Enhanced Forensic Speaker Verification Using A Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions

Ahmed Kamil Hasan AL-ALI, David Dean, Bouchra Senadji, Vinod Chandran, and Ganesh R. Naik

***Abstract*—Environmental noise and reverberation conditions severely degrade the performance of forensic speaker verification. Robust feature extraction plays an important role in improving forensic speaker verification performance. This paper investigates the effectiveness of combining features, Mel frequency cepstral coefficients (MFCC) and MFCC extracted from the discrete wavelet transform (DWT) of the speech, with and without feature warping for improving modern identity-vector (i-vector) based speaker verification performance in the presence of noise and reverberation. The performance of i-vector speaker verification was evaluated using different feature extraction techniques: MFCC, feature-warped MFCC, DWT-MFCC, feature-warped DWT-MFCC, a fusion of DWT-MFCC and MFCC features and fusion feature-warped DWT-MFCC and feature-warped MFCC features. We evaluated the performance of i-vector speaker verification using the Australian Forensic Voice Comparison (AFVC) and QUT-NOISE databases in the presence of noise, reverberation, and noisy and reverberation conditions. Our results indicate that the fusion of feature-warped DWT-MFCC and feature-warped MFCC is superior to other feature extraction techniques in the presence of environmental noise under the majority of signal to noise ratios (SNRs), reverberation, and noisy and reverberation conditions. At 0 dB SNR, the performance of the fusion of feature-warped DWT-MFCC and feature-warped MFCC approach achieves a reduction in average equal error rate (EER) of 21.33%, 20.00%, and 13.28% over feature-warped MFCC, respectively, in the presence of various types of environmental noises only, reverberation, and noisy and reverberation environments. The approach can be used for improving the performance of forensic speaker verification and it may be utilized for preparing legal evidence in court.**

***Index Terms*—Discrete wavelet transform, environmental noise and reverberation conditions, forensic speaker verification, feature warped-MFCC.**

I. Introduction

The goal of speaker verification is to accept or reject the identity claim of a speaker by analyzing their speech samples [1], [2]. Speaker verification can be used in many applications such as security, access control, and forensic applications [3]. For many years, lawyers, judges, and law enforcement agencies have wanted to use forensic speaker verification when investigating a suspect or confirming the judgment of guilt or innocence [4]. Forensic speaker verification compares speech samples from a suspect (speech trace) with a database of speech samples of known criminals to prepare legal evidence for the court [5].

Automatic speaker recognition systems are often developed and tested under clean conditions [5]. However, in real forensic applications, the speech traces provided to the system are often corrupted by various types of environmental noise such as car and street noises [5]. The performance of speaker verification systems reduces dramatically in the presence of high levels of noise [6], [7].

The police often record speech from the suspect in a room where reverberation is often present. In reverberation environments, the original speech signal is often combined with a multiple reflection version of the speech due to the reflection of the original speech signals from the surrounding room [8]. The reverberated speech can be modeled by the convolution impulse response of the room with the original speech signal. The amount of reverberation can be characterized by reverberation time (T_{20} or T_{60}), which describes the amount of time for the direct sound to decay by 20 dB or 60 dB, respectively [9]. The presence of reverberation distorts feature vectors and degrades the speaker verification performance because of mismatched conditions between trained models and test speech signals [10].

For speaker verification systems, it is important to extract the features from each frame which captures the essential characteristics of the speech signals. There are various feature extraction techniques used in speaker verification algorithms such as mel frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), and perceptual linear predictive coefficients (PLPC) [11], [12]. The MFCC is the most widely used as the feature extraction techniques for modern speaker verification systems and it achieves high performance under clean conditions [13], [14]. However, the performance of the MFCC features drops significantly in the presence of noise and reverberation conditions [8], [14].

A number of techniques, such as cepstral mean subtraction (CMS) [15], cepstral mean variance normalization (CMVN) [16], and RASTA processing [17], have been used to extract features by reducing the effect of noise directly from speaker-specific information. However, these techniques are less effective for non-stationary additive distortion and reverberation environments [8], [18]. Pelecanos *et al.* [19] introduced a

A. K. H. AL-ALI, D. Dean, B. Senadji and V. Chandran are with the Queensland University of Technology, 2 George Street, GPO Box 2434, Brisbane, Queensland 4001 e-mail: ahmedkamilhasan.alali@hdr.qut.edu.au, ddean@ieee.org, b.senadji@qut.edu.au, vinod.chandran2@bigpond.com

G. R. Naik is with MARCS Institute, Western Sydney University, Sydney, Australia e-mail: Ganesh.Naik@westernsdydney.edu.au

feature warping technique to speaker verification to compensate the effect of additive noise and linear channel mismatch in the feature domain. This technique maps the distribution of the cepstral features into a standard normal distribution. Feature warping provides a robustness to noise, while retaining the speaker-specific information that is lost when using other channel compensation techniques such as CMS, CMVN, and RASTA processing [20].

Multiband feature extraction techniques were used in [21]–[24] as the feature extraction of noisy speaker recognition systems. These techniques achieved better performance than traditional MFCC features. Multiband feature techniques are based on combining MFCC features of the noisy speech signals and MFCC extracted from the discrete wavelet transform (DWT) in a single feature vector.

The fusion of MFCC and DWT-MFCC features of the speech signal improves speaker verification performance under noisy and reverberation conditions for two main reasons. Firstly, reverberation affects low frequencies more than high-frequency subbands, since the boundary materials used in most rooms are less absorptive at low frequency subbands [25]. The DWT can be used to extract more features from the low frequency subbands. These features add some important features to the full band of the MFCC. Thus, fusion of MFCC and DWT-MFCC features of the reverberated signals may achieve better forensic speaker verification performance than full band cepstral features in the presence of reverberation conditions. Secondly, the MFCC features extracted from the DWT add more features to the features extracted from the MFCC of the noisy speech signals, thereby assisting in improving speaker recognition performance in the presence of noise. [14].

In this paper, we investigate the effectiveness of combining the features of MFCC and DWT-MFCC of speech signal with and without feature warping for improving i-vector speaker verification performance under noise, reverberation, and noisy and reverberation conditions. We used different individual and concatenative feature extraction techniques for evaluating the modern i-vector forensic speaker verification performance in the presence of various types of environmental noise and different reverberation conditions.

Although the combination of MFCC and DWT was used as the feature extraction technique in [14], [24] to improve the performance of speaker identification systems, the effectiveness of combining *feature warping* with DWT-MFCC and MFCC features individually or concatenative fusion of these features has not been investigated yet for state-of-the-art i-vector forensic speaker verification in the presence of environmental noise only, reverberation, and noisy and reverberation conditions. This is the original contribution of this research.

The remainder of the paper is organized as follows. Section II provides a brief introduction to speech and noise data sets used in this paper. Section III presents feature extraction techniques. The i-vector based speaker verification is described in Section IV. Section V describes the experimental methodology. The results and discussion are presented in Section VI, and Section VII concludes the paper.

II. Speech and noise data sets

This section will briefly outline the Australian Forensic Voice Comparison (AFVC) and QUT-NOISE databases which will be used to construct the noisy and reverberation corpora described in this section.

A. *AFVC database*

The AFVC database [26] consists of 552 speakers. Each speaker was recorded in three speaking styles: informal telephone conversation, information exchange over the telephone, and pseudo-police styles. Informal telephone conversations and information exchange over the telephone were recorded between two speakers using a telephone. For the pseudo-police style, each speaker was interviewed by an interviewer and the speech signals were recorded using a microphone. The clean speech signals were sampled at 44.1 kHz and 16 bit/sample resolution [27]. The AFVC database will be used in this paper because this database contains different speaking style recordings for each speaker, and these speaking styles are often found in casework and police investigations.

B. *QUT-NOISE database*

The QUT-NOISE database [28] consists of 20 noise sessions. The duration of each session is approximately 30 minutes. QUT-NOISE was recorded in five common noise scenarios (CAFE, HOME, CAR, STREET, and REVERB). The noise was sampled at 48 kHz and 16 bit/sample resolution.

For most forensic speaker verification approaches, the clean speech signals from existing speech databases are corrupted with short periods of environmental noise collected separately at a certain noise level. However, while the large number of speakers in the speech databases available to researchers through these approaches allows a wide variety of speakers to be evaluated for speaker verification systems, most existing noise databases such as the NOISEX92 database [29], freesound.org [30], and AURORA-2 [31] have limited conditions and short recordings (less than five minutes). The limited duration of noise databases has lacked the ability to evaluate test speaker recognition systems in a wide range of environmental noise conditions in forensic situations. Therefore, in this paper, we mixed a random session of noise from the QUT-NOISE database with clean forensic audio recordings to achieve a closer approximation to forensic situations.

C. *Construction of noisy and reverberation corpora*

The forensic audio recordings available from the AFVC database [26] cannot be used to evaluate the robustness of forensic speaker verification in the presence of environmental noise and reverberation conditions, because this database contains only clean speech signals. In order to evaluate the performance of the speaker verification systems in the presence of environmental noise and reverberation conditions, we designed two corpora. First, a noisy forensic (QUT-NOISE-AFVC) database, which combined noise from the QUT-NOISE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2728801, IEEE Access

database with clean speech from the AFVC database. Second, the reverberation noisy forensic (QUT-NOISE-AFVC-REVERB) corpus, which combined noise from the QUT-NOISE database with clean speech from the AFVC database in the presence of reverberation. A brief description of each corpus is provided in this section.

1) QUT-NOISE-AFVC database: The objective of designing the QUT-NOISE-AFVC database was to evaluate the robustness of forensic speaker verification under environmental noise conditions. We extracted full duration utterances from 200 speakers using pseudo-police style and short duration utterances (10 sec, 20 sec, and 40 sec) using informal telephone conversation styles. These data can be used as enrolment and test speech signals, respectively. Voice activity detection (VAD) based on Sohn's statistical model [32] was used to remove silence from the enrolment and test speech signals. It was necessary to remove the silent portions from the test clean speech signals before adding the noise because the silence would artificially increase the true short-term active speech signal to noise ratio (SNR) compared to that of the desired SNR. The voice activity detection was applied to clean speech instead of noisy speech signals in this paper because manual segmentation of speech activity segments or speech labelling may be implemented in a forensic scenario when encountering noisy speech [5]. A random session of STREET, CAR, and HOME noises from the QUT-NOISE database [28] was chosen and down-sampled from 48 kHz to 44.1 kHz to match the sampling frequency of the test speech signal. These noises were used in this paper because these types of environmental noise are more likely to occur in real forensic situations. The average noise power was scaled in relation to the reference speech signal after removing the silent region according to the desired SNR. The noisy test speech signals were obtained by sample summing of the test speech signal and the scaled environmental noise at SNRs, ranging from -10 dB to 10 dB.

2) QUT-NOISE-AFVC-REVERB database: The aim of designing the QUT-NOISE-AFVC-REVERB corpus was to investigate the effect of different reverberation conditions on the performance of i-vector forensic speaker verification systems.

Training room impulse responses were computed from the fixed room dimension $3 \times 4 \times 2.5 \ (m)$ using the image source described in [33]. Table I and Figure 1 show reverberation room parameters and a diagram of the room. We extracted full duration utterances from 200 speakers using a pseudo-police interview style. The VAD algorithm [32] was used to remove the silent portions from the speech signals. These data can be used as enrolment speech signals. Each of the enrolment speech signals was convolved with the impulse room response to generate the reverberated speech with the same duration as the clean enrolment speech signal.

In order to investigate the effect of the duration of utterance on noisy speaker verification, the test speech signals were extracted from random sessions of 10 sec, 20 sec, and 40 sec duration from 200 speakers, using the informal telephone conversation style after removing the silent portions using the VAD algorithm [32]. The test speech signals were corrupted with different segments of CAR, STREET, and HOME noises from the QUT-NOISE database [28] at various SNR values ranging from -10 dB to 10 dB.

TABLE I: *Reverberation test room parameter*

Configuration	Source 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)	(2.4, 1, 1.3)
3	(2, 1, 1.3)	(2.8, 1 ,1.3)
4	(2, 1, 1.3)	(2.8, 2.5 ,1.3)

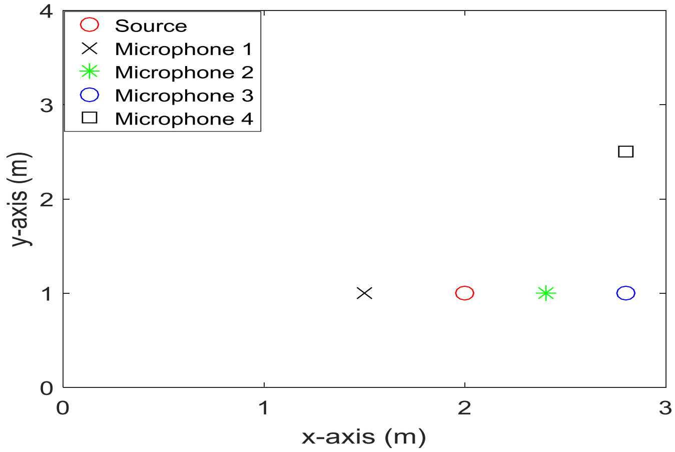

Fig. 1: *Diagram of the room.*

III. Feature extraction techniques

The feature extraction approach can be defined as the process of converting raw speech signals into a small sequence of feature vectors. These feature vectors carry essential characteristics of the speech signal to identify the speaker by their voice [34]. This section describes a brief introduction to the feature extraction techniques which are used in this paper.

A. MFCC feature warping

MFCCs have been widely used as the feature extraction techniques for speaker recognition systems. They are extracted features from the speech signals using cepstral analysis. The human speech production process consists of an excitation source and the vocal tract. The concept of the cepstral features is based on separation of the excitation source and the vocal tract [14]. The basic block diagram of extracting the MFCC features is described in Figure 2.

The first step is to divide the speech signals into frames using an overlapped window. In this research, the speech signal was framed into 30 msec and 10 msec shifts by using a Hamming window. Then, the discrete Fourier transform (DFT) was used to convert the frame of the speech signals from the time domain to the frequency domain. The MFCC can be obtained using a triangular mel filterbank of 32 channels

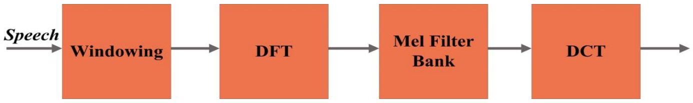

Fig. 2: *A block diagram of extracting the MFCC features*

followed by a transformation to the cepstral domain using discrete cosine transform (DCT). The 13-dimensional MFCC is extracted from each frame of the speech signals. The first and second derivatives of the cepstral coefficients were appended to MFCC features to capture the dynamic properties of the speech signal [12].

Since additive noise and channel distortion corrupt the log-energy of the cepstral features, the distribution of the cepstral features over time undergoes nonlinear distortion [35]. Feature warping [19] was used to compensate this nonlinearity by mapping the distribution of a feature to standard normal distribution. The process of the feature warping is described in the following steps. Firstly, the characteristics of the speech signal can be extracted by using MFCC features. Each cepstral feature can be treated independently over a sliding window (typically three seconds) [19]. Then, the values of the cepstral features are sorted in descending order in a given sliding window. The lookup table can be used to map the rank of the sorted cepstral features into a warped feature using warping normal distribution. The process is repeated by shifting the sliding window for a single frame each time [19].

Given an N points analysis window and the rank R of the middle cepstral feature in the current sliding window, the lookup table (or feature warped components) can be determined by finding m. [19]

$$\frac{N+\frac{1}{2}-R}{N}=\int_{z=-\infty}^{m}\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz \tag{1}$$

where m is the feature warped components.

The warped value m can be estimated initially by putting the rank to $R=N$, solving m by numerical integeration and then repeating for each decremented value of R.

B. Wavelet transform

The wavelet transform is a tool for analyzing the speech signals. It was used to solve the issues related to time and frequency resolution properties in short time Fourier transform (STFT) [36]. Unlike, the STFT that uses fixed window size for all frequency bands, the wavelet transform uses an adaptive window which provides high-time resolution in high-frequency subbands and high-frequency resolution in low-frequency subbands. In that respect, the human auditory system exhibits similar time-frequency resolution properties to the wavelet transform [36].

The DWT is a type of the wavelet transform that can be represented as

$$W(j,k)=\sum_{j}\sum_{k}x(k)2^{\frac{-j}{2}}\psi(2^{-j}n-k) \tag{2}$$

where ψ is the mother wavelet function with finite energy and fast decay, *j* is the number of the level, $x(k)$ is the speech sample, *n* and *k* are integer values. The DWT can be performed using a pyramidal algorithm [37]. Figure 3 shows the block schematic of the dyadic wavelet transform. The speech signal (x) is split into various frequency subbands by using a dyad of finite impulse response (FIR) filters, h and g, which are a low-pass and high-pass filter respectively. The ($\downarrow$ 2) is a down-sampling operator used to discard half of the speech sequences after the filter is performed. The approximation coefficients (CA1) can be obtained by convolving the speech signal with a low-pass filter. The detail coefficients (CD1) can be computed by convolving the speech signals with a high pass filter. The decomposition of the speech signals can be repeated by applying the DWT to the approximation coefficients (CA1).

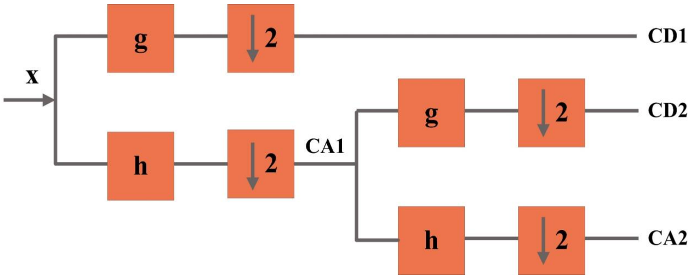

Fig. 3: *Block schematic of the dyadic wavelet transform*

C. Combination of DWT and MFCC feature warping techniques

The technique for extracting the features is based on the multiresolution property of the discrete wavelet transform. The MFCC features were computed over Hamming windowed frames of 30 msec size and a 10 msec shift to discard the discontinuities at the edges of the frame. The MFCC was obtained using a mel filterbank of 32 channels followed by a transformation to the cepstral domain. The 13-dimensional MFCC features, with appended delta (Δ) and double delta ($\Delta\Delta$) coefficients, were extracted from the full band of the noisy speech. Feature warping with a 301 frame window was applied to the features extracted from the MFCC. The DWT was applied to decompose the noisy speech into two frequency subbands: the approximation (low-frequency sub-band) and the detail (high frequency sub-band) coefficients. The approximation and detail coefficients were combined into a single vector. The feature-warped MFCC was then used to extract features from the single feature vector of the DWT.

In this paper, we investigate the effect of feature warping on DWT-MFCC and MFCC features, both individually and in a concatenative fusion of these features in the presence of various types of environmental noise, reverberation, and noisy and reverberation conditions, as shown in Figure 4.

To clarify the feature extraction labels used in Figure 4, the two branches in Figure 4 are labelled 1 and 2. Each branch can also be subdivided into two sub-branches labelled A and B. The output from each sub-branch represents a label of the feature extraction technique and these feature extraction techniques can be combined to generate fusion feature techniques. Tables II and III give a summary of feature extraction labels and a description of the number of the features extracted corresponding to each feature extraction label. The symbol (FW) in Tables II and III represents the acronym of feature warping. The feature extraction techniques described in Table II can be used to train the state-of-the-art i-vector probabilistic linear discriminant analysis (PLDA) speaker verification systems, which will be described in the next section.

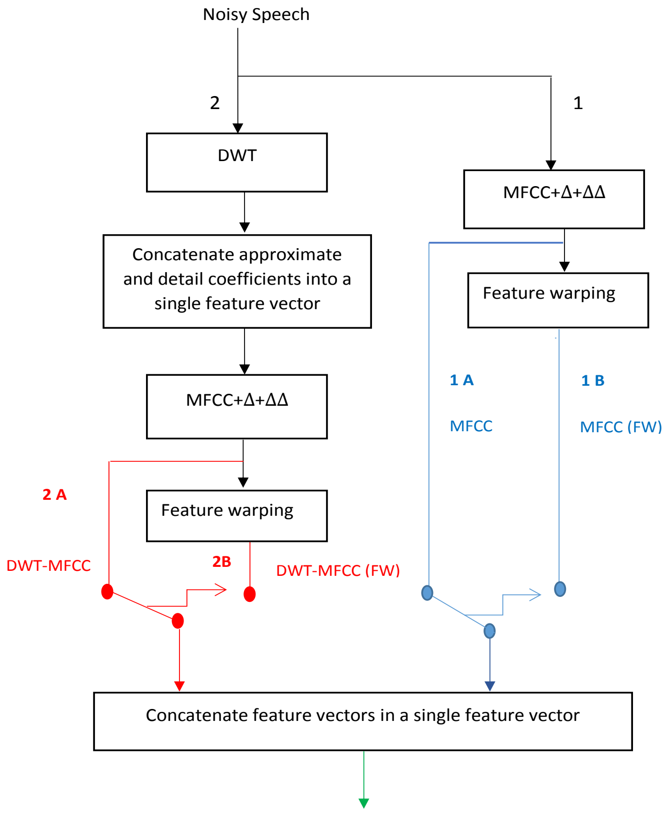
Noisy Speech
2
1
DWT
MFCC+Δ+ΔΔ
Concatenate approximate and detail coefficients into a single feature vector
Feature warping
1 A
MFCC
1 B
MFCC (FW)
MFCC+Δ+ΔΔ
2 A
Feature warping
DWT-MFCC
2B
DWT-MFCC (FW)
Concatenate feature vectors in a single feature vector

Fig. 4: *Extraction and fusion of DWT-MFCC and MFCC features with and without feature warping (FW).*

TABLE II: *Summary of feature extraction labels*

Sub-branch label	Label feature extraction
1 A	MFCC
1 B	MFCC (FW)
2 A	DWT-MFCC
2 B	DWT-MFCC (FW)
Fusion 1 A and 2 A	Fusion (no FW)
Fusion 1 B and 2 B	Fusion both (FW)

TABLE III: *Description of the number of features extracted from each feature extraction labels*

Label feature extraction	Number of features
MFCC	39
MFCC (FW)	39
DWT-MFCC	39
DWT-MFCC (FW)	39
Fusion (no FW)	78
Fusion both(FW)	78

IV. I-VECTOR BASED SPEAKER VERIFICATION

The i-vector was proposed by Dehak *et al.* [38] and it has become a common technique for speaker verification systems. The i-vector can be used in a length normalized Gaussian PLDA (GPLDA) classifier. The i-vector and length normalized GPLDA classifier are outlined in the following sections.

A. I-vector feature extraction

The i-vector represents the Gaussian mixture model (GMM) super-vector by using a single low-dimensional total variability space that contains both speaker and channel variability. This single-subspace was motivated by the discovery that the channel variability space of joint factor analysis (JFA) [39] contains speaker information which could be used in recognizing speakers more efficiently. An i-vector speaker and session dependent GMM super-vector, $\mathbf{s}$, can be represented as [38]

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{3}$$

where $\mathbf{m}$ is the super-vector of the mean from the universal background model (UBM), $\mathbf{T}$ is the low-rank matrix representing the major variability across a large number of development data, and $\mathbf{w}$ is the i-vector which has a standard normal distribution. The i-vectors can be extracted by computing the Baum-Welch zero-order, $\mathbf{N}$, and centralized first-order, $\mathbf{F}$, statistic of the cepstral coefficients extracted from the speech utterances. The statistic is calculated for a given utterance with respect to the number of UBM components (C) and the dimensions of the feature extraction (F). The i-vectors for a given utterance are extracted as in [38]

$$\mathbf{w} = (\mathbf{I} + \mathbf{T}^T\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{\Sigma}^{-1}\mathbf{F} \tag{4}$$

where $\mathbf{I}$ is an identity matrix that has a dimension $CF \times CF$, $\mathbf{N}$ is the $F \times F$ diagonal matrix, and $\mathbf{F}$ is performed through concatenating of the centralized first-order statistics. The covariance matrix $\mathbf{\Sigma}$ is the residual variability matrix. The method for estimating the total variability subspace is described in [38], [40].

The total variability matrix should be trained in both telephone and microphone environments to exploit the useful speaker variability obtained from both sources. McLaren *et al.* [41] investigated the effect of using different types of total variability matrix, such as pooled and concatenated on i-vector speaker verification systems. For the pooled technique, microphone and telephone speech utterances are combined and an individual total variability matrix is used to train this combination of speech signals. For the concatenated total-variability technique, two total-variability matrices for microphone and telephone are trained separately using speech from those sources, then both subspaces are combined to generate a single total-variability space. McLaren *et al.* [41] found that the pooled technique achieved better representation of i-vector speaker verification than the concatenated total variability technique. Thus, the pooled total variability technique will be used in this paper.

B. Length normalized GPLDA classifier

The PLDA was first proposed by Prince *et al.* [42] for face recognition systems, and was later introduced to model i-vector speaker verification by Kenny *et al.* [43]. Kenny investigated two PLDA models: GPLDA and heavy-tailed PLDA

(HTPLDA). They found that HTPLDA improved speaker verification performance significantly compared with the GPLDA model because the distribution of the i-vectors is heavy-tailed [43]. Garcia-Romero *et al.* [44] proposed the length normalized GPLDA technique to transform the behavior of the i-vectors from the heavy-tailed to Gaussian behavior. The results in [44] have indicated that the length normalized GPLDA gives a similar performance with less computational complexity than HTPLDA. Thus, the length normalized GPLDA was used in this paper.

The length normalized GPLDA consists of two steps (a) whitening i-vectors (b) length normalization. The whitening process of i-vector, $\mathbf{w}_{wht}$, can be computed as

$$\mathbf{w}_{wht} = \mathbf{d}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{w} \tag{5}$$

where $\mathbf{\Sigma}$ is the covariance matrix which can be estimated from the development i-vector, $\mathbf{U}$ is an orthogonal matrix including the eigenvectors of the covariance matrix, and $\mathbf{d}$ is the diagonal matrix containing the corresponding of the eigenvalues. The length normalized of i-vector, $\mathbf{w}^{norm}$, can be computed as

$$\mathbf{w}^{norm} = \frac{\mathbf{w}_{wht}}{\|\mathbf{w}_{wht}\|} \tag{6}$$

The length normalization i-vector, $\mathbf{w}^{norm}$, can be represented in the GPLDA model as follows,

$$\mathbf{w}_r^{norm} = \bar{\mathbf{w}}^{norm} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{y}_r + \epsilon_r \tag{7}$$

where $r = 1, 2, 3, \cdots, R$ represents the number of the recordings for each speaker, $\bar{\mathbf{w}}^{norm}$ is the speaker-independent mean of all i-vectors, $\mathbf{U}_1$ and $\mathbf{U}_2$ are the eigenvoice and eigenchannel matrices, respectively. The speaker factors $\mathbf{x}_1$ are assumed to have standard normal distribution and the vector ϵ_r represents the residual term assumed to be a standard normal distribution with a zero mean and covariance matrix ($\mathbf{\Lambda}^{-1}$). The GPLDA model consists of two parts: the speaker part $\bar{\mathbf{w}}^{norm} + \mathbf{U}_1\mathbf{x}_1$ with covariance matrix $\mathbf{U}_1\mathbf{U}_1^T$ and represents between speaker variability. The channel part $\mathbf{U}_2\mathbf{y}_r + \epsilon_r$ with covariance matrix $\mathbf{\Lambda}^{-1} + \mathbf{U}_2\mathbf{U}_2^T$, which represents within speaker variability.

In our experiment, the precision matrix ($\mathbf{\Lambda}$) is assumed to be a full rank and the eigenchannel matrix ($\mathbf{U}_2$) is removed from Equation 7. It was found that removing the eigenchannel did not show significant improvement in speaker verification performance and removing the eigenchannel matrix is useful for decreasing the computational complexity [43], [44]. The modified GPLDA can be represented by

$$\mathbf{w}_r^{norm} = \bar{\mathbf{w}}^{norm} + \mathbf{U}_1\mathbf{x}_1 + \epsilon_r \tag{8}$$

The details of the estimation model parameter $\{\mathbf{U}_1, \mathbf{x}_1, \mathbf{\Lambda}\}$ are given in [43]. The scoring was conducted using the batch likelihood ratio between the normalization i-vector of the target $\mathbf{w}_{target}^{norm}$ and test $\mathbf{w}_{test}^{norm}$ and it can be represented as [43]

$$score = \ln \frac{P(\mathbf{w}_{target}^{norm}, \mathbf{w}_{test}^{norm}|H_1)}{P(\mathbf{w}_{target}^{norm}|H_0)P(\mathbf{w}_{test}^{norm}|H_0)} \tag{9}$$

where H_1 is the hypothesis that the i-vectors come from the same speaker and H_0 is the hypothesis that they do not.

V. Experimental methodology

The i-vector based experiments were evaluated using the AFVC database. A universal background model with 256 Gaussian components was used in our experimental results. The UBMs were trained on telephones and microphones from 348 speakers from the AFVC database. These UBMs were used to compute the Baum-Welch statistics before training a total-variability subspace of dimension 400. These total variabilities were used to compute the i-vector speaker representation. The i-vector dimension was reduced to 200 i-vectors using linear discriminant analysis (LDA). The i-vectors length normalization was used before GPLDA modelling using centering and whitening of the i-vectors [44]. The performance of the i-vector PLDA speaker verification systems was evaluated using the Microsoft Research (MSR) identity toolbox [45].

VI. Results and discussion

This section describes the effectiveness of fusion features of MFCC and DWT-MFCC with and without feature warping on the speaker verification performance under noisy, reverberation, and noisy and reverberation conditions. The modern i-vector PLDA was used as a classifier in all results throughout this paper. The performance of speaker verification systems was evaluated using the equal error rate (EER).

A. Noisy conditions

This section will describe the performance of fusion features of MFCC and DWT-MFCC with and without feature warping in the presence of STREET, CAR, and HOME noises only. The effect of level decomposition and duration utterances on the performance of fusion feature warping with MFCC and DWT-MFCC based speaker verification systems will also be described in this section.

1) Effect of level decomposition: This experiment evaluated the effect of level decomposition used in the performance of fusion feature warping with MFCC and DWT-MFCC features. The full duration of enrolment speech signals was kept in clean conditions, while 10 sec of the test speech signals were corrupted with a random session of STREET, CAR, and HOME noises at SNRs ranging from -10 dB to 10 dB. The enrolment and noisy test speech signals were decomposed into 2, 3, and 4 levels using Daubechies 8 DWT.

Figure 5 shows the effect of the decomposition levels on the performance of fusion feature warping with MFCC and DWT-MFCC features in the presence of various types of environmental noise at SNRs ranging from -10 dB to 10 dB. Lower EER in Figure 5 indicates better performance of noisy forensic speaker verification. It was found that increasing the number of levels to more than three over the majority of SNR values degraded the speaker verification performance in the presence of various types of environmental noise. In this case, the number of samples in the lowest frequency subbands was so low that the essential characteristics of the speech signals could not be estimated accurately by the classifier [23]. Thus, level 3 is used in the feature extraction based on DWT in the presence of noise in the next section.

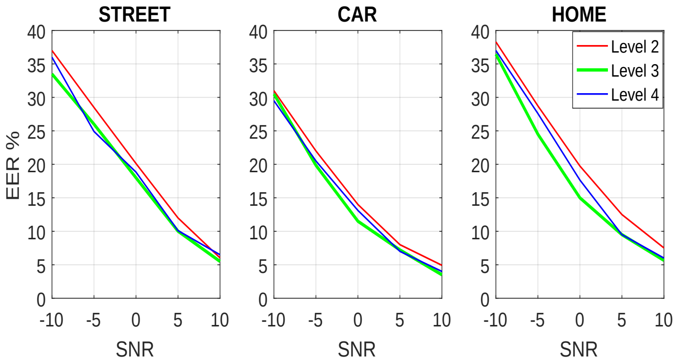

Fig. 5: *Effect of the decomposition levels on the performance of fusion feature warping with MFCC and DWT-MFCC in the presence of noise.*

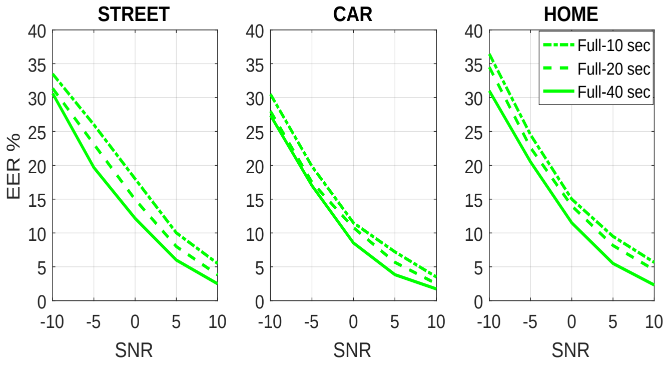

Fig. 6: *Effect of the utterance length on the performance of fusion feature warping with MFCC and DWT-MFCC in the presence of noise.*

2) Effect of utterance length: In real forensic applications, long speech samples from a suspected speaker are recorded in an interview scenario under clean conditions, while the test speech signal is corrupted by environmental noises and the duration of the test speech signals is uncontrolled [5], [46]. Thus in this paper, the full duration of the enrolment speech signals was kept in a clean condition, while the duration of the test speech signals was changed from 10 sec to 40 sec. The test speech signals were corrupted with random segments of STREET, CAR, and HOME noises at SNRs ranging from -10 dB to 10 dB.

Figure 6 shows the effect of the utterance length on the performance of fusion feature warping with MFCC and DWT-MFCC features in the presence of environmental noise. It is clear that increasing the utterance duration improved the performance of forensic speaker verification systems in the presence of STREET, CAR, and HOME noises.

The reduction in EER, when the duration of the test speech signal increases from 10 sec to 40 sec, can be computed as

$$EER_{red} = \frac{EER_{10\ sec} - EER_{40\ sec}}{EER_{10\ sec}} \tag{10}$$

where $EER_{10\ sec}$ and $EER_{40\ sec}$ are the EER of fusion feature-warped DWT-MFCC and feature-warped MFCC features when the duration of the test speech signals is 10 sec and 40 sec respectively. The average reduction in EER can be computed by calculating the mean of EER_{red} for various types of environmental of noise at each noise level. Figure 7 shows the average reduction in EER for fusion feature warped DWT-MFCC and feature-warped MFCC features when the duration of the test speech signal increased from 10 sec to 40 sec. In 0 dB SNR, the peformance of fusion feature-warped with DWT-MFCC and feature-warped MFCC features achieved an average reduction in EER of 17.92% when the duration of the test speech signals increased from 10 sec to 40 sec.

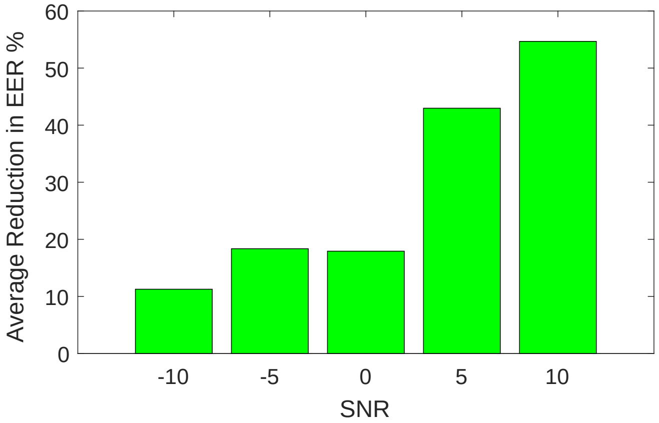

Fig. 7: *Average reduction in EER for fusion feature-warped DWT-MFCC and feature-warped MFCC features when the duration of the test speech signals increased from 10 sec to 40 sec. Higher average reduction in EER indicates better performance*

3) Comparison of feature extraction techniques under noisy conditions : This experiment evaluated the performance of combining MFCC and DWT-MFCC features with and without feature warping in the presence of various levels of environmental noise. The full length of enrolment speech signals was used, while 10 sec of the test speech signals was mixed with random sessions of STREET, CAR, and HOME noises at SNRs ranging from -10 dB to 10 dB. Figure 8 shows a comparison of speaker verification systems using different feature extraction techniques in the presence of environmental noise at various SNR values. We conclude the following points from this figure:

- The performance of forensic speaker verification systems achieves significant improvements in EER over the majority SNR values when applying feature warping to the MFCC features in the presence of various types of environmental noises (blue solid vs blue dash).
- Fusion of feature warping with MFCC and DWT-MFCC features achieves greater improvements in EER than fusion without any feature warping in the presence of various levels of environmental noises (green solid vs green dash).
- Fusion feature warping with MFCC and DWT-MFCC achieves significant improvements in EER over traditional MFCC features in the presence of various types and levels

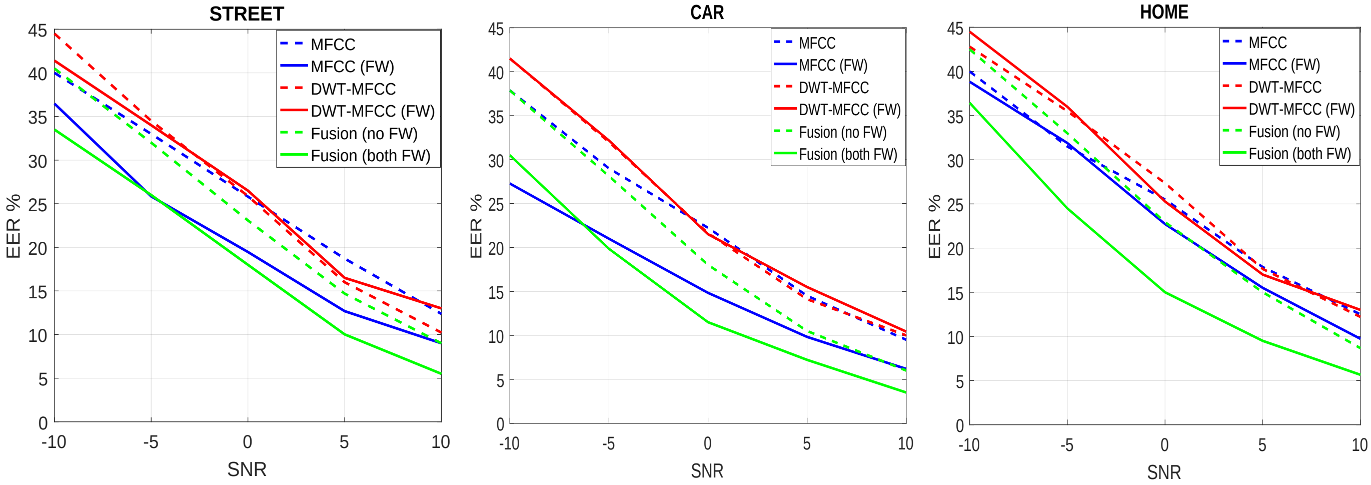

Fig. 8: *Comparison of speaker verification system using different feature extraction techniques in the presence of noise.*

of environmental noises (green solid vs blue dash). The reduction in EER for fusion feature warping with MFCC and DWT-MFCC at 0 dB SNR is 48.28%, 30.27%, and 41.17%, respectively, over MFCC features in the presence of random sessions of CAR, STREET, and HOME noises.

- Feature warping did not improve the performance of the forensic speaker verification system when DWT-MFCC was used as the feature extraction. However, the performance of speaker verification improves by applying feature warping to MFCC features (red solid vs blue solid). The major drawback of using DWT-MFCC (FW) as the feature extraction is that it lost some important correlation information between subband features. The lack of correlation information between subband features decreases the performance of speaker verification systems [47].

The reduction in EER for the fusion of feature warping with MFCC and DWT-MFCC features over feature-warped MFCC , EER_{red}, can be computed as

$$EER_{red} = \frac{EER_{MFCC(FW)} - EER_{fusion}}{EER_{MFCC(FW)}} \tag{11}$$

where $EER_{MFCC(FW)}$ is the equal error rate for feature-warped MFCC and EER_{fusion} is the equal error rate for fusion feature warping with MFCC and DWT-MFCC features. The average reduction in EER can be computed by calculating the mean of EER_{red} for various types of environmental noise at each noise level.

Figure 9 shows average reduction in EER for fusion feature warping with MFCC and DWT-MFCC over feature-warped MFCC features in the presence of various types of environmental noise for each noise level. The results show that fusion feature warping with MFCC and DWT-MFCC achieves a reduction in average EER over feature-warped MFCC features in the presence of various types of environmental noise at SNRs ranging from -10 dB to 10 dB. At 0 dB SNR, the average reduction in EER for fusion feature-warping with MFCC and DWT-MFCC over feature-warped MFCC is 21.33%.

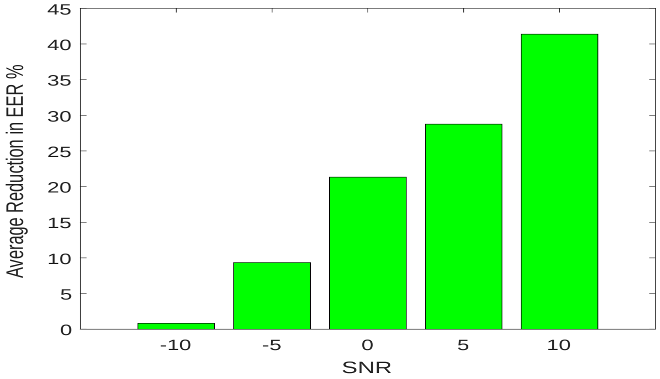

Fig. 9: *Average reduction in EER for fusion feature warping with MFCC and DWT-MFCC over feature-warped MFCC in the presence of various types of environmental noise. Higher average reduction in EER indicates better performance*

B. *Reverberation conditions*

This section will describe the performance of speaker verification based on the fusion features of MFCC and DWT-MFCC with and without feature warping under reverberation conditions only. The effect of decomposition level, utterance length, reverberation time, and position of source and microphone on the performance of forensic speaker verification will also be presented in this section.

1) Effect of decomposition level : The effect of the decomposition level on the performance of fusion feature warping with MFCC and MFCC-DWT was evaluated by using different decomposition levels. We computed the impulse response of a room by using reverberation time (T_{20}= 0.15 sec). The T_{20} was used instead of T_{60} in this paper because T_{20} reduces the computational time when computing the time reverberation in a simulated room impulse response [9]. Each of the enrolment speech signals was convolved with the impulse room response to generate the reverberated speech, while a 10 sec duration of the test speech signals was kept in a clean condition. The first configuration of the room is used in this experiment, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2728801, IEEE Access

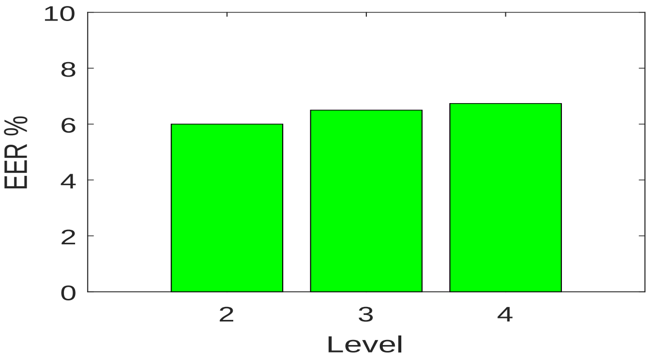

Fig. 10: *Effect of level decomposition on the performance of fusion feature warping with MFCC and DWT-MFCC under reverberation conditions only.*

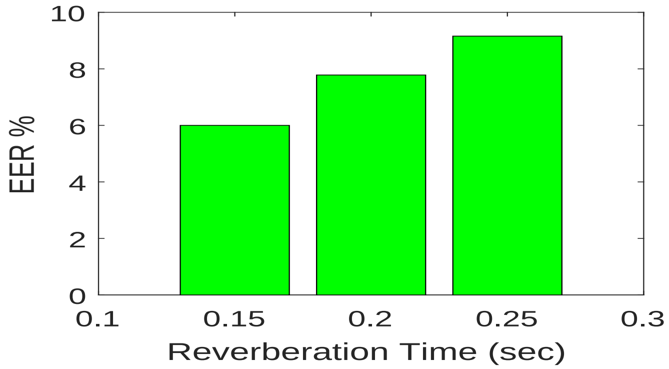

Fig. 11: *Effect of reverberation time on the performance of fusion feature warping with MFCC and DWT-MFCC.*

shown in Table I and Figure 1.

In this experiment, we used Daubechies 8 of the DWT and different decomposition levels (2, 3, and 4) to investigate the effect of the decomposition levels on the performance of fusion feature warping with MFCC and DWT-MFCC under reverberation conditions only. Figure 10 shows the effect of level decomposition on the performance of fusion feature warping with MFCC and DWT-MFCC under reverberation conditions only.

It was found from Figure 10 that level 2 achieves better improvement in performance than other decomposition levels. Reverberation often affects low frequencies more than high frequencies, since the materials used in the most popular rooms are less absorptive at low frequencies, leading to longer reverberation times and more distortion of the spectral information at those frequencies [25]. Thus, the performance of speaker verification in reverberation environments improved by increasing the number of coefficients at a low frequency using two levels of decomposition.

2) Effect of reverberation time: This experiment evaluated the effect of reverberation time on the performance of fusion feature warping with MFCC and DWT-MFCC (level 2) by using different reverberation times. We computed the impulse response of the room by using the following reverberation times: T_{20}= 0.15 sec, 0.20 sec, and 0.25 sec. Each impulse room response matrix was convolved with enrolment speech data to generate reverberated enrolment data at different reverberation times, while a 10 sec duration of the test speech signals was maintained in a clean condition. The first configuration of the room was also used in this experiment, as shown in Table I and Figure 1.

Figure 11 shows the effect of reverberation time on the performance of fusion feature warping with MFCC and DWT-MFCC. The performance of speaker verification was degraded by increasing the reverberation time. There was a degradation of 34.42% in the performance of fusion feature warping with MFCC and DWT-MFCC when the reverberation time increased from 0.15 sec to 0.25 sec. The reverberation adds more inter-frame distortion to the cepstral features when the reverberation time was increased. Therefore, increasing the reverberation time leads to decreased speaker verification performance [48].

3) Comparison of feature extraction techniques under reverberation conditions: The performance of i-vector speaker verification was evaluated using various feature extraction techniques in the presence of reverberation, as shown in Figure 12. The enrolment of the speech signals was reverberated at 0.15 sec reverberation time, while a 10 sec portion of the test speech signals was kept in a clean condition. The first configuration of the room was used in this experiment, as shown in Table I and Figure 1. It was found from Figure 12 that fusion feature warping with MFCC and DWT-MFCC features (level 2) improves the performance of speaker verification over other feature extraction techniques and it achieves a reduction in EER of 20.00% over feature-warped MFCC. The performance of forensic speaker verification under reverberation conditions achieved significant improvements in EER when feature warping was applied to MFCC features. The performance of speaker verification based on the subband features (DWT-MFCC and DWT-MFCC (FW)) degraded in the presence of reverberation because of subband features lost some important information between subband features.

4) Effect of utterance duration: We investigated the effect of varying utterances duration on the i-vector PLDA speaker

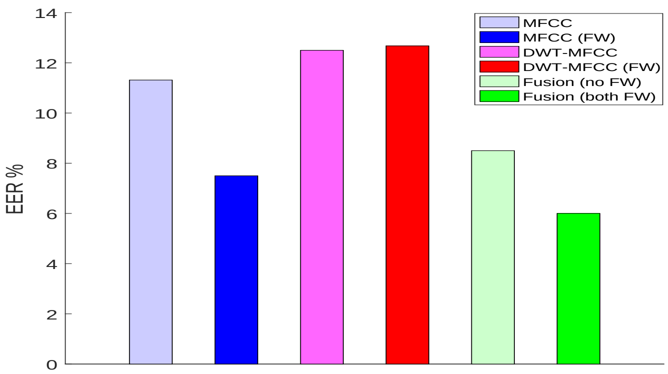

Fig. 12: Comparison of speaker verification performance using different feature extraction techniques in the presence of reverberation.

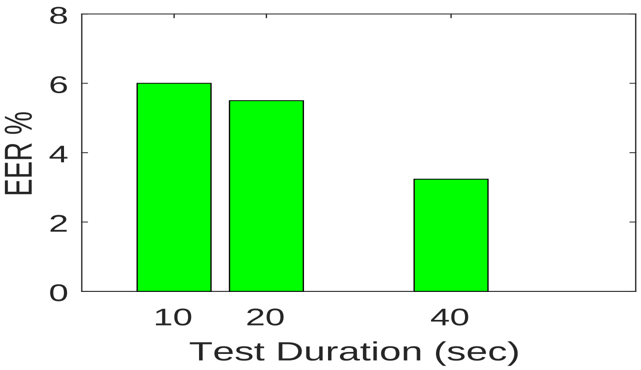

Fig. 13: Effect of test utterance duration on the performance of fusion feature-warping with MFCC and DWT-MFCC under reverberation conditions.

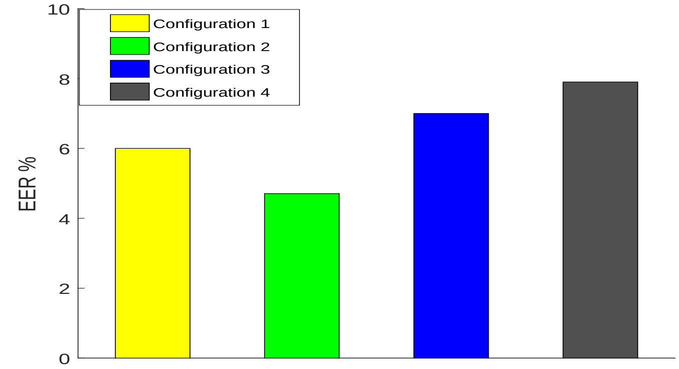

Fig. 14: Effect of configuration microphone and source positions on the performance of fusion feature warping with MFCC and DWT-MFCC features

verification systems in the presence of reverberation conditions only. In this experiment, we reverberated the full duration of the enrolment speech signal at 0.15 sec using the first configuration of the room described in Table I and Figure 1, while the duration of the test speech signals was changed from 10 sec to 40 sec.

Figure 13 shows the effect of test utterance duration on the performance of fusion feature warping with MFCC and DWT-MFCC (level 2) in the presence of reverberation conditions only. The results show that as the utterance length increases, the performance of fusion feature warping with MFCC and DWT-MFCC improves. The reduction in EER is approximately 46.04$\%$ when the duration of the test speech signals increased from 10 sec to 40 sec.

5) Effect of source and microphone position : In this experiment, the enrolment speech signals reverberated at 0.15 sec, while 10 sec of test speech signals was kept in clean conditions. The position of the source signals was not changed and four different positions of the microphone were used to investigate the effect of source/ microphones position on the performance of fusion feature warping with MFCC and DWT-MFCC (level 2). The configuration of source/ microphones used in these experimental results is shown in Table I and Figure 1.

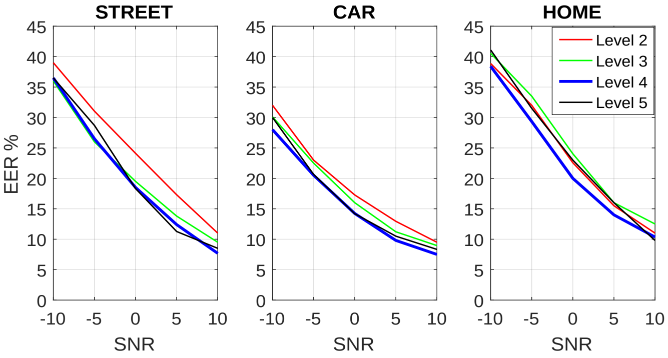

Fig. 15: *Effect of the decomposition levels on the performance of fusion feature warping with MFCC and DWT-MFCC in the presence of reverberation and various types of environmental noises.*

Figure 14 shows the effect of microphone/ source positions on the performance of fusion feature warping with MFCC and DWT-MFCC. The results demonstrate that changing the distance between the source and microphone affects the performance of fusion feature warping with MFCC and DWT-MFCC. Configuration 2, which has the shortest distance between the source and microphone, achieved the highest improvement in EER compared with other configurations. The performance of fusion feature warping with MFCC and DWT-MFCC decreased when the distance between the source and microphone increased.

C. Noisy and reverberation conditions

The performance of fusion feature warping with MFCC and DWT-MFCC was evaluated and compared with speaker verification based on traditional MFCC and feature-warped MFCC under noisy and reverberation conditions. The effect of level decomposition and utterance length will also be discussed in this section.

1) Effect of decomposition level on noisy and reverberation conditions: The effect of the decomposition level on the performance of fusion feature warping with MFCC and DWT-MFCC was evaluated using Daubechies 8 of DWT and different levels (2, 3, 4, and 5). The full duration of the enrolment speech signals reverberated at 0.15 sec. Ten seconds of the test speech signals was corrupted with different segments of CAR, STREET, and HOME noises from the QUT-NOISE database [28] at SNRs ranging from -10 dB to 10 dB.

Figure 15 shows the effect of the decomposition levels on the performance of fusion feature warping with MFCC and DWT-MFCC in the presence of reverberation and various types of environmental noises. It is clear that level 4 achieves better performance in EER over the majority of SNR values and different types of environmental noises.

2) Comparison of feature extraction techniques under noisy and reverberation conditions: This section compares the performance of fusion feature warping with MFCC and DWT-MFCC (level 4) with traditional MFCC and feature-warped MFCC in the presence of reverberation and different

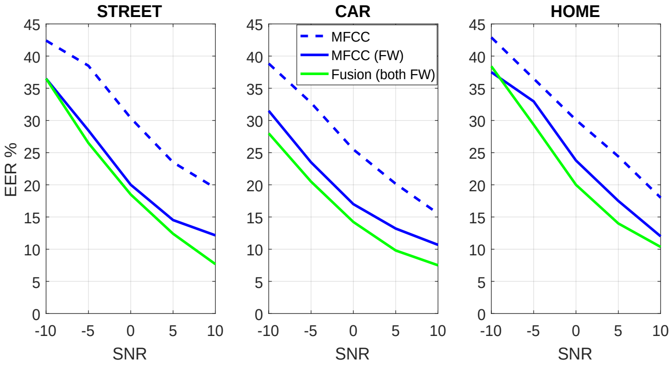

Fig. 16: Comparison of speaker verification performance using different feature extraction techniques in the presence of environmental noise and reverberation conditions

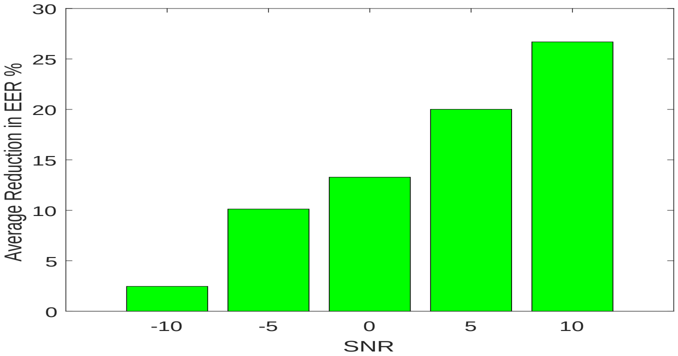

Fig. 17: Average reduction in EER for fusion feature warping with MFCC and DWT-MFCC over feature-warped MFCC in the presence of various types of environmental noise and reverberation conditions. Higher average reduction in EER indicates better performance

types of environmental noise. In these experimental results, the enrolment speech signals reverberated at 0.15 sec and 10 sec of the test speech signals was mixed with different sessions of CAR, STREET, and HOME noises at SNRs ranging from -10 dB to 10 dB. The first configuration of the room is used in this experiment, as shown in Table I and Figure 1.

Figure 16 shows comparison of speaker verification performance using different feature extraction techniques in the presence of environmental noise and reverberation conditions. Overall, the results show that fusion feature warping with MFCC and DWT-MFCC achieves improvements in EER over feature-warped MFCC, when the test speech signals were corrupted with random segments of STREET, CAR, and HOME noises at various SNR values. The results also demonstrate that feature-warped MFCC achieved significant improvements in EER compared with traditional MFCC.

The average reduction in EER for fusion feature warping with MFCC and DWT-MFCC over feature-warped MFCC features was computed by calculating the mean of the EER reduction for various types of environmental noise at each noise level in the presence of reverberation, as shown in Figure 17. The results demonstrate that the performance of fusion feature warping with MFCC and DWT-MFCC outperforms feature-warped MFCC in average reduction of EER at SNRs ranging from -10 dB to 10 dB. At 0 dB SNR, the average reduction in EER of fusion feature warping with MFCC and DWT-MFCC is 13.28% over feature-warped MFCC in the presence of various types of environmental noise and reverberation conditions.

3) Effect of utterance length: In order to evaluate the effect of utterance length on the performance of fusion feature warping with MFCC and DWT-MFCC in the presence of environmental noise and reverberation conditions, we mixed random sessions of STREET, CAR, and HOME noises from the QUT-NOISE database [28] with 10, 20, and 40 seconds from the test speech signals. The full duration of the enrolment speech signals was reverberated at 0.15 sec without adding environmental noises.

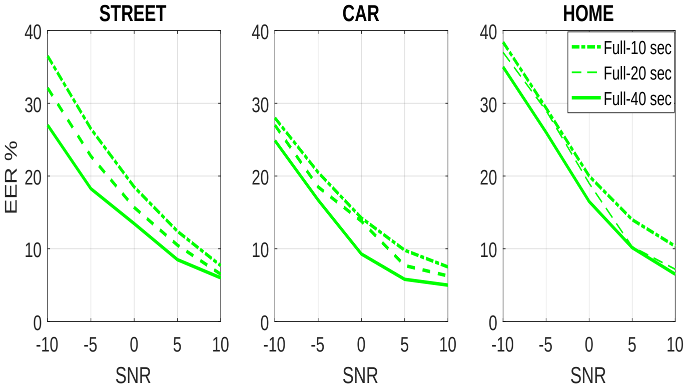

Fig. 18: *Effect of utterance length on the performance of fusion feature warping with MFCC and DWT-MFCC in the presence of noise and reverberation conditions*

Figure 18 shows the effect of utterance length on the performance of fusion feature warping with MFCC and DWT-MFCC features (level 4) in the presence of noise and reverberation environments. It was found that the performance of speaker verification under noisy and reverberation conditions improved when the duration of the test speech signal increases from 10 sec to 40 sec at various types and levels of environmental noise. The average reduction in EER for fusion feature-warped DWT-MFCC and feature-warped MFCC features was 26.51 % when the duration of the test speech signals increased from 10 sec to 40 sec in the presence of reverberation and various types of environmental noise at 0 dB SNR as shown in Figure 19 .

VII. Conclusion

This paper introduced the use of DWT-based MFCC features and their combination with traditional MFCC features for forensic speaker verification. It evaluated the performance of these features with and without feature warping. A state-of-the-art i-vector PLDA based speaker verification was used as a classifier in this paper. The performance of i-vector speaker verification has been evaluated in the presence of environmental noise only, reverberation, and noisy and reverberation conditions. Experimental results indicate that the fusion feature

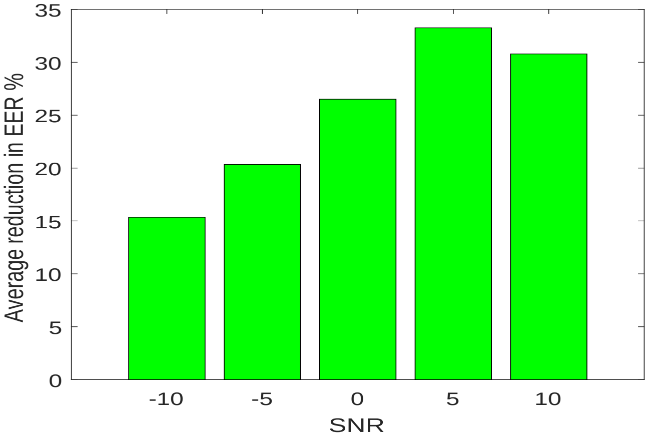

Fig. 19: *Average reduction in EER for fusion feature-warped DWT-MFCC and feature-warped MFCC features when the duration of the test speech signals increased from 10 sec to 40 sec in the presence of reverberation and various types of environmental noises.*

warping DWT-MFCC and feature-warped MFCC approach achieved better performance under most environmental noise, reverberation, and noisy and reverberation environments. The robustness in the performance of the fusion feature approach could be used in forensic applications. In future work, we will evaluate the performance of the fusion feature approach using other databases such as NIST 2010 and the performance will also be evaluated using reverberation used in the QUT-NOISE database.

References

[1] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 859–872, 1997.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2002, pp. IV– 4072–IV– 4075.

[4] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, 2009.

[5] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4249–4252.

[6] S. Kim, M. Ji, and H. Kim, "Robust speaker recognition based on filtering in autocorrelation domain and sub-band feature recombination," *Pattern Recognition Letters*, vol. 31, no. 7, pp. 593–599, 2010.

[7] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[8] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature normalization for speaker verification in room reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4836–4839.

[9] E. A. Lehmann, A. M. Johansson, and S. Nordholm, "Reverberation-time prediction method for room impulse responses simulated with the image-source model," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 159–162.

[10] N. R. Shabtai, Y. Zigel, and B. Rafaely, "The effect of room parameters on speaker verification using reverberant speech," in *25th IEEE Convention of Electrical and Electronics Engineers*, 2008.

[11] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proceedings of the SPECOM*, vol. 1, 2005, pp. 191–194.

[12] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

[13] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2012.

[14] A. Shafik, S. M. Elhalafawy, S. Diab, B. M. Sallam, and F. A. El-Samie, "A wavelet based approach for speaker identification from degraded speech," *International Journal of Communication Networks and Information Security*, vol. 1, no. 3, pp. 52–58, 2009.

[15] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[16] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, 1998.

[17] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[18] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A robust speaker identification system using the responses from a model of the auditory periphery," *PloS one*, vol. 11, no. 7, pp. 1–21, 2016.

[19] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[20] A. Kanagasundaram, "Speaker verification using i-vector features," Ph.D. dissertation, Queensland University of Technology, 2014.

[21] S. Kim, M. Ji, and H. Kim, "Noise-robust speaker recognition using subband likelihoods and reliable-feature selection," *ETRI Journal*, vol. 30, no. 1, pp. 89–100, 2008.

[22] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proceedings IEEE International Conference Acoustics, Speech and Signal Processing*, vol. 2, 1998, pp. 641–644.

[23] W.-C. Chen, C.-T. Hsieh, and E. Lai, "Multiband approach to robust text-independent speaker identification," *Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 63–76, 2004.

[24] M. I. Abdalla, H. M. Abobakr, and T. S. Gaafar, "DWT and MFCCs based feature extraction methods for isolated word recognition," *International Journal of Computer Applications*, vol. 69, no. 20, 2013.

[25] N. Mirghafori and N. Morgan, "Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers." in *International Conference Spoken Language Processing*, 1998, pp. 743–746.

[26] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow, "Forensic database of voice recordings of 500+ Australian English speakers," *URL: http://databases. forensic-voice-comparison. net*, 2015.

[27] G. S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," *Australian Journal of Forensic Sciences*, vol. 44, no. 2, pp. 155–167, 2012.

[28] 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 *Proceedings of Interspeech, Makuhari, Japan*, 2010.

[29] A. Varga and H. J. 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[30] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of NIST 2011 workshop*, 2011.

[31] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions." in *Sixth International Conference on Spoken Language Processing*, 2000, pp. 29–32.

[32] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[33] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model,"

Journal of the Acoustical Society of America, vol. 124, no. 1, pp. 269–277, 2008.

[34] S.-H. Chen and Y.-R. Luo, "Speaker verification using MFCC and support vector machine," in *International MultiConference of Engineers and Computer Scientists*, vol. 1, 2009.

[35] R. J. Vogt, "Automatic speaker recognition under adverse conditions," Ph.D. dissertation, Queensland University of Technology, 2006.

[36] G. Tzanetakis, G. Essl, and P. Cook, "Audio analysis using the discrete wavelet transform," in *Proceeding Conference in Acoustics and Music Theory Applications*, 2001.

[37] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[38] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transaction Audio, Speech, Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[39] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification." in *Interspeech*, vol. 9, 2009, pp. 1559–1562.

[40] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[41] M. McLaren and D. van Leeuwen, "Improved speaker recognition when using i-vectors from multiple speech sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5460–5463.

[42] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *11th IEEE International Conference on Computer Vision.*, 2007, pp. 1–8.

[43] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey Speaker and Language Recogntion Workshop*, 2010.

[44] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech Conference on Speech Communication and Technology*, 2011, pp. 249–252.

[45] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1. 0: A MATLAB toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, 2013.

[46] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application." in *Interspeech*, 2011, pp. 21–24.

[47] J. McAuley, J. Ming, D. Stewart, and P. Hanna, "Subband correlation and robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 956–964, 2005.

[48] N. R. Shabtai, Y. Zigel, and B. Rafaely, "The effect of GMM order and CMS on speaker recognition with reverberant speech," in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 144–147.