Received: 1 December 2023 | *Revised: 12 March 2024* | *Accepted: 27 March 2024*
DOI: 10.37917/ijeee.22.1.8 *Vol. 22 | Issue 1 | June 2026*

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Open Access
IJEEE University of Basrah College of Engineering

Speaker Verification Based on Mel Frequency Cepestral Coefficients and Correlation

Abdalem A. Rasheed
Dept. of Electrical Eng., College of Engineering, University of Mosul, Iraq

Correspondance
Abdalem A. Rasheed
University of Mosul, Mosul, Iraq
Email: alem12@uomosul.edu.iq

Abstract
Speaker recognition refers to identifying the speaker by his or her voice. People talk in a variety of tones and each speaking voice has features that distinguish one person from another. Speaker verification (SV)involves comparing a set of measures of the speaker's utterances with a reference for the person whose identification is being asserted to accept or reject the speaker's identity claim. An identity claim is made during speaker verification which consists of two steps: extraction of feature and matching of feature. In this work, the analysis of correlations of Mel-scale coefficients for the voice of utterance to identify the intended speaker is presented. Short text-dependent word and other text-independent word is represented in this study. The correlation accuracy ranged from 98% to 99% for user1 (same speaker) for text-dependent. whereas 83% and 61% for user1 correlation with other speakers for text-dependent and independent respectively. Furthermore, the MFCC feature extraction approach based on distributed Discrete Cosine Transform (DCT) is provided in this research. SV tests are carried out using the MFCC feature extractions method where close variance for the target speaker and away variance for other speakers is obtained. Additionally, the principle component analysis (PCA) is provided to improve the discriminative system performance. Where the PCA chooses the optimal path between every pair of extremely confusing speakers. The results obtained from PCA were similar to the correlation finding from the Mel-scale results with enhancing the discriminative information and with lowering the dimension of MFCCs data..

Keywords
Speaker Recognition, Speaker Verification, MFCC, Correlation, FFT, PCA.

I. Introduction

The acoustic patterns and traits of the human voice vary from person to person based on speaking status, accents, and vocal tract anatomical anatomy. One can recognize the speaker by using these distinguishing characteristics. The technique of automatically extracting, classifying, and recognizing a speaker's identity from the data contained in their speech signal is known as speaker recognition. Users can confirm their identification using automatic voice recognition systems, which have potential usefulness in a variety of applications, particularly in security and identity management. For instance, speech biometrics and voice activated commands are two common security features used by many modern devices for remote authentication and access management. In criminal investigations and law enforcement, speaker recognition might be helpful (for example, by comparing the speech of an attacker to a database of the suspects to discover the closest match). Speaker recognition goes via numerous processes, including data collecting, data pre-processing, feature extraction, and feature matching, like most pattern recognition systems. There are various studies describing the use of the speaker recognition [1–7].

For audio and voice signals, the Mel coefficients model and Mel-frequency cepestral coefficients (MFCC) method is probably the most widely utilized feature extraction technique to date [8]. According to published research, the majority

of cutting-edge speaker recognition systems use MFCC as a feature extractor and feed their extracted features into GMM-based methods for building speaker models [9]. However, the strategies the researchers first presented are those that can handle a lot of speech during training and testing [10]. As a result, speaker recognition is a tough field when taking into account short speech utterances. In the proposed work, the Mel scale coefficients correlation and MFCC extraction from brief speech utterances are investigated. The experimental findings demonstrate the comparison for both text-dependent and independent examples for speaker recognition where the results offer satisfactory findings.

The performance of MFCC degrades by the speech datasets and in noise environments [11]. For instance [12] conclude that denoising the input signal can improve the result more and when the highest MFCCs is adopted. In this work, the samples were taken in the clean room (Lab) to avoid the noise problem and concentrate on the Mel-coefficients correlation. Where the noise can be degraded by FIR filter, low pass filter, or by adding pre-emphasize unit at the input signal.

In this work, the SV aims to achieve a high correlation between training data sets and strong variance between the MFCC of the targeted speaker and those of other speakers. Whereas, for speaker identification (SI) a high variance is required between the intended speakers. Consequently, can be ensured that the targeted speaker model will be able to effectively take use of discriminations. In order for the vision to be clearer about the MFCCs variance and to discover robust characteristics features for modeling of speaker, the PCA is applied on the MFCCs features where the PCA reduces the huge data of training samples, reduce the time of analysis which represents the important factor for recognition, and then reduce the required memory space and cost of computations.

The brief text-dependent words like ”GUD” are far away from high frequency words like ”S” and ”sh” noise letters and repetitive letters like ”o”. Where the concentration of Mel-Scale is in low frequency to simulate the real human hearing. Throughout the utterance duration, several windows are not necessary for the text-dependent short duration. In addition, there is low data size, low memory needed, and fast processing speed. As a result, the intended speaker's robust model for speaker verification was obtained.

The paper is organized as follow: The SV and identification are described in section II. . The speaker recognition system is presented in section III. . The MFCC feature extraction is presented in section IV. . Experimental results are given in section V. . Section VI. shows the Mel-scale coefficients correlation, while section VII. presents the MFCC variance analysis and the PCA concept for reducing non-essential features. The comparison with other works is offered in section VIII. .

II. Speaker Verification and Identification

A speaker's identity is verified by a speaker recognition system based on the speaker's voice, which is one of the most practical biometric characteristics for human machine communication. Speaker recognition extends into two categories: SI and SV [13], [14]. SV aims to confirm whether an input speech conforms to a claimed identity, whereas SI seeks to identify an input speech by choosing one model from a group of enrolled speaker models. Three crucial components make up the SI and SV system: feature extraction, speaker modeling, and matching. The purpose of the feature extraction is to identify speakers by extracting key elements from input speech. The speaker modeling process involves statistically simulating the characteristics of the speakers who have signed up. The matching is done to match different speaker model input features [15].

III. Speaker Recognition System

The voice recognition system, as shown in Fig. 1. [9], contains from: sampling input signal, windowing, FFT, Mel-scale filters, Take log, Discrete Cosine Transform (DCT), and finally MFCC extraction.

Data acquisition is the initial step, which comprises recorded voice signals. Threshold is used to distinguish speech from background noise. Only high powered frames, which include speech, are left after the signal above the computed threshold is maintained and the remaining signals are deleted. Also, Correct threshold calculation is necessary for effective end point detection.

A. Windowing

Since the speech period is very short, not exceeding 200ms, and the endpoint is detected, a single rectangle type of window was used, Eq.(1):

$$\mathrm{h[n]} = \begin{cases} 1 \\ \text{otherwise} \end{cases} \tag{1}$$

The output of rectangle window is

$$Y(n) = h(n) \cdot X(n) \tag{2}$$

X(n) is the magnitude of the sample, ($0 \leq n < N$).

Where n is the number of samples and N is the total number of samples per frame.

B. FFT transform

After that, the Fast Fourier Transform of Eq.(1) is applied to the sampled speech signal. This equation converts to frequency domain operation.

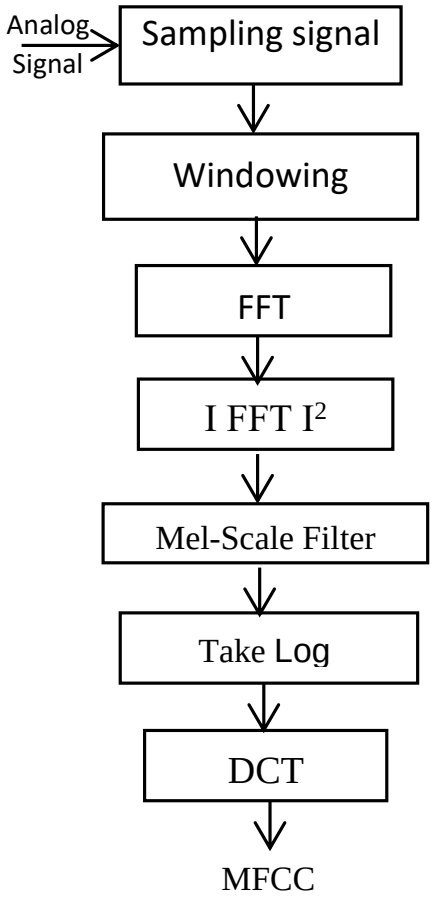

Fig. 1. Voice recognition system [9]

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \quad 0 \le k < N-1 \tag{3}$$

In general X(k)'s are complex numbers and we consider only their absolute values.

C. Mel Scale Coefficients

To make the features extracted more closely resemble what human hears, the Mel scale is employed to connect perceived frequency (pitch) to the actual measured frequency. The formula for forward changing a frequency measurement into a Mel sale is:To make the features extracted more closely resemble what human hears, the Mel scale is employed to connect perceived frequency (pitch) to the actual measured frequency. The formula for forward changing a frequency measurement into a Mel sale is [16]:

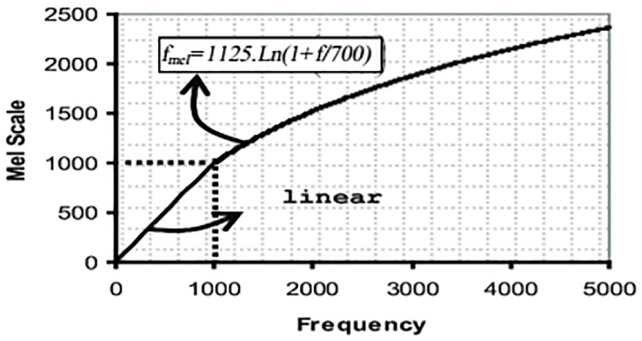

Fig. 2. Mel scale versus frequency [16]

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{4}$$

But for reverse changing Mel scale frequency is :

$$f_{mel} = 700 \,.\, (e^{M/1125} - 1) \tag{5}$$

Fig.2. shows the Mel scale versus frequency. [16].

The frequency range from 0Hz to 8KHz that contains the majority of characteristics of human speech is considered. In this work, the 10 Mel-coefficients are utilized because is the most common number. To present the spectrum in Mel-scale format, it is first essential to compute a comb of filters. A triangular window known as the Mel filter adds the energy within its frequency range and determines the Mel coefficients. A set of ten filters is created as shown in Fig, 3. The number of windows is greatest in the low frequency region where this region of the frequency spectrum is most interesting which result in good resolution. This has the potential to greatly enhance recognition quality. The signal spectrum is multiplied by the triangular windows function to determine the signal energy, from which the coefficients vector are derived.

The frequency range is from 0Hz to 8KHz. This range corresponds to a Mel-scale value of 0 to 2835. To create 10

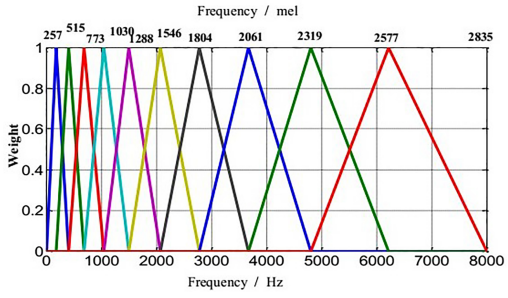

Fig. 3. Mel scale Filter Bank

triangular filters (scaled in the Mel-scale M[i] and in Hertz H[i]), eleven points (number of coefficients +1) of control are constructed as shown in the matrix (6) and (7). Where the number of ten Mel-scale coefficients produces high accuracy speech recognition system [17].(i) is a number of coefficients:

$$M[i] = [0;257;515;773;1030;1288;1546;1804;2061; 2319;2577;2835] \quad (6)$$

$$H[i] = [0;180;406;692;1048;1500;2066;2780;3672; 4800;6217;8000] \quad (7)$$

. By applying the created Mel-filters, Eq.8, the filters are applied to the spectrum's energy.

$$W(f) = \begin{cases} \frac{f-f_{cm-1}}{f_{cm}-f_{cm-1}} & \text{for } f_{cm-1} \leq f < f_{cm} \\ 1 - \frac{f-f_{cm}}{f_{cm+1}-f_{cm}} & \text{for } f_{cm} \leq f < f_{cm+1} \\ 0 & \text{else} \end{cases} \quad (8)$$

Where fcm-1 and fcm+1 are the positive and negative slope region of the triangle coefficient respectively and fcm is the center frequency of the coefficient. As a result, the findings are condensed, the contribution of the first frequencies is enhanced, and the contribution of the remaining frequencies is reduced.

IV. MFCC Feature Extraction

The energy of the spectrum undergoes to Mel-filters, and the resulting values are then taken as the logarithm and apply the discrete cosine transformation DCT) to generate the MFCC coefficients, Eq.9 [18].

$$CepstralCoefficients = DCT(log(abs(FFT(window)))) \quad (9)$$

Using a discrete cosine transform, the signal is represented as MFCC (Mel-Fequency Cepestral Coefficients). MFCC, which indicates the energy of the signal spectrum, is typically a vector of ten real values. This approach considers the wave nature of the signal.The mel-scale assigns the frequencies that the individual perceives as being most important, and the number of MFCC coefficients may be adjusted to any number, allowing for frame compression and a reduction in the amount of information processed [19], [20]. For instance, if the word is uttered differently by two persons or at a different tempo, the set of MFCC coefficients for that word varies.

V. Experimental Results

The first step is data acquisition, which consists of a band-limited record voice signal between 0Hz and 8KHz, where each recording file is formatted as .WAV type. The sample frequency that is used is 44.1KHz. The filtered signal was then properly denoised by noise cancellation. The down sampling is used to denoise the input signal.

The aim was to gather data on two recognition words: (Gud and Go). The findings are collected by the Matlab program with the help of the Excel program during an experimental examination technique for detecting voice signals. Ten different persons said the same word ten times. These are the words: GUD and GO. The effective technique of the standard of correlation principle is used in this research to identify patterns. The most crucial component for pattern recognition is the database. The goal of this study was to use a database to create an easy way to test your system. This approach can be extended to include a bigger database and more computing.

Voice recognition is a crucial biometrics task that may be split into two categories: text-dependent and text-independent. The user must use the same terms for training and testing sessions in text-dependent systems. The samples used for the training and test sessions in text-independent systems are different.

The power spectrum of the FFT in the text-dependent scenario for saying the word "GUD" three times for the same user is shown in Fig. 4. It is clear that the training times for uttering are much the same. Where the clear power spectral properties of this term can be easily discriminated. Fig. 5., shows the power spectrum of uttering the word "GUD" by different users. It is obvious that the power spectrum for each speaker has unique criteria, and the power spectrum parameters have vanished.

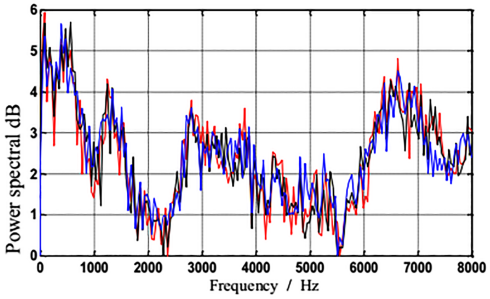

Fig. 4. Power spectrum of three times of saying "GUD" by the same user (user1)

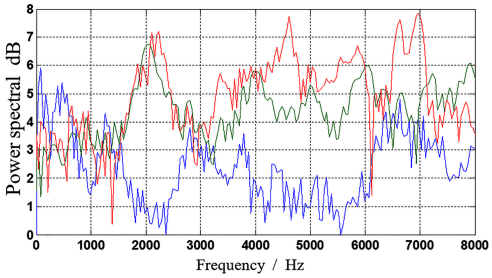

Fig. 5. Power spectrum of saying ”GUD” by different three user

VI. MEL-SCALE COEFFICIENTS CORRELATION

The database consists of ten training of male user1 (claimed person) for the dependent word ”GUD” with other male users uttering the same word. The voice is recorded in the normal classroom environment without external noise. Figure 6 depicts the average ”GUD” word detection for user1 Mel-scale coefficients correlation that ranges from 0.98 to 0.99. Therefore, this indicates a high coincidence between the extracted Mel-scale features and the received voice of the user1. On the other side, Fig. 6 shows the correlation of usear1 with other male users for utterance same word. The correlation is ranged from 0.82 to 0.78 which indicates that can be easily discriminate between the user1 and other users.

Fig. 7 shows the Mel-scale coefficients correlation between user1’s uttering the ”GUD” word with the same and different users uttering the ”GO” word which is a more near voice to the ”GUD”. The selection of the ”GO” word to obtain the worst case for discriminating. The outcomes will be improved if different-sounding terms are used that far from the ”GUD” sound. For the same user (user1) the correlation is 0.81, which shows a high level of user discrimination. User1 and other users who use distinct ”GO” words have a low correlation between them, with values between 0.58 and 0.63. That demonstrates the large difference and strong verification between user1 and other users.

VII. MFCC VARIANCE ANALYSIS

For the objectives of comparing MFCC coefficients between user1 for ten training trials and other users who uttered the same ”GUD” utterance sound, findings were established. The scatter plots in Fig. 8 compare the values of the ”GUD” utterance for ten trainings of user1 (x-axis) to the values of the same utterance for ten other users (y-axis) in the data set for each of the ten MFCC dimensions. The 10 MFCC coefficients are highly distributed on the y-axis and constrained in a small range along the x-axis. To provide a clear comparison, set the x-axis range to be the same as the y-axis range. Therefore, these figures demonstrate the applicability of the regression model and the ability to successfully predict values with success.

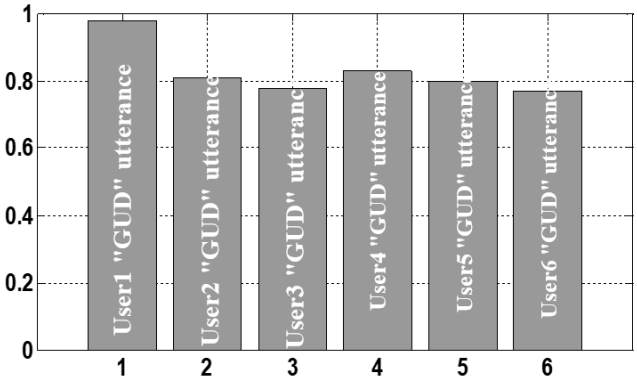

Fig. 6. Average of 30 training correlation of Mel-scale for user 1 saying ”GUD” and comparing it to five additional users saying the same phrase

Furthermore, a useful metric for this variance, which evaluates how widely apart a bunch of numbers are from their mean value. Low variance is required to identify the targeted speaker (Sp.1), whereas high variance is required for other users (Sp.s). Table I shows that the ten MFCC coefficients variance for the intended speaker is very low with high variance for other speakers. To enhance the clarity of the MFCC discrimination between the intended speaker and other speakers, the 10-dimensional MFCCs, for 30 training sets, are reduced to a two-dimensional feature space using the principle Component Analysis (PCA) algorithm. Through the use of PCA, dimensionality reduction can be achieved by determining a set of orthogonal axes, or principle compo-

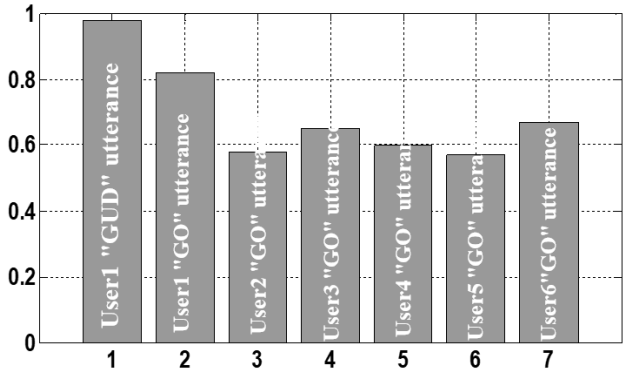

Fig. 7. Average of 30 training correlation of Mel-scale for user 1 saying ”GUD” and comparing it to user 1 with five additional users saying ”GO”

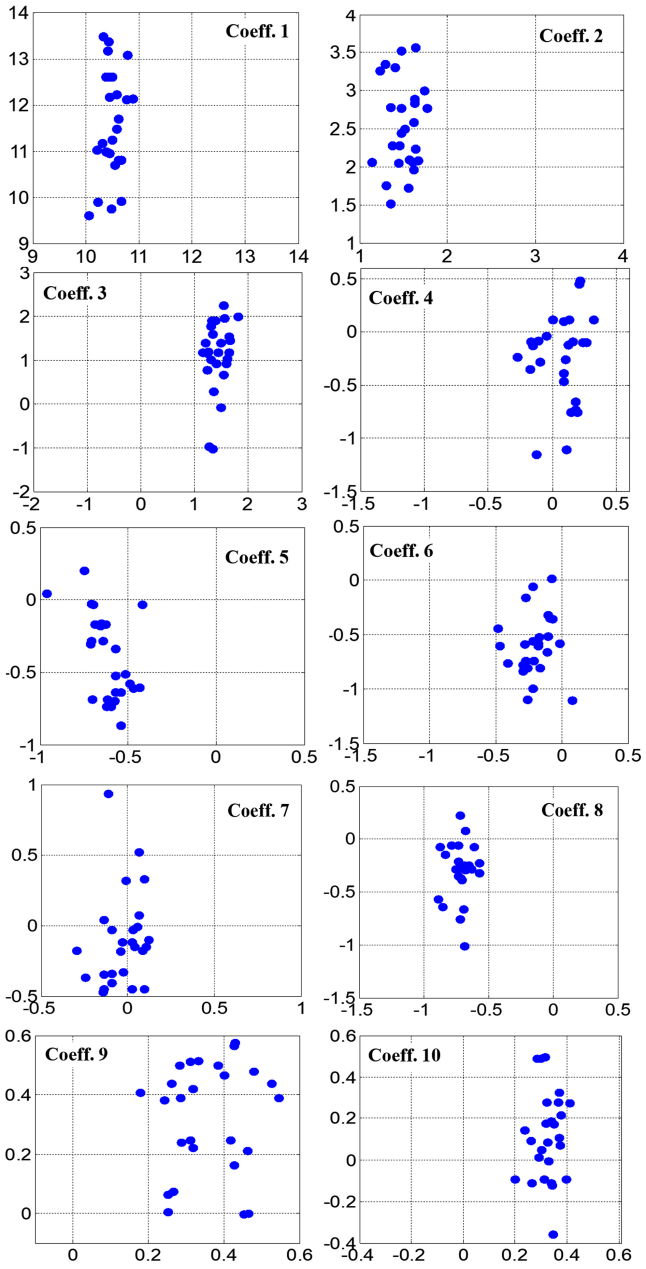

Fig. 8. MFCC scatter plots created by 30-training of user1 (x-axis) and the other 30- users (y-axis) for the utterance sound ”GUD”.

TABLE I.
DISPLAYS THE VARIANCE FOR EACH MFCC COEFFICIENT FOR INTENDED SPEAKER1 (SP1 TEN TRIALS) AND THE VARIANCE OF OTHER SPEAKERS (SPS TEN USERS)

Coeff. No.	Sp.1	Sp.s	Sp.s/Sp.1
Coeff. 1	0.034	1.307	38.4
Coeff. 2	0.025	0.334	13.4
Coeff. 3	0.028	0.662	46
Coeff. 4	0.025	0.171	23.6
Coeff. 5	0.013	0.087	6.7
Coeff. 6	0.016	0.08	5
Coeff. 7	0.012	0.11	9.2
Coeff. 8	0.0065	0.073	11.2
Coeff. 9	0.0096	0.034	3.5
Coeff. 10	0.0024	0.044	18.3

nent, that typically represent the majority of the variance in the data but on the contrary, where the correlation depends on how the coefficients converge. Fig. 9 demonstrates the two-dimensional PCA feature distribution of MFCCs of both the target speaker (utterance ”GUD” text) and other speakers (utterance ”GUD” and ”GO” text). The sum of the first and second eigenvalues (PC1 and PC2) is 54.7%, where PC1 has the most variation of the data (38.2%), but PC2 has the second most variation of the data (16.5%).

It is clear from Fig. 9 that the PCA characteristics are distributed more widely in the PC1 dimension than the PC2 dimension where the blue dots (target speaker) and the red dots) do not overlap, and the green dots (”GO” text uttered by the other speaker) are farther away than the blue dots and the main variance is in the PC1 direction. It is also important to keep in mind that not all the speakers have the most variation

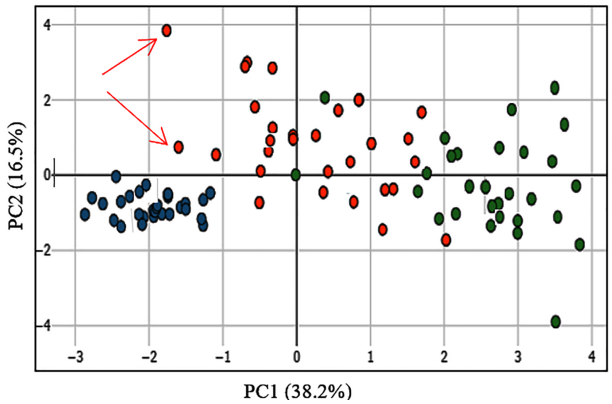

Fig. 9. PCA feature distribution of MFCCs of target speaker (blue dots) and other speakers red and green dots for ”GUD” text and ”GO” text respectively.

on the PC1 axis. For instance, the dots indicated by red arrows have the most variation on the PC2 (for target speaker blue dots) and nearly do not have a variation on the PC1 axis.

This suggests that there is a significant target speaker discrimination in the MCC feature space. That would therefore ensure that the target speaker model could effectively take of the discrimination by relying on the discriminatory information in the feature space.

VIII. Comparison With Published Works

The MFCC is one of the chosen characteristics, that can be useful in measuring and characterizing the performance of speaker recognition. It can also be used to assess various features in an effort to increase the SV and SI effectiveness. The published research and the proposed are compared, and the results are displayed in Table II. The accuracy of the proposed system is high, where the location of tests is in the lab of my department which has relatively low noise. This demonstrates that, in comparison to the outcomes of published studies, our findings are encouraging.

TABLE II.
Comparison Between Published Research and the Suggested MFCC System

Ref.	System type	Number of speaker	Voice type	Accuracy
[21]	Text Indepen dent-SV and SI	80	Universi ty office (Differe nt noise level)	89.2% /90%
[22]	Text depende nt and indepen dent/ SI	15	Microph On	97.9% and 94.4%
[23]	Text depende nt SV	15	Standar d protocol	EER 0.18
[24]	SV	80	Lab	EER 2.5%
[25]	SV	7302	Low SNR	EER 3.98
[26]	SI	30	Indoor (low noise)	96.67%
[27]	SV And SI	22	-	96.9%, 90%
Pro pose d wor k	Text depend ent SV	30	Lab (low noise)	98%– 99%

EER=false rejected rate=false acceptance rate=no. accepted imposter/total no. of imposter=no. rejected true speech/total no. of true speaker [28].

The sound still has many challenges, including noise, emotion, age, psychological state, reverberation level, etc. The sound is still complicated even more than imaging subjects which has limited and nearly unvaried data. In future work, the speaker recognition system can be placed in adverse circumstances by including the audio noise which is represented by those from generators, cars, trains, air conditioners, reverberation, etc. In the future, all the observations gathered from this research will be evaluated in our broad database and map those features into a low-dimensional space using methods like PCA, linear discriminant Analysis (LDA), or neural networks. Another significant field is the effect of the FFT phase which can be included with the absolute value of FFT results, on which not many researchers have focused.

IX. Conclusion

This study attempted to investigate characteristics of the short-time text-dependent and independent speech words between the intended user and others. A detailed comparison and analysis to obtain a correlation between the Mel-scale coefficients of speakers and MFCC coefficients extraction for the intended user was presented. For text-dependent, the correlation accuracy for Mel scale coefficients of user1 (the same person voice) ranged from 98% to 99%. For text-dependent and independent speakers, the user1 Mel scale coefficients correlation with other speakers was 83% and 61%, respectively. Additionally, the ten MFCC coefficients variance for the intended speaker is very low compared to significant variance for other speakers, which implies the intended user has a high level of discrimination ability. Furthermore, to enhance the performance of the discriminative system, PCA is offered where each pair of distorted speakers is chosen by the PCA to take the best possible route. The PCA results, which improve the discriminative information that is often used to reduce the dimensionality of large MFCCs data, were comparable to the correlation finding from the Mel-scale results.

The foundations show that the intended speaker identity information extracted from the short utterance was distinguishable from other speakers despite the uttered selected words being very nearly voice to others. In light of this, future works will concentrate on recording a wider corpus while taking into account the use of speaker's customized password messages and employing their own style.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

REFERENCES

[1] S. Hizlisoy and R. S. Arslan, "Text independent speaker recognition based on mfcc and machine learning," *Selcuk University Journal of Engineering Sciences*, vol. 20, no. 3, pp. 73–78, 2021.

[2] U. Ayvaz, H. Gürüler, F. Khan, N. Ahmed, T. Whangbo, and A. A. Bobomirzaevich, "Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning.," *Computers, Materials & Continua*, vol. 71, no. 3, 2022.

[3] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Applied Sciences*, vol. 11, no. 8, p. 3603, 2021.

[4] X. Liu, M. Sahidullah, and T. Kinnunen, "Learnable mfccs for speaker verification," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, IEEE, 2021.

[5] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxceleb speaker recognition challenge 2020 system description," *arXiv preprint arXiv:2010.12468*, 2020.

[6] N. D. Minh, "Dsp mini-project: An automatic speaker recognition system," 2012.

[7] B. Kari and S. Muthulakshmi, "Real time implementation of speaker recognition system with mfcc and neural networks on fpga," *Indian Journal of Science and Technology*, vol. 8, no. 19, p. 1, 2015.

[8] J.-C. Liu, F.-Y. Leu, G.-L. Lin, and H. Susanto, "An mfcc-based text-independent speaker identification system for access control," *Concurrency and Computation: Practice and Experience*, vol. 30, no. 2, p. e4255, 2018.

[9] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271, 2017.

[10] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2018.

[11] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 397–406, 2012.

[12] F. K. Faek and A. K. Al-Talabani, "Speaker recognition from noisy spoken sentences," *International Journal of Computer Applications*, vol. 70, no. 20, 2013.

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

[14] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 277–288, 2005.

[15] M. Jin and C. D. Yoo, "Speaker verification and identification," in *Behavioral Biometrics for Human Identification: Intelligent Applications*, pp. 264–289, IGI Global, 2010.

[16] A. Buono, W. Jatmiko, and B. Kusumoputro, "Mel-frequency cepstrum coeffficients as higher order statistics representation to characterize speech signal for speaker identification system in noisy environment using hidden markov model," in *Self Organizing Maps-Applications and Novel Algorithm Design*, IntechOpen, 2011.

[17] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of mfcc feature extraction accuracy using pca in indonesian speech recognition," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, pp. 379–383, IEEE, 2018.

[18] A. Sahoo and A. Panda, "Study of speaker recognition systems," *National Institute of Technology, Rourkela*, 2011.

[19] R. Gupta and G. Sivakumar, "Speech recognition for hindi language," *IIT BOMBAY*, 2006.

[20] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[21] K. Daqrouq and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Applied Soft Computing*, vol. 27, pp. 231–239, 2015.

[22] A. Antony and R. Gopikakumari, "Speaker identification based on combination of mfcc and umrt based features," *Procedia computer science*, vol. 143, pp. 250–257, 2018.

[23] A. Books, "Template-matching for text-dependent speaker verification, dey, subhadeep, motlicek, petr, madikeri, srikanth and ferras, marc, idiap-rr-32-2017," *Speech Communication*, 2017.

[24] M. Athulya and P. Sathidevi, "Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers," *Digital Investigation*, vol. 25, pp. 70–77, 2018.

[25] W. Lin and M.-W. Mak, "Robust speaker verification using population-based data augmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7642–7646, IEEE, 2022.

[26] S. Hidayat, M. Tajuddin, S. A. A. Yusuf, J. Qudsi, and N. N. Jaya, "Wavelet detail coefficient as a novel wavelet-mfcc features in text-dependent speaker recognition system," *IIUM Engineering Journal*, vol. 23, no. 1, pp. 68–81, 2022.

[27] N. Chauhan, T. Isshiki, and D. Li, "Text-independent speaker recognition system using feature-level fusion for audio databases of various sizes," *SN Computer Science*, vol. 4, no. 5, p. 531, 2023.

[28] S. Sreedharan and C. Eswaran, "A review on speaker verification: Challenges and issues," *Int. J. Sci. Technol. Res.*, vol. 8, no. 8, pp. 956–960, 2019.