Journal of Engineering and Sustainable Development
http://jeasd.uomustansiriyah.edu.iq
https://doi.org/10.31272/jeasd.24.4.7

JEASD

Vol. 24, No. 04, July 2020
ISSN 2520-0917

ROBUST HYBRID FEATURES BASED TEXT INDEPENDENT SPEAKER IDENTIFICATION SYSTEM OVER NOISY ADDITIVE CHANNEL

*** Ali Muayad Jalil[1]** **Dr. Fadhel Sahib Hasan[2]** **Dr. Hesham Adnan Alabbasi[3]**

1) MSc. student, Computer Engineering Department, Mustansiriyah University, Baghdad, Iraq.
2) Assistant Prof., Electrical Engineering Department, Mustansiriyah University, Baghdad, Iraq.
3) Lecturer, Computer Engineering Department, Mustansiriyah University, Baghdad, Iraq.

Received 19/6/2019 ***Accepted in revised form 20/11/2020*** ***Published 1/7/2020***

Abstract: Robustness of speaker identification systems over additive noise is crucial for real-world applications. In this paper, two robust features named Power Normalized Cepstral Coefficients (PNCC) and Gammatone Frequency Cepstral Coefficients (GFCC) are combined together to improve the robustness of speaker identification system over different types of noise. Universal Background Model Gaussian Mixture Model (UBM-GMM) is used as a feature matching and a classifier to identify the claim speakers. Evaluation results show that the proposed hybrid feature improves the performance of identification system when compared to conventional features over most types of noise and different signal-to-noise ratios.

Keywords: *Robust speaker identification, robust feature extraction, PNCC, GFCC, FW, UBM-GMM.*

1. Introduction

Speaker recognition is a task of identifying the speaker based on his\her voice information that is extracted from speakers underlying speech [1]. Speaker recognition is divided into two parts:

1. Speaker identification: is determining the target speaker among a group of speakers and answering the question (whose voice is that?).
2. Speaker verification: is determining if the voice belongs to the claimed speaker and answering the question (is that the claimed speaker's voice?)[2].

Furthermore, Speaker Identification (SID) system can be divided into two classes: text-dependent and text-independent systems, in text-dependent, the speaker is required to speak a password sentence while in text-independent the speaker is not concentrating to a specific sentence meaning the speaker is free to say any sentence in his\her mind [3]. It has several applications such as remote access to services, banking operations through a telephone line, authentication and forensic applications [1].

For SID systems, the features related to each frame of speech are very important factors to implement a good SID system, the SID system works well in clean environments, but its performance may be degraded on real-life environments, where noise is being around the speaker [4], so that the features extracted from noisy speech in testing phase are no longer match the distribution model of clean training data [5].

The researchers use different approaches to overcome this problem, one approach is to use

**Corresponding Author*: alianengineer@live.com

speech enhancement techniques such as spectral subtraction [6], iterative Wiener filtering [7], Ephraim-Malah filtering [8], adaptive bionic wavelet shrinkage [5], deep neural networks [9], Empirical mode decomposition [10].

Another approach is to extract strong features that are robust to noise, Mel-Frequency Cepstral Coefficients (MFCC) are a very known and widely used features for SID systems [11]. Kim and Stern [12] presents an algorithm for feature extraction called Power Normalized Cepstral Coefficients (PNCC), that use power nonlinearity x^a instead of log nonlinearity used in MFCC features, the results show that PNCC features outperform MFCC features in clean and noisy environments. Hong and Pan [13] use spectrum mean normalization and cepstral mean normalization with MFCC features, to produce more robust features called modified MFCC. Wang *et al* [14] used wavelet octave coefficients residues, to provide complementary information to MFCC features, which give a noticeable improvement in mismatched spoken contents. Regularized linear prediction (RLP) is proposed by Hanilci *et al* [15] to decrease the mismatch of training and testing samples. RLP is spectral modeling that gives smoothed spectra without changing the positions of the formants, by correcting rapid changes in all-pole spectral envelops. This technique gives better results when compared with linear prediction methods. Khaled and Khalooq [16] consider the use of average framing linear prediction code and wavelet transform based feature extraction method, where the wavelet transform is used to decompose the speech signal first, then, Linear Prediction Code coefficients are calculated for each subband signal, finally, a dimension reduction is used by averaging the considered frames, experiments show improved recognition rate in white noise. Ganapathy *et al* [17] develop a frequency domain linear prediction features, based on the two dimensional auto-regressive model on the high energy peaks of the input speech signal, in time-frequency domain. The results show 30% improvement in noisy environments when compared to baseline MFCC features. Zhao *et al* [18] introduce new features called Gammatone Frequency Cepstral Coefficients (GFCC), the work is based on the auditory peripheral model, the paper uses the gammatone filter bank instead of the Mel-frequency filter bank, which improves the performance hen compared to MFCC features. Shantha *et al* [19] propose a feature extraction technique named inverted Mel-frequency cepstral coefficients, that captures complementary information of MFCC that presents in the high frequency part of the spectrum, with MFCC features and fused score Gaussian mixture model, 93.88 % identification rate is obtained on the TIMIT dataset, with 120 test speakers. Turner and Joseph [20] propose an improvement to the MFCC features, by replacing the discrete Fourier transform with the wavelet packet transform and discrete wavelet transform in computing the spectrum of speech signal at a variety of wavelet types and levels. Mean Hilbert Envelop Coefficients (MHEC) is proposed by Sadjadi and Hansen [21] to extract features by using smoothed Hilbert envelop of gammatone filter bank, the results show that MHEC features are less prone to noise than MFCC features. Islam *et al* [1] propose the use of neurogram, which is resulted by applying the speech signal to the auditory nerve model; their work achieves good results when features are extracted from narrowband frequency less than 1 kHz. Shi *et al* [22] try to improve the GFCC features by normalizing gammatone filter bank, and adding dynamic features, and then use an autoregressive moving average filter, they achieve better results than conventional GFCC features. Kim and Stern [23] try to improve PNCC features by using power nonlinearity

instead of log nonlinearity, medium time processing, and asymmetric nonlinear filtering to estimate the level of background noise for each individual frame and frequency bin and temporal masking. Guo *et al* [24] combine subtotal resonance (SGRs) with PNCC and LPCC features to get more robust features because of subglottal acoustics spectral characteristics varying less than corresponding speech signal spectral characteristics for the same speaker and that SGRs estimation algorithm is reliable even at low signal-to-noise ratio. Ahmed *et al* [25] proposed a feature extraction method by using discrete wavelet transform and MFCC features with feature warping to extract robust features, the proposed work gives good results for additive noise and presence of reverberation. Kobra *et al* [26] proposed to use mean and variance normalization and then applying auto-regression moving-average (ARMA) filter (MVA) to MFCC features, the new features give 28% accuracy improvement comparing with MFCC features at 5db SNR level.

The rest of this paper organized as follows. In section 2, the proposed speaker identification system described. Methodology was described in section 3. Simulation results and discussion described in section 4. Finally, the conclusion was in section 5.

2. Speaker Identification System

Fig. 1 shows the proposed SID system, where the speech signal is pre-processed first, after that, the features are extracted by the proposed extraction algorithm from clean and noisy pre-processed speech signals. The extracted features from the clean utterances are used to train the classifier and models are saved. For the testing stage, the extracted features from each test utterance used as an input to the model of each speaker. The model with maximum probability result is identified as the target speaker. The detail description of each component in the system is explained in the next subsections.

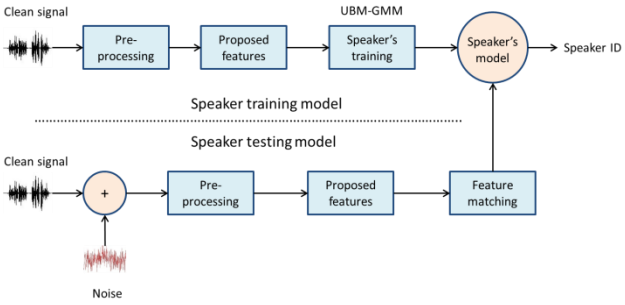

Figure 1. Block diagram of SID system.

2.1. Pre-Processing

In pre-processing stage, the input speech signal is pre-processed through three sub-stages, pre-emphasis filter, framing and windowing. The pre-emphasis filter is a high pass filter used to emphasize high frequencies and recompense for human speech production which usually has a tendency to attenuate high frequencies [27]. A simple high order filter with a value of 0.97 is given in the equation [27]:

$$y(t) = x(t) - 0.97x(t-1) \quad (1)$$

Where *x(t)* is the input and *y(t)* is the output. Pre-emphasis is only applied to PNCC features as in [23], but not to GFCC features because it has a significant impact on GFCC energy-related features which lead to performance dropping [4].

Framing is to divide the speech signal into small segments, because usually, the speech signal length is very high, so it's better to divide it into small frames typically 20-30 milliseconds length to assure continuity of speech signal. Discontinuity in speech signal may lead to wrong extracted features, and may affect the accuracy of SID system and to ensure that the feature vectors are evenly spaced in time-domain [28].

The Hamming window is used to reduce framed speech signal discontinuity in the beginning and end of the frame, and increase signal continuity between neighboring frames [29]. Given a speech signal $s(n)$, n=1,2,…,N-1 where N is the length of the framed signal, then the resulted frame signal after invoking hamming window, $s_w(n)$ [29]:

$$s_w(n) = s(n) * w(n) \quad ,0 \le n < N \qquad (2)$$

Where $w(n)$ is the Hamming window function and It is defined by [29]:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) , 0 \le n < N \ (3)$$

2.2. Proposed Feature Extraction Algorithm

In this stage, the clean and noisy pre-processed speech signal is prepared to extract the proposed features that are used to build a model for each speaker in the training phase and to match them in the testing phase. The proposed algorithm is shown in Fig. 2, where each stage of the extraction algorithm is described below. Δ represents dynamic features.

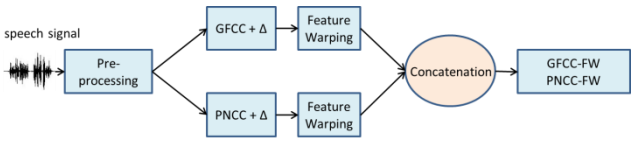

Figure 2. Proposed feature extraction algorithm.

2.2.1. *Power Normalized Cepstral Coefficients*

PNCC is a very accurate feature that outperforms many existing features in clean and noisy environments, with slight increases in computational cost compared with conventional features [30]. The high identification accuracy rate comes by using power-law nonlinearity, which gives a close approximation of how humans can hear [30]. Fig. 3 shows the block diagram of processes stages to produce PNCC features as described in [12].

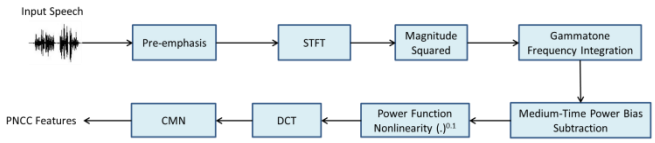

Figure 3. Block diagram of PNCC algorithm.

Step1*:* Sort Time Fourier Transform (STFT) is applied by using Discrete Fourier Transform (DFT) to convert the speech signal from time domain to frequency domain, to extract the cepstral coefficients [31].

Step2: The magnitude is taken to calculate the frame power [27].

Step3: Gammatone filter bank is calculated using Equivalent Rectangular Bandwidth (ERB) of the band-pass filter.

Step4: Three noise reduction techniques were applied: Asymmetric noise suppression, temporal masking and weight smoothing [23] to suppress the noise and channel variation.

Step5: Applying power function nonlinearity because the output behavior does not critically rely on the amplitude of the input, just like the human auditory system, when the input level is below the threshold; the output level is zero.

Step6*:* Discrete cosine transform is used to de-correlate the cepstral features which were highly correlated as spectral features [32].

Step7*:* Cepstral Mean Normalization (CMN) is a simple feature normalization technique, where each cepstral vector x_t is subtracted by the mean value μ_x to produce the normalized cepstral vector $\hat{x}_t$ as [33]:

$$\hat{x}_t = x_t - \mu_x \qquad (4)$$

$$\mu_{x_t} = \frac{1}{T}\sum_{t=0}^{T-1} x_t \qquad (5)$$

When the normalization is done, the cepstral sequence mean is zero [33]. CMN is good in removing the channel distortion and improving the recognition rate in noisy environments [33].

2.2.2. *Gammatone Frequency Cepstral Coefficients*

MFCC and Perceptual Linear Prediction features are widely used techniques for constructing SID systems [34]. The gammatone filter bank creates a series of overlapping band-pass filters as a model of the human auditory system [35]. The implementations of a composition of gammatone filter bank, ERB and cubic root lead to increase the resulted GFCC features robustness in clean and noisy environments [4], which made it successfully replace for MFCC [18][34]. The block diagram of GFCC feature is depicted in Fig. 4 [36].

The GFCC features extraction process is summarized as [18]:

Step1: Preprocessed speech signal passed through 64 channel gammatone filter bank whose center frequencies ranging from 50 – 8000 Hz.

Step2: Fully rectify the response of the filter (i.e. take absolute value) at each channel then decimate into 100 Hz which yields a 10 ms frame rate.

Step3: Take the absolute value for the decimated to create T-F representation that is a variant of cochleagram.

Step4: Take cubic root for the decimated outputs magnitudes to loudness-compressed as in the following equation [18]:

$$G_m[i] = ||g|_{decimate}[i,m]|^{1/3} \quad (6)$$

Where $i=0,1,\ldots,N-1$, N is number of filters, $m=0,1,\ldots,M-1$, M is number of time frames taken after decimation.

Step5: Apply DCT to de-correlate the components and reduce dimensionality.

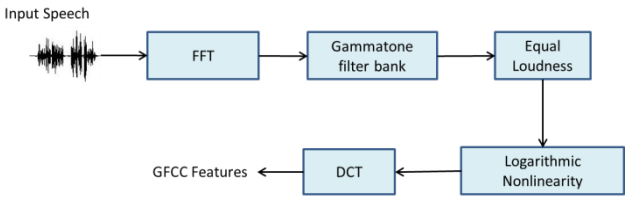

Figure 4. Block diagram of GFCC algorithm .

2.2.3. Dynamic Features

Dynamic features can capture and model temporal information between frames, and concatenate it with the cepstral features in speaker recognition because it helps in identifying the speaker style and speaking durations more accurately [27]. The first order derivative (delta) features Δ at time t are calculated from a set of cepstral neighboring feature vectors Z [32]:

$$\Delta(t) = \frac{\sum_{\omega=1}^{W} \omega(Z(t+\omega)-Z(t-\omega))}{2\sum_{\omega=1}^{W} \omega^2} \quad (7)$$

Where ω is window index, W is the half-window length and is set to 2. These temporal derivatives were then concatenated with the cepstral features to result the augmented feature vector [37].

2.2.3. *Feature Warping (FW)*

The purpose of feature warping is to gain more robust features by making the features following a specific distribution target. Feature warping processing steps can be summarized in the following steps [38]:

Step 1: select a target distribution.

Step 2: extract cepstral coefficients (PNCC and GFCC in this paper).

Step 3: create a lookup table to map the rank of sorted cepstral features to target warped features using the desired distribution.

Step 4: isolate a window of N (3 seconds) features and sort their values in descending order and give a rank for that sorted features,

where a rank of 1 is for the most positive value and rank N for the most negative value, this rank is used as an index in the lookup table created in step 3.

Step 5: move the sliding window by 1 frame.

Step 6: repeat step (4) for each sliding window frame shift.

The lookup table can be determined by finding m [25]:

$$\frac{N+\frac{1}{2}-R}{N} = \int_{z=-\infty}^{m} h(z)dz \quad (8)$$

If a normal distribution is chosen then [25]:

$$h(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad (9)$$

Where *m* is the feature warped component, *N* is analysis window length and *R* is the rank. The warped value *m* can be calculated by initially making *R=N* and solving *m* by numerical integration method for each decremented value of *R* [25].

2.2.4. *Feature Concatenation*

The final step in the proposed feature extraction algorithm is to concatenate the extracted features to form the PNCCFW-GFCCFW features as shown in Fig. 5.

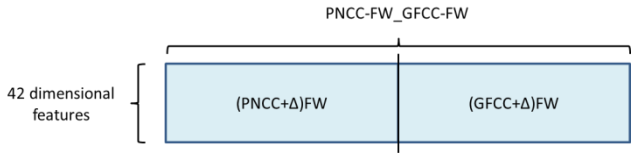

Figure 5. Final features matrix.

2.3. Universal Background Model Gaussian Mixture Model

UBM-GMM is the dominant classifier implemented for the speaker recognition system [1], [2]. In this paper, UBM-GMM is used to test the robustness of the proposed features. Fig. 6 shows block diagram of UBM-GMM for testing and training phases.

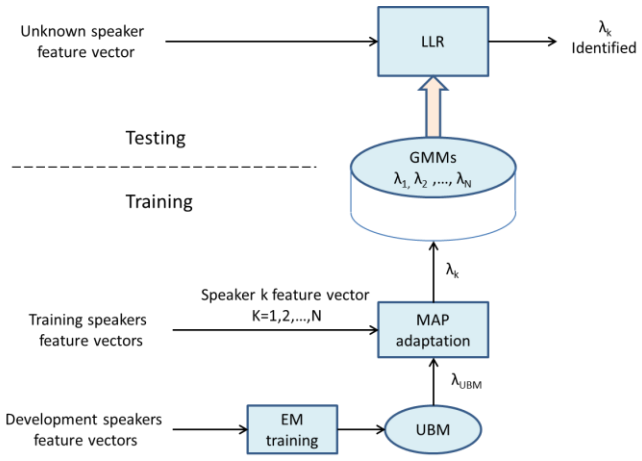

Figure 6. Block diagram of UBM-GMM.

2.3.1. *Gaussian Mixture Model (GMM)*

GMM is a model that gives the distribution probabilities of feature vectors resulting from each individual speaker [30]. A GMM for speaker *j* is a weighted sum of *M* components densities and it is expressed by [27]:

$$p(\vec{x}, \lambda_j) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (10)$$

where $\vec{x}$ is D-dimensional random vector, $b_i(\vec{x})$, *i=1,2,...,M* are the component densities, and p_i , *i=1,2,...,M* are mixture weights such that

$$\sum_{i=1}^{M} p_i = 1 \quad (11)$$

Each component density $b_i(\vec{x})$ is the D-variate random vector $\vec{x}$ represented by [39]:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}} Exp(s_i(\vec{x})) \quad (12)$$

$$s_i(\vec{x}) = -\frac{1}{2}(\vec{x} - \overrightarrow{\mu_i})^T \sum_i{}^{-1} (\vec{x} - \overrightarrow{\mu_i}) \quad (13)$$

Where $\overrightarrow{\mu_i}$ is mean vector and $\sum_i$ is the covariance matrix. For SID system, each speaker is represented by GMM mean vectors,

covariance matrices, and mixture weights and they are denoted by his\her model λ [39]:

$$\lambda = \{p_i, \overrightarrow{\mu_i}, \sum_i\} \quad (14)$$

2.3.2. *Expectation Maximization (EM)*

This is a two steps training technique; initialization step and expectation maximization step, the initialization step gives initial estimates for Gaussian components, while EM step used to re-compute means, covariance and weights for GMM components iteratively [40]. For UBM mixture *i*, the posterior probability $Pr(i \mid x_t)$ is computed as [41]:

$$Pr(i \mid x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \quad (15)$$

New estimation of weights w_i [41]:

$$w_i = \sum_{t=1}^{T} Pr(i \mid x_t) \quad (16)$$

New estimation of mean μ_i [41]:

$$\mu_i = \frac{\sum_{t=1}^{T} Pr(i \mid x_t)\, x_t}{\sum_{t=1}^{T} Pr(i \mid x_t)} \quad (17)$$

New estimation of covariance σ_i^2 [41]:

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} Pr(i \mid x_t)\, x_t{}^2}{\sum_{t=1}^{T} Pr(i \mid x_t)} \quad (18)$$

2.3.3. *Universal Background Model (UBM)*

UBM is produced by speech samples of all the speakers except the speakers to be tested [30] to compute the probabilities of the data that not belongs to the target speaker [39]. The reason behind using UBM is that it is trained using hundreds of speaker's data, which means, it does not suffer from the insufficient training and unseen data. With that in mind, UBM is trained more reliably than any speaker GMM model, and the speakers models can be estimated with small amounts of data by using maximum a posteriori adaptation to find a model for each speaker [27].

2.3.4. *Maximum A Posteriori (MAP) adaptation*

The speaker model is derived by adapting UBM parameters from training utterances of the speaker and MAP adaptation. The basic idea of adaptation is to extract the speaker's model from UBM parameters. This provides a better coupling between UBM and speaker's model, which produces better performance and allows for a fast scoring technique [41].

2.3.5. *Log-Likelihood Ratio (LLR)*

The final decision of matching between adapted models, λ_{GMM}, *UBM*, λ_{UBM}, that resulted from the training stage and the testing utterances feature vectors X is done by using log likelihood ratio (LLR) [30]:

$$LLR(X) = \log P(X|\lambda_{GMM}) - \log P(X|\lambda_{UBM}) \quad (19)$$

Where $X = [\, x_1, \ldots, x_T]$, T is the number of feature vectors.

3. Experimental Methodology

Experiments are done on TIMIT [42] dataset which consists of 630 speakers, 10 utterances per speaker, 530 speakers are randomly chosen to calculate UBM and 100 speakers used for performance evaluation of the proposed feature, 9 of 10 utterances are chosen randomly to train the GMM and the 1 left utterance is used for

testing. The proposed feature extraction algorithm robustness is tested in noisy and clean conditions, 15 noise types are chosen from the Noisex-92 [43] noise dataset which are artificially added to the test utterances with a signal to noise ratio levels 0,5,10 and 15 db. The noise types description are listed in Table 1. All utterances are framed into overlapping frames with a Hamming window of 25 milliseconds frame length, and 10 milliseconds window shifts. GFCC with 42 (21 GFCC and 21 ΔGFCC) features are extracted, with 64 gammatone filters and dropping the 0th coefficient, and 42 PNCC (21 PNCC and 21 ΔPNCC) features are extracted with 40 filters, and applying pre-emphasizing filter with 0.97, and dropping the 0th coefficient, from each frame, then applying feature warping with window length of 301 frames (3 sec) [38], to each cepstral features (GFCC and PNCC), to produce GFCC-FW and PNCC-FW, after that, a concatenation of resulted features is taken place to obtain the final proposed features. UBM-GMM is used to evaluate the results, 256 Gaussian mixtures and 10 expectation maximization iterations are used.

Table 1. noise types description of noisex-92 noise

Noise Type	Description
Babble	Multi talkers
Buccaneer 1	Fighter jet noise
Buccaneer 2	Fighter jet noise
Destroyerengin	Military destroyer engine room
Destroyerops	Military destroyer operations room
F16	Fighter jet noise
Factory 1	Factory noise
Factory 2	Factory noise
Hfchannel	High radio frequency noise
Leopard	Military tank
M109	Military tank
Machinegun	Machine gun shoots
Pink	An audible noise that has high power in low frequencies and low power in high frequencies
Volvo	Car interior
White	Additive white Gaussian noise

4. Simulation Results and Discussion

In this section, the proposed features in both clean and noisy environments are tested and compared with similar studies to show the robustness of the features.

4.1. Simulation Results for Baseline and Proposed Features

Table 2 shows the results of baseline features (PNCC, PNCC-FW, GFCC, and GFCC-FW) and the proposed features in clean and noisy test utterances with similar extraction parameters such as frame length, frame shift, features length, and dynamic features. The identification accuracy is calculated based on the following formula:

$$Acc = \frac{correctly\ identified\ speakers}{total\ test\ speakers} \times 100\% \quad (20)$$

In this table, high numbers indicating better accuracy. As can be seen that the proposed feature extraction algorithm achieves a higher identification rate in almost every noise type except for *machinegun* noise type, where PNCC feature has better results than the proposed feature. GFCC and GFCCFW showed a poor performance compared with other features. PNCC and PNCCFW have better than GFCC based features, due to the robustness in their extraction algorithm.

Table 2. Identification rate results of baseline and proposed features

Noise type \ Features	PNCC	PNCC-FW	GFCC	GFCC-FW	Proposed features
Clean	*99*	*99*	96	98	*99*
babble					
0db	69	75	56	51	*80*
5db	86	87	77	77	*89*
10db	*96*	93	90	86	93
15db	95	95	94	93	*96*
buccaneer1					
0db	41	45	19	42	*53*
5db	65	63	49	62	*76*
10db	76	82	73	78	*86*
15db	89	89	87	92	*93*
buccaneer2					
0db	35	34	5	30	*40*
5db	59	55	30	51	*65*
10db	64	73	58	75	*78*
15db	81	83	81	83	*91*
destroyerengine					
0db	55	64	23	48	*69*
5db	74	73	50	72	*82*
10db	89	87	77	88	*90*
15db	89	92	88	*93*	91
destroyerops					
0db	54	60	31	51	*65*
5db	70	76	67	70	*82*
10db	83	83	82	78	*87*
15db	91	89	88	91	*93*
f16					
0db	53	58	32	42	*60*
5db	74	76	60	63	*82*
10db	87	89	80	81	*89*
15db	91	91	87	91	*94*
factory1					
0db	49	54	46	53	*63*
5db	75	*79*	75	74	*79*
10db	88	*89*	86	85	*89*
15db	92	94	89	92	*96*
factory2					
0db	84	*85*	73	76	*85*
5db	92	*93*	87	86	90
10db	*95*	*95*	92	89	*95*
15db	94	95	95	93	*97*
hfchannel					
0db	64	59	30	58	*72*
5db	*84*	*84*	56	75	80
10db	91	92	85	88	*95*
15db	93	96	88	92	*97*
leopard					
0db	*96*	94	57	80	92
5db	*98*	96	81	89	97
10db	97	97	88	93	*98*
15db	98	95	95	96	*99*
m109					
0db	72	76	66	64	*77*
5db	*87*	85	82	81	86
10db	87	89	85	91	*92*
15db	95	94	92	92	*96*
machinegun					
0db	*98*	96	90	89	97
5db	*98*	*98*	93	89	97
10db	*99*	98	95	94	98
15db	99	98	97	96	*100*
Pink					
0db	39	30	13	38	*45*
5db	52	54	29	66	*74*
10db	70	75	62	80	*85*
15db	82	81	81	89	*93*
Volvo					
0db	98	97	93	94	*99*
5db	*99*	96	93	96	*99*
10db	*99*	98	96	98	*99*
15db	*99*	97	97	*99*	*99*
White					
0db	52	44	25	46	*55*
5db	67	68	55	67	*76*
10db	75	80	79	84	*86*
15db	89	87	88	93	*94*
Average	80.5	81.13	70.7	77.89	*85.15*

4.2. Effect of Feature Length

Figure 7 shows the effect of features length on the identification rate of the proposed features. The feature length is the number of features that are extracted from each frame of the speech signal. The feature vector maximum length is equal to the number of filters in the filterbank. We can choose the desired number of features

per frame to be extracted. In our work, we chose these numbers randomly and tested it, for comparison, three features lengths are tested 13, 21 and 30 to see the effect of features length on the identification rate. The results listed in Fig. 7 represents the average accuracy over all the noise types and levels used in this paper. As can be seen from the table, 21 and 30 feature length gives almost similar results but 21 features length gives the best performance in average and 13 features length gives poor results among them.

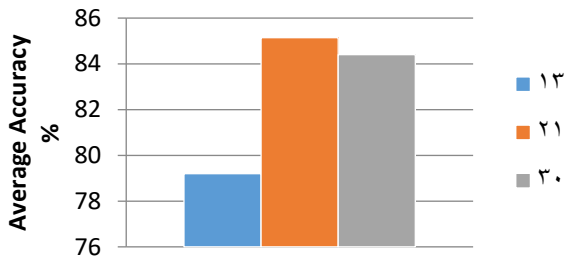

Figure 7. Identification rate for different lengths of the features.

4.3. Effect of Dynamic Features

To see the effect of adding dynamic features to the proposed features on the identification rate, three tests were done. The first one was done without adding dynamic features. In the second test, the first-order derivative (Δ or Delta) is added to the proposed features. A second order derivative (Δ^2 or Delta-Delta) is extracted and added on the third test. The test results are listed in Fig. 8, where the results indicating that adding the first order derivative (Delta) gives a slightly better identification rate than second order derivative (Delta-Delta). Again, The results listed in Fig. 8 represents the average accuracy over all the noise types and levels used in this paper.

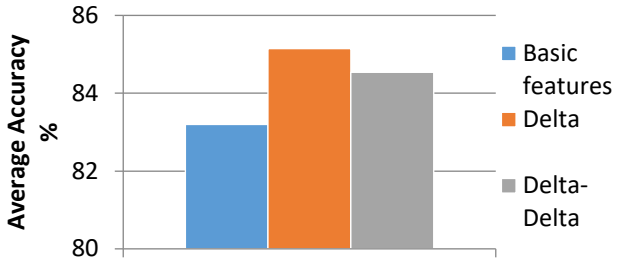

Figure 8. Effect of dynamic features on identification rate.

4.4. Effect of Frame Length

The effect of frame length is shown in Fig. 9 with frame lengths 16, 25, and 32 ms are presented respectively. The results show 25 ms is the best frame length for the proposed feature extraction algorithm which gives the best identification rate among tested lengths. The results listed in Fig. 8 represents the average accuracy over all the noise types and levels used in this paper.

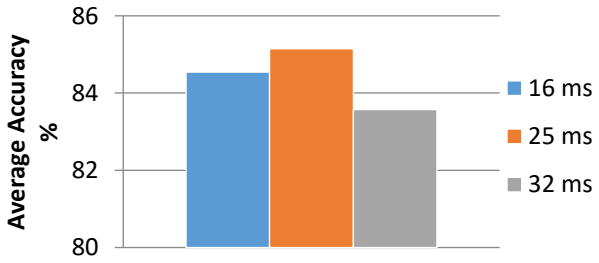

Figure 9. Effect of Frame length on identification rate.

4.5. Comparison with Other Studies

To test the effectiveness of the proposed features in clean and noisy environments, the proposed features results are compared with several studies that briefly described in Table 3; the results indicate that the proposed features outperform all of the works compared with.

Fig. 10,11,12 and 13 shows a comparison between the proposed work and the works proposed by [1], [24], [26] and [44] respectively. The noise types and levels used to compare are based on the authors works.

Table 3. Brief description of the systems used in the comparison.

Work	Proposed features	No. of testing speakers	No. of test utterances	No. of features	Frame length	Frame shift	Pre-emphasis	Evaluation system	No. of mixtures
Work proposed by [1]	Neurogram	100	2	25	42 ms	25.2 ms	No	UBM-GMM	128
Work proposed by [24]	1. PNCC+SGRs 2. LPCC+SGRs	630	1	.. PNCC: 60 ?. LPCC: 24	Not stated	Not stated	No	UBM-GMM	128
Work proposed by [26]	Combining MFCC and MVA	100	2	20	15 ms	10 ms	0.97	GMM	64
Work proposed by [44]	Auto-regressive with MFCC (AR-MFCC)	200	1	64	20 ms	10 ms	0.97	GMM	64

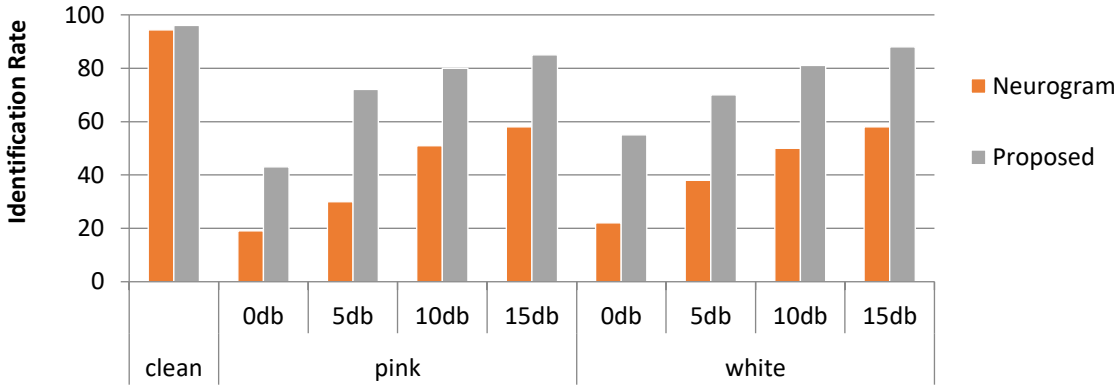

Figure 10. Compare proposed work with work proposed by [1].

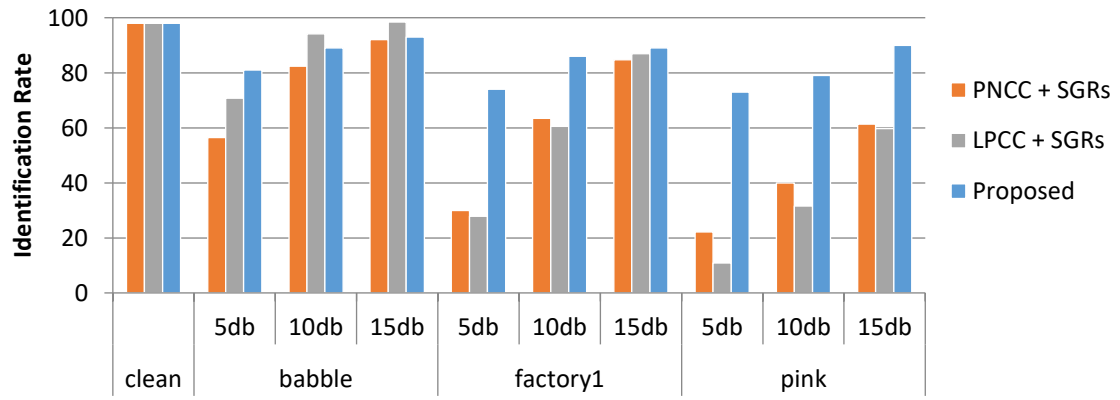

Figure 11. Compare proposed work with work proposed by [24].

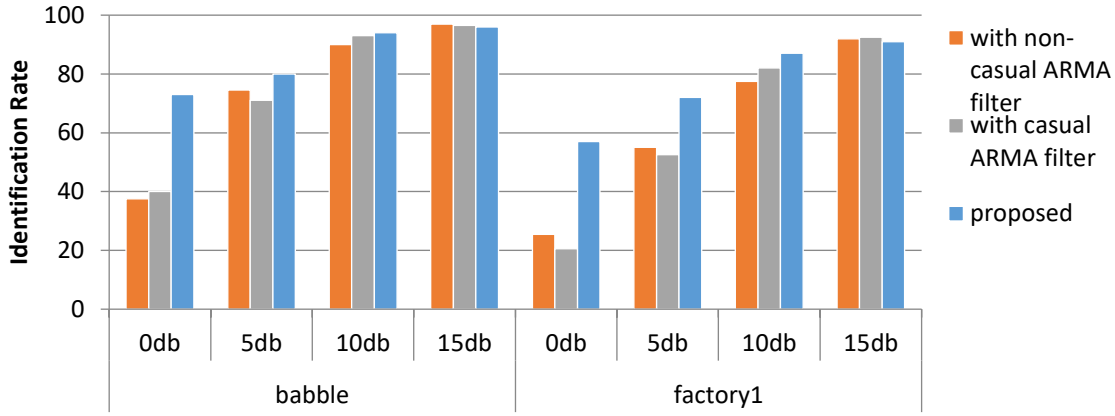

Figure 12. Compare proposed work with work proposed by [26].

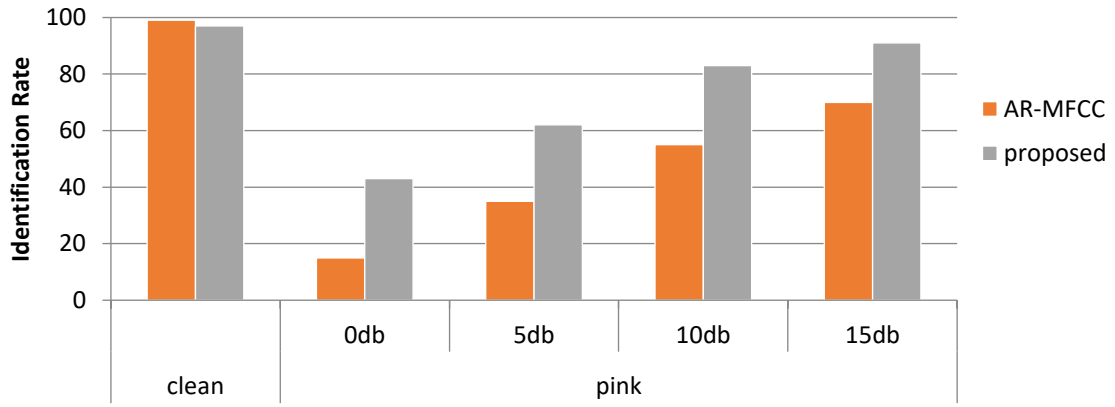

Figure 13. Compare proposed work with work proposed by [44].

5. Conclusions

In this work, new concatenation feature based on PNCC and GFCC are studied for robust SID system over noisy channel. UBM-GMM is used as feature matching with 256 Gaussian mixtures, and 10 expectation maximization iterations. Experiments are done on the TIMIT dataset with 100 speakers used to test the performance of the proposed feature, 9 of 10 utterances are chosen randomly to train the GMM and 1 utterance used for testing. The testing is done on both clean and noisy conditions to test the robustness of the proposed features, 15 noise types are chosen from the Noisex-92 noise dataset that are added to the test utterances with a signal to noise ratio levels of 0,5,10 and 15 db. The performance results show that the proposed features outperforms baseline features (PNCC and GFCC) and other proposed works [1], [24], [26] and [44], feature warping technique even increased the identification rate.

Acknowledgements

This work is supported by Computers Department/ College of Engineering/ Mustansiriyah University.

Conflict of interest

The authors declare that the publication of this article cause no conflict of interest.

6. References

1. Islam , M. A., Jassim, W. A., Cheok , N. S., Zilany , M. S., (2016). "*A robust speaker identification system using the responses from a model of the auditory periphery*," PLoS ONE, Vol. 11, No. 7, pp.1-21.
2. Hsieh, C. T., Lai, E., and Wang, Y. C., (2003). "*Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model*," Journal of Information Science and Engineering, Vol. 19, No. 2, pp. 267-282.
3. Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J., (2014). "*Deep neural networks for small footprint text-dependent speaker verification*," Proc. International Conference on Acoustics, Speech and Signal Processing ICASSP, Italy, pp. 4080-4084.
4. Zhao, X. and Wang, D., (2013). "*Analyzing noise robustness of mfcc and gfcc features in speaker identification*," Proc. International Conference on Acoustics, Speech and Signal Processing ICASSP, Canada, pp. 7204–7208.
5. Govindan, S. M., Duraisamy, P., and Yuan, X., (2014). "*Adaptive wavelet shrinkage for noise robust speaker recognition*," Digital Signal Processing, Vol. 33, pp. 180–190.
6. Ganchev, T., Potamitis, I., Fakotakis, N., and Kokkinakis, G., (2004). "*Text-independent speaker verification for real fast-varying noisy environments*," International Journal of Speech Technology, Vol. 7, No. 4, pp. 281-292.
7. El-Fattah, M. A. A., Dessouky, M. I., Abbas, A. M., Diab, S. M., El-Rabaie, E. S. M., Al-Nuaimy, W., and El-Samie, F. E. A, (2014). "*Speech enhancement with an adaptive Wiener filter*," International Journal of Speech Technology, Vol. 17, No. 1, pp. 53-64.
8. Brajević, Z., and Petošić, A., (2012). "*Signal denoising using STFT with Bayes prediction and Ephraim–Malah estimation*," in: Proceedings of the 54th International Sympo-sium ELMAR, IEEE, Croatia, pp. 183–186.
9. Xu, Y., Du, J., Dai, L. R., and Lee, C. H., (2015). "*A regression approach to speech enhancement based on deep neural networks*," IEEE/ACM *Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 23, No. 1, pp. 7-19.
10. Hasan, T., and Hansen, J. H., (2011). "*Robust speaker recognition in non-stationary room environments based on empirical mode decomposition*," INTERSPEECH, Italy, pp. 2733-2736.
11. Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., and Shamma, S., (2011). "*Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition*," IEEE Automatic Speech Recognition and Understanding Workshop, USA, pp. 559-564.
12. Kim, C., and Stern, R. M., (2009). "*Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction*," INTERSPEECH, UK, pp. 28-31.
13. Hong, W., and Jin'gui, P., (2010). "*Modified MFCCs for robust speaker recognition*," in Proceedings of the IEEE International Conference on Intelligent Computing and Intelligent Systems, China, Vol. 1, pp. 276-279.
14. Wang, N., Ching, P. C., Zheng, N., and Lee, T., (2011). "*Robust speaker recognition using denoised vocal source and vocal tract features," IEEE transactions on audio,*

speech, and language processing, Vol. 19, No. 1, pp. 196-205.

15. Hanilçi, C., Kinnunen, T., Ertas, F., Saeidi, R., Pohjalainen, J., and Alku, P., (2012). "*Regularized all-pole models for speaker verification under noisy environments*," IEEE Signal Processing Letters., Vol. 19, pp. 163–166.
16. Daqrouq, K., and Al Azzawi, K. Y., (2012). "*Average framing linear prediction coding with wavelet transform for text-independent speaker identification system*," Computers and Electrical Engineering, Vol. 38, No. 6, pp. 1467–1479.
17. S. Ganapathy, S., Thomas, S., and Hermansky, H., (2012). "*Feature extraction using 2-D autoregressive models for speaker recognition*," Odyssey 2012-The Speaker and Language Recognition Workshop, Singapore, pp. 229-235.
18. Zhao, X., Shao, Y., and Wang, D., (2012). "*CASA-Based Robust Speaker Identification*," IEEE Transactions on Audio, Speech and Language Processing, Vol.20, No.5, pp.1608-1616.
19. Kumari, R. S. S., and Nidhyananthan, S. S., (2012). "*Fused MEL feature sets based text-independent speaker identification using Gaussian mixture model*," Procedia Engineering, Vol. 30, pp. 319-326.
20. Turner, C., and Joseph, A., (2015). "*A wavelet packet and mel-frequency cepstral coefficients-based feature extraction method for speaker identification*," Procedia Computer Science, Vol. 61, pp. 416–421.
21. Sadjadi, S. O., and Hansen, J. H., (2015). "*Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification*," Speech Communication, Vol. 72, pp. 138-148.
22. Shi, X., Yang, H., and Zhou, P., (2017). "*Robust speaker recognition based on improved GFCC*," IEEE International Conference on Computer and Communications, China, pp. 1927–1931.
23. Kim, C., and Stern, R. M., (2012). "*Power-normalized cepstral coefficients (PNCC) for robust speech recognition*," IEEE/ACM Transactions on Audio, Speech and Language Processing, pp. 4101-4104.
24. Guo, J., Yang, R., Arsikere, H., and Alwan, A., (2017). "*Robust speaker identification via fusion of subglottal resonances and cepstral features*," The Journal of the Acoustical Society of America, Vol. 141, No. 4, pp. 420-426.
25. Al-Ali, A. K. H., Dean, D., Senadji, B., Chandran, V., and Naik, G. R., (2017). "*Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions*," IEEE Access, Vol. 5, pp. 15400-15413.
26. Korba, M. C. A., Bourouba, H., and Rafik, D., (2018). "*Text-Independent Speaker Identification by Combining MFCC and MVA Features*," IEEE International Conference on Signal, Image, Vision and their Applications (SIVA), Algeria, pp. 1-5.
27. Togneri, R., and Pullella, D., (2011). "*An overview of speaker identification: Accuracy and robustness issues*," IEEE circuits and systems magazine, Vol. 11, No. 2, pp. 23-61.
28. Liu, J. C., Leu, F. Y., Lin, G. L., and Susanto, H., (2018). "*An MFCC-based text-independent speaker identification system for access control*," Concurrency and Computation: Practice and Experience, Vol. 30, No. 2, 2018: e4255.
29. AboElenein, N. M., Amin, K. M., Ibrahim, M., and Hadhoud, M. M., (2016). "*Improved text-independent speaker identification system for real time applications*," 2016 Fourth International Japan-Egypt Conference on Electronics, Communications

and Computers (JEC-ECC), Egypt, pp. 58-62.

30. Nayana, P. K., Mathew, D., and Thomas, A., (2017). "*Comparison of text independent speaker identification systems using GMM and i-vector methods*," Procedia computer science, Vol. 115, pp. 47-54.
31. Sengupta, S., Yasmin, G., and Ghosal, A., (2017). "*Speaker Recognition Using Occurrence Pattern of Speech Signal*," Recent Trends in Signal and Image Processing, Springer, Singapore, pp. 207-216.
32. Shao, Y., and Wang, D., (2008). "*Robust speaker identification using auditory features and computational auditory scene analysis*," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, USA, pp. 1589-1592.
33. Droppo J., Acero A. (2008) "*Environmental Robustness*," In: Benesty J., Sondhi M.M., Huang Y.A. (Eds) Springer Handbook of Speech Processing. Springer Handbooks. Springer, Berlin, Heidelberg, pp. 653-680.
34. Dua, M., Aggarwal, R. K., and Biswas, M., (2019). "*GFCC based discriminatively trained noise robust continuous ASR system for Hindi language*," Journal of Ambient Intelligence and Humanized Computing, Vol. 10, No. 6, pp.2301-2314.
35. Jeevan, M., Dhingra, A., Hanmandlu, M., and Panigrahi, B. K., (2017). "*Robust speaker verification using GFCC based i-Vectors*," Proceedings of the International Conference on Signal, Networks, Computing, and Systems, New Delhi, Vol. 395, pp.85-91.
36. Jayanth, M., and Reddy, B., (2016). "*Speaker Identification based on GFCC using GMM-UBM*," International Journal of Engineering Science Invention, Vol. 5, No. 5, pp. 62-65.
37. Qi, J., Wang, D., Jiang, Y., and Liu, R., (2013). "*Auditory features based on gammatone filters for robust speech recognition*," 2013 IEEE International Symposium on Circuits and Systems (ISCAS2013), China, pp. 305-308.
38. Pelecanos, J., and Sridharan, S., (2001). "*Feature warping for robust speaker verification*," Proc. Speaker Odyssey, Greece, pp. 213-218.
39. Scheffer, N., and Bonastre, J. F., (2006). "*Ubm-Gmm driven discriminative approach for speaker verification*," 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop, Puerto Rico, pp. 1-7.
40. Taif A. M., (2009). "*Text-independent speaker identification system*," M.Sc. Thesis, Department of Electrical Engineering Mustansiriyah Uni., Baghdad, Iraq.
41. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., (2000). "*Speaker verification using adapted Gaussian mixture models*," Digital signal processing, Vol. 10, pp.19-41.
42. TIMIT dataset, available online at: https://catalog.ldc.upenn.edu/LDC93S1. last accessed at 28 May 2019.
43. NOISEX-92 noise dataset available online on: http://spib.linse.ufsc.br/noise.html. last accessed at 28 May 2019.
44. AJGOU, R., Salim, S. B. A. A., GHENDIR, S., CHEMSA, A., and TALEB-AHMED, A., (2016). "*Robust speaker identification system over AWGN channel using improved features extraction and efficient SAD algorithm with prior SNR estimation*," International Journal of Circuits, Systems and Signal Processing, Vol. 10, pp. 108-118.