Multimedia Tools and Applications (2021) 80:22231–22249
https://doi.org/10.1007/s11042-021-10767-6

Improving short utterance speaker verification by combining MFCC and Entrocy in Noisy conditions

Khamis A. Al-karawi[1] · Duraid Y. Mohammed[2]

Received: 18 May 2020 / Revised: 22 December 2020 / Accepted: 25 February 2021/
Published online: 25 March 2021

Abstract

Short utterance and background noise represent great challenging for speaker verification due to the mismatch and limited training and/or retrieve data. A remarkable performance using matched training and testing conditions generally could be achieved in automatic speaker verification. However, mismatched noisy and short utterances conditions attend to drop the results significantly. Furthermore, the performance is significantly affected by the features extraction. The most common features in this field of the study are Mel-Frequency Cepstral Coefficients (MFCCs). With a noise presents in the background and short utterances, MFCC performance could not be reliable without a support feature. To address this, a new feature 'Entrocy' for accurate and robust speaker verification under limited data and noisy environments is proposed and employed to support MFCC coefficients. Entrocy feature represents the Fourier Transform of the Entropy that calculates the fluctuation of the information in the sound segments over time. The resulting Entrocy features are combined with MFCC functionality to generate a composite feature, which is tested using the Gaussian Mixture Model (GMM) recognition method. The suggested method was conducted out over a range of signal/noise ratios and utterances were truncating into shorts (2, 3, 4, 5, 6, 8, and 10s) for verification. The proposed method has shown strong robustness in the challenging of background noise and limited testing data and they consistently perform better than the well-known MFCC.

Keywords Speaker verification · Short utterance · I-vector · GMM · Robustness · MFCC · Noisy environment · Entrocy

✉ Khamis A. Al-karawi
alkasi_68@yahoo.com

Duraid Y. Mohammed
duraidyehya19@gmail.com

[1] Diyala University, Baqubah, Diyala, Iraq

[2] School of education for women, Al-Iraqia University, Baghdad, Iraq

1 Introduction

Real-world speech recording samples when overlapping with acoustic conditions such as adding noise and room reverberation in addition to the short utterances which are represented a major challenges against the robustness of automatic speaker recognition (ASR). The verification of speech segments is important for robust speaker recognition but becomes relatively difficult in noisy environments [4]. Many studies have been reported the aforementioned challenges. Some of which in feature domain, such as cepstral mean subtraction approach [9], relative spectral processing method [10], and combine MFCC's feature with Gammatone Frequency Cepstral Coefficients (GFCC) [1] used to reduce additive and convolutional distortions of the channel. Another approach used to reduce the effect of these challenges is using training and retrieval speaker models. For example, Al-Karawi et al. used noisy samples, thus reducing the discrepancy between the developed model as reference and retrieval samples [3]. Additional approach to reduce the effect of additive noise is using multi-condition training. In this method, each speaker samples has been mixed with different noisy conditions. During the recognition phase, the reference model which is closest to the features of the input speech sample is then selected [2, 20]. Regarding short utterance challenge, Zhao et al. has used spectral energy that calculated through short segments has been as dominant features to discriminate the speech samples from other soundtracks [30]. Additional effort is proposed by [8, 14] to increase the authentication accuracy utilizing short speech signal (2 words in maximum) in retrieval stage, and dealing with the condition where both training and retrieval signal is short. However, the robustness and reliability of these features in a noisy environment are affected negatively especially in the case of overlapping with sound artefacts and non-stationary noise such as heavy breathing, and mouth clicks, etc. Furthermore, the speech samples quality represents one of the main consideration that affecting performance [31] which makes the using of Voice Activity Detection VAD technique in the samples pre-processing step crucial for removing the silence frames [13]. The significant results and performance of the MFCCs besides low estimation algorithm complexity is considered the main reasons behind the widespread of employing it for ASR tasks in clean, matched conditions [10]. However, the MFCC fails to achieve adequate accuracy in the case of reverberation or noise are presence [30]. The inadequate performance of MFCC's coefficients in the presence of noisy, reverberant or mismatched conditions was the main motivation to develop and investigate robust extraction methods [30]. Accordingly, new technique that employing noise adaptive threshold have been suggested in [11], but the presence of sound artefacts and relatively high noise levels makes the performance drops significantly. That is why it is suggested that the Entrocy-based algorithm be combined with MFFC feature in this work to overcome the mismatched and noisy conditions. An Entropy-based method has been proposed and developed by the authors as a hybrid feature for tackling the overlapped audio classes challenge in [19] and has shown good improvement in the detection of the music segments. Therefore, the developed technique based on the entropy in time-frequency domain, here referred to as the Spectral Entropy, is combined with MFCC coefficients in this study. According to the spectral Entrocy, the spectrum probability density function (pdf) is firstly computed for each single frame of the input speech sample. The findings indicate the effectiveness of the suggested method in discriminating the segments of speech from the non-speech in a continuously recorded utterance, particularly in unclean speech background. In this paper, we propose a new system using combine feature in order to improve robustness with limited speech data duration. We demonstrate the usefulness of these coefficients

compared to the well-known features with speakers taken from different databases recorded under different conditions. Remainder of this study is structured according to the following. Section-2 describes the rationale of the study, describes the calculation of the features in Section-3, and explains the experimental setup in Section-4, presents the experimental outcomes in Section−5 and as a final point, we discussed conclusions in Section-6.

2 Rationale and system architecture

Most of the Existing classification features have been mainly calculated and constructed on non-overlapping audio frames or segments that are artificially configured. While the soundtracks in the real world could be speech, music, audio events, or a combination of them. Mohammed et al. have been therefore suggested alternative audio attributes for enhancing the discrimiminitaing of the speech from non-speech and it was shown significant results for speech detection regardless the speech was pure or non-pure [19]. The developed feature is called Entrocy based on the calculation methodology that depends on the computations of entropy-frequency combination. The main concept in the measurement of frequencies is to measure the degree of uncertainty in many succeeding frames. The developed feature is called Entrocy based on the calculation methodology that depends on the computations of entropy-frequency combination. The main concept in the measurement of frequencies is to measure the degree of uncertainty in many succeeding frames. Entropy theory was introduced by Shannon to indicate the level of information via estimation and representing the probability density function (pdf) of every single sample in the sequence and thus reflecting the random distribution of data [11]. Entropy has demonstrated the ability to estimate the signal complexity and this was through employed it in a range of diverse research problems. The domain of Entropy application varies from speech handling, signal processing, healthiness applications, ecology, etc. The study of Reynolds et al. was calculating entropy for audio spectral to discriminate clean speech from non-clean speech and it was suggested as feature ASR [24]. Entropy also was adapted to STFT subband combined with the coefficients of some MFCCs for improving the automatic speaker recognition result in a noisy environment; the adopted feature is referred to called spectral entropy [15]. In [15] presents another application of Entropy is a maxent technique and refers to the maximum entropy model. The study was by Berger et al. and it first proposed to be a statistical module for the processing of natural languages [24]. The maxent technique has in turn since been employed in a broad range of fields. To sum up, we calculate the Fourier Transform (DCT) of the Entropy over several consecutive frames and use the coefficients to form the feature ‘Entrocy’ that is used to improve speech utterance. Thus, speaker recognition is done based on depends on the aforementioned enhancement speech.

2.1 Speaker recognition system

Identity toolbox for assessing speaker recognition has been developed by Microsoft Research (MSR) [25]. The developed toolbox applies Gaussian Mixture Model (GMM) and Universal Background Model (UBM) machine recognition and provides paradigms for i-vector analysis. The proceeding of speaker recognition systems is performed through two main stages named front end and back end. The functionality of the first stage is feature extraction from speech

signals of each enrolled speaker and transformation to acoustic features. The Cepstral features, such as the MFCCs are most commonly used with speaker recognition systems in consideration of the Mel-scale in MFCC is a scale that represents the base of converting the frequency and the perceived pitch to the features coefficients equivalent human auditory system, which is not linear system [15]. By contrast, in the second phase (back-end) the reference models for the enrolled speaker is generating following the extracted features from the front-end phase. It should be noted that both the Gaussian Mixture Model (GMM) and the Gaussian Mixture Model-Universal Background Model (GMM-UBM) are regarded as the basis for ASR systems. GMM parameters are acquired in the (GMM-UBM) framework utilizing the expectation-maximization (EM) algorithm. Speaker modules are acquired during enrolment using the adaptation Maximum a Posteriori (MAP) [24]. Thresholding of the log-likelihood is utilized to estimate scoring and take decisions. The UBM methodology is to collect speech samples from a huge group of speakers that are collective to train the universal background model as a speaker-independent module. Figure 1 illustrates the block diagram of the GMM-UBM framework.

2.2 Baseline ASV: I-vector

I-vector approach has been largely used in several speech classification tasks (speaker, language, dialect recognition, speaker diarization, speech recognition, clustering …).

i-vector is a compact representation that summarizes what happens in a given speech recording, furthermore, the classical i-vectore approach is based on Gaussian Mixture model (GMM) that means applying the subspace approaches to model neurons activation. I-Vector represents the superior vector of GMM with a total change subspace. The one-part space approach was driven by the detection that the JFA channel area covers data that can be utilized to differentiate amid speakers [6]. The GMM-compatible i-vector speaker can be represented by

$$\mu = m + \mathrm{T}w \tag{1}$$

where

m denotes the mean supervector take out from the Universal Background Model UBM, T refers to a rectangular low-rank matrix and w denote the random vector with normal distribution, i.e., the supposedly i-vectors [7]. Within the i-vector frame, the result procedure consists of calculating the matching amid the factors of the target and the test speaker. To realize this point, a few post-processing stages have been recommended, comprising linear discriminant analysis (LDA) to maximize the between-class variance and minimize the within-class variance, in addition to Gaussian probabilistic linear discriminant analysis (PLDA) [23]. Results are then prepared to depend on log-likelihood thresholding of PLDA hyperparameters. The total variability factors, w, are an independent normally distributed random vector. It is presumed that i-vectors (w) are usually dispersed with parameters N(0,1). Take out i-vector from the total-variability subspace is fundamentally a maximum aposteriori adaptation (MAP) of w in the subspace clear by T. The total-variability subspace is accountable for essential an appropriate subspace from which i-vectors is extracted. As the total variability space signified by T covers together speaker and channel variability, i-vector approaches require extra intersession recompense approaches before scoring to attenuate the effects of channel variability. Figure 2 illustrates the block diagram of the i-vector framework.

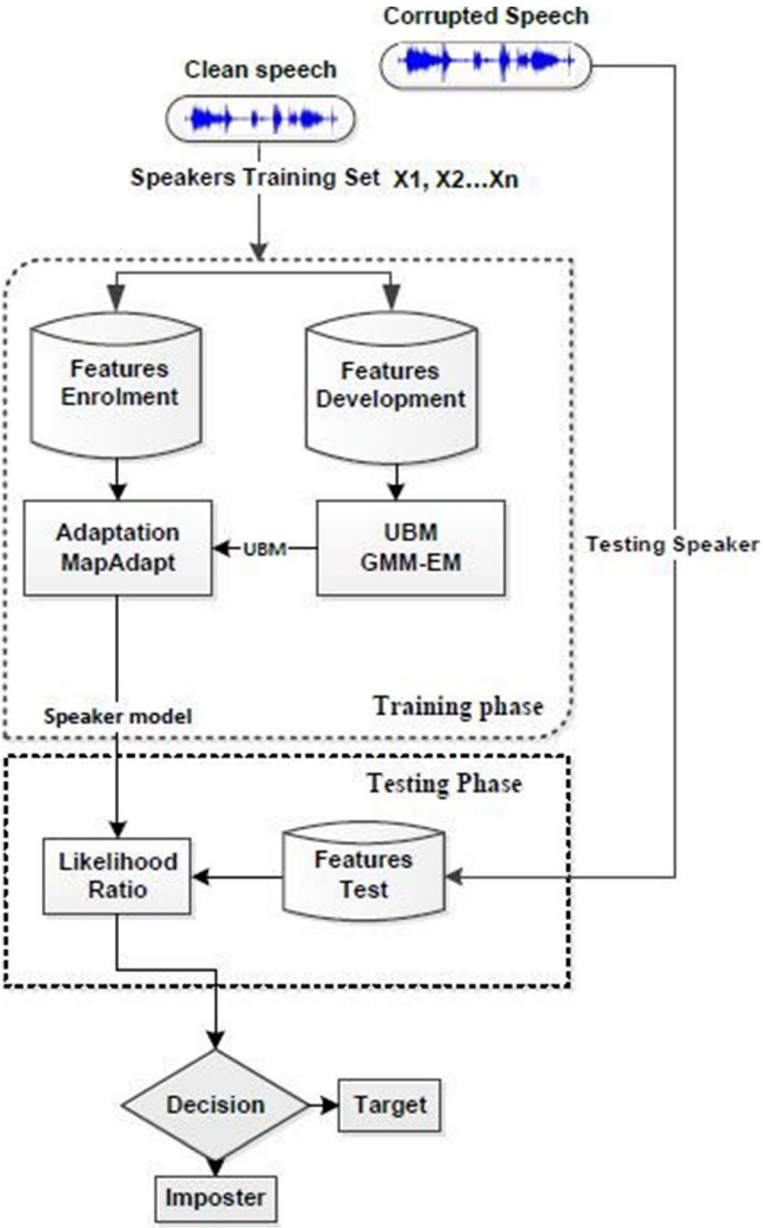

Fig. 1 Block diagram of the GMM-UBM framework

2.3 Short utterances challenge

The short duration speech issue is commonly identified inside the society of speaker recognition [22]. Despite a significant achievement, the current speaker recognition systems do not perform well unless training and retrieval samples are long enough. However, in various applications, the speakers are unwilling to deliver many utterances data, especially during the

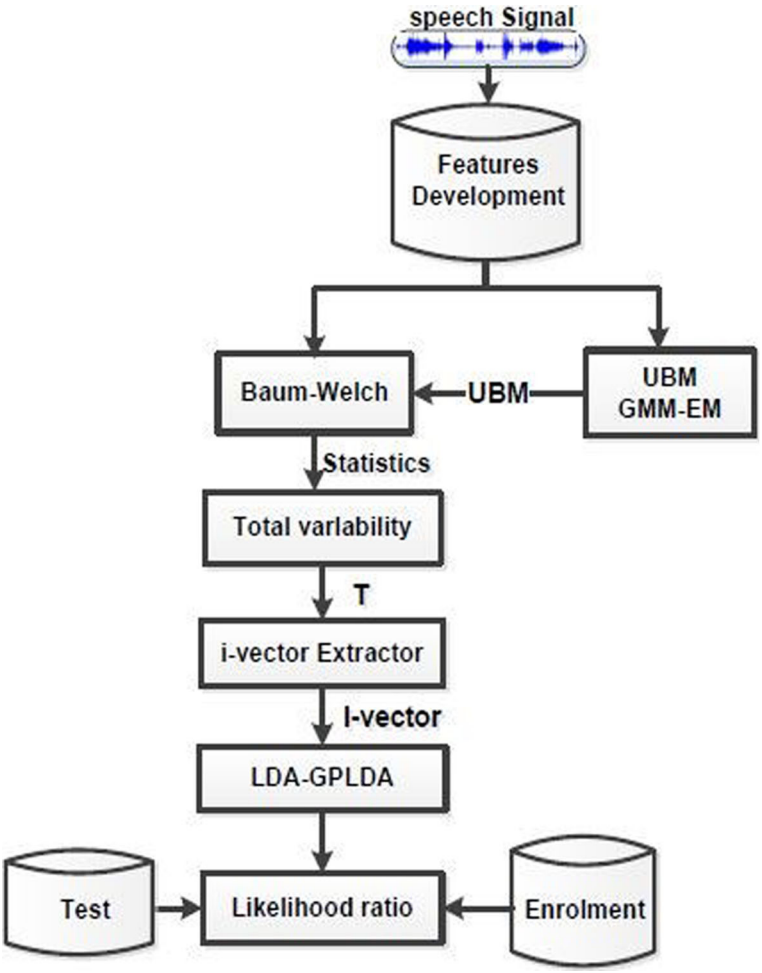

Fig. 2 Block diagram of the i-vector framework

testing phase, for example in services of telephone banking. However, in other application like forensic applications, it is not simple to collect adequate speech data. Clearly, degradation in the presentation of the text independent speaker recognition, in case the training and testing data are not long sufficient as has been specified in numerous past researches. As an example, Vogt stated that if the testing utterances were less than 20 s to 2 s, then the system performance will be highly degraded in terms of the percentage of EER from 6.34% to 23.89% [29]. Moreover, when the testing signal length is shorter than 2 s, the equal error rate EER raised to over 35.00% [16]. A notable improvement in the speaker recognition performance with short speech using the JFA framework which is models speaker and channel variability's in two separate subspaces [28]. This work is lengthy in [12] that shows the i-vector model can amplify speaker data more actively so it is more appropriate for speaker recognition system. Furthermore, score-based segment selection method has been described in [21] that assesses the dependability of every test speech signal depending on a set of cohort models and scores the testing speech with the dependable signals only. The decrease in the percentage of EER of

22% was described by the researchers on a recognition task when the speech signal used in the testing phase is less than 15 s in length. The results that described in these works are depending on the testing signals (5-10s) long. This length is still long in several situations. In case of too small speech segment, i.e., (1,2 s) in length, there are no satisfactory solutions yet. Furthermore, the recognition will be more challenging if the training speech signal is also short, for which very slight research has been lead.

3 Feature calculation

3.1 Mel frequency Cepstral coefficients

Recognition both in speech and speaker, MFCC has proven to be an effective feature extraction technique. That is because MFCC has the advantage and ability to capture the phonetically important features of recording speech. MFCC feature is designed to mimic the main physical temperament of the human hearing system. Its interprets that crucial attributes of the speech and all other information are de-emphasized [30]. MFCC thus reflects more important audio characteristics that time-domain features. AL-KARAWI study has been proved that MFCC profitable more efficient in the clean environment compared with the other recognition methods. [1]. However, the minor drawback is that MFCC performance in noisy environments can significantly deteriorate. That it is why this work has been suggested to combine MFCC with Entrocy feature. MFCC estimation is composed of five phases. The first step in the MFCC feature extraction process is pre-processing in which the signals are pre-processed before extracting features extraction stage. Then, the given speech is framed into small frames with a size that preserves the information periodicity. Windowing process is applied in the next step by multiplying each frame by Hamman window to reduce the frames discontinuities at the beginning and end of each frame. Next, the time domain frames are converted to the spectral frequency domain using a Discrete Fourier Transform (DFT). The output spectrum magnitude is subjected to a Log function and then to the inverse DFT to produces the Mel-Cepstrum coefficients. For each frame, a set of coefficients (vectors) is extracted and processed as a multidimensionality feature. The output-calculated matrix is referred to as Mel-Frequency Cepstrum Coefficients. Figure 3 demonstrates the calculation procedure.

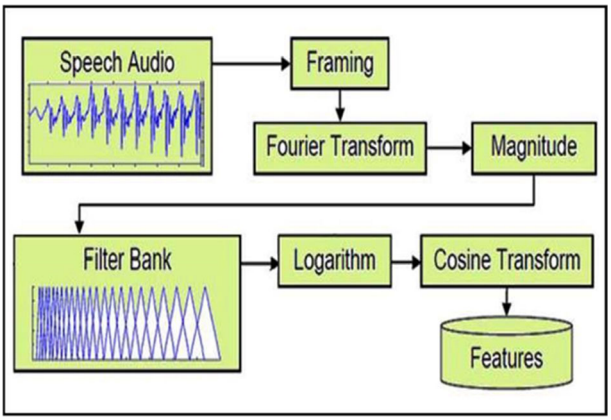

Fig. 3 MFCC Extraction Procedure

3.2 Entrocy calculation

The Enrocy feature was proposed for the first time and used to improve results in the overlapped music classification [29]. The computation scheme started by resampling each speech sample to 22.05 kHz as standard sample rate, 16-bit resolution. Each sample was divided into frames of 50 ms with a 25 ms overlap. It is worth to note that the overlap size represents a trade-off with increasing the frequency resolution. Then, the calculation of Entropy for each resulted frame. Entropy estimation could be summed up as follows: firstly, the probability calculation for every single sample. Stewart demonstrates the probability calculation as given in Eq. 2 [27].

$$Pr_{f(n)}(x_i) = Pr(s \in S | f(s) = x_i), i = 1, 2, 3, \ldots, \tag{2}$$

Where, S denotes the symbol domain of the i^{th} frame, the sum over the probability of all samples that related to the tested frame must be equal to 1.Let H be a vector of entropy features (1... NF) extracted from NF frames; then Entropy of each frame is calculated using Eq. 8 [26].

$$H_i = -\frac{1}{\log_2(L)} \left[\sum_{n=1}^{L} Pr(f_i(n)) \log_2(Pr(f_i(n))) \right] \tag{3}$$

The stages for calculating the Entrocy are as follows:

- The normalization is conducted on the calculated Entropy the logarithm of the frame size, which is denoted by L, thus the affection on the frame size is eliminated. Consequently, the entropy domain value bounded in the interval [0, 1], the maximum randomness represented by 1. The empirical outcomes indicate that most speech frames have lower randomness (entropy) than music frames.
- We then segmented the entropy vector H into small segments with 32 samples size, thus the frequency is calculated for each segment. To be clearer the behavior of variations across multiple consecutive frames is used in the recognition decision making (sound visualization). For example, babble noise, engines sound, vehicles moving, opening and shutting the doors of buses all these sounds together refers to be a bus station.
- Framing technique is done by windowing the calculated Entropy vector to split it into several segments. The moving window was one sample each time. If $H = \{x1, x2, \ldots, xn\}$. Then, The first and second segments will be as shown in Fig. 4:

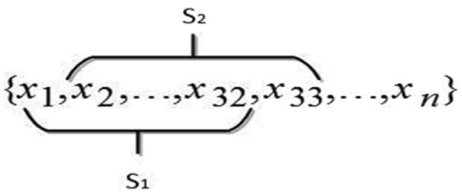

Fig. 4 Entropy Segmentation

- We multiplied each segment firstly by the Hanning window for spectral analysis purposes.
- Then we applied the DCT for each segment, the DCT method is used to calculate the variance of 32 adjacent entropy values of each set.

From the experimental results and depends on the calculation of the feature importance determined by the Random Forest RFs, we selected the two most important DCT coefficients and omitted the remaining coefficients [19]. The experimental results show that the 3rd and 5th coefficient were the most significant coefficients of the calculated Entrocy and this conclusion was confirmed by both of the RFs and PCA techniques see [17] for more details. Furthermore, to add a glance that reflects the spectral shape of the i^{th} entropy segment, the center of gravity or Spectral Centroid (SC), was also computed using Eq. 4 of the first coefficients part (16-DCT coefficients).

$$SC(i) = \frac{\sum_{k=1}^{N_{FT}/2} (kP(k))}{\sum_{k=1}^{N_{FT}/2} (P(k))} \tag{4}$$

Where *P(k)* represents the squared magnitude that captured for the audio spectrum while *k* refers to the frequency bin index of each frame. Finally, Entrocy feature is only expressed by three coefficients (3rd and 5th DCT coefficients, measured based on entropy plus the Spectral Centroid defined by the measured SC). The process for the entrocy feature calculation is illustrated in Fig. 5. The calculation of the proposed feature, which is simple and mathematically efficient, is carried out at any of the above calculation stages without any computationally expensive optimizations. The suggested and developed method could be applied and evaluated in different audio-information retrieval works such as music/speech discrimination, segmentation, retrieval or classification of music information due to its general an computationally efficient.

3.3 MFCC combined with Entrocy

To clarify, for each reconstructed frame the features are extracted on a short time scale. That is, we segmented the input signal into a sequence of successive analytical frames with a 50% overlap size and a feature value is calculated for each of those frames. The output feature dimension is denoted by an M × 25 matrix of feature coefficients; say *C*, with every single column representing a particular feature vector and every single row, represents the time sequences of a given coefficient. The first 23 coefficients correspond to the MFCC feature, while the last three are the Entrocy features.

$$C = \begin{bmatrix} MFCC(1,1) & \cdots & MFCC(23,1) & En(1,1) & \cdots & En(3,1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ MFCC(1,N) & \cdots & MFCC(23,N) & En(1,N) & \cdots & En(3,N) \end{bmatrix} \tag{5}$$

As illustrated the calculated matrix *C* is a concatenation of both MFCC and Entrocy coefficients, its size is *N* (*N* no of frames) × 25. It is worth noting that the aim here is to decide on the whole speech sample (classify the speaker into either imposter or target). Furthermore, the MFCCs are short term features, which are extracted from a small size frame window, whereas

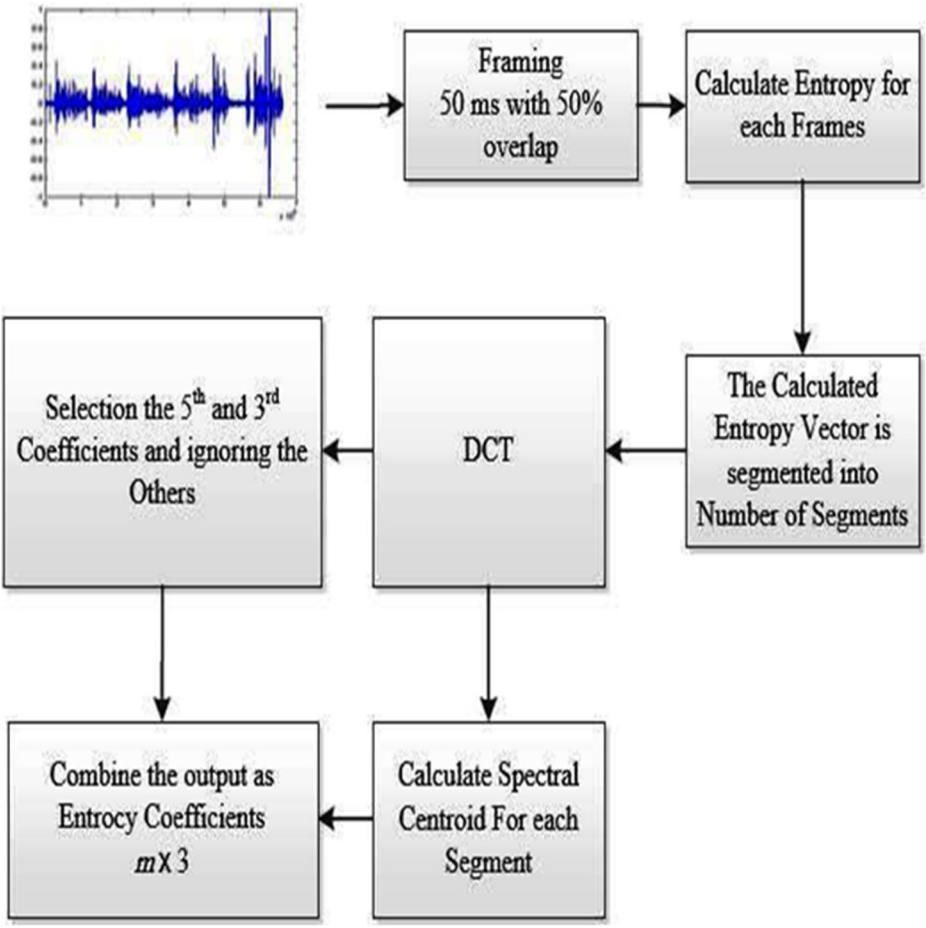

Fig. 5 Entrocy calculation procedure

the Entrocy is calculated over a longer timescale. Therefore, the calculated MFCC coefficient vectors will be longer than Entrocy vectors. Consequently, to combine the two features in one matrix with size $N \times 25$, where N represents the frame number, we have padded the end of entrocy vectors with zeroes to make it equivalent to the MFCC's coefficients length. This is shown in Fig. 6.

4 Experimental setup

4.1 Speech datasets

The speech content utterances used for the problem assessment were obtained from the SALU-AC database. This database was collected in the anechoic chamber of Salford University. The Salford anechoic chamber is characterized as one of the noiseless chambers with approximately −12.4 dB signal to the noise level. The data used in this experiment consists of 80 volunteer speakers with 40 male and 40 female speakers. The sampling-rate of the recorded

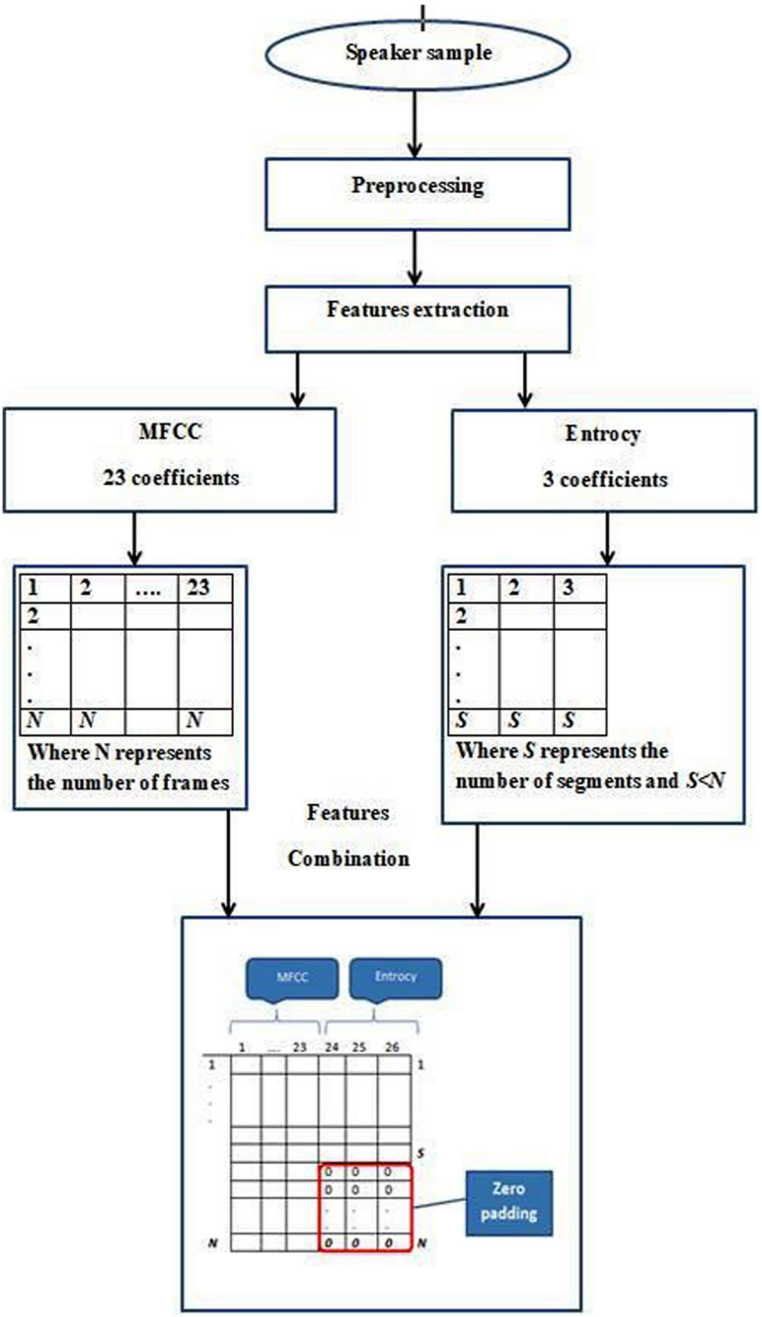

Fig. 6 Feature space calculation strategy

samples was 16 kHz and those samples with length 10 s for each recorded sample. Speaker Recognition, evaluation utterances are being truncated into shorts (2, 3, 4, 5 and 8, and 10s) for verification. The recording was truncated into 8-s excerpts for training purposes and 2

utterances divided into (2, 3, 4, 5, 8 and 10s) from different record to each speaker utilized for testing to create a test database. In this paper, the purpose is to test talker recognition versus the environment with short utterances. The voice activity detection was used to remove the salience portion. The sample rate used for the speech signal is 16 kHz.

4.2 Noisy data

The recorded samples explained in the previous section were mixed with different noise levels to generate noisy speech with different signal-to-noise (SNR) ratios ranging from 20 dB to 0 dB, for this study. In this experiment, babble is in the type of used noise. An audio mixer was developed using the MATLAB code, the blends pre-recorded speech segments with noise according to their signal intensity. The mixing strategy used was empirically verified and published to emulate the best mix of soundtracks [18]. The sound mixer procedure described as follows: firstly, the issue of normalization is addressed in such a way that speech and noise are added in the wanted proportion to avoid misinterpretation. The normalization is done by normalizing the mixed or compared signals to the same perceived level. In this point, the mixed samples are handled to have the same (RMS). Next, 160 samples from the mentioned overall samples obtained from 80 speakers are mixed with babble noise at 5-difference SNR ratios ranging from −5 dB to 20 dB in steps of five then the noisy speech and truncating into shorts (2, 3, 4, 5 and 8, and 10s) for verification.

4.3 Evaluation methods

For the system error evaluation, The test scores are determined as the log-likelihood ratio between the speaker models and Universal Background Model test observations There are two kinds of errors in the assumption of the statistical testing, these errors are the false match rate (FMR) and false non-match rate (FNMR). A false match rate (FMR) refers to a percentage of the falsely confirms an impostor speaker as the target through the impostor verification stage. However, a false non-match rate (FNMR) represents defining the target speaker as an impostor through the verification target trials. Moreover, the DET (Detection Error Trade-off) curve is a very useful way to assessing the accuracy of the system in a linear plot of bit error rates on a standard scale, referred by the NIST [5]. The critical area of the curve where the false match rate (FMR) and false non-match rate (FNMR) are equal is called the EER (Equal Error Rate). For speaker recognition and other biometric security systems, the EER (Equal Error Rate) is often used as a combined single measure for error. Generally, the lower EER %, the higher the reliability of the biometric system. In the evaluation stage, each test speech signal scored against the background model to accept/discard the claimed speaker.

5 Experiments results

To evaluate the universal approach to channel matching and short utterance is not an easy task, since the featuring selection, machine-learning algorithms, and training approaches can all have implications on system performance. The lack of a standardized benchmark regime makes a comparison to other's works difficult. In this work, the proposed method is validated, as on a text-independent speaker recognition using the built-in i-vector and two features. Figure 8 shows the boxplot acquired with the percentage of EER and the length of the utterance (s). The shorter utterances seem to have

Table 1 System performance with Entrocy features based on EER%

Utterance Length (training-testing)	Entrocy-EER%						
	clean Sig.	SNR(dB)					
		20%	15%	10%	5%	0%	−5%
full - 2s	**31.2**	**32.3**	**34.0**	**36.2**	**39.2**	**44.6**	**47.4**
full - 3s	**29.1**	**30.1**	**32.3**	**33.1**	**37.5**	**42.1**	**46.5**
full - 4s	**27.2**	**24.2**	**26.4**	**27.2**	**32.4**	**40.7**	**44.4**
full - 5s	**8.1**	**17.3**	**19.2**	**21.4**	**27.8**	**37.2**	**41.3**
full - 6s	**7.3**	**14.5**	**16.1**	**18.5**	**25.3**	**35.4**	**40.4**
full - 8s	**4.2**	**7.2**	**10.2**	**13.3**	**22.4**	**33.3**	**38.3**
full - 10s	**2.7**	**4.6**	**6.6**	**8.3**	**20.3**	**30.6**	**37.2**

bigger variability than the bigger utterances. Furthermore, it is showed that the bigger utterances yield the greatest performance as it shows a lesser maximum and median. This boxplot is depending on the standard deviation of the equal error rate. A Tables 1, 2 and 3 furthermore, shows the result of authentication any one speaker against the remaining 79 speakers, with different utterance length and signal to noise ratio. The first column in the table represents the length of the utterance while the row refers to the SNR level. These tables illustrate the impact of different utterances length and SNR on the performance of speaker verification systems using Entrocy, MFCC and the combination of both features based on the percentage of EER. The Tables clearly show that MFCC features provide good results for the various SNR and utterance length compared with the Entrocy feature. However, combined MFCC and Entrocy features help us to establish a greater degree of accuracy and robustness on this matter for a variety of SNR and sample length than both features, when they are used separately, especially for low SNRs such as 20 and 15 dB and long utterance. Consequently, there is a noteworthy enhancement in the authentication, presentation when the test speech segment was longer than 5 s and SNR is below 10 dB. An example for more explanation can be given, using sample length 10s, the combined features gave 3.3% with 10 dB and 17.3% ERR for 5 dB respectively. While the results of EER using the MFCC are 6.3% with 10 dB, and 18.3% EER for 5 dB. However, the results of the Entrocy feature in the same SNRs are 8.3% with 10 dB and 20.3% with 5 dB. Figures 7, 8, 9, 10, 11, 12 and 13 show the DET graphs for the system performance in different scenarios for the recognition phase with different sample length and 20, 10, 15 and 5 dB SNR. The combined features, performance has a noticeably better rate for both

Table 2 System performance with MFCC features based on EER%

Utterance Length (training-testing)	MFCC-EER %						
	clean Sig.	SNR(dB)					
		20%	15%	10%	5%	0%	−5%
full - 2s	24.2	26.3	27.0	29.2	37.2	43.6	45.4
full - 3s	**21.1**	**23.1**	**25.3**	**27.1**	**35.5**	**40.1**	**44.5**
full - 4s	**15.2**	**17.2**	**20.4**	**22.2**	**28.4**	**38.7**	**42.4**
full - 5s	**6.1**	**8.3**	**11.2**	**13.4**	**25.8**	**36.2**	**40.3**
full - 6s	**3.3**	**5.5**	**9.1**	**11.5**	**23.3**	**35.4**	**39.4**
full – 8s	**2.0**	**4.2**	**7.2**	**9.3**	**21.4**	**33.3**	**38.3**
full - 10s	**0.94**	**2.2**	**4.6**	**6.3**	**18.3**	**31.6**	**36.2**

Table 3 System performance with both features based on EER%

Utterance Length (training-testing)	Combined-EER%						
	clean Sig.	SNR(dB)					
		20%	15%	10%	5%	0%	-5%
full - 2s	**21.2**	**21.3**	**23.0**	**24.2**	**31.2**	**40.6**	**44.4**
full - 3s	**14.1**	**17.3**	**19.3**	**21.1**	**28.5**	**39.1**	**43.5**
full - 4s	**10.8**	**11.2**	**13.4**	**16.2**	**26.4**	**37.7**	**41.4**
full - 5s	**3.1**	**8.3**	**10.2**	**12.4**	**24.8**	**35.9**	**40.3**
full - 6s	**2.3**	**5.5**	**7.1**	**9.5**	**21.7**	**33.4**	**38.4**
full – 8s	**1.0**	**3.2**	**4.2**	**6.3**	**19.4**	**32.3**	**37.3**
full - 10s	**0.32**	**1.3**	**2.3**	**3.3**	**17.0**	**30.2**	**35.0**

FPR, FNR than the MFCC and entrocy feature. As DET graphs exhibit, there is a significant improvement in the recognition, trend when the combined features are used rather than using each one of the aforementioned features alone. The DET graphs in obvious illustration the false acceptance rate (FAR) and false rejected rate (FRR) for the suggested way Indeed, in a few cases, the false match rate is closed. Moreover, to validate the proposed methods in real conditions, and to demonstrate the models' generalization,. The performance of the diagnostics of the implemented model is evaluated by applying K-fold cross-validation to gain the strength of the SVM classifier 5-fold cross-validation used in the training data set to mitigate over fitting which is mean when a classifier does not generalize well from our training data to unseen data experiments results. From the result of experiment and from the literature, it has been shown that the MFCC feature is sensitive to background noise and reverberation conditions (especially with increasing SNR). Therefore, by combined matrix of MFCC and Entrocy, The elements of combined features represent the low frequency vectors, which are more noise-resistant than vectors of matrix of MFCC. We refer to this method extraction by the designation MFCC + MFCC_Entrocy. Therefore, the performance increased by using the combination. Combining them at the score level resulted in a good

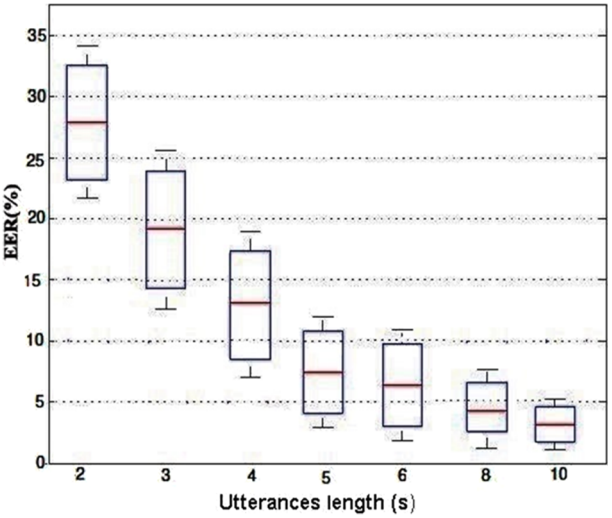

Fig. 7 Boxplots of using different speech sample length

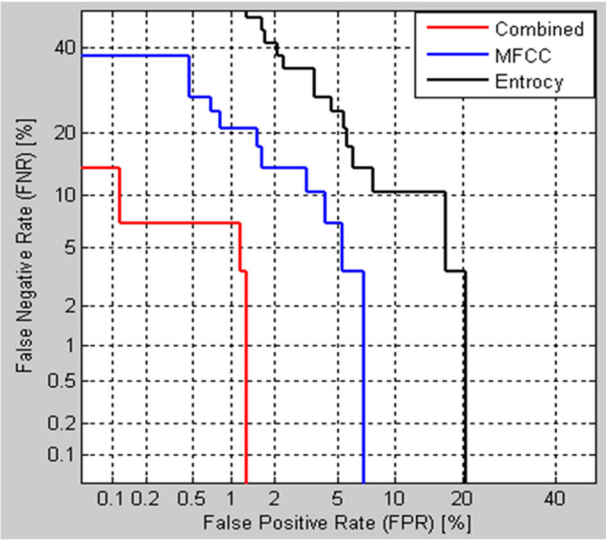

Fig. 8 DET graphs for features based in 20 dB SNR and length = 8

improvement of the speaker verification performance, that because the combining vector is more noise-resistant than vectors of matrix of MFCC with different utterances duration. Furthermore, the extracted feature using combing lead to reduce the distortion in features that are extracted.

6 Concluding remarks

Producing robust speaker verification remains a great challenge in the wide-ranging implementation of speaker verification systems, especially for applications using short speech utterances in noisy condition. The work we have successfully investigated the challenge of

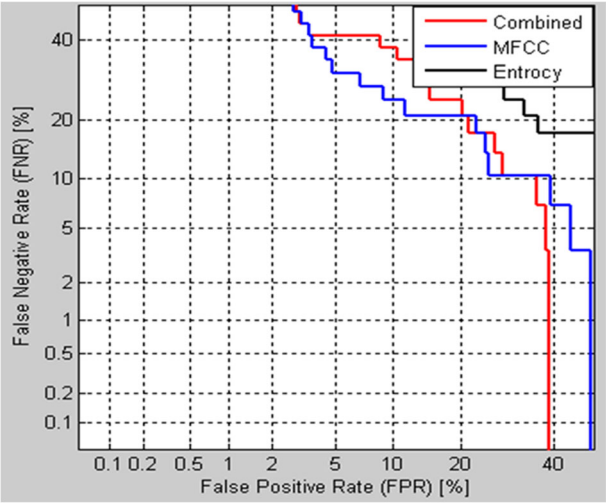

Fig. 9 DET graphs for features based in 20 dB SNR and length = 2

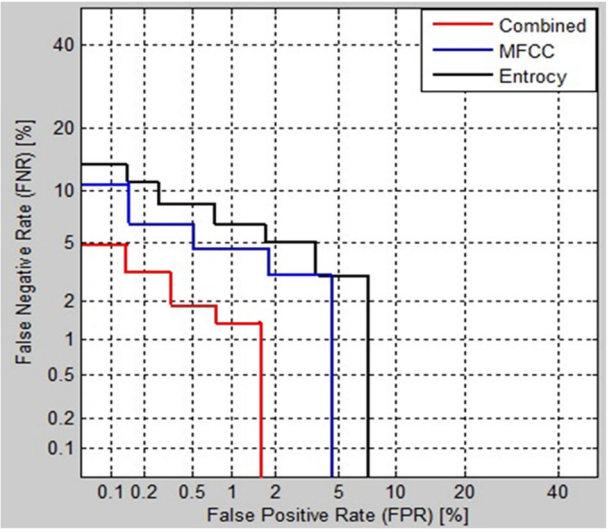

Fig. 10 DET graphs for features based in 20 dB SNR and length = 2 s

short utterances and obtainable research on the impacts of limited speech data in the first stage of this experiment. While the effect of short utterances corrupted with different SNR level of babble noise on system performance was conducted in the second stage. In this experimental study, a robust combined feature set has been implemented, evaluated and compared to the baseline. The high noise signals and limited testing data are challenging as the noise is distributed over all frequencies in the segments in different ratios and the limited feature can be extracted from the short utterance. Thus, a speaker can be reliably verified in a noisy condition using information-rich features that can identify the speaker based on the speech frequency spectrum. In other words, Speaker Recognition in this proposed work can be carried out on noisy and short speaker samples employing i-vector technique with Mel-frequency

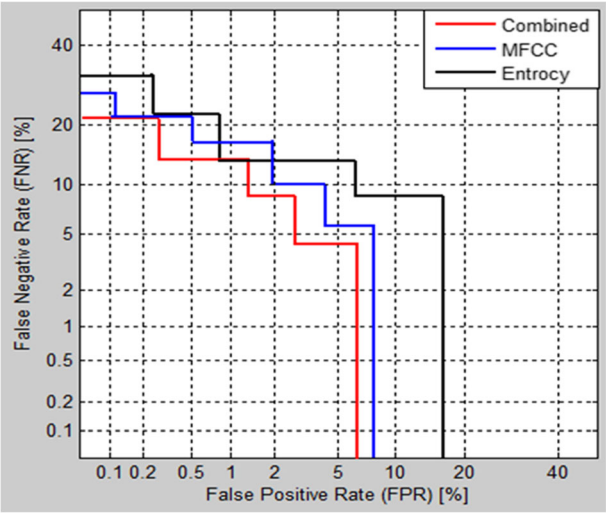

Fig. 11 DET graphs for features based in 10 dB SNR and length = 10s

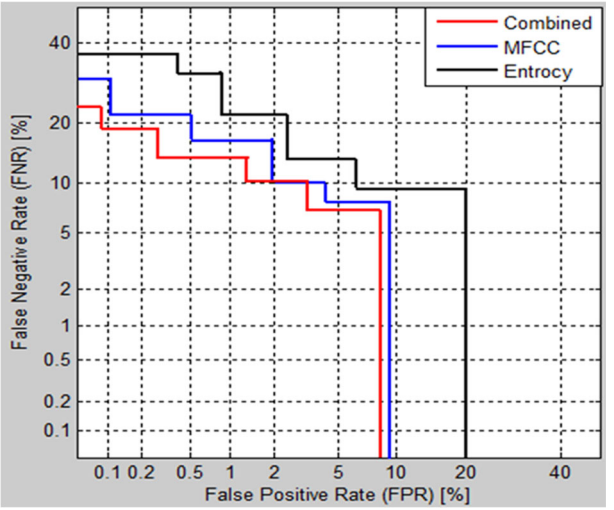

Fig. 12 DET graphs for features based in 15 dB SNR and length = 8 s

Cepstral coefficients and the Entrocy feature, which has been developed for overlapped speech/music/audio feature data. It has been shown in the literature that the MFCC feature is sensitive to background noise and reverberation conditions (especially with increasing SNR). Consequently, the illustrated results using the MFCC showed better performance than Entrocy under long utterance and low noise environment as demonstrated in Tables 1 and 2. However, It is observed that the speaker verification performance, reduces as the noise level increases and the sample length decrease. While the experiment for different SNR level results shows that using Entrocy combined with the MFCC feature is more robust than using the MFCC feature alone as shown in Table 3. Therefore, combining them at the score level resulted in a good improvement of the speaker verification performance, that because the

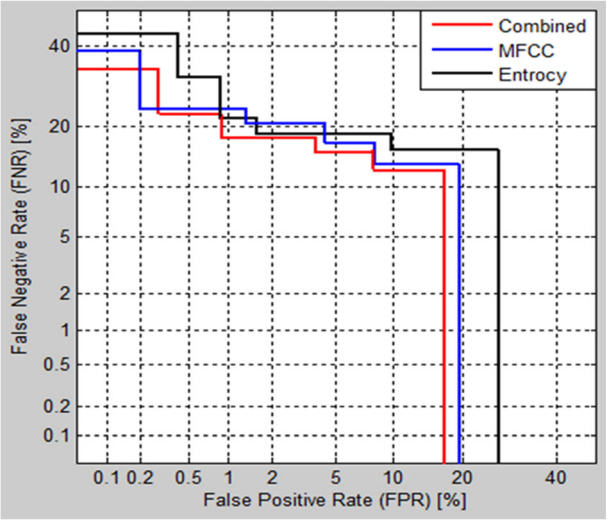

Fig. 13. graph for features based in 5 dB SNR and length = 10s

combining vector is more noise-resistant than vectors of matrix of MFCC with different utterances duration. Furthermore, the extracted feature using combing lead to reduce the distortion in features that are extracted. The investigation demonstrated the dependences of the score and the evaluation system performance (EER) terms of the length and level of the proposed utterance of noise. The enhancement in the authentication, presentation is clear when the test speech segment was longer than 5 s and SNR is below 10 dB.

References

1. Al-Karawi k A (2019) Robustness speaker recognition based on feature space in clean and Noisy condition. Int J Sens Wireless Commun Control 9:1–10
2. Al-Karawi KA (2020) Mitigate the reverberation effect on the speaker verification performance using different methods. Int J Speech Technol, pp. 1–11
3. Al-Karawi KA, Li F (2017) Robust speaker verification in reverberant conditions using estimated acoustic parameters—A maximum likelihood estimation and training on the fly approach, in 2017 Seventh International Conference on Innovative Computing Technology (INTECH), pp 52–57
4. Al-Karawi KA, Al-Noori AH, Li FF, Ritchings T (2015) Automatic speaker recognition system in adverse conditions–implication of noise and reverberation on system performance. Int J Inform and Electron Eng 5: 423–427
5. Chen Y-W, Lin C-J (2006) Combining SVMs with various feature selection strategies, in feature extraction. Springer, pp 315–324
6. 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 Tenth Annual conference of the international speech communication association
7. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification Audio, Speech, and Language Processing. IEEE Trans 19:788–798
8. Fatima N, Zheng TF (2012) Short utterance speaker recognition a research agenda. International Conference on Systems And Informatics (ICSAI2012) IEEE
9. Furui S (1981) Cepstral analysis technique for automatic speaker verification, Acoustics, Speech and Signal Processing. IEEE Trans on 29:254–272
10. Hermansky H, Morgan N (Oct 1994) RASTA processing of speech. IEEE Transactions on Speech and Audio Processing 2(4)
11. Junqua J-C, Reaves B, Mak B (1991) A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer. In: Second European Conference on Speech Communication and Technology
12. Kanagasundaram A, Vogt R, Dean DB, Sridharan S, Mason MW (2011) I-vector based speaker recognition on short utterances, in Proceedings of the 12th Annual Conference of the International Speech Communication Association, pp 2341–2344
13. Kinnunen T, Li H (January 2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Communication Journal 52(1):12–40
14. Li L, Wang D, Zhang C, Zheng TF (June 2016)Improving short utterance speaker recognition by modeling speech unit classes. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(6)
15. Logan B (2000) Mel frequency cepstral coefficients for music modeling in Ismir, pp 1–11.
16. Mak M-W, Hsiao R, Mak B (2006) A comparison of various adaptation methods for speaker verification with limited enrollment data. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings
17. Mohammed DY 2017 Overlapped speech and music segmentation using singular spectrum analysis and random forests," Salford University
18. Mohammed DY, Duncan PJ, Al-Maathidi MM, Li FF (2015) A system for semantic information extraction from mixed soundtracks deploying MARSYAS framework. 2015 IEEE 13th International Conference on Industrial Informatics (INDIN)
19. Mohammed K Al-Karawi A, Duncan P, Li FF (2019) Overlapped Music segmentation using a new Effective Feature and Random Forests," International Journal Of artificial intelligence (IN-IA), vol 8
20. Duraid Y, Al-Karawi KA, Husien IM, Ghulam MA (2020) Mitigate the reverberant effects on speaker recognition via multi-training. In: Applied computing to support industry: innovation and technology.

International Conference on Applied Computing to Support Industry: Innovation and Technology ACRIT 2019, Cham, pp 95–109
21. Nosratighods M, Ambikairajah E, Epps J, Carey MJ (2010) A segment selection technique for speaker verification. Speech Comm 52:753–761
22. Poddar A, Sahidullah M, Saha G (2017) Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics 7:91–101
23. Prince SJ, Elder JH (2007) Probabilistic linear discriminant analysis for inferences about identity. In: 2007 IEEE 11th International Conference on Computer Vision
24. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digital Signal process 10:19–41
25. Sadjadi SO, Slaney M, Heck L (2013) MSR identity toolbox v1. 0: a MATLAB toolbox for speaker-recognition research. Speech and Language Processing Technical Committee, Newsletter
26. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
27. Stewart WJ (2009) Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling. Princeton University Press
28. Vogt R, Sridharan S, Mason M (2010) Making confident speaker verification decisions with minimal speech. IEEE Trans Audio Speech Lang Process 18(6)
29. Vogt R, Sridharan S, Mason M (2009) Making confident speaker verification decisions with minimal speech. IEEE Trans Audio Speech Lang Process 18(6):1182–1192
30. Zhao X, Wang D (2013) Analyzing noise robustness of MFCC and GFCC features in speaker identification, " in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 7204–7208
31. Zhao XY, Wang D (2014) Robust speaker identification in Noisy and reverberant conditions. IEEE/ACM Trans Audio Speech Lang Process 22(4):836–845