ICA mixtures

where z(n) is the clean speech, e(n) is environmental noise, and A is the mixing matrix. The observed noisy surveillance signals can be recorded by using a mixing matrix. As the parameters of the mixing matrix are based on the configuration of the sources and the microphones, the amplitude of the source signal is proportional to the inverse of the distance from the source to the microphone. Thus, the inverse of each parameter of the mixing matrix is proportional to the distance between each source and microphone [35]. The relationship between the distance of sources and microphones d_{ij} and parameters of the mixing matrix a_{ij} can be expressed by the following equation:

$$d_{ij} = \frac{1}{a_{ij}}. \tag{9}$$

In most real forensic scenarios, the police often record surveillance speech signal from the criminal using hidden microphones. The distance between the microphones and surveillance speech signal should be less than or equal to the distance between the microphones and noise signals to obtain the noisy speech signals that have less effective to environmental noise. These noisy surveillance speech signals can be used as the input signals to speaker verification system based on the ICA algorithm in real forensic applications. Therefore, the values of the mixing matrix are chosen according to Eq. (8).

4.2 Simulating reverberant interview data

The aim of designing the reverberant interview data was to investigate the effect of different reverberation environments on forensic speaker verification performance based on the ICA-EBM algorithm. The impulse response computed from the fixed room dimensions 3 × 4 × 2.5 (m) using the image source algorithm described in [36]. The impulse response of a room was computed by using reverberation time (T_{20} = 0.15 s). The T_{20} was used instead of T_{60} in this work because T_{20} reduces the computational time when computing the time reverberation in a simulated room impulse response [4]. Table 1 and Fig. 2 show the reverberation test room parameters and position of suspect and microphone in a room, respectively. In adding reverberation, the distance between the microphone and the suspect could be in horizontal level in the room to achieve a closer approximation to forensic situations, as shown in Table 1 and Fig. 2 because in most real forensic scenarios, the police often put the microphone on the table in a room and the distance between the microphone and the suspect could be in horizontal level to record the speech from the suspect. The interview speech signals were obtained from the full-length utterances of 200 speakers using the pseudo-police interview style. Silent regions from the interview data were removed using the VAD algorithm

Table 1 Reverberation test room parameter

Configuration	Suspect position (x_s, y_s, z_s)	Microphone position (x_m, y_m, z_m)
1	(2, 1, 1.3)	(1.5, 1, 1.3)
2	(2, 1, 1.3)	(2.6, 1, 1.3)
3	(2, 1, 1.3)	(2.9, 1, 1.3)

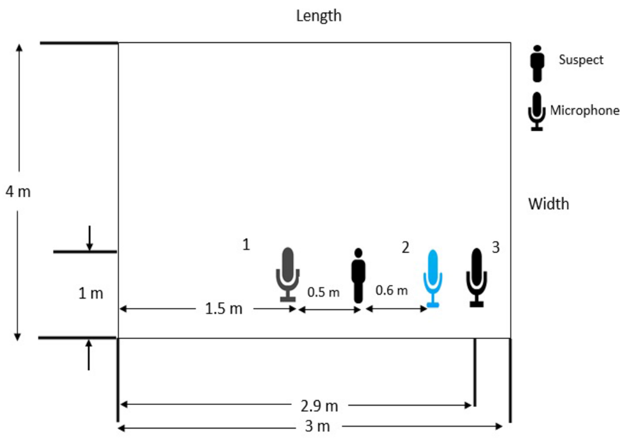

Fig. 2 Position of suspect and microphone in a room. Microphone and suspect are at 1.3 m height and the height of the room is 2.5 m

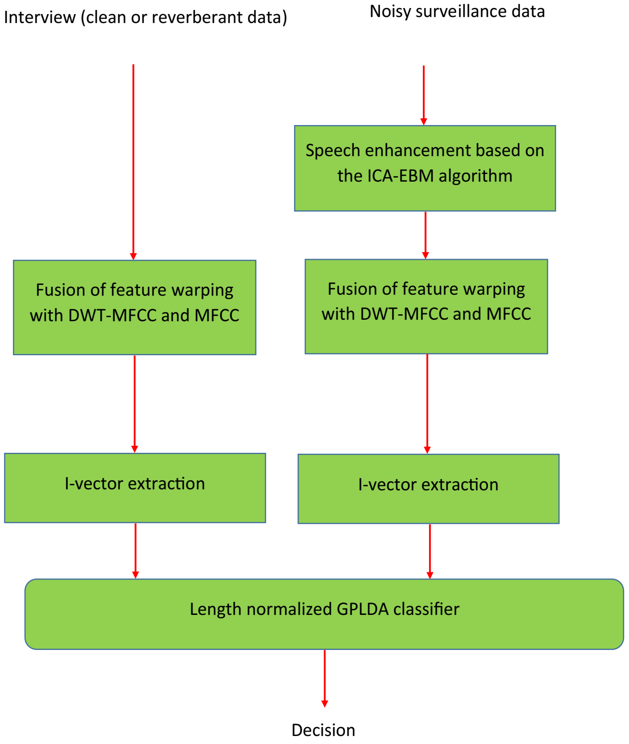
Interview (clean or reverberant data)
Noisy surveillance data
Speech enhancement based on the ICA-EBM algorithm
Fusion of feature warping with DWT-MFCC and MFCC
Fusion of feature warping with DWT-MFCC and MFCC
I-vector extraction
I-vector extraction
Length normalized GPLDA classifier
Decision

Fig. 3 Flowchart of the proposed approach of forensic speaker verification based on the ICA-EBM algorithm

[33]. The room impulse response was convolved with each of the interview data to produce the reverberated speech at 0.15 s reverberation time.

5 Proposed approach

The proposed approach of forensic speaker verification based on the ICA-EBM algorithm will be presented in this section when interview recordings are kept under clean or reverberant conditions and surveillance recordings are mixed with different types of environmental noise, as shown in Fig. 3.

5.1 Speech enhancement based on the ICA-EBM algorithm

The ICA-EBM algorithm is used in the proposed system as a multiple channel speech enhancement algorithm because this algorithm achieves better separation performance than other ICA algorithms due to its tighter bound and superior convergence behavior [13]. By using a small class of nonlinear contrast functions, the ICA-EBM algorithm performs source separations, which are sub- or super-Gaussian, unimodal or multimodal, symmetric or skewed. The algorithm uses the entropy bound estimator to approximate the entropies of different types of distribution. The ICA-EBM algorithm minimizes the mutual information of the estimated source signals to estimate the unmixing matrix. The algorithm uses a line search procedure, which forces the unmixing matrix to be orthogonal for better convergence behavior. The mutual information cost function can be defined as:

$$I(y_1, y_2, \dots, y_N) = \sum_{n=1}^{N} H(y_n) - \log|det(W)| - H(x) \tag{10}$$

where $H(y_n)$ is the differential entropy of the nth separated sources y and entropy of the observations $H(x)$ is a term independent with respect to the unmixing matrix W which can be treated as a constant C.

Minimization of the mutual information among the estimated source signals is related to the maximization of the log-likelihood cost function as long as the model of the probability density function (PDF) matches the PDF of the true latent source signal [37]. The bias is introduced in the estimation of the unmixing matrix due to the deviation of the model from the true PDF of the source signal. This bias can be removed by integrating a flexible density model for each source signal into the ICA algorithm to minimize the bias of the unmixing matrix providing separated source signals from a wide range of the PDF accurately [38].

To achieve this, the cost function and its gradient can be rewritten with respect to each row vector of the unmixing matrix $w_n, n = 1, 2, 3, \dots, N$. This can be done by expressing the volume of the N-dimensional parallelepiped, spanned by the row vectors of W, as the inner product of the nth row vectors and unit Euclidian length vector h_n, that is perpendicular to all row vectors of the unmixing matrix except w_n [13]. Therefore, the mutual information cost function in Eq. 10 can be rewritten as a function of only w_n as:

$$J_n(w_n) = \sum_{n=1}^{N} H(y_n) - \log\left|h_n^T w_n\right| + C. \tag{11}$$

The gradient of Eq. 11 can be computed as:

$$\frac{\partial J_n(w_n)}{\partial w_n} = -V'_{k(n)}\{E[G_{k(n)}(y_n)x]\}E[g_{k(n)}(y_n)x] - \frac{h_n}{h_n^T w_n} \tag{12}$$

where $V'(.)$ and $g_{k(n)}(.)$ are the first order derivative of the negentropy $V(.)$ and the kth contrast functions $G_{k(n)}(.)$, respectively, and E is the expected operator.

The line search algorithm for the orthogonal ICA-EBM can be obtained by the following equations:

$$w_n^+ = \frac{E[xg_{k(n)}(y_n)] - E[g'_{k(n)}(y_n)]w_n}{E[g_{k(n)}(y_n)y_n] - E[g'_{k(n)}(y_n)]} \tag{13}$$

$$w_n^{new} = \frac{w_n^+}{\|{w_n}^+\|} \tag{14}$$

where $g'_{k(n)}(.)$ is the second derivative of the kth contrast functions $G_{k(n)}(.)$.

The line search algorithm for ICA-EBM given in Eqs. 13 and 14 repeats over different row vectors of the unmixing matrix until convergence. The criteria of the convergence satisfies the following equation:

$$1 - max(abs[diag(W^{new}W^T)]) \le \epsilon \tag{15}$$

with a typical value of ϵ is 0.0001. After each row vector of the unmixing matrix W has been updated, the symmetrical decorrelation method is performed to remain the unmixing matrix orthogonal and it can be obtained by the following equation:

$$W^{new} = (WW^T)^{\frac{-1}{2}}W. \tag{16}$$

5.2 Feature warped MFCC

MFCCs are the most popular technique for extracting features in speaker verification systems. A block diagram of the MFCC feature extraction is shown in Fig. 4. The first stage of extracting MFCC features involves framing the speech signals into several segments using Hamming window. Then, the fast Fourier transform (FFT) can be used to transform a frame of the acoustic speech signals from the time domain into the frequency domain. The shape of the magnitude spectrum contains information about the resonance properties of the vocal tract which is considered a better feature for speaker verification systems. After that, a triangular filter bank is applied to capture the subband energies. The MFCC features are computed using a psychoacoustically motivated filter bank, followed by using discrete cosine transform (DCT). The MFCC can be represented as:

$$c_n = \sum_{m=1}^{M} \left[\log(Y(m)\right] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right], \tag{17}$$

where m is the number of mel filter banks, $Y(m)$ is the output of M- channel filter bank, and n is the index of the cepstral coefficients. The first 10–20 cepstral coefficients are typically extracted from each frame. In order to capture the dynamic characteristics of the speech signals, the first and second time derivatives of the MFCC are usually appended to each feature [39].

Pelecanos and Sridharan [40] proposed feature warping technique to decrease the effect of noise and channel distortion by transforming the distribution of the cepstral features into standard normal distribution. The warping process executes as follows:

1. Extracting the MFCC features from the acoustic speech signals $i = 1, 2, \dots, D$ where D is the number of feature extraction dimension.
2. Ranking features in dimension i in ascending order for a given sliding window (typically 3 s).
3. Warping the cepstral feature z in dimension i of the central frame to its warped value

$$\frac{(N + 1/2 - R)}{N} = \int_{z=-\infty}^{\hat{x}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)dz \tag{18}$$

where $\hat{x}$ is the warped feature. Suppose the raw cepstral feature z has rank R and window size N. The value of $\hat{x}$ can be estimated by putting $R = N$ and solving $\hat{x}$ using the numerical integration and then repeating for each decremented value of R.

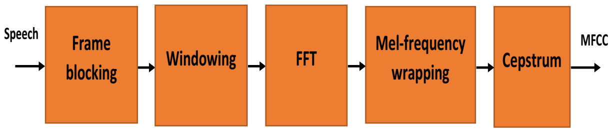

Fig. 4 A block diagram of the MFCC feature extraction

4. The warped value can be found by lookup in a standard normal table.
5. Repeating the process by shifting the sliding window for a single frame each time

Feature warping achieves robustness to additive noise and channel mismatch which retains the speaker-specific information that is lost by using cepstral mean subtraction (CMS), cepstal mean variance normalization (CMVN), and relative spectral (RASTA) processing [40]. Thus, feature warping will be used in the proposed system.

5.3 Wavelet transform

The wavelet transform is a technique for analyzing speech signals. It uses an adaptive window that provides low time resolution at low-frequency subbands and high time resolution at high-frequency subbands [41]. The DWT is a type of wavelet transform, and it can be defined as

$$D(j,K) = \sum_{j} \sum_{K} \mathbf{V}(K) 2^{\frac{-j}{2}} \psi(2^{-j}N - K) \tag{19}$$

where ψ is the time function with fast decay and finite energy called the mother wavelet, j is the level number, $\mathbf{V}(K)$ is the speech signal, and N and K are scaling and translation parameters, respectively. The DWT can be performed using a pyramidal algorithm [42].

The block diagram of the DWT is shown in Fig. 5. The speech signal ($\mathbf{V}$) is decomposed into different frequency sub-bands by using two filters, **g** and **h** which are a high-pass and low-pass filters, respectively. The half coefficients of the speech signals are discarded by using the down-sampling operator (↓ 2) after applying the filter. The approximation coefficients (CA1) can be obtained by convolving the low-pass filter with the speech signal and applying the down-sampling operator to the output of the filter **h**. The detailed coefficients (CD1) can be obtained by convolving the high pass filter with the speech signals and applying the down-sampling to the output of the filter **g**. The speech signals can also be decomposed by applying the DWT to the approximation coefficients (CA1).

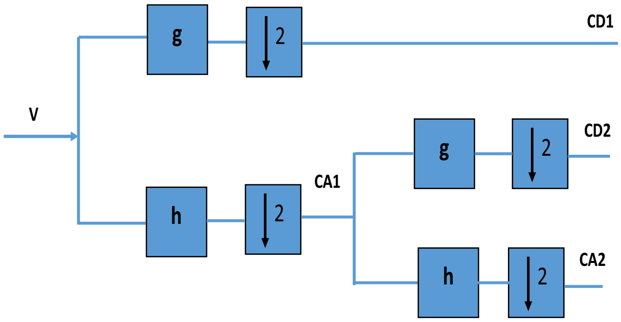

Fig. 5 Block diagram of the DWT

5.4 Fusion of feature warping with MFCC and DWT-MFCC

The approach to extracting the features from the interview and enhanced surveillance data is based on the DWT technique. The interview and enhanced surveillance data were framed into several segments using a Hamming window, with 30 ms size and 10 ms shift. The frame of the interview and enhanced surveillance data were split into low-frequency sub-band (approximation coefficients) and high-frequency sub-band (detail coefficients). The approximation and detail coefficients were concatenated into a single feature vector. The MFCCs were used to extract the features from the DWT of the interview/enhanced surveillance data. The first and second derivatives of the cepstral coefficients were added to the MFCCs. Feature warping technique with a 301 frame window was applied to the features extracted from the DWT-MFCC. The feature warped MFCC was used to extract the features from the full band of the interview/enhanced surveillance data. Finally, the fusion of feature warping with MFCC and DWT-MFCC was obtained by concatenating the features extracted from the feature warped DWT-MFCC and feature warped MFCC of the full band interview/enhanced surveillance data into a single feature vector, as shown in Fig. 6.

5.5 I-vector based speaker verification

I-vector through length-normalized GPLDA has become the modern technique for speaker verification systems [43]. Such systems consist of two stages: i-vector feature extraction and length-normalized GPLDA.

5.5.1 I-vector representation

The introduction of i-vector based speaker verification was inspired by the discovery that session variability in joint factor analysis contains speaker information which can be used to distinguish between speakers more efficiently [44]. In an i-vector based speaker verification system, the speaker and channel-dependent Gaussian mixture model (GMM) supervector, μ, can be represented by

$$\mu = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{20}$$

where **m** is the speaker and session independent mean of the universal background model (UBM) super-vector, **T** is a low-rank total variability matrix, and **w** represents i-vectors which have a standard normal distribution.

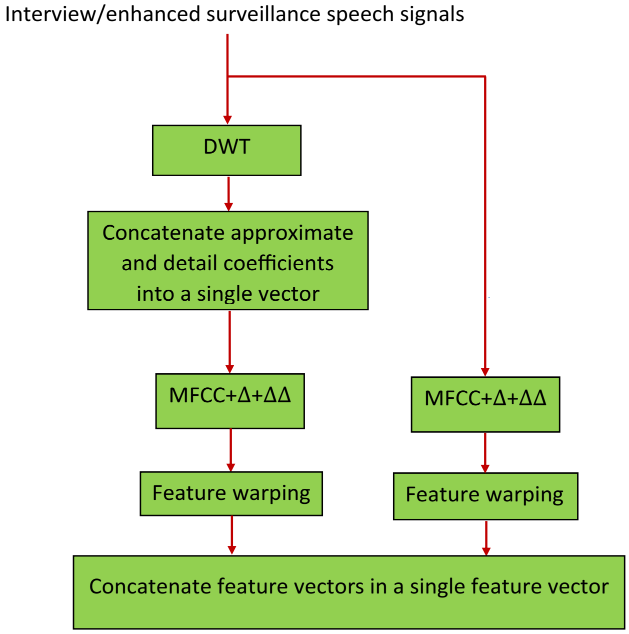
Interview/enhanced surveillance speech signals
DWT
Concatenate approximate and detail coefficients into a single vector
MFCC+Δ+ΔΔ
MFCC+Δ+ΔΔ
Feature warping
Feature warping
Concatenate feature vectors in a single feature vector

Fig. 6 Flowchart of fusion of feature warping with MFCC and DWT-MFCC approach

The description of training the total variability matrix is available in [45] and [46]. McLaren and van Leeuwen [47] investigated different types of total variability representations, such as concatenated and pooled techniques with i-vector system. For the concatenated total-variability technique, the total-variability of telephone and microphone subspaces are trained separately using speech from those sources, then both subspaces transformation are combined to create a single total-variability space. For the pooled technique, the total variability is trained on microphone and telephone speech utterances. Their studies found that the pooled technique provides a better representation of i-vector speaker verification than the concatenated total variability technique. Thus, the pooled total variability technique is used in this work.

5.5.2 Length-normalized GPLDA classifier

Kenny has introduced GPLDA and heavy-tailed PLDA (HTPLDA) for i-vector speaker verification systems [48]. Kenny found that HTPLDA achieved significant improvement in speaker verification performance over GPLDA because the heavy-tailed behavior showed a better match for i-vector distribution [48]. The length-normalized GPLDA was proposed by Garcia-Romero and Espy-Wilson [49] to convert the heavy-tailed i-vector into Gaussian distribution. In the proposed system, we used length-normalized GPLDA classifier because it achieves similar performance with efficient computation when compared with the HTPLDA approach [49]. The length normalization i-vector, $\mathbf{w}_r^{norm}$, can be defined as

$$\mathbf{w}_r^{norm} = \bar{\mathbf{w}}^{norm} + \mathbf{U}_1\mathbf{z}_1 + \epsilon_r \tag{21}$$

where r is the number of recordings for a given speaker, $\mathbf{U}_1$ is the eigenvoice matrix, $\mathbf{z}_1$ is the speaker factor, and ϵ_r is the residual. The between-speaker variability $\bar{\mathbf{w}}^{norm} + \mathbf{U}_1\mathbf{z}_1$ can be represented by a low rank of covariance matrix $\mathbf{U}_1\mathbf{U}_1^T$. The within-speaker variability can be represented by Λ^{-1} and assumes that the precision matrix (Λ) is full rank.

The details of the length-normalized GPLDA technique and the estimation of the model parameters are described in [48] and [49]. The scoring can be calculated using the batch likelihood ratio [48]. Given the i-vectors of the interview $\mathbf{w}_{interview}^{norm}$ and surveillance $\mathbf{w}_{surveillance}^{norm}$, the score can be computed as follows,

$$score = \ln \frac{P(\mathbf{w}_{interview}^{norm}, \mathbf{w}_{surveillance}^{norm}|H_1)}{P(\mathbf{w}_{interview}^{norm}|H_0)P(\mathbf{w}_{surveillance}^{norm}|H_0)} \tag{22}$$

where H_1: the speakers are the same and H_0 the speakers are different.

5.6 Experimental set-up

The proposed approach was evaluated using the length-normalized GPLDA based speaker verification framework. In the development phase, we extracted 78-dimensional fusion of feature warping with DWT-MFCC and MFCC from the development speech signals. The UBM containing 256 Gaussian components were used through our experimental results. We kept the number of the UBM in low values in order to reduce the computational complexity cost, and it's easy to adapt to real forensic applications. The UBMs were trained on microphone and telephone data using 348 speakers from the AFVC database [23]. The UBMs were used to estimate the Baum-Welch statistics before training the total variability subspace of dimension 400, which was used to compute the i-vector speaker representations. The dimension of the i-vectors reduced to 200 using linear discriminant analysis (LDA). The i-vectors length normalization was used before GPLDA modeling using centering and whitening of the i-vectors [49]. In the interview and verification phases, the interview and surveillance speaker models were created from the interview and enhanced surveillance speech signals to represent them in i-vector subspace. The hidden parameters of the PLDA were estimated using variational posterior distribution. Scoring in the length-normalized GPLDA was conducted using the batch likelihood ratio to calculate the similarity score between the interview and surveillance speaker models [48]. We used the Microsoft Research (MSR) identity toolbox [50] to evaluate length-normalized

GPLDA speaker verification performance. Table 2 shows a summary of the experimental set-up used for simulation of the proposed approach.

6 Simulation results

In this section, we investigate the effectiveness of using the ICA-EBM as a speech enhancement algorithm to improve forensic speaker verification performance under noisy, as well as noisy and reverberant environments. The performance of the proposed forensic speaker verification based on the ICA-EBM approach was evaluated using the EER.

6.1 Baseline speaker verification systems

In order to investigate the effectiveness of the ICA-EBM algorithm for improving forensic speaker verification performance under noisy environments, as well as noisy and reverberant environments, we compared speaker verification performance based on the ICA-EBM algorithm with clean interview-noisy surveillance and reverberant interview-noisy surveillance speaker verification baselines. These baselines do not use speech enhancement algorithms (ICA-EBM or traditional ICA) to improve forensic speaker verification performance under noisy and reverberant environments, but they use the same feature extraction and classifier techniques as in the proposed approach of forensic speaker verification system. The interview data for these baselines were obtained from full-length utterance of 200 speakers using the pseudo-police style. Silent regions were removed using the VAD algorithm [33]. The surveillance data were obtained from 10 s duration from 200 speakers using the informal telephone conversation style after removing silent regions using the VAD algorithm. Table 3 shows a description of the data used in baseline speaker verification systems. A brief description of the two baselines is described below

Table 3 A description of the data used in baseline speaker verification systems

Type of style used in interview data	Pseudo-police style
Duration of interview data	Full length
Number of speakers used in interview data	200
Type of style used in surveillance data	Informal telephone conversation style
Number of speakers used in surveillance data	200
Duration of surveillance data	10 s

6.1.1 Clean interview-noisy surveillance speaker verification baseline

The clean interview-noisy surveillance speaker verification baseline is obtained by keeping interview recordings under clean conditions and surveillance recordings were corrupted by a random session of HOME, CAR, and STREET noises from the QUT-NOISE corpus using a single microphone (x_1) at SNRs ranging from − 10 to 10 dB. A fusion of feature warping with MFCC and DWT-MFCC was used to extract the features from the interview and surveillance recordings. According to our previous experimental results [51], the fusion of feature warping with MFCC and DWT-MFCC improved forensic speaker verification performance under noisy environments compared with traditional MFCC or other combination of MFCC and DWT-MFCC with and without feature warping. Level 3 and Daubechies 8 were used in the fusion of feature warping with MFCC and DWT-MFCC because our previous experimental results [51] demonstrated that level 3 achieved improvements in speaker verification performance than other levels under noisy environments over the majority of SNR values. The interview and surveillance speaker models were created from the speech signals to represent them in i-vector subspace. The

Table 2 A summary of the experimental set-up used for simulation of the proposed approach

Type of feature used	Fusion of feature warping with DWT-MFCC and MFCC
Number of features used	78
Type of classifier used	Length-normalized GPLDA
Number of Gaussian components used in the UBM	256
Types of data used for training UBM	Microphone and telephone data from the AFVC database
Number of speakers used for training UBM	348
Dimension of total variability matrix	400
Dimension of i-vector after using LDA	200

length-normalized GPLDA and batch likelihood ratio were used to compute the similarity score between those speaker models, as shown in Fig. 7.

6.1.2 Reverberant interview-noisy surveillance speaker verification baseline

The interview recordings were convolved with the room impulse response at 0.15 s reverberation time using the image source algorithm [36]. The first configuration of the room was used in this simulation, as shown in Table 1 and Fig. 2. The surveillance recordings were mixed with different levels and types of noise using a single microphone (x_1). Fusion of feature warping with MFCC and DWT-MFCC approach was used to extract the features from the interview and surveillance recordings because our previous experimental results [25] demonstrated that the fusion approach improved forensic speaker verification performance under noisy and reverberant environments compared with other feature extraction techniques. Level 4 and Daubechies 8 were used in a fusion of feature warping with MFCC and DWT-MFCC to extract the features from the interview and surveillance recordings because our previous experimental results [51] demonstrated that level 4 achieved improvements in noisy and reverberant speaker verification performance than other levels under different types of noise at SNRs ranging from − 10 to 10 dB. The i-vector through length-normalized GPLDA and batch likelihood ratio were used to calculate the similarity score between the i-vectors of interview and surveillance recordings, as shown in Fig. 8.

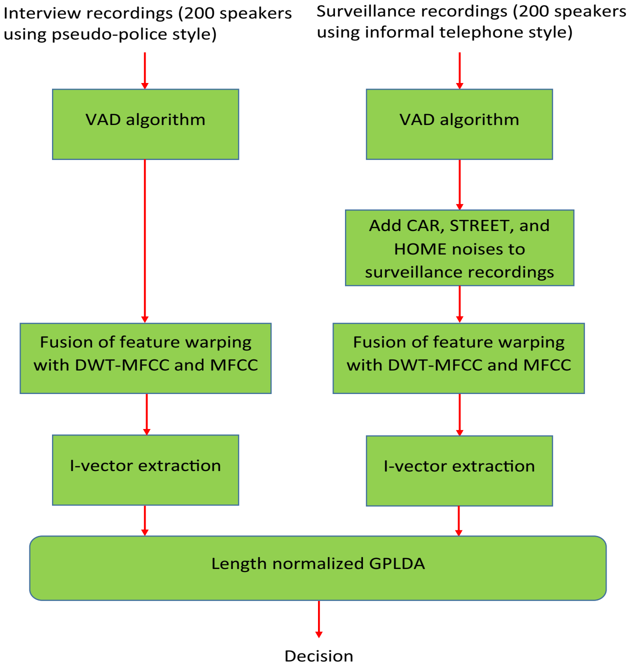
Interview recordings (200 speakers using pseudo-police style)
Surveillance recordings (200 speakers using informal telephone style)
VAD algorithm
VAD algorithm
Add CAR, STREET, and HOME noises to surveillance recordings
Fusion of feature warping with DWT-MFCC and MFCC
Fusion of feature warping with DWT-MFCC and MFCC
I-vector extraction
I-vector extraction
Length normalized GPLDA
Decision

Fig. 7 Flowchart of the clean interview-noisy surveillance speaker verification baseline

6.2 Clean interview-noisy surveillance conditions

In this section, we describe forensic speaker verification performance based on the ICA-EBM when interview recordings are kept under clean conditions and surveillance recordings are mixed with different types and levels of noise. In the simulation results of noisy surveillance conditions, we chose level 3 and Daubechies 8 of DWT to extract the features from the noisy surveillance recordings in order to provide a fair comparison with the baseline of clean interview-noisy surveillance speaker verification system. The performance of forensic speaker verification system was evaluated using the EER and minimum decision cost function (mDCF), calculated using $C_{miss} = 10, C_{fa} = 1, P_{target} = 0.01$.

6.2.1 The performance improvement from using ICA-EBM

Speaker verification performance based on the ICA-EBM was evaluated and compared with the clean interview-noisy surveillance speaker verification baseline and the traditional ICA (Fast ICA), as shown in Table 4. The SNRs on the Table 4 were calculated from the first microphone (x_1). The results show that using the ICA-EBM algorithm achieved significant improvements in speaker verification performance than the clean interview-noisy surveillance speaker verification baseline when surveillance recordings

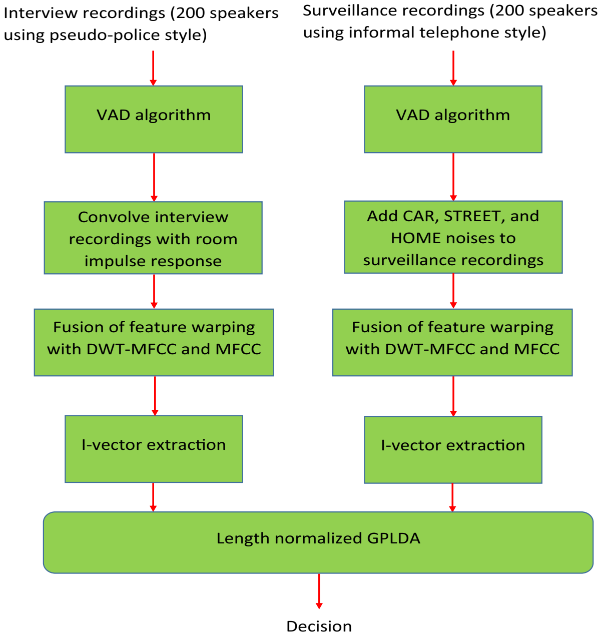
Interview recordings (200 speakers using pseudo-police style)
Surveillance recordings (200 speakers using informal telephone style)
VAD algorithm
VAD algorithm
Convolve interview recordings with room impulse response
Add CAR, STREET, and HOME noises to surveillance recordings
Fusion of feature warping with DWT-MFCC and MFCC
Fusion of feature warping with DWT-MFCC and MFCC
I-vector extraction
I-vector extraction
Length normalized GPLDA
Decision

Fig. 8 Flowchart of the reverberant interview-noisy surveillance speaker verification baseline

Table 4 EER for speaker verification when interview recordings are kept under clean conditions and surveillance recordings are corrupted by different types of environmental noise

Methods	Type of noise	SNR (dB)				
		−10	−5	0	5	10
Clean interview-noisy surveillance speaker verification baseline	STREET	33.500	26.00	18.000	10.020	5.5000
Traditional ICA	STREET	12.547	11.407	13.000	11.467	11.5000
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	STREET	**11.000**	**10.650**	**10.900**	10.894	11.000
Clean interview-noisy surveillance speaker verification baseline	CAR	30.500	19.854	11.500	7.213	3.500
Traditional ICA	CAR	10.957	10.5000	10.080	11.500	10.246
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	CAR	**10.000**	**10.211**	**9.874**	10.500	9.794
Clean interview-noisy surveillance speaker verification baseline	HOME	36.447	24.500	15.000	9.500	5.658
Traditional ICA	HOME	13.856	12.500	11.500	12.400	12.326
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	HOME	**11.500**	**10.643**	**10.942**	11.334	11.354

Lower EER indicates better performance

Bold value indicates that the lowest EER values for proposed approach of forensic speaker verification based on the ICA-EBM algorithm when interview recordings are kept under clean condition and surveillance recordings are corrupted by different types of environmental noise

are corrupted by different types of noise at low SNRs ranging from − 10 to 0 dB. The improvement in EER of speaker verification based on the ICA-EBM decreased when SNR increased. The performance of the ICA-EBM degraded compared with the clean interview-noisy surveillance baseline when surveillance recordings are corrupted by different types of environmental noise at SNRs ranging from 5 dB to 10 dB.

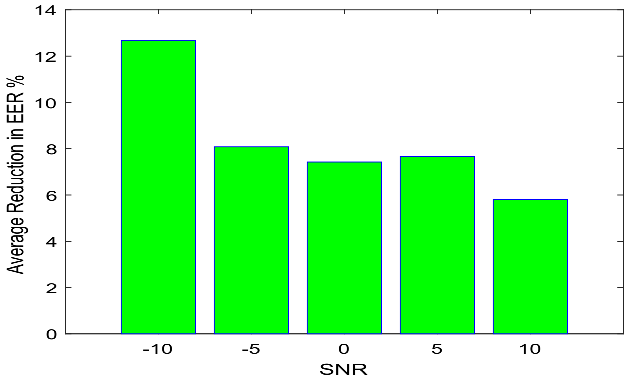

Fig. 9 Average EER reduction for the ICA-EBM algorithm compared with the traditional ICA when interview recordings are kept under clean conditions and surveillance recordings are mixed with STREET, CAR, and HOME noises at SNRs ranging from − 10 to 10 dB

The ICA-EBM algorithm improved speaker verification performance over the traditional ICA when surveillance recordings are corrupted by different levels and types of noise. The reduction in EER for ICA-EBM over ICA, EER_{red}, can be computed as

$$EER_{red} = \frac{EER_{ICA} - EER_{(ICA-EBM)}}{EER_{ICA}} \tag{23}$$

where EER_{ICA} and $EER_{(ICA-EBM)}$ are the equal error rates for conventional ICA and ICA-EBM approaches, respectively. The average EER reduction can be calculated by computing the mean in EER_{red} for different types of noise at each noise level, as shown in Fig. 9. The average EER reduction for the ICA-EBM algorithm ranges from 12.68% to 7.42% compared with conventional ICA when surveillance recordings are mixed with different types of noise at SNRs ranging from − 10 to 0 dB. From the above results, it is clear that ICA-EBM is more desirable than traditional ICA due to its superior separation performance than traditional ICA algorithm.

Table 5 shows a comparison mDCFs for speaker verification based on ICA-EBM algorithm and clean interview-noisy surveillance speaker verification baseline under different types of environmental noise at SNRs ranging from − 10 to 10 dB. It is clear from Table 5 that speaker verification based on the ICA-EBM algorithm significantly improved mDCF at low SNR (− 10 to 0 dB) in the presence of STREET, CAR and HOME noise. There was a degradation in the performance of speaker verification based on the ICA-EBM algorithm than the clean interview-noisy surveillance speaker verification baseline over the majority of SNRs ranging from 5 dB to 10 dB.

6.2.2 Time performance

In this section, the computation time of forensic speaker verification based on the ICA-EBM algorithm was tested

Table 5 Comparison mDCFs for speaker verification based on ICA-EBM algorithm and clean interview-noisy surveillance speaker verification baseline

Methods	Type of noise	SNR (dB)				
		−10	−5	0	5	10
Clean interview-noisy surveillance speaker verification baseline	STREET	0.0996	0.0974	0.0801	0.0590	0.0339
ICA-EBM algorithm	STREET	**0.0575**	**0.0573**	**0.0570**	**0.0586**	0.0579
Clean interview-noisy surveillance speaker verification baseline	HOME	0.0991	0.0924	0.0803	0.0511	0.0325
ICA-EBM algorithm	HOME	**0.05682**	**0.05603**	**0.0572**	0.0554	0.0572
Clean interview-noisy surveillance speaker verification baseline	CAR	0.0995	0.0884	0.0669	0.0389	0.0240
ICA-EBM algorithm	CAR	**0.0563**	**0.05631**	**0.05633**	0.0570	0.05631

Bold value indicates that the lowest mDCFs values for ICA-EBM algorithm when interview recordings are kept under clean condition and surveillance recordings are corrupted by different types of environmental noise

and compared with the computation time of clean interview-noisy surveillance speaker verification baseline and traditional ICA using a processor Intel(R) Core (TM) i7-4600U CPU 2.70 GHz and MATLAB 2017 a. The interview speech signals for different speaker verification methods were extracted from full duration utterances from 200 speakers using the pseudo-police style. The VAD algorithm [33] was used to remove silent regions from the interview speech signals. The surveillance recordings were obtained from random sessions of one utterance of 10 s duration from 200 speakers using the informal telephone conversation style after removing silent regions using the VAD algorithm.

The combination time for different speaker verification methods is obtained by keeping interview data under clean conditions and surveillance recordings were mixed with a random session of CAR, STREET and HOME noises from the QUT-NOISE database at SNRs ranging from −10 to 10 dB, resulting in a two-channel noisy surveillance recordings, according to Eqs. 7 and 8. Fusion of feature warping with MFCC and DWT-MFCC was used to extract the features from the interview and surveillance recordings. Level 3 and db8 were used in a fusion of feature warping with MFCC and DWT-MFCC. The interview and surveillance speaker models were created from the speech signals to represent them in i-vector subspace. Then, the length normalized GPLDA and batch likelihood ratio were used to compute the similarity score between those speaker models.

Table 6 shows the computation time (s) for different speaker verification methods when interview recordings are kept under clean conditions and surveillance recordings are corrupted with different types of environmental noise. The SNRs on the Table 6 were calculated from the first microphone (x_1). It is clear from this table that the proposed method takes a longer time than the other methods when interview recordings are kept under clean conditions and surveillance recordings are mixed with different types and levels of environmental noise.

6.2.3 Effect of utterance duration

In these simulation results, the interview recordings were obtained from the full duration utterances of 200 speakers using pseudo-police style under clean conditions. However,

Table 6 The computation time (s) for different speaker verification methods when interview recordings are kept under clean conditions and surveillance recordings are corrupted with different types of environmental noise

Methods	Type of noise	SNR (dB)				
		−10	−5	0	5	10
Clean interview-noisy surveillance speaker verification baseline	STREET	12381	12275	12213	12178	12093
Traditional ICA	STREET	16275	16195	16122	16098	16038
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	STREET	21576	21492	21423	21392	21388
Clean interview-noisy surveillance speaker verification baseline	CAR	12295	12265	12210	12167	12081
Traditional ICA	CAR	16266	16187	16101	16088	16025
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	CAR	21563	21483	21408	21381	21367
Clean interview-noisy surveillance speaker verification baseline	HOME	12322	12308	12286	12253	12185
Traditional ICA	HOME	16331	16285	16222	16187	16124
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	HOME	21667	21573	21520	21483	21426

we extracted the surveillance recordings from 10 s, 20 s and 40 s using informal telephone conversation style. Random sessions of HOME, STREET, and CAR noises from the QUT-NOISE corpus were mixed with the surveillance recordings at different SNR values using two microphones, as in Eqs. 7 and 8. Since the performance of forensic speaker verification based on the ICA-EBM algorithm decreased EER compared with other techniques when surveillance recordings mixed with various types of environmental noise at SNRs ranging from − 10 to 0 dB, as described in Sect. 6.2.1, the effect of utterance length was evaluated on the performance of forensic speaker verification based on the ICA-EBM algorithm in this section.

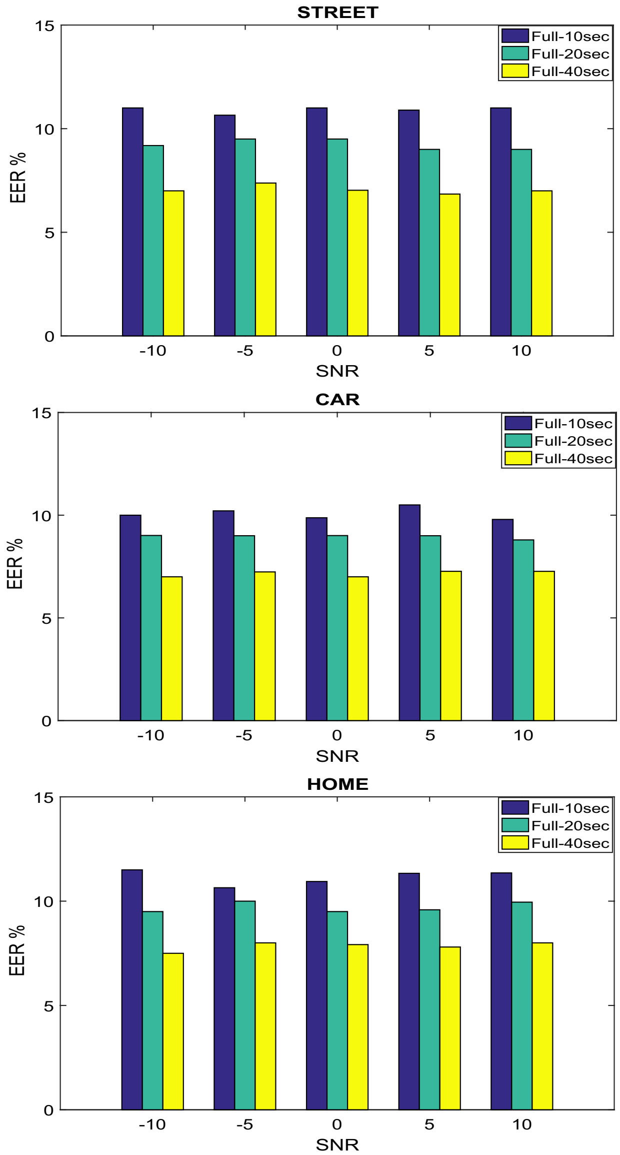

Fig. 10 Effect of utterance surveillance duration on speaker verification performance based on the ICA-EBM algorithm when interview recordings are kept under clean condition and surveillance recordings are mixed with different levels and types of noise

Figure 10 shows the effect of utterance surveillance duration on speaker verification performance based on the ICA-EBM algorithm when interview recordings are kept under clean conditions and surveillance recordings are mixed with different levels and types of noise. It is clear that speaker verification performance based on the ICA-EBM algorithm improved when surveillance utterance length increased from 10 to 40 s under noisy environments.

6.3 Reverberant interview-noisy surveillance conditions

Forensic speaker verification based on the ICA-EBM algorithm was evaluated and compared with the reverberant interview-noisy surveillance speaker verification baseline and the traditional ICA algorithm. The effect of reverberation time, surveillance utterance duration, and suspect/microphone position on the performance of speaker verification based on the ICA-EBM algorithm is also discussed in this section. In the simulation results of noisy surveillance and reverberant interview conditions, we chose level 4 and Daubechies 8 of DWT to extract the features from the noisy surveillance and reverberant interview speech signals in order to provide a fair comparison with the baseline of reverberant interview-noisy surveillance speaker verification system. The performance of forensic speaker verification system was evaluated using the EER and mDCF, calculated using $C_{miss} = 10$, $C_{fa} = 1$, $P_{target} = 0.01$. Part of the simulation results in this section have been published in [52].

6.3.1 The performance improvement from using ICA-EBM

In these simulation results, the reverberation interview signals are obtained by convolving the interview recordings with the room impulse response at 0.15 s reverberation time using the image source algorithm described in [36]. The first configuration of the room was used in these simulation results, as shown in Table 1 and Fig. 2. The surveillance recordings were corrupted by one random segment of environmental noises (STREET, HOME and CAR) at SNRs ranging from − 10 to 10 dB, resulting in a two-channel noisy surveillance recordings, according to Eqs. 7 and 8. The noisy surveillance recordings were kept without reverberation conditions because the noisy surveillance data are usually recorded in open areas in most real forensic scenarios [34].

Table 7 shows EER for speaker verification when interview recordings reverberated at 0.15 s and surveillance recordings are mixed with different types of noise. The SNRs on the Table 7 were computed from the first microphone (x_1). The results show that speaker verification performance based on the ICA-EBM decreased EER over the

Table 7 EER for speaker verification when interview recordings reverberated at 0.15 s and surveillance recordings are mixed with different types of noise

Methods	Type of noise	SNR (dB)				
		−10	−5	0	5	10
Reverberant interview-noisy surveillance speaker verification baseline	STREET	36.500	26.500	18.500	12.400	7.680
Traditional ICA	STREET	14.600	14.032	15.062	14.379	14.300
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	STREET	**12.932**	**12.733**	**13.045**	13.012	13.500
Reverberant interview-noisy surveillance speaker verification baseline	CAR	28.000	20.500	14.230	9.800	7.500
Traditional ICA	CAR	13.673	13.758	13.900	13.500	13.410
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	CAR	**12.859**	**12.826**	**12.771**	13.000	13.410
Reverberant interview-noisy surveillance speaker verification baseline	HOME	38.410	29.340	20.000	14.000	10.350
Traditional ICA	HOME	15.000	14.500	14.400	14.575	14.500
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	HOME	**14.341**	**14.032**	**14.000**	14.477	14.266

Bold value indicates that the lowest EER values for proposed approach of forensic speaker verification based on the ICA-EBM algorithm when interview recordings reverberated at 0.15 s reverberation time and surveillance recordings are mixed with different types of environmental noise

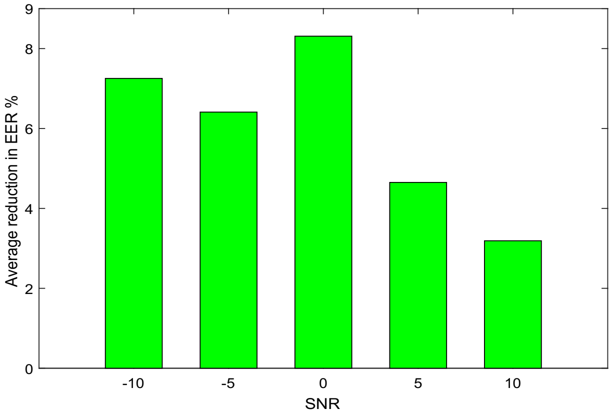

Fig. 11 Average EER reduction for the ICA-EBM algorithm compared with the traditional ICA when interview recordings reverberated at 0.15 s reverberation time and surveillance recordings are mixed with STREET, CAR, and HOME at SNRs ranging from − 10 to 10 dB

reverberant interview-noisy surveillance speaker verification baseline when surveillance recordings are mixed with different types of noise at low SNRs ranging from − 10 to 0 dB. The improvement in the performance decreased when SNR increased. The performance of the ICA-EBM algorithm degraded compared with the reverberant interview-noisy surveillance speaker verification baseline when surveillance recordings are mixed with different types of noise at SNRs ranging from 5 to 10 dB.

Figure 11 shows average EER reduction for the ICA-EBM algorithm compared with the traditional ICA algorithm for different types of noise and reverberation. When interview recordings reverberated at 0.15 s and the surveillance recordings were corrupted with different types of noise at SNRs ranging from − 10 to 0 dB, the performance of speaker verification based on the ICA-EBM achieved average EER reduction ranging from 7.25 to 8.31% compared with the traditional ICA.

Table 8 comparison mDCFs for speaker verification based on the ICA-EBM algorithm and reverberant interview-noisy surveillance speaker verification baseline

Methods	Type of noise	SNR (dB)				
		−10	−5	0	5	10
Reverberant interview-noisy surveillance speaker verification baseline	STREET	0.0994	0.0951	0.0852	0.0685	0.0497
ICA-EBM algorithm	STREET	**0.0691**	**0.0677**	**0.0680**	0.0686	0.0682
Reverberant interview-noisy surveillance speaker verification baseline	HOME	0.100	0.0953	0.0844	0.0665	0.0527
ICA-EBM algorithm	HOME	**0.0706**	**0.0705**	**0.0704**	0.0701	0.0705
Reverberant interview-noisy surveillance speaker verification baseline	CAR	0.0989	0.0904	0.0697	0.0545	0.0425
ICA-EBM algorithm	CAR	**0.0666**	**0.0672**	**0.0679**	0.0680	0.0689

Bold value indicates that the lowest mDCFs values for the ICA-EBM algorithm when interview recordings reverberated at 0.15 s reverberation time and surveillance recordings are mixed with different types of environmental noise

Table 8 shows comparison mDCFs for speaker verification based on the ICA-EBM algorithm and reverberant interview-noisy surveillance speaker verification baseline. It is clear that speaker verification performance based on the ICA-EBM algorithm improved mDCF over the reverberant interview-noisy surveillance speaker verification baseline when interview recordings reverberated at 0.15 s reverberation time and the surveillance recordings were mixed with various types of environmental noise at low SNR values (− 10 to 0 dB). Forensic speaker verification performance based on the ICA-EBM algorithm degraded compared with the reverberant interview-noisy surveillance speaker verification baseline at SNRs ranging from 5 to 10 dB.

6.3.2 Time performance

In this section, the computation time of forensic speaker verification based on the ICA-EBM algorithm was tested and compared with the computation time of reverberant interview-noisy surveillance speaker verification baseline and traditional ICA using a processor Intel(R) Core (TM) i7-4600U CPU 2.70 GHz and MATLAB 2017 a. The full duration of the interview recordings was obtained from 200 speakers using the pseudo-police style. Silent regions from the interview recordings were removed using the VAD algorithm. The interview recordings were convolved with the impulse response of the room at 0.15 s reverberation time to generate the reverberated speech. The first configuration of the room is used in these experimental results, as shown in Table 1. The surveillance recordings were obtained from 10 s duration of one utterance from 200 speakers using the informal telephone conversation style after removing silent regions using the VAD algorithm. The surveillance recordings were mixed with one random session of the environmental noises (CAR, STREET and HOME noises) from the QUT-NOISE database at SNRs ranging from − 10 to 10 dB, resulting in two-channel noisy speech signal according to Eqs. 7 and 8. Fusion of feature warping with MFCC and DWT-MFCC was used to extract the features from the interview and surveillance recordings. Level 4 and db8 were used in a fusion of feature warping with MFCC and DWT-MFCC. The interview and surveillance speaker models were created from the speech signals to represent them in i-vector subspace. Then, the length normalized GPLDA and batch likelihood ratio were used to compute the similarity score between those speaker models.

Table 9 shows the computation time (s) for different speaker verification methods when interview recordings reverberated at 0.15 s reverberation time and surveillance recordings are mixed with different types of environmental noise. The SNRs on the Table 9 were calculated from the first microphone (x_1). It is clear from this table that the proposed method takes a longer time than the other methods when interview recordings reverberated at 0.15 s reverberation time and surveillance recordings are mixed with different types and levels of environmental noise.

6.3.3 Effect of reverberation time

In these simulation results, the reverberated interview speech was obtained by convolving the interview signals with room impulse response using different reverberation times (T_{20}= 0.15 s, 0.20 s, and 0.25 s). The first configuration of the room was used in this simulation results, as shown in Table 1. The surveillance recordings were corrupted with different levels and types of noise, resulting in a two-channel noisy surveillance recording at SNRs ranging from − 10 to 10 dB, according to Eqs. 7 and 8. Since the performance of speaker verification based on the ICA-EBM algorithm decreased EER compared with other techniques when interview recordings reverberated at 0.15 s and surveillance recordings corrupted by different types of noise at SNRs ranging from − 10 to 0 dB, as described in Sect. 6.3.1,

Table 9 The computation time (s) for different speaker verification methods when interview recordings reverberated at 0.15 s reverberation time and surveillance recordings are corrupted with different types of environmental noise

Methods	Type of noise	SNR (dB)				
		−10	−5	0	5	10
Reverberant interview-noisy surveillance speaker verification baseline	STREET	13533	13508	13476	13423	13397
Traditional ICA	STREET	18355	18321	18273	18246	18193
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	STREET	23756	23738	23686	23639	23591
Reverberant interview-noisy surveillance speaker verification baseline	CAR	13487	13449	13396	13363	13325
Traditional ICA	CAR	18287	18233	18185	18146	18117
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	CAR	23747	23710	23653	23624	23545
Reverberant interview-noisy surveillance speaker verification baseline	HOME	13685	13623	13587	13525	13488
Traditional ICA	HOME	18763	18746	18710	18683	18654
Proposed approach of forensic speaker verification based on the ICA-EBM algorithm	HOME	23973	23879	23834	23769	23715

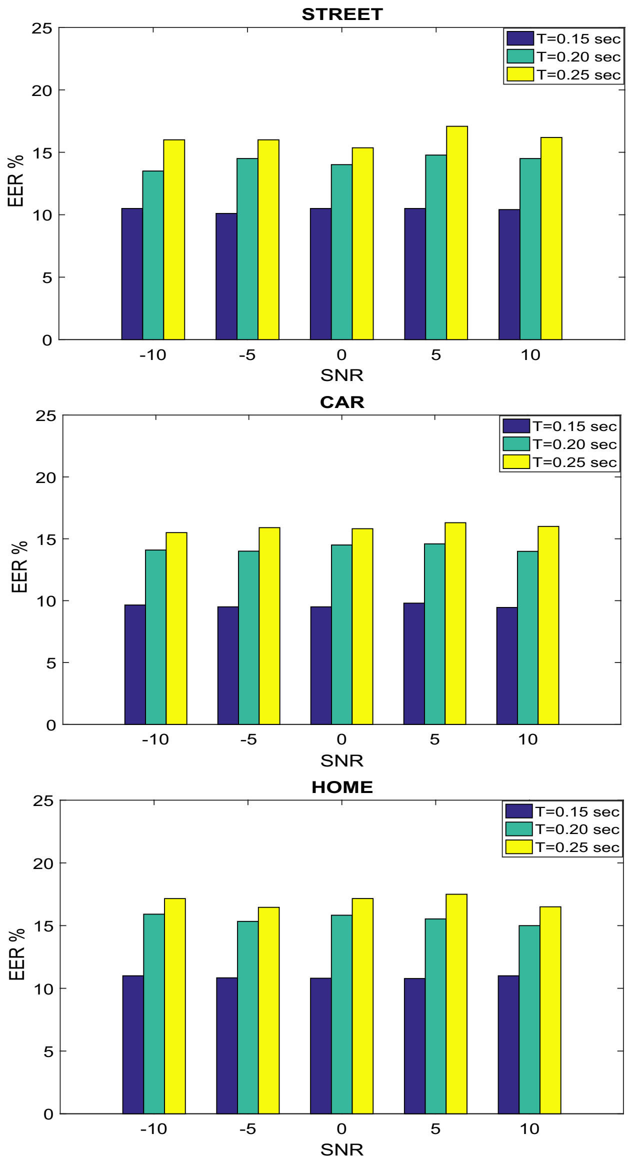

Fig. 12 Effect of the reverberation time on speaker verification based on the ICA-EBM algorithm when interview recordings reverberated at different reverberation times ranging from 0.15 to 0.25 s and surveillance recordings are mixed with different types and levels of environmental noise

the effect of reverberation time was evaluated on speaker verification performance based on the ICA-EBM algorithm in this section.

Figure 12 shows the effect of reverberation time on speaker verification performance based on the ICA-EBM algorithm when interview recordings reverberated at different reverberation times ranging from 0.15 to 0.25 s and surveillance recordings are mixed with different types and levels of environmental noise. The SNRs on the x-axis in Fig. 12 were calculated from the first microphone (x_1). The results show that noisy forensic speaker verification performance based on the ICA-EBM degrades as the reverberation time increases.

The EER degradation for the ICA-EBM when the reverberation time increased from 0.15 to 0.25 s, EER_{deg}, can be calculated

$$EER_{deg} = \frac{EER_{(T=0.25\ s)} - EER_{(T=0.15\ s)}}{EER_{(T=0.25\ s)}} \qquad (24)$$

where $EER_{(T=0.15\ s)}$ and $EER_{(T=0.25\ s)}$ are the equal error rates for the ICA-EBM algorithm when interview recordings reverberated at 0.15 s and 0.25 s, respectively and surveillance speech signals are corrupted by different types of environmental noise. The average EER degradation can be calculated by computing the mean in EER_{deg} for different types of noise at each noise level. At − 10 dB SNR, the average EER degradation of the ICA-EBM is 16.40%, 19.17%, and 17.07%, when the time of reverberation varied from 0.15 to 0.25 s and surveillance recordings were corrupted by HOME, STREET, and CAR noises, respectively. The reverberation time is a parameter that represents the length of the room impulse response. High reverberation time leads to increased distortion in the feature vectors [53]. Therefore, speaker verification performance decreases when the reverberation time increases.

6.3.4 Effect of utterance duration

In real forensic scenarios, the full-length interview utterance of the speech signals from a suspect is often recorded in a police room where reverberation is usually present. However, the surveillance recordings are often mixed with different types of noise, and the utterance length of the surveillance recordings is uncontrolled (typically ranged from few seconds to 40 s) [54]. Thus in this work, the full duration of the interview recordings was convolved with the room impulse response at 0.15 s to produce the reverberated interview signals using the first configuration of the room, as shown in Table 1. The length of the surveillance recordings varied from 10 to 40 s. We mixed the surveillance recordings with different levels and types of environmental noise, resulting in a two-channel noisy surveillance recording at SNRs ranging from − 10 to 10 dB, according to Eqs. 7 and 8.

Figure 13 shows the effect of utterance surveillance duration on speaker verification performance based on the ICA-EBM algorithm when interview recordings reverberated at 0.15 s and surveillance recordings are mixed with different levels and types of noise. The SNRs on the x-axis in Fig. 13 were calculated from the first microphone (x_1). It demonstrates that the EER for the ICA-EBM algorithm reduced when the utterance length of the surveillance signals

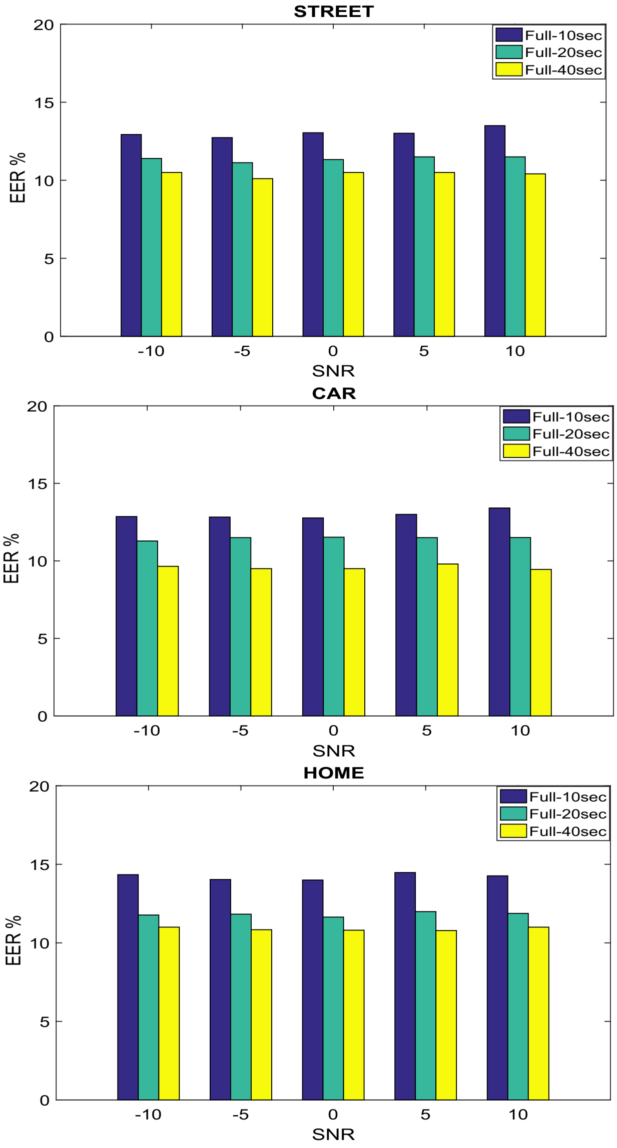

Fig. 13 Effect of utterance surveillance duration on speaker verification performance based on the ICA-EBM when interview recordings reverberated at 0.15 s reverberation time and surveillance recordings are mixed with different types and levels of environmental noise

increased in the presence of different types of environmental noise.

6.3.5 Effect of changing position between microphone and suspect

In these simulation results, the interview recordings reverberated at 0.15 s, and 10 s of the surveillance recordings is corrupted by a random session of HOME, CAR, and STREET noises at SNRs ranging from − 10 to 10 dB using two microphones as in Eqs. 7 and 8. In order to investigate the effect of changing position between microphone and suspect on the performance of speaker verification based on the ICA-EBM algorithm, we used three different configurations of microphone/suspect position, as shown in Table 1 and Fig. 2. Since the performance of speaker verification based on the ICA-EBM decreased EER compared with other techniques when interview recordings reverberated at 0.15 s and surveillance recordings mixed with different types of noise at SNRs ranging from − 10 to 0 dB, as described in Sect. 6.3.1, the effect of microphone and suspect position was evaluated on the speaker verification performance based on the ICA-EBM algorithm in this section.

Table 10 shows the effect of microphone and suspect positions on the speaker verification performance based on the ICA-EBM algorithm when interview recordings reverberated using different configurations and surveillance recordings are corrupted by different levels and types of noise. The SNRs on the Table 10 were calculated from the first microphone (x_1). The simulation results show that changing microphone/suspect position affects the speaker verification performance based on the ICA-EBM. Configuration 1, which has the shortest distance between the microphone and suspect decreased EER compared with other configurations. The performance of speaker verification based on the ICA-EBM algorithm decreased when the distance between the suspect and microphone increased. The impulse response of the room consists of early and late reflections. The characteristics of the early reflections, typically 50 ms after the

Table 10 Effect of microphone and suspect positions on the speaker verification performance based on the ICA-EBM when interview recordings reverberated using different configurations and surveillance recordings are mixed with different types and levels of noise

Configuration	Type of noise	SNR (dB)				
		−10	−5	0	5	10
1	STREET	12.932	12.733	13.045	13.012	13.500
2	STREET	13.251	13.189	13.567	13.550	14.321
3	STREET	13.542	13.398	14.153	14.140	14.672
1	CAR	12.859	12.826	12.771	13.000	13.410
2	CAR	13.186	13.165	13.105	13.467	13.851
3	CAR	13.457	13.389	13.365	13.761	14.223
1	HOME	14.341	14.032	14.000	14.477	14.266
2	HOME	14.615	14.354	14.348	14.731	14.589
3	HOME	14.891	14.769	14.690	14.950	14.896

arrival of the direct sound, depends strongly on the suspect/microphone positions [55]. The duration of the early reflections could increase and leads to increased spectral alteration of the original speech signal when the distance between the suspect and microphone increases. Thus, the performance of speaker verification based on the ICA-EBM degrades when the distance between the suspect and microphone increases.

7 Conclusion

In this paper, we present a new approach to improve forensic speaker verification performance under noisy and reverberant environments. This approach is based on using the ICA-EBM to reduce the effect of noise from noisy surveillance speech signals. Features extracted from the enhanced surveillance speech signals were obtained by using a fusion of feature-warping with MFCC and DWT-MFCC. The i-vector length-normalized GPLDA framework was used as a classifier. Forensic speaker verification performance based on the ICA-EBM algorithm was evaluated under conditions of noise only, as well as noise and reverberation.

Simulation results demonstrate that the proposed speaker verification based on the ICA-EBM improved EER over the traditional ICA algorithm when interview recordings are kept under clean or reverberant conditions and surveillance recordings are mixed with different types of environmental noise at SNRs ranging from − 10 to 10 dB. The ICA-EBM algorithm is better suited to noisy speech separation applications due to its good convergence behavior. Speech/audio signals are usually either super-Gaussian or slightly skewed in nature and, hence, they perform well with ICA-EBM, compared with traditional ICA methods, due to its tighter bounds and superior convergence properties.

Although the proposed method has superior performance for noisy surveillance and reverberant interview recordings, it does not perform as well as the clean interview-noisy surveillance and reverberant interview-noisy surveillance baselines when interview recordings reverberated or kept under clean conditions and surveillance recordings are corrupted by different types of environmental noise at SNRs ranging from 5 dB to 10 dB. Further work is required to use an SNR estimation before the proposed speaker verification based on the ICA-EBM algorithm to determine whether or not the ICA-EBM is used as a speech enhancement algorithm. The effectiveness of using the convolutive ICA algorithm to separate clean interview speech signals from the reverberation can also be investigated in future work.

References

1. Campbell JP, Shen W, Campbell WM, Schwartz R, Bonastre J-F, Matrouf D (2009) Forensic speaker recognition. IEEE Signal Process Mag 26:95–103
2. Mandasari MI, McLaren M, van Leeuwen DA (2012) The effect of noise on modern automatic speaker recognition systems. In: IEEE international conference on acoustic, speech and signal processing, pp 4249–4252
3. Ganapathy S, Pelecanos J, Omar MK (2011) Feature normalization for speaker verification in room reverberation. In: 2011 IEEE international conference on acoustics, speech and signal processing, pp 4836–4839
4. Lehmann EA, Johansson AM, Nordholm S (2007) Reverberation-time prediction method for room impulse responses simulated with the image-source model. IEEE workshop on applications of signal processing to audio and acoustics, pp 159–162
5. Al-Ali AKH, Dean D, Senadji B, Chandran V (2016) Comparison of speech enhancement algorithms for forensic applications. In: 16th Australian international speech science and technology conference, pp 169–172
6. Ribas D, Vincent E, Calvo JR (2015) Full multicondition training for robust i-vector based speaker recognition. In: Proceedings of interspeech, pp 1057–1061
7. Rosca J, Balan R, Beaugeant C (2003) Multi-channel psychoacoustically motivated speech enhancement. In: Proceedings of international conference on multimedia and expo, pp I84–I87
8. González-Rodríguez J, Ortega-García J, Martín C, Hernández L (1996) Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays. In: 4th international conference on spoken language, pp 1333–1336
9. Gannot S, Burshtein D, Weinstein E (2001) Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans Signal Process 49(8):1614–1626
10. Buckley K, Griffiths L (1986) An adaptive generalized sidelobe canceller with derivative constraints. IEEE Trans Antennas Propag 34(3):311–319
11. Borowicz A (2014) A robust generalized sidelobe canceller employing speech leakage masking. Adv Comput Sci Res 11:17–29
12. Jin YG, Shin JW, Kim NS (2014) Spectro-temporal filtering for multichannel speech enhancement in short-time Fourier transform domain. IEEE Signal Process Lett 21(3):352–355
13. Li X-L, Adali T (2010) Independent component analysis by entropy bound minimization. IEEE Trans Signal Process 58(10):5151–5164
14. Sedlák V, Ďuračková D, Záluskỳ R (2012) Investigation impact of environment for performance of ICA for speech separation. IEEE ELEKTRO, pp 89–93
15. Lee SC, Wang JF, Chen MH (2018) Threshold-based noise detection and reduction for automatic speech recognition system in human–robot interactions. Sens J 18(7):1–12
16. Shanmugapriya N, Chandra E (2016) Evaluation of sound classification using modified classifier and speech enhancement using ICA algorithm for hearing aid application. ICTACT J Commun Technol 7(1):1279–1288
17. Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13(4):411–430
18. Hyvärinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 10(3):626–634
19. Bell AJ, Sejnowski TJ (1995) An information–maximization approach to blind separation and blind deconvolution. Neural Comput 7(6):1129–1159

20. Koldovskỳ Z, Málek J, Tichavskỳ P, Deville Y, Hosseini S (2009) Blind separation of piecewise stationary non-Gaussian sources. Signal Process 89(12):2570–2584
21. Al-Ali AKH, Senadji B, Naik GR (2017) Enhanced forensic speaker verification using multi-run ICA in the presence of environmental noise and reverberation conditions. In: IEEE international conference on signal and image processing applications, pp 174–179
22. Comon P (1994) Independent component analysis, a new concept? Signal Process 36(3):287–314
23. Morrison GS, Zhang C, Enzinger E, Ochoa F, Bleach D, Johnson M et al (2015) Forensic database of voice recordings of 500+ Australian English speakers. http://databases.forensic-voice-comparison.net/#australian_english_500
24. Morrison GS, Rose P, Zhang C (2012) Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Aust J Forensic Sci 44(2):155–167
25. Al-Ali AKH, Senadji B, Chandran V (2017) Hybrid DWT and MFCC feature warping for noisy forensic speaker verification in room reverberation. In: IEEE international conference on signal and image processing applications, pp 434–439
26. Dean DB, Sridharan S, Vogt RJ, Mason MW (2010) The QUT-NOISE- TIMIT corpus for the evaluation of voice activity detection algorithms. In: Proceedings of interspeech
27. Novotny O, Plchot O, Glembek O, Cernocky JH, Burget L (2018) Analysis of DNN speech signal enhancement for robust speaker recognition. arXiv preprint arXiv:1811.07629, pp 1–16
28. Lee M, Chang JH (2018) Deep neural network based blind estimation of reverberation time based on multi-channel microphones. Acta Acust United Acust 104(3):486–495
29. Plinge A, Gannot S (2016) Multi-microphone speech enhancement informed by auditory scene analysis. In: 2016 IEEE sensor array and multichannel signal processing workshop, pp 1–5
30. Varga A, Steeneken HJM (1993) 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 12(3):247–251
31. Ferrer L, Bratt H, Burget L, Cernocky H, Glembek O, Graciarena M et al (2011) Promoting robustness for speaker modeling in the community: the PRISM evaluation set. In: Proceedings of NIST 2011 workshop, pp 1–7
32. Pearce D, Hirsch, HG (2000) The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: 6th international conference of spoken language processing, pp 181–188
33. Sohn J, Kim NS, Sung W (1999) A statistical model-based voice activity detection. IEEE Signal Process Lett 6(1):1–3
34. Al-Ali AKH, Dean D, Senadji B, Baktashmotlagh M, Chandran V (2017) Speaker verification with multi-run ICA based speech enhancement. In: 11th international conference on signal processing and communication systems, pp 1–7
35. Taddese BT (2006) Sound source localization and separation, Mathematics and Computer Science. Macalester College
36. Lehmann EA, Johansson AM (2008) Prediction of energy decay in room impulse responses simulated with an image-source model. J Acoust Soc Am 124(1):269–277
37. Adali T, Anderson M, Fu G-S (2014) Diversity in independent component and vector analyses: identifiability, algorithms, and applications in medical imaging. IEEE Signal Process Mag 31(3):18–33
38. Boukouvalas Z, Mowakeaa R, Fu G-S, Adali T (2016) Independent component analysis by entropy maximization with kernels. arXiv preprint, pp 1–6
39. Reynolds DA (1994) Experimental evaluation of features for robust speaker identification. IEEE Trans Speech Audio Process 2(4):639–643
40. Pelecanos J, Sridharan S (2001) Feature warping for robust speaker verification. In: Proceedings of speaker odyssey-speaker recognition workshop, pp 1–6
41. Tzanetakis G, Essl G, Cook P (2001) Audio analysis using the discrete wavelet transform. In: Proceedings conference in acoustic and music theory applications, pp 1–6
42. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11(7):674–693
43. Kanagasundaram A, Dean D, Sridharan S, Gonzalez-Dominguez J, Gonzalez-Rodriguez J, Ramos D (2014) Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Commun 59:69–82
44. Dehak N, Dehak R, Kenny P, Brümmer N, Ouellet P, Dumouchel P (2009) Support vector machines versus fast scoring in the low dimensional total variability space for speaker verification. In: Proceedings of interspeech, pp 1559–1562
45. Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798
46. Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P (2008) A study of interspeaker variability in speaker verification. IEEE Trans Audio Speech Lang Process 16(5):980–988
47. McLaren M, van Leeuwen D (2011) Improved speaker recognition when using i-vectors from multiple speech sources. In: IEEE international conference on acoustic, speech and signal processing, pp 5460–5463
48. Kenny P (2010) Bayesian speaker verification with heavy-tailed priors. Odyssey speaker and language recognition workshop, pp 1–10
49. Garcia-Romero D, Espy-Wilson CY (2011) Analysis of i-vector length normalization in speaker recognition systems. In: Proceedings of interspeech, pp 249–252
50. Sadjadi SO, Slaney M, Heck L (2013) MSR identity toolbox v1. 0: a MATLAB toolbox for speaker-recognition research. Speech Lang Process Tech Comm Newsl 1(4):1–32
51. Al-Ali AKH, Dean D, Senadji B, Chandran V, Naik GR (2017) Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. IEEE Access 5(99):15400–15413
52. Al-Ali AKH (2019) Forensic speaker recognition under adverse conditions. PhD Thesis. Queensland University of Technology, Australia
53. Shabtai NR, Zigel Y, Rafaely B (2008) The effect of GMM order and CMS on speaker recognition with reverberant speech. In: Proceedings of hands-free speech communication and microphone arrays, pp 144–147
54. Mandasari MI, McLaren M, van Leeuwen DA (2011) Evaluation of i-vector speaker recognition systems for forensic application. In: Proceedings of interspeech, pp 21–24
55. Yoshioka T, Sehr A, Delcroix M, Kinoshita K, Maas R, Nakatani T et al (2012) Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process Mag 29(6):114–126