**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 10, No. 5, October 2020, pp. 5208~5216
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i5.pp5208-5216

# Adaptive wavelet thresholding with robust hybrid features for text-independent speaker identification system

**Hesham A. Alabbasi[1], Ali M. Jalil[2], Fadhil S. Hasan[3]**
[1]Department of Computers Science, Faculty of Education, Mustansiriyah University, Iraq
[2]Department of Computer Engineering, Faculty of Engineering, Mustansiriyah University, Iraq
[3]Department of Electrical Engineering, Faculty of Engineering, Mustansiriyah University, Iraq

**Article Info**

***Article history:***

Received Aug 31, 2019
Revised Apr 6, 2020
Accepted Apr 17, 2020

***Keywords:***

FW
GFCC
PNCC
Robust speaker identification
Wavelet thresholding

## ABSTRACT

The robustness of speaker identification system over additive noise channel is crucial for real-world applications. In speaker identification (SID) systems, the extracted features from each speech frame are an essential factor for building a reliable identification system. For clean environments, the identification system works well; in noisy environments, there is an additive noise, which is affect the system. To eliminate the problem of additive noise and to achieve a high accuracy in speaker identification system a proposed algorithm for feature extraction based on speech enhancement and a combined features is presents. In this paper, a wavelet thresholding pre-processing stage, and feature warping (FW) techniques are used with two combined features named power normalized cepstral coefficients (PNCC) and gammatone frequency cepstral coefficients (GFCC) to improve the identification system robustness against different types of additive noises. Universal background model Gaussian mixture model (UBM-GMM) is used for features matching between the claim and actual speakers. The results showed performance improvement for the proposed feature extraction algorithm of identification system comparing with conventional features over most types of noises and different SNR ratios.

***Corresponding Author:***

Ali Muayad Jalil,
Department of Computer Engineering,
Mustansiriyah University, Baghdad, Iraq.
Email: alianengineer@live.com

## 1. INTRODUCTION

Speaker recognition is a task of identifying the speaker by the voice information that is extracted from speakers underlying speech [1] and its divided into speaker identification system which define the speaker from a group of speakers, and speaker verification system which determining if the voice is really the claimed speaker's voice [2]. Furthermore, it falls in two classes, text-dependent, in which the speaker should utter a password, and text-independent, which letting the speaker free to say any words in mind [3]. There are many applications for speaker identification system, such as remote access to services, banking operations through a telephone line, authentication and forensic applications [1].

In speaker identification (SID) systems, the extracted features from each speech frame are a crucial factor for building a reliable identification system. In clean environments, the identification system performs well, but in noisy environments, the distribution models of the features that extracted from the noisy speech will not matches the clean features distribution model that built in training phase [4]. To overcome this problem, the researchers applied many approaches to achieve this goal. In this section, we presents some related works; Speech enhancement is one of these approaches, where the noisy speech signal pre-processed first to suppress the noise. Spectral subtraction (SS) speech enhancement technique [5] depends on the correlation absence between the clean speech and the noise in which they are additive in time domain.

The noise is assumed to change very slow compared with speech so, the noise spectrum can be estimated during silence periods, and the clean speech spectrum can be estimated by subtracting the estimated noise spectrum from noisy speech signal spectrum. Improved spectral subtraction [6] based on two steps, speech activity detection (SAD) and noise amplitude spectral estimation; it was adopted based on frequency band variance to detect speech endpoints to calculate the noise power spectrum. Brajevic *et al* [7] proposed to use Ephraim-Malah estimation and short time Fourier transform to suppress stationary noise by reducing spectral coefficients.

Abd El-Fattah *et al* [8], used Adaptive wiener filter in time domain to estimate noise from speech signal. Ahmet M. and Aydin A. [9] used empirical mode decomposition (EMD) for speech signal decomposition with detrended fluctuation analysis (DFA) technique to threshold the noisy intrinsic mode functions (IMFs) and drop them, the experiments showed good results on Gaussian noise at 0 db. S. Abd El-Moniem, *et al.* [10] proposed the use of EMD and SS as a pre-processing stage to enhance the noisy speech, SS was used to estimate and suppress noise spectrum on each IMF before reconstructing the input signal which would be enhanced. S.M. Govidan *et al* [4] used Adaptive bionic wavelet shrinkage (ABWS) which is a speech enhancement technique that's used to suppress the additive noise and increase the accuracy of the speaker recognition system, a double threshold is computed and applied based on estimated noise on each sub-band decomposed by adaptive bionic wavelet coefficients, a good results was reported in variety of noise types and levels. Y. Xu *et al* [11] obtained clean speech signal from noisy one with deep neural network (DNN) by calculating log-power spectra of noisy speech signal then mapping noisy to clean data using a well-trained DNN, the mapping function was trained with DNN over 104 noise types with 2500 hour of training.

Another approach is to extract a noise robust features that achieve a high identification rate without suppressing noise. H. Hermansky and N.Morgan [12] used Relative spectral perceptual linear prediction (RASTA-PLP) was built on the assumption that the human auditory system is sensitive for stimulus that are slowly varying, and the performance can be improved by eliminating the very slowly changing components comparing to the speech signal change. RASTA filtering ensuring that the output signal is much less to the stimuli that varying very slowly. Kim and Stern [13] proposed new features called power normalized cepstral coefficients (PNCC) in which the power nonlinearity was used instead of log nonlinearity in MFCC features and power-bias subtraction technique to suppress the additive noise. Wang, *et al*. [14] propose to use wavelet octave coefficients residues (WOCOR) that complements MFCC features information, the results state that this technique enhanced the system accuracy in mismatched spoken contents. Zhao et al [15] introduced new method for extracting features called gammatone frequency cepstral coefficients (GFCC), the work was based on the human auditory peripheral model where gammatone filter bank was used as a replacement of Mel-frequency filter bank which made it performs better than MFCC features. Mean Hilbert envelop coefficients (MHEC) proposed by Sadjadi and Hansen [16] to extract features by using smoothed Hilbert envelop of gammatone filter bank, the results showed that MHEC features are less prone to noise than MFCC features. Satyanand Singh and Pragya Singh [17] proposed to extract speaker specific features based on statistical modeling techniques of the speaker, the authors used TIMIT dataset with 1000 utterances and the results showed that using GMM gives the best recognition accuracy of 99.1%. Kobra *et al* [18] proposed to use mean and variance normalization and then applying auto-regression moving-average filter (MVA) to MFCC features, the new features give 28% accuracy improvement comparing with MFCC features at 5db SNR level.

To achieve a high accuracy in speaker identification by exclude the problem of additive noise, a proposed algorithm for feature extraction based on speech enhancement and roust combined features is used. The speech enhancement is based on wavelet thresholding as a pre-processing stage to remove the noise from the input speech signal first, after that, two cepstral features (PNCC and GFCC) are extracted from the estimated clean speech signal, feature warping is applied to the extracted features, and finally, a concatenation of the resulted features was applied to produce the final proposed robust features.

The rest of the paper is organized as follow. Section (2) describes the proposed feature extraction algorithm. Section (3) presents the experimental methodology. Section (4) presents simulation results and discussion. Finally, conclusion is in section (5).

## 2. PROPOSED FEATURE EXTRACTION ALGORITHM

Figure 1 shows the proposed feature extraction algorithm, where, the input speech signal is denoised by implementing discrete wavelet transform (DWT) semisoft thresholding, then extracting PNCC and GFCC features followed by applying feature warping technique, and finally concatenating them. The extraction algorithm steps are describe in the proceeding subsections.

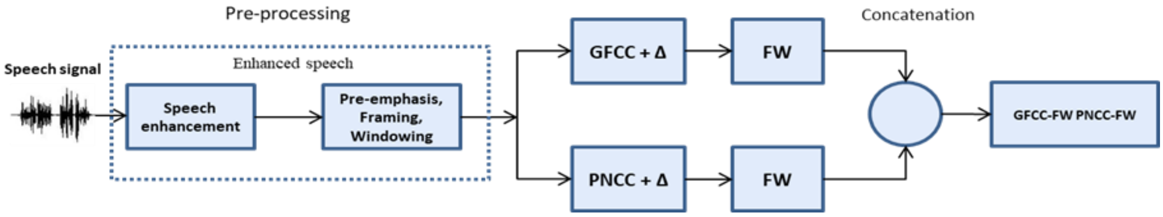

Figure 1. Block diagram of the proposed feature extraction algorithm

### 2.1. Speech enhancement

Wavelet transform used to analyze speech signals and DWT is a type of wavelet transform where the speech signal is decomposed to detail coefficients (CD) and approximation coefficients (CA) at several frequency subband levels with a finite impulse response (FIR) filter [19] as shown in Figure 2. The CA produced by convolving the speech signal with low-pass filter and the CD produced by convolving the speech signal with high-pass filter. Each decomposition level is done by applying DWT to the approximation coefficients [20].

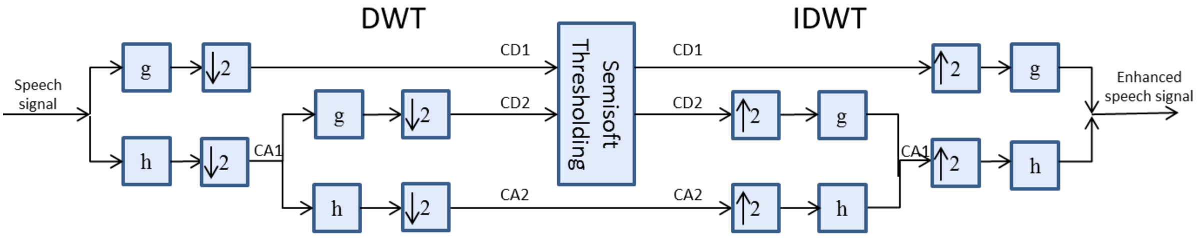

Figure 2. Block diagram of speech enhancement using DWT thresholding
(h is low-pass filter, g is a high-pass filter, ↓ 2 is down sampling that discarding half of signal data, and ↑ 2 is up sampling that doubles signal data)

After the speech signal decomposition, adaptive thresholding is applied to each resulted sub-band except for last approximation sub-band. Semisoft thresholding function is given by [21]:

$$D(Y,\lambda_1,\lambda_2) = \begin{cases} 0 & ,\lambda_1 < |Y| \leq \lambda_2 \\ sgn(Y)\dfrac{\lambda_2(|Y|-\lambda_1)}{\lambda_2-\lambda_1} & ,|Y| \leq \ \lambda_1 \\ Y & ,|Y| > \lambda_2 \end{cases} \tag{1}$$

where: $D(Y,\lambda_1,\lambda_2)$ is the output value after thresholding, $Y$ is the DWT subband frame, $\lambda_1,\lambda_2$ is the upper and lower thresholds respectively.

The thresholding value $\lambda_1$ is very important to the denoising performance, if it’s too low, the noise won’t be removed, and if it’s too high, part of the speech signal will be lost [22], Donoho [23] suggested the following estimation for the $\lambda_1$ threshold value:

$$\lambda_1 = \ \sigma_k\sqrt{2\log N_k} \tag{2}$$

where: $N_k$ is the signal length at subband level k, $\sigma_k$ is the standard deviation at subband level $k$ and given by [23]:

$$\sigma_k = \frac{Median\ (|CD_k|)}{0.6745} \tag{3}$$

where: $Median\ (|CD_k|)$ is the median absolute deviation detail coefficients at sub-band level $k$. $\lambda_2$ is calculated as [20]:

$$\lambda_2 = \sqrt{2}\lambda_1 \tag{4}$$

To recover the enhanced speech signal, Inverse DWT (IDWT) is applied; the de-noising procedure is repeated for each frame.

## 2.2. Robust features extraction

### 2.2.1. Pre-processing

After speech enhancement stage, the enhanced speech is used to extract the proposed features, the second step of the pre-processing is the pre-emphasis filter that is applied first to the speech signal to intensify high frequencies [24], Pre-emphasis is applied to PNCC features only as in [25] but not applied to GFCC features because it leads to performance dropping [26]. Framing is the third step in pre-processing stage where the enhanced speech signal is to be cut into short overlapping frames of 20-30 ms to overcome the discontinuity problem of the speech signal that may lead to wrong extracted features and performance dropping [27]. The fourth step is to apply a hamming window to every frame to increase signal continuity of the start and end of the frame [28].

### 2.2.2. Power normalized cepstral coefficients (PNCC)

PNCC features are powerful features that outperform conventional features in noisy and clean environments [25]. The high identification accuracy is resulted by the using of power-law nonlinearity that gives a close approximation of human auditory system [29]. Figure 3 illustrate the processing stages block diagram of PNCC features as described in [13].

After the pre-processing stage, the cepstral features are extracted from frequency domain by STFT. Frame power is calculated, then, gammatone filter bank is used with Equivalent Rectangular Bandwidth (ERB). To suppress the channel noise, Asymmetric Noise Suppression (ANS), temporal masking and weight smoothing are used. Power function nonlinearity is used because the output behavior does not critically rely on the amplitude of the input [30]. Discrete Cosine Transform (DCT) is applied then to de-correlate the highly correlated spectral features [31]. Finally, implement cepstral mean normalization to produce the normalized cepstral vector to remove channel distortion and improve recognition rate in noisy environments [32].

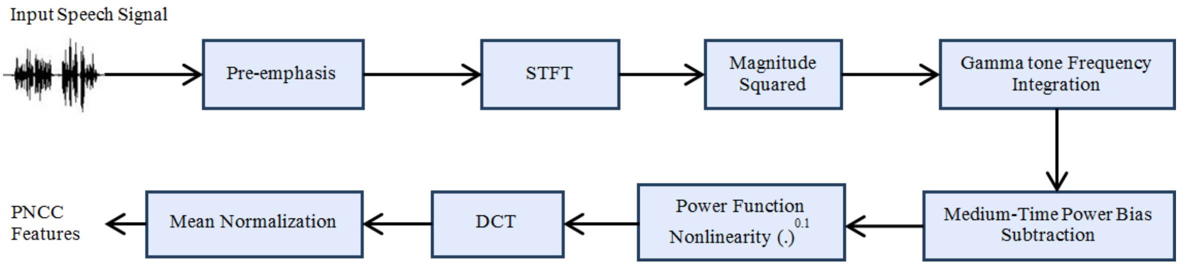
Input Speech Signal
Pre-emphasis
STFT
Magnitude Squared
Gamma tone Frequency Integration
Medium-Time Power Bias Subtraction
Power Function Nonlinearity (.)0.1
DCT
Mean Normalization
PNCC Features

Figure 3. Block diagram of PNCC algorithm

### 2.2.3. Gammatone frequency cepstral coefficients (GFCC)

The gammatone filter bank is series of overlapping band-pass filters that models the human auditory system [33]. The combination of gammatone filter bank (GF), cubic root and equivalent rectangular bandwidth (ERB) gives the robustness of GFCC features in noisy environments [34]. The block diagram of GFCC features processing stages is depicted in Figure 4 as described in [15].

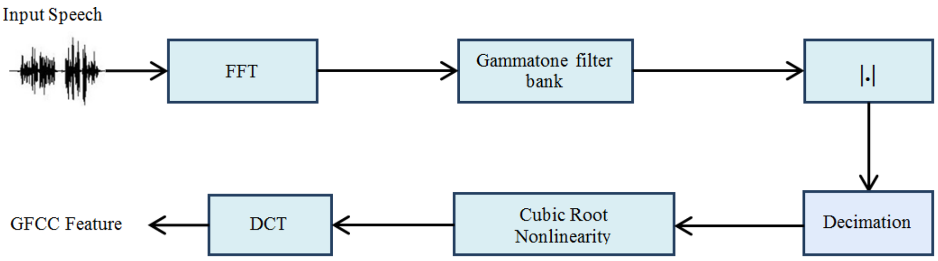

Figure 4. Block diagram of GFCC feature extraction algorithm

Preprocessed speech signal passed through 64-channel gamma tone filter bank whose center frequencies ranging from 50 – 8000 Hz, then, 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. The absolute value is calculated to create T-F representation that is a variant of cochlea-gram. After that, implement cubic root for the decimated outputs magnitudes. Finally, apply DCT to de-correlate the cepstral coefficients and reduce dimensionality [15].

### 2.2.4. Feature warping (FW)

Feature warping is letting the cepstral features following a distribution target to increase the robustness of the resulted features. FW processing steps can be summarized as following [35]:

a. Select a target distribution.
b. Extract cepstral coefficients.
c. Create a lookup table to map the rank of sorted features to target warped features.
d. Isolate a window of $N$ features (typically 3 seconds) and sort the values in descending order then give a rank of 1 for the maximum value and rank $N$ for the minimum value, to be used as an index in the lookup table created in step c.
e. Move the sliding window by 1 frame.
f. Steps (d) and (e) are repeated for each frame shift.

The lookup table can be created by calculating the value of $m$ by numerical integration method for each $R$ value by initially making $N = R$ [35]:

$$\frac{N+\frac{1}{2}-R}{N} = \int_{z=-\infty}^{m} h(z)dz \tag{5}$$

If a normal distribution is chosen, then:

$$h(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{6}$$

where: $m$ is the feature warped component, $N$ is the window length and $R$ is the rank.

## 3. EXPERIMENTAL METHODOLOGY

Experiments are done on TIMIT [36] dataset, which consists of 630 speakers, each speaker has 10 utterances. To train the UBM-GMM classifier, 530 speakers are chosen randomly (i.e. 5300 utterances) to train UBM and 100 speakers are left for testing. The GMM is trained with 9 utterances from each speaker and the last utterance is left for testing. To test the robustness of the proposed algorithm presented here, 4 noise types are chosen from the Noisex-92 [37] noise dataset which are artificially added to the test utterances with a signal to noise ratio levels 0,5,10 and 15 db. For speech enhancement stage, all utterances are framed into non-overlapping frames with 16 ms. length and decomposed with 4 levels using DWT wavelet decomposition, scaling function Daubechies 8 techniques and pruned using semisoft thresholding, and then reconstruct each frame and recombine them to produce the enhanced speech signal. The feature extraction stage includes framing the speech signal into an overlapping 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 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 0th coefficient from each frame, then applying feature warping with window length of 301 frames (3 sec) 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.

## 4. SIMULATION RESULTS AND DISCUSSION

In this section, the proposed features robustness tested with both clean and noisy environments, then compared with similar studies.

### 4.1. Speech enhancement technique analysis

To select the best parameters settings for the speech enhancement pre-processing stage, number of factors are selected and used in the test, such as, frame length, number of decomposition levels, filter function, and number of filters, for their effect on the average identification accuracy. The average results are listed in Table 1. The average identification accuracy results that are shown in Table 1, indicates that 4 levels of DWT decomposition with db8 and frame length of 16 ms gives the best identification accuracy.

Table 1. Average identification accuracy on different parameters settings for speech enhancement pre-processing stage

| Effect of number of DWT decomposition levels on the average identification accuracy | | | | Effect of choosing the filter function on the average identification accuracy | | |
|---|---|---|---|---|---|---|
| 2 levels | 3 levels | 4 levels | 5 levels | DB. | Symlet | Coiflet |
| 83.86 | 85.01 | **85.24** | 85.21 | **85.24** | 84.31 | 85.17 |
| Effect of frame size on the average identification accuracy | | | | Effect of the number of filters on the average identification accuracy | | |
| 16 ms | 25 ms | 32 ms | No Framing | 5 | 8 | 13 |
| **85.24** | 84.73 | 83.49 | 82.22 | 83.00 | **85.24** | 84.96 |

### 4.2. Comparison between baseline and the proposed features

Table 2 shows a comparison between baseline and proposed features with and without DWT speech enhancement technique and its effect on the identification accuracy rate. The results obtained in Table 2 shows that DWT with semisoft thresholding and the proposed features give a noticeable improvement in identification rate except for the clean speech signal where PNCC features gives the top identification rate.

Table 2. Comparison between baseline and the proposed features

| Noise Type | Noise Level | PNCC | GFCC | PNCC-FW-GFCC-FW | DWT(PNCC-FW-GFCC-FW) |
|---|---|---|---|---|---|
| Clean | | 99 | 96 | 98 | 98 |
| Babble | 0db | 69 | 56 | 74 | *80* |
| | 5db | 86 | 77 | 88 | *92* |
| | 10db | 96 | 90 | 92 | *98* |
| | 15db | 95 | 94 | 95 | *97* |
| Factory 1 | 0db | 49 | 46 | 64 | *70* |
| | 5db | 75 | 75 | 78 | *81* |
| | 10db | 88 | 86 | 88 | *94* |
| | 15db | 92 | 89 | 93 | *97* |
| Pink | 0db | 39 | 13 | 46 | *50* |
| | 5db | 52 | 29 | 72 | *78* |
| | 10db | 70 | 62 | 85 | *91* |
| | 15db | 82 | 81 | 91 | *95* |
| White | 0db | 52 | 25 | 54 | *62* |
| | 5db | 67 | 55 | 72 | *78* |
| | 10db | 75 | 79 | 89 | *92* |
| | 15db | 89 | 88 | 93 | *96* |
| Average | | 75 | 67.12 | 80.71 | 85.24 |

### 4.3. Comparison with similar studies

The proposed feature extraction algorithm is compared with similar studies to show the effectiveness of the algorithm. Table 3 describes briefly the systems of the studies used in the comparison. Figure 5 shows the comparison results with other studies and the proposed feature extraction algorithm outperforms the other studies results. The same parameters used in the comparison, which are frame length, frame shift, the number of gaussian mixtures, and noise types and SNR levels. The comparison results shows that the proposed feature extraction algorithm outperforms all the compared studies with a grate identification accuracy.

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

| Work | Proposed features | Speaker's dataset | No. of testing speakers | No. of features | Frame length | Frame shift | Pre-emphasis | Evaluation system | No. of mixtures |
|---|---|---|---|---|---|---|---|---|---|
| Proposed | Proposed | TIMIT | 100 | 42 (21 + 21Δ) | 25 mSec | 10 mSec | 0.97 for PNCC features | UBM-GMM | 256 |
| [1] | Neurogram | TIMIT | 100 | 25 | Not mentioned | Not mentioned | No | UBM-GMM | 128 |
| [18] | Combining MFCC and MVA | TIMIT | 100 | 20 | 15 mSec | 10 mSec | 0.97 | GMM | 64 |
| [38] | 1. PNCC+SGRs<br>2. LPCC+SGRs | TIMIT | 630 | 1. 60 for PNCC+SGRs (20 + 20Δ + 20 ΔΔ)<br>2. 24 for LPCC+SGRs | Not mentioned | Not mentioned | No | UBM-GMM | 128 |
| [39] | Auto-regressive with MFCC (AR-MFCC) | TIMIT | 200 | 64 (32 MFCC and 32 AR) | 20 mSec | 10 mSec | 0.97 | GMM | 64 |

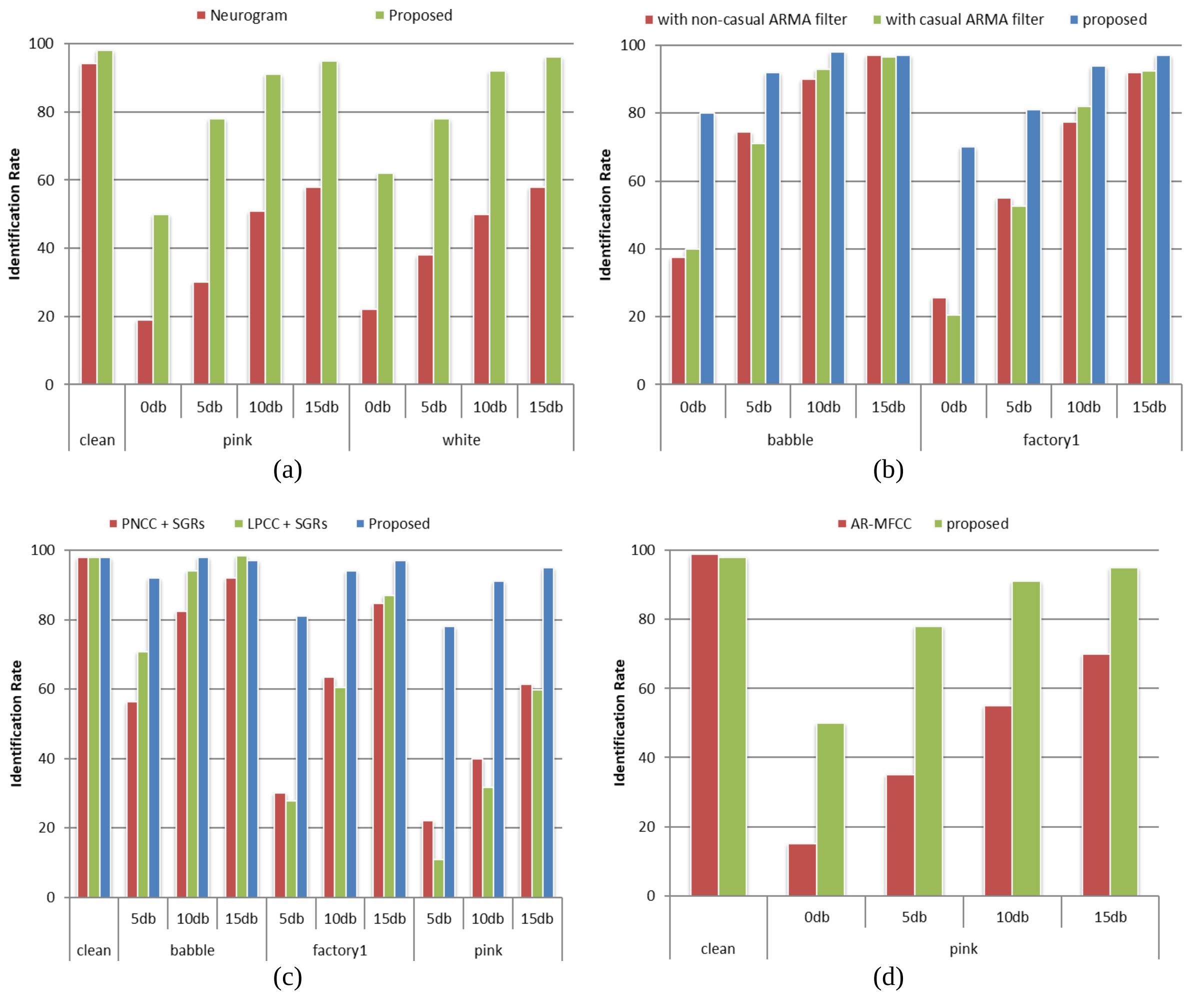

Figure 5. Comparison with other studies: (a) with work proposed by [1], (b) with work proposed by [18], (c) with work proposed by [38], and (d) with work proposed by [39]

## 5. CONCLUSION

In this work, new feature extraction algorithm is presented, it consist of two stages, first stage is speech enhancement with DWT semisoft thresholding. The second stage is concatinate two extracted features named power normalized cepstral coefficients (PNCC) with feature warping (FW) and gammatone frequency cepstral coefficients (GFCC) with FW that are studied for robust speaker identification system over noisy channel. UBM-GMM is used as feature. Experiments are done on TIMIT dataset where 100 speakers are used for test. The testing is done on clean and noisy conditions to test the robustness of the proposed feature extraction algorithm, 4 noise types are chosen from the Noisex-92 noise dataset (babble, factory 1, pink and white) that are added to the test utterances with SNR levels 0, 5, 10 and 15 db. The results showed that the proposed features outperforms baseline features (PNCC and GFCC) and other proposed works Islam et al., in 2016, Korba et al., in 2018, Guo et al., in 2017 and Ajgou et al. in 2016, so it's a promising approach for extracting robust features and increasing speaker identification rate.

ACKNOWLEDGEMENTS

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad Iraq for its support in the present work.

## REFERENCES

[1] M. A. Islam, W. A. Jassim, N. S. Cheok, M. S. A. Zilany, "A robust speaker identification system using the responses from a model of the auditory periphery," *PLoS ONE*, vol. 11, no. 7, pp.1-21, 2016.
[2] C. Hsieh, E. Lai and Y. Wang, "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, 2003.
[3] E. Variani, *et al*., "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, pp. 4080-4084, 2014.
[4] S. M. Govindan, P. Duraisamy, and X. Yuan, "Adaptive wavelet shrinkage for noise robust speaker recognition," *Digital Signal Processing*, vol. 33, pp. 180-190, 2014.
[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans Audio, Speech, Signal Process*., vol. 27, no. 2, pp. 113–120, 1979.
[6] L. Gao, Y. Guo, S. Li, F. Chen, "Speech Enhancement Algorithm Based on Improved Spectral Subtraction," in *IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2009, pp.140-143.
[7] Z. Brajevic and A. Petosic, "Signal denoising using STFT with Bayes prediction and Ephraim–Malah estimation," in *Proceedings of the 54th International Symposium ELMAR*, Zadar, Croatia, pp.183-186, 2012.
[8] M. A. Abd El-Fattah, *et al*., "Speech enhancement with an adaptive Wiener filter," *Int. J. Speech Technol*. vol. 17, no. 1, pp. 53-64, 2013.
[9] A. Mert and A. Akan. "Detrended fluctuation thresholding for empirical mode decomposition based denoising," *Digital signal processing*, vol. 32, pp. 48-56, 2014.
[10] S. A. El-Moneim, *et al*., "Hybrid speech enhancement with empirical mode decomposition and spectral subtraction for efficient speaker identification," *Int J Speech Technol*, vol. 18, no. 4, pp. 555-564, 2015.
[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7-19, 2015.
[12] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.
[13] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech*, 2009, pp. 28-31.
[14] N. Wang, P. C. Ching, N. Zheng, and T. Lee, "Robust speaker recognition using denoised vocal source and vocal tract features*," IEEE Trans. Audio Speech Lang. Process*., vol. 19, no. 1, pp. 196-205, 2011.
[15] X. Zhao, Y. Shao, and D. L. Wang, "CASA-Based Robust Speaker Identification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1608-1616, 2012.
[16] S. O. Sadjadi and J. H. L. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification," *Speech Commun.*, vol. 72, pp. 138-148, 2015.
[17] S. Singh and P. Singh, "High level speaker specific features modeling in automatic speaker recognition system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 2, pp. 1859–1867, 2020.
[18] M. C. A. Korba, H. Bourouba, and D. Rafik, "Text-Independent Speaker Identification by Combining MFCC and MVA Features," *2018 International Conference on Signal, Image, Vision and their Applications (SIVA)*, Guelma, Algeria, pp. 1-5, 2018.
[19] R. Aggarwal, et al, "Noise Reduction of Speech Signal using Wavelet Transform with Modified Universal Threshold," *International Journal of Computer Applications*, vol. 20, no. 5, pp. 14-19, 2011.
[20] F. S. Hassen, "Performance of Discrete Wavelet Transform (DWT) Based Speech Denoising in Impulsive and Gaussian Noise," *Journal of Engineering and Development*, vol. 10, no. 2, pp. 175-193, 2006.
[21] X. Liu, "A New Wavelet Threshold Denoising Algorithm in Speech Recognition," in *Asia-Pacific Conference on Information Processing*, vol. 2, pp. 310-303, 2009.

[22] J. N. Gowdy and Z. Tufekci. “Mel-scaled discrete wavelet coefficients for speech recognition,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2000, pp. 1351-1354.

[23] D. L. Donoho, “Denoising by soft thresholding,” *IEEE Trans. on Information Theory*, vol. 41, no. 3, pp. 613-627, 1995.

[24] R. Togneri and D. Pullella. “An overview of speaker identification: Accuracy and robustness issues,” *IEEE circuits and systems magazine*, vol. 11, no. 2, pp. 23-61, 2011.

[25] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315-1329, July 2016.

[26] 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, pp. 7204-7208, 2013.

[27] J. C. Liu, *et al*., “An MFCC-based text-independent speaker identification system for access control,” *Concurrency and Computation: Practice and Experience*, vol. 30, no. 2, e4255, 2018.

[28] N. M. AboElenein, K. M. Amin, M. Ibrahim and M. M. Hadhoud, “Improved text-independent speaker identification system for real time applications,” *2016 Fourth International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC)*, Cairo, pp. 58-62, 2016.

[29] P. K. Nayana, D. Mathew, and A. Thomas, “Comparison of text independent speaker identification systems using GMM and i-vector methods,” *Procedia computer science*, vol. 115, pp. 47-54, 2017.

[30] C. Kim and R. M. Stern, “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1315-1329, 2016.

[31] Y. Shao and D. Wang, “Robust speaker identification using auditory features and computational auditory scene analysis,” *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, pp. 1589-1592, 2008.

[32] J. Droppo and A. Acero, “Environmental robustness,” in *Handbook of Speech Process*., J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York, NY, USA: Springer, Ch. 33, 2008.

[33] M. Jeevan, *et al*., “Robust speaker verification using GFCC based i-Vectors,” *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*, Springer, New Delhi, pp. 85-91, 2017.

[34] M. Jayanth and B. Roja Reddy, “Speaker Identification based on GFCC using GMM-UBM,” *International Journal of Engineering Science Invention*, vol. 5, no. 5, pp. 62-65, 2016.

[35] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. Speaker Odyssey*, pp. 213-218, 2001.

[36] *TIMIT dataset*, accessed at 28 May 2019. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1

[37] *NOISEX-92 noise dataset*, accessed at 28 May 2019. [Online]. Available: http://spib.linse.ufsc.br/noise.html

[38] J. Guo, R. Yang, H. Arsikere, and A. Alwan, “Robust speaker identification via fusion of subglottal resonances and cepstral features,” *The Journal of the Acoustical Society of America*, no. 4, pp. 420-426, 2017.

[39] R. Ajgou, *et al*., “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, 2016.