Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 4, August 2021, pp. 2310~2319
ISSN: 2302-9285, DOI: 10.11591/eei.v10i4.2957

Robust speaker verification by combining MFCC and entrocy in noisy conditions

Duraid Y. Mohammed[1], Khamis Al-Karawi[2], Ahmed Aljuboori[3]
[1]College of Engineering, Al-Iraqia University, Baghdad, Iraq
[2]University of Diyala, Diyala, Baqubah, Iraq
[3]College of Education for Pure Science Ibn-Al-Haitham, University of Baghdad, Baghdad, Iraq

Article Info

Article history:

Received Nov 30, 2020
Revised Feb 24, 2021
Accepted May 30, 2021

Keywords:

MFCC
Noisy environment
Speaker recognition
Speaker verification

ABSTRACT

Automatic speaker recognition may achieve remarkable performance in matched training and test conditions. Conversely, results drop significantly in incompatible noisy conditions. Furthermore, feature extraction significantly affects performance. Mel-frequency cepstral coefficients MFCCs are most commonly used in this field of study. The literature has reported that the conditions for training and testing are highly correlated. Taken together, these facts support strong recommendations for using MFCC features in similar environmental conditions (train/test) for speaker recognition. However, with noise and reverberation present, MFCC performance is not reliable. To address this, we propose a new feature 'entrocy' for accurate and robust speaker recognition, which we mainly employ to support MFCC coefficients in noisy environments. Entrocy is the fourier transform of the entropy, a measure of the fluctuation of the information in sound segments over time. Entrocy features are combined with MFCCs to generate a composite feature set which is tested using the gaussian mixture model (GMM) speaker recognition method. The proposed method shows improved recognition accuracy over a range of signal-to-noise ratios.

Corresponding Author:

Duraid Y. Mohammed
Department of Computer Engineering
Al-iraqia University, Baghdad, Iraq
Email: duraidyehya19@gmail.com

1. INTRODUCTION

Samples of real-world speech recording when overlapping with acoustic conditions such as additive noise, room reverberation constitute a major challenge to automatic speaker recognition (ASR) robustness. For robust speaker recognition, the verification of speech segments is important, but it becomes relatively difficult in noisy environments [1]. Many studies have been reported the aforementioned challenges. Some of which in feature domain [2], cepstral mean subtraction approach [3], relative spectral processing method [4], feature mapping [5], and combine MFCC's feature with gammatone frequency cepstral coefficient (GFCC) [6] used to reduce additive and convolutional distortions of the channel. Al-for training and building speaker recognition models, Karawi *et al.* used noisy samples, thus reducing the discrepancy between the developed model as reference and testing samples [7]. Each speaker sample has been convoluted 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 [8], [9]. Zhao *et al.* has used spectral energy calculated through short segments has been as dominant features to discriminate the speech samples from other soundtracks [10]. However, the robustness and reliability of these features in a noisy environment are affected negatively especially in the

case of overlapping with sound artifacts and non-stationary noise such as heavy breathing, and and mouth clicks. Furthermore, the speech samples quality represents one of the main considerations that affecting performance [11] which makes the use of the voice activity detection (VAD) technique in the sample preprocessing step crucial for removing the silence frames [12]. This usually depends on the number of parameters such as signal energy, mel-frequency cepstrum coefficients (MFCCs), long-term spectral divergence, or entropy [12], [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 [14]. Notwithstanding, the MFCC fails to achieve adequate accuracy in the case of reverberation or noise is presence [10]. 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 feature and extraction methods [10]. Accordingly, a new technique that employing noise adaptive threshold has been suggested [15], but the presence of sound artifacts and relatively high noise levels makes the performance drops significantly. To overcome the mismatched and noisy conditions, it is therefore suggested that the entrocy-based algorithm be combined with the MFFC feature in this work. We proposed and developed an entropy-based method as a hybrid feature to address the challenge of overlapping audio classes in [15] and demonstrated significant improvements in the detection of music segments. Therefore, in this study, the developed technique based on the time-frequency entropy domain, here referred to as the spectral entrocy, is combined with the coefficients of MFCC. The spectrum probability density function (pdf) is computed first for each single frame of the input speech sample, according to the spectral entry. The results show the efficacy of the suggested method in discriminating in a continuously recorded utterance, especially in unclean speech background, the segments of speech from the non-speech. Remainder of this study is structured according to; 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 DEFINITION

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 discriminating of the music from speech and it was shown significant detection results regardless the speech was pure or non-pure D. Y. Mohammed [16]. 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 [17]. Entropy has demonstrated the ability to estimate the signal complexity and this was through employed in a range of diverse research problems. The domain of entropy application varies from speech handling, signal processing, healthiness applications, and ecology. 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 [18]. Misra reveals that pure speech samples have a lower entropy spectrum than non-pure speech, because abrupt changes average higher in a noisy environment, thus clean speech is proven to have lower spectral entropy levels than speech with noisy background [18]. Entropy also was adapted to the 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 called [19]. A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra [20] 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 was first proposed to be a statistical module for the processing of natural languages. 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 based on MFCC-GMM

The identity toolbox for assessing speaker recognition has been developed by microsoft research (MSR) [21]. 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 [22]. By contrast, in the second phase (back-end) the reference models for the enrolled speaker are 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) [23]. 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. Since our process was constructed following the use of retraining models therefore in our experiments, we do not use UBM models and rely primarily on the use of the baseline system MFCC-GMM. Figure 1 illustrates MSR toolbox framework.

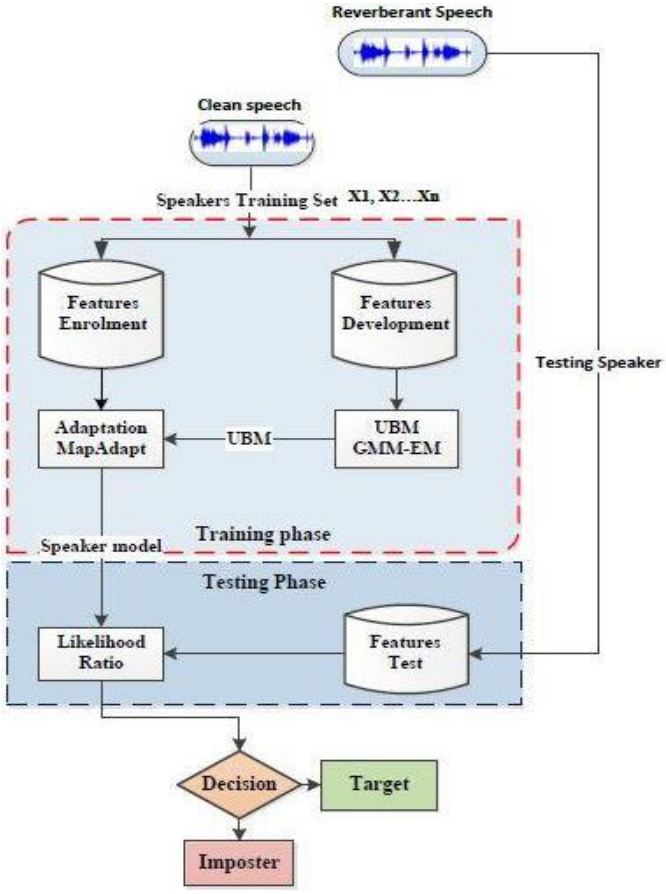

Figure. 1 MSR toolbox framework

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. It interprets that crucial attributes of the speech and all other information are de-emphasized [10]. MFCC, thus reflects more important audio characteristics than time-domain features. Al-Karawi study has been proved that MFCC profitable more efficiently in a clean environment compared with the other

recognition methods [6]. However, the minor drawback is that MFCC performance in noisy environments can significantly deteriorate. That 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. The windowing process is applied in the next step by multiplying each frame by Hamman window to reduce the frame's discontinuities at the beginning and end of each frame. Next, the time domain frames are converted to the spectral frequency domain using 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 referring to as mel-frequency cepstrum coefficients. Figure 2 demonstrates the calculation procedure.

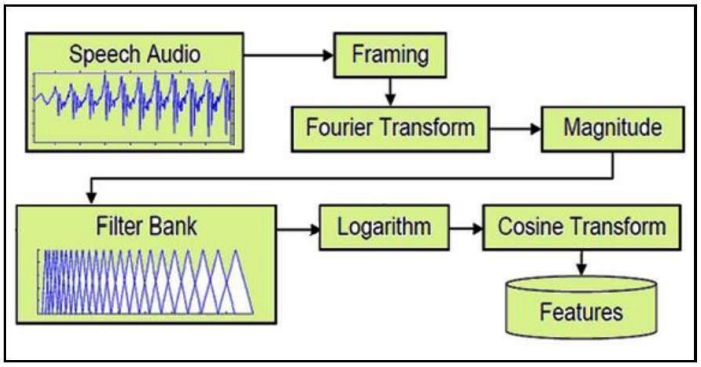

Figure 2. MFCC extraction procedure

3.2. Entrocy calculation

The Enrocy feature was proposed for the first time and used to improve results in the overlapped music classification [16]. 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 noting 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 is being as: firstly, the probability calculation for every single sample. Stewart [24] demonstrates the probability calculation as given in (1).

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

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 (2) (Shannon, 1948).

$$\mathbf{H}_i = -\frac{1}{\log_2(L)}\left[\sum_{n=1}^{L} Pr(f_i(n)) \log_2\left(Pr(f_i(n))\right)\right] \quad (2)$$

The stages for calculating the entrocy is being as:

- 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.
- The framing technique is done by windowing the calculated Entropy vector to split it into the number of

segments. The moving window was one sample each time. if $H = \{x1, x2, \dots, xn\} H = \{x1, x2, \dots, x_n\}$. Then, the first and second segments will be as shown in Figure 3.

– 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.

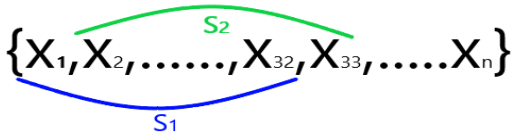

Figure 3. Entropy segmentation

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 [16]. The experimental results show that the 3rd and 5th coefficients were the most significant coefficients of the calculated entrocy and this conclusion was confirmed by both of the RFs and PCA techniques see [16] 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 (3) of the first coefficients part (16-DCT coefficients).

$$SC(i) = \frac{\sum_{k=1}^{N_{FT}/2} (kP(k))}{\sum_{k=1}^{N_{FT}/2} (P(k))} \quad (3)$$

P(k) represents the squared magnitude that is 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 Figure 4. 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 and computationally efficient.

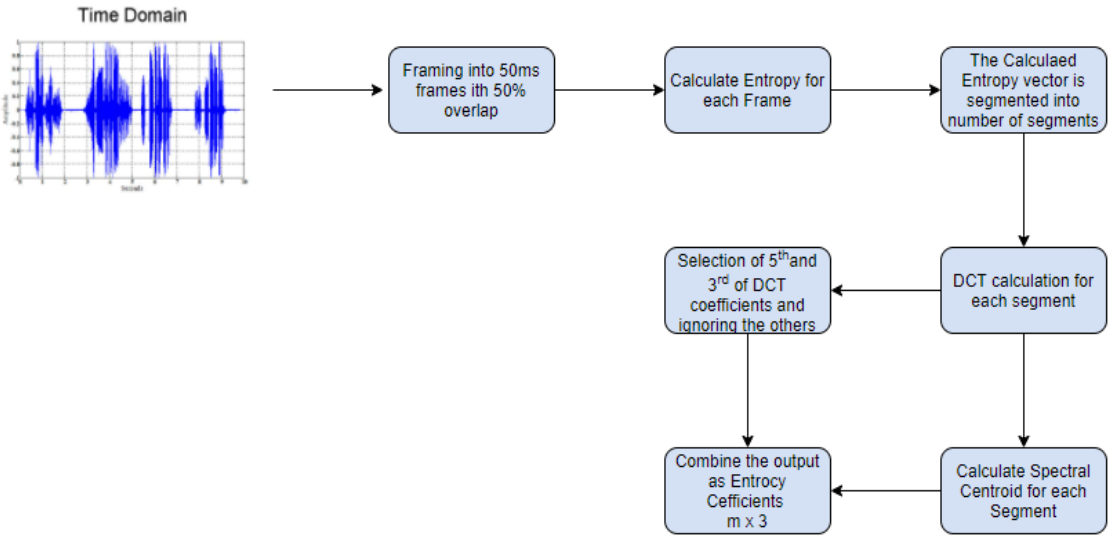

Figure 4. Entrocy calculation procedure

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 percent 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.

As illustrated the calculated matrix C is a concatenation of both MFCC and entrocy coefficients, its size is 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 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 the size N×25, where N represents the frame's 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 Figure 5.

$$C = \begin{bmatrix} MFCC(1,1) & \cdots & MFCC(23,1) & En(1,1) & \cdots & En(3,1) \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ MFCC(1,N) & \cdots & MFCC(23,N) & En(1,N) & \cdots & En(3,N) \end{bmatrix} \quad (4)$$

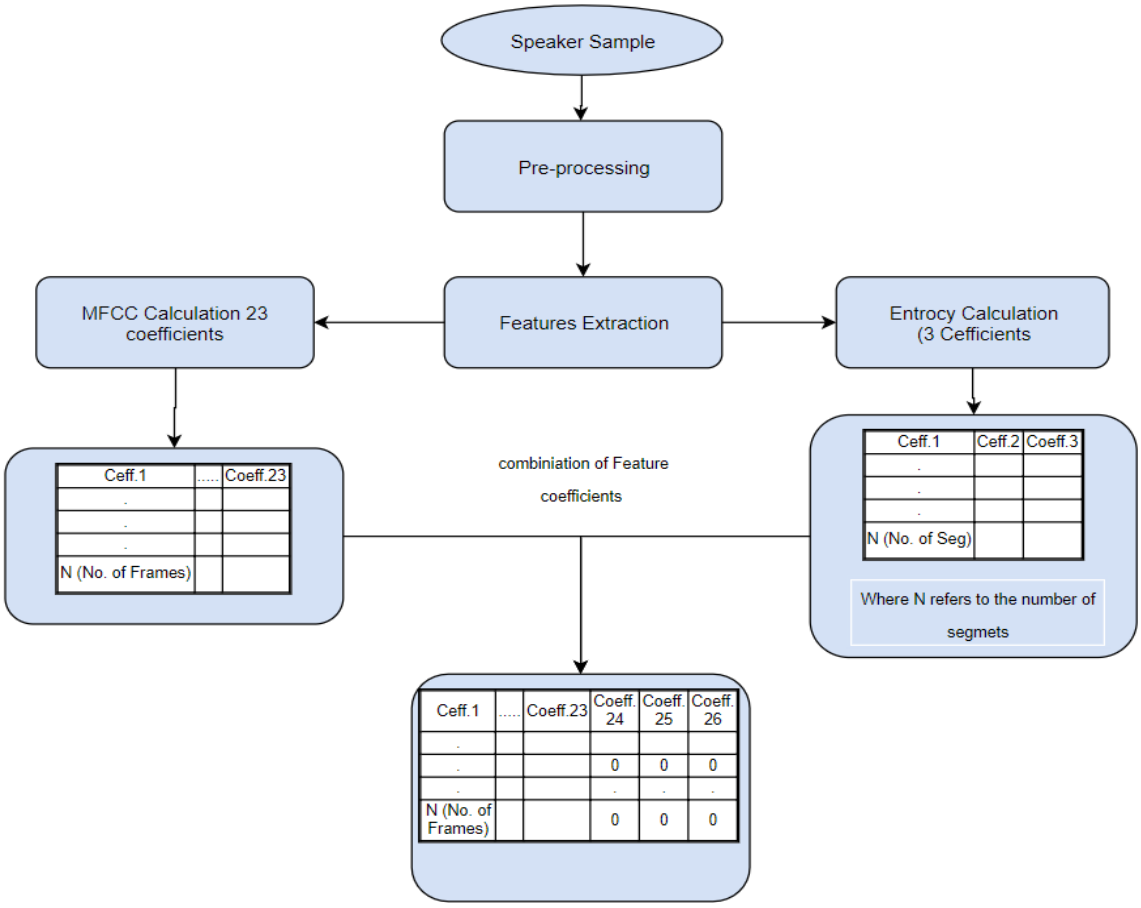

Figure 5. Feature space calculation strategy

4. EXPERIMENTAL SETUP

4.1. Speech datasets

The speech content utterances used for the problem assessment were collected in the anechoic chamber of salford university (SALU-AC). Such a tailored sample set is specifically beneficial in this work. The data collected consists of 100 volunteer speakers with 50 male and 50 female speakers. The sampling-

rate of the recorded samples was 16 kHz and those samples with length 5seconds for each recorded sample. The Salford anechoic chamber is characterized as one of the noiseless chambers with approximately -12.4 dBA signal to the noise level. The gathered utterances are therefore clean and by any background noise. Each volunteering speaker was recorded. Volunteering speakers recorded 10 speech samples in the English language. It is worth noting that each sample was recorded to provide text-independent samples, without any constraint for the volunteers with specific sentences.

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 the type of used noise. An audio mixer was developed using the MATLAB code, which 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 [25], [26]. The sound mixer procedure described is being as: 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. At this point, the mixed samples are handled to have the same (RMS). Next, 400 samples from the mentioned overall samples obtained from 40 speakers (10 utterances for each speaker (20 male and 20 female)) are mixed with babble noise at 5-difference SNR ratios ranging from 0 dB to 20 dB in steps of five then the noisy speech samples were used to validate the suggested speaker recognition method.

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 [27]. 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 detection error trade-off (DET) 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 to by the NIST [28]. The critical area of the curve where the false match rate (FMR) and false non-match rate (FNMR) are equal is called the equal error rate (EER). For speaker recognition and other biometric security systems, the EER is often used as a combined single measure for error. Generally, the lower the percentage of EER is 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

Table 1 and Figure 6 illustrate the impact of different SNR on the performance of speaker verification systems using entrocy, MFCC besides the combination of both features based on the percentage of EER. The x-axis of this Figure refers to the SNR in dB (clean, 20, 15, 10, 5 and 0dB) and the y-axis denotes the percentage of the error equal rate. Plotted in Figure 6 clearly show that MFCC features provide good results for the various SNR 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 than both features, when they are used separately, especially for low SNRs such as 20 and 15 dB. For example, in speech samples, the combined features gave 4.6% with 10 dB and 7.9% ERR for 5dB respectively. While the results are 5.5% with 10 dB, and 8.4% EER for 5 dB using the MFCC, and 7.6% with 10 dB, and 9.3% with 5 dB for the entrocy feature in the same SNRs. Figures 7, 8, 9 and 10 show the DET graphs for the system performance in different scenarios for the recognition phase in clean speech and 20,10 and 5 dB SNR. The combined features performance has a noticeably better rate for both FPR, FNR than the MFCC 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.

Table 1. System performance with both features based on EER%

Features	Clean Speech	Noisy speech (SNR)				
		20	15	10	5	0
MFCC	0	1.8	2.5	5.5	8.4	15.8
Entrocy	0	2.1	3.8	7.6	9.3	16.3
Combined	0	0.5	1.2	4.6	7.9	15.4

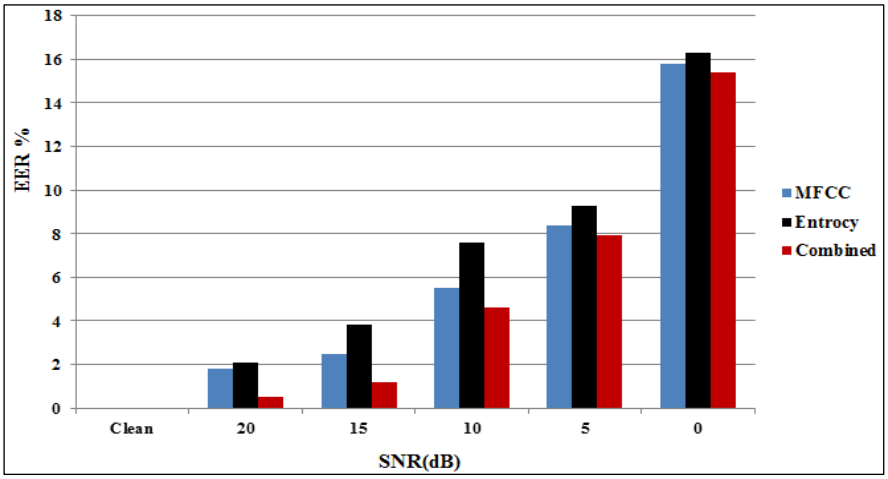

Figure 6. System performance based on different SNR

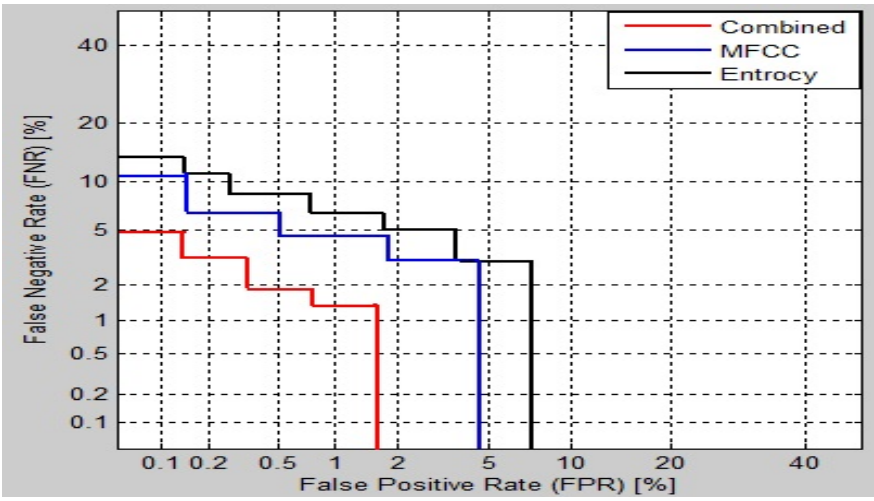

Figure 7. DET graphs for features based on 15 dB SNR

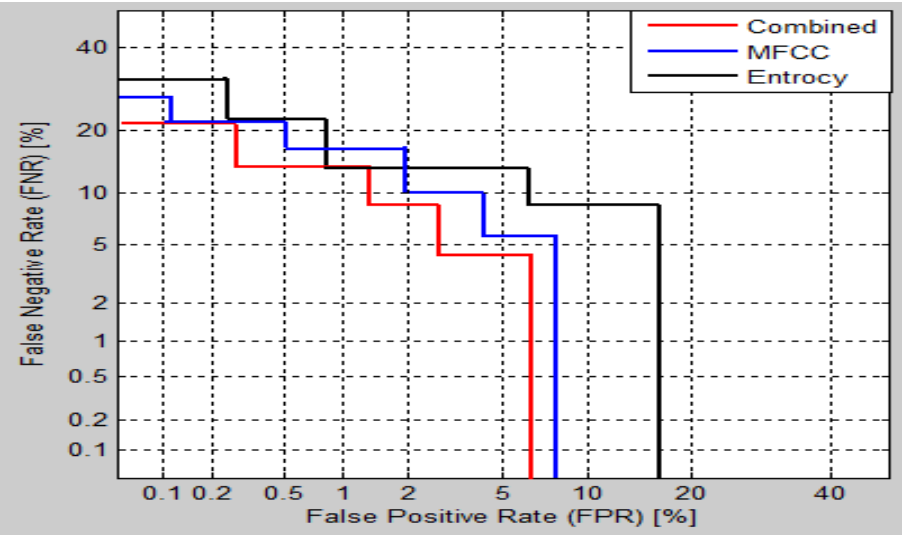

Figure 8. DET graphs for features based on 10 dB SNR

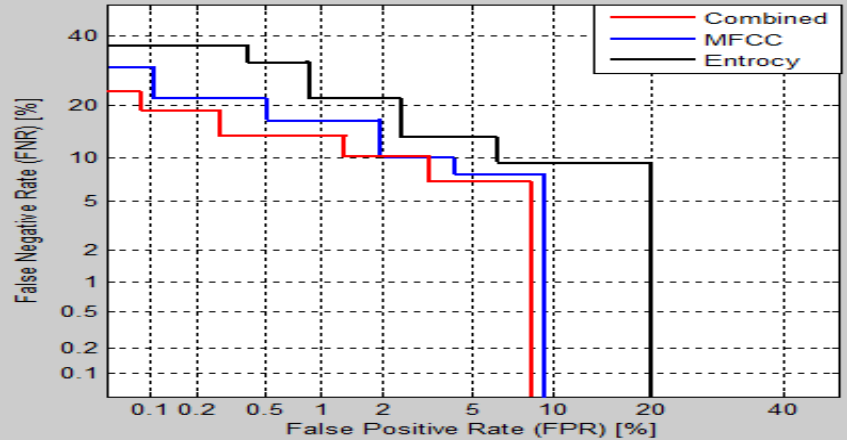

Figure 9. DET graphs for features based on 5 dB SNR

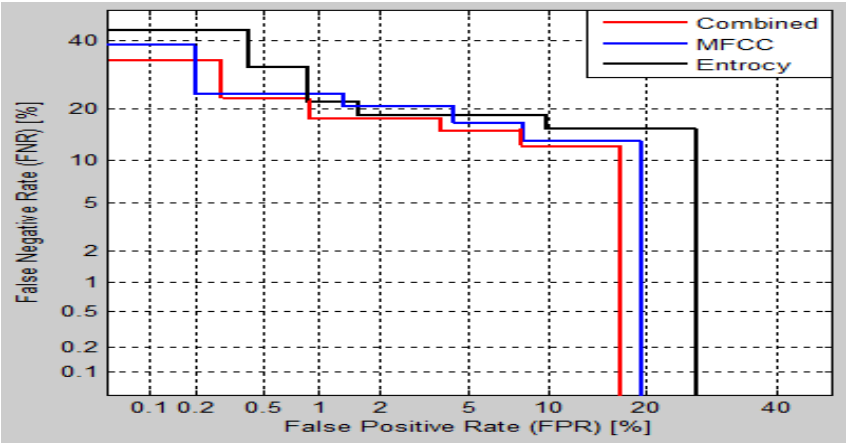

Figure 10. DET graphs for features based in 0 dB SNR

6. CONCLUSION

In this experimental study, a robust combined feature set has been implemented, evaluated, and compared to the baseline. The high noise signals are challenging as the noise is distributed over all frequencies in the segments in different ratios. 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 speaker samples employing the GMM technique with Mel-frequency 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 a clean and low noise environment as demonstrated in Figure 8 and Figure 9. However, it is observed that the speaker verification performance reduces as the noise level increases. 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.

REFERENCES

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254-272, April 1981, doi: 10.1109/TASSP.1981.1163530.

[2] R. Kaluri and P. R. CH, "Optimized feature extraction for precise sign gesture recognition using self-improved genetic algorithm," *International Journal of Engineering and Technology Innovation*, vol. 8, no. 1, pp. 25-37, 2018.

[3] Al-Karawi, K.A., *et al.*, Automatic Speaker Recognition System in Adverse Conditions--Implication of Noise and Reverberation on System Performance. International Journal of Information and Electronics Engineering, 2015. 5(6): p. 423.
[4] H. Hermansky and N. Morgan, "RASTA processing of speech," in *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, Oct. 1994, doi: 10.1109/89.326616.
[5] D. A. Reynolds, "Channel robust speaker verification via feature mapping," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, Hong Kong, China, 2003, pp. II-53, doi: 10.1109/ICASSP.2003.1202292.
[6] K. A. Al-Karawi, "Robustness Speaker Recognition Based on Feature Space in Clean and Noisy Condition," *International Journal of Sensors, Wireless Communications and Control*, vol. 9, no. 4, pp. 1-10, 2019, DOI: https://doi.org/10.2174/2210327909666181219143918.
[7] K. A. Al-Karawi and F. Li, "Robust speaker verification in reverberant conditions using estimated acoustic parameters -A maximum likelihood estimation and training on the fly approach," *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, Luton, 2017, pp. 52-57, doi: 10.1109/INTECH.2017.8102427.
[8] A. H. Al-Noori, K. A. Al-Karawi and F. F. Li, "Improving Robustness of Speaker Recognition in Noisy and Reverberant Conditions via Training," *2015 European Intelligence and Security Informatics Conference*, Manchester, UK, 2015, pp. 180-180, doi: 10.1109/EISIC.2015.20.
[9] D. Y. Mohammed, K. A. Al-Karawi, I. M. Husien, and M. A. Ghulam, "Mitigate the Reverberant Effects on Speaker Recognition via Multi-training," *Cham*, 2020, pp. 95-109, DOI: https://doi.org/10.1007/978-3-030-38752-5_8.
[10] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 7204-7208, doi: 10.1109/ICASSP.2013.6639061.
[11] X. Zhao, Y. Wang, and D. Wang, "Robust Speaker Identification in Noisy and Reverberant Conditions," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836-845, April 2014, doi: 10.1109/TASLP.2014.2308398.
[12] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12-40, 2010, https://doi.org/10.1016/j.specom.2009.08.009.
[13] M. Asgari, A. Sayadian, M. Farhadloo, and E. a. Mehrizi, "Voice Activity Detection Using Entropy in Spectrum Domain," *2008 Australasian Telecommunication Networks and Applications Conference*, Adelaide, SA, Australia, 2008, pp. 407-410, doi: 10.1109/ATNAC.2008.4783359.
[14] Al-Karawi, K.A., "Mitigate the reverberation effect on the speaker verification performance using different methods," *International Journal of Speech Technology*, p. 1-11, 2020.
[15] J.-C. Junqua, B. Reaves, and B. Mak, "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*, 1991.
[16] D. Y. Mohammed, K. A. Al-Karawi, P. Duncan, and F. F. Li, "Overlapped music segmentation using a new effective feature and random forests," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 2, pp. 181-189, 2019, DOI: 10.11591/ijai.v8.i2.pp181-189.
[17] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, pp. 3-55, 1948.
[18] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, 2004, pp. I-193, doi: 10.1109/ICASSP.2004.1325955.
[19] A. M. Toh, R. Togneri, and S. Nordholm, "Spectral entropy as speech features for speech recognition," *Proceedings of PEECS*, vol. 1, p. 92, 2005.
[20] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39-71, 1996, doi: https://dl.acm.org/doi/abs/10.5555/234285.234289.
[21] 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*, 2013,
[22] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Ismir*, 2000, pp. 1-11.
[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19-41, 2000, doi: https://doi.org/10.1006/dspr.1999.0361.
[24] W. J. Stewart, "Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling," *Princeton university press*, 2009.
[25] D. Y. Mohammed, "Overlapped speech and music segmentation using singular spectrum analysis and random forests," *Salford University*, 2017.
[26] D. Y. Mohammed, P. J. Duncan, M. M. Al-Maathidi, and F. F. Li, "A system for semantic information extraction from mixed soundtracks deploying MARSYAS framework," *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, Cambridge, UK, 2015, pp. 1084-1089, doi: 10.1109/INDIN.2015.7281886.
[27] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature extraction*, Springer, 2006, pp. 315-324, DOI: https://doi.org/10.1007/978-3-540-35488-8_13.
[28] A. S. Aljuboori, " Performance of case-based reasoning retrieval using classification based on associations versus Jcolibri and FreeCBR: a further validation study " *Journal of Physics: Conference Series* vol. 1003, NO. 1, 0. 012130, May 2018.