ESTEEM Academic Journal
Vol. 19, March 2023, 101-112

Comparative Analysis of LPC and MFCC for Male Speaker Recognition in Text-Independent Context

Mohamad Khairul Najmi Zailan[1], Yusnita Mohd Ali[2], Emilia Noorsal[3], Mohd Hanapiah Abdullah[4], Zuraidi Saad[5], and Adni Mat Leh[6]
[1,2,3,4,5,6]*Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Cawangan Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang, Malaysia*
**corresponding author: [2]yusnita082@uitm.edu.my*

ARTICLE HISTORY

Received
20 January 2023

Accepted
14 March 2023

Available online
31 January 2023

ABSTRACT

Speech is the utmost communication medium for human beings which conveys rich and valuable information such as accent, gender, emotion and unique identity. Therefore, automatic speaker recognition can be developed based on unique characteristics of one's speech and utilized for applications such as voice dialing, online banking, and telephone shopping to verify the identity of its users. However, retrieving salient features which are capable of identifying speakers is a challenging problem in speech recognition systems since speech contains abundant information. In this study, a total of 438 audio data obtained from speakers uttering speech in text-independent context is proposed using speech data elicited from three Malay male speakers. The performance of two popularly used feature extraction techniques namely, linear prediction coefficients (LPC) and Mel-frequency cepstral coefficients (MFCC) were compared using discriminant analysis model. Although both features yielded impressive outcomes, the MFCC features surpassed that of LPC by achieving a higher accuracy rate of 99.09%, which was 4.34% higher than the latter.

Keywords: *speaker recognition; biometric; linear prediction coefficients; mel-frequency coefficients; discriminant analysis*

1. INTRODUCTION

Automatic speaker recognition or abbreviated as ASpkR is the process of recognizing the identity of a speaker from his/her speech signal through the use of appropriate speech analysis and machine learning techniques by the implementation of computer algorithms and hardware. It is one of the important topics in speech recognition system. The realm of ASpkR system includes either speaker verification or speaker identification [1] depending on the required tasks. The motivation for a system to identify a speaker from speech signals arises from the fact that each individual has unique characteristics such as accent, gender and emotion [2-3], the goal of automatic speaker identification is to extract, characterize, and acknowledge the speaker specific voiceprint for identification purpose.

ASpkR can be considered one of the successful biometric technologies that can be easily implanted in various mobile devices or other security equipment due to its convenience and ease of accessibility, namely, it requires only a microphone to capture the speech signals as inputs to the systems. ASpkR has significantly changed the way humans interact with computers via these spoken technologies in more natural ways rather than using keypad and mouse. Designing an ASpkR system requires two stages namely, the front-end and back-end. The front-end includes pre-processing and feature extraction, whilst the back-end comprises a

p-ISSN 1675-7939; e-ISSN 2289-4934

classifier or speaker model [4]. Feature extraction is crucial in ASpkR since the function is to discriminate between speaker identity, and past studies have shown numerous feature extractors [5] with different success rates. However, extracting salient features which are capable of discriminating the speakers is a challenging task and when combined with a good classifier, it provides accurate results of speaker identification. There exists specific feature extractors for speech signal such as linear prediction coefficients (LPC), Mel-frequency cepstrum coefficient (MFCC), linear prediction cepstral coefficients, real cepstral coefficients and many more [6].

Nevertheless, speaker recognition remains an open question in terms of determining the features and classifiers that could produce the best results. In previous study by Chauhan, et al. [7] a fusion of MFCC, LPC and zero-crossing rate features using artificial neural network yielded 92.8 % while for support vector machine, the best accuracy rate of 80.6 % was obtained by MFCC and LPC fusion for 10 speaker recognition. In another study, Swedia, et al. [8] proposed speech digit recognition utilizing 12-LPC coefficients and 12-MFCC coefficients with LSTM model outperformed hidden markov model which yielded the accuracy rates of 96.6 % for MFCC and 93.8 % LPC respectively. For robust speaker recognition, Salvati, et al. [9] proposed a late fusion deep neural network using raw time-domain features and gammatone cepstral coefficients (GTCC) and yielded an improved accuracy rates of 2.2 % to 12.5 % over the tested baseline features for TSP speech database. MFCC performed the worst as compared to raw waveforms and GTCC under noisy conditions.

Due to the popularity, this study aims to apply appropriate speech pre-processing, extract LPC and MFCC features, and compare their performance using simple discriminant analysis (DA) as the classifier for text-independent context. Compared to a text-dependent, a text-independent ASpkR is more challenging and should work better if the system is trained using long utterances to suppress vast lexicon variability adverse effects [10-11]. For that purpose, this study combined a list of isolated words (short utterances) and sentences (long utterances) in the development of the system for Malaysian English accents database [12]. In addition, this study is limited to only Malay male speakers to avoid other factors such as gender and accent which may have influences on the results. The objective of this study is to provide design solutions for correlating speech unique features to speaker identity under these limitations.

2. METHODOLOGY

This system consists of front-end and back-end parts. The front-end consists of pre-processing and feature extraction while the back-end consists of feature matching or classification for the decision making of the speaker label. The overview of the system is represented in Figure 1.

2.1 Speech Database

The database of voice signals used for this research was collected from Malaysian English Accent database, UniMAP [12]. This study only used 19 isolated words from Sections A (consisted of 52 isolated words) and all 17 sentences/phrases from Section B. The elicited materials are tabulated in Table 1 and Table 2. Each isolated word from Section A was replicated five times while the sentences from Section B were replicated three times for each speaker. The collection of the dataset amounted to 438 speech samples taken from only three Malay speakers out of 103 speakers. The signals were recorded in a semi-anechoic acoustic chamber with noise level of approximately 22 dB. The sampling rate and bit resolution were set to 16 KHz and 16 bit for high quality speech recognition purposes.

p-ISSN 1675-7939; e-ISSN 2289-4934

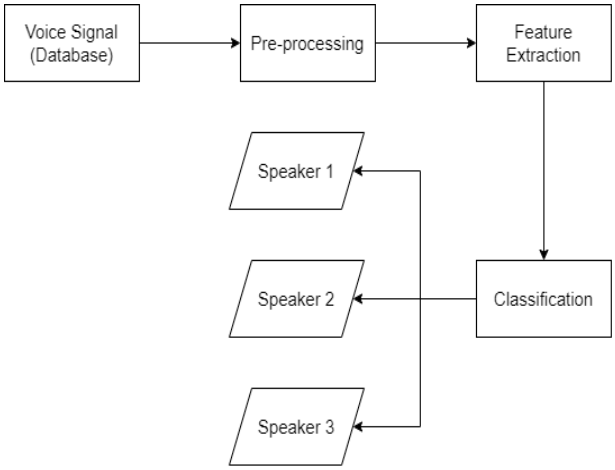

Figure 1: Automatic Speaker Recognition (ASpkR) Block Diagram

Table 1: List of Isolated Word in Section A

No	Isolated Word	No	Isolated Word	No	Isolated Word
1	Aluminium	8	Destination	15	Target
2	Better	9	Five	16	Thirsty
3	Bottom	10	Girl	17	Time
4	Boy	11	Pleasure	18	Would
5	Bringing	12	Station	19	Zero
6	Brother	13	Stella		
7	Communication	14	Student		

Table 2: Phrases in Section B

No	Sentences	Word Count
1	This is my mother.	4
2	He took my book.	4
3	How old are you?	4
4	Where are you going?	4
5	It would be better if a boy and a girl have more time for communication.	15
6	Look! Catch that bird! It goes to south.	8
7	Three businessmen pump their money into this project in bringing up the profit as their target.	16
8	Aluminium is not white. Your teeth are.	7
9	We must hear the expert before we change our mind.	10
10	Root anchors the plant to the ground.	7
11	The student drew a line at the bottom of the map.	11
12	It is my pleasure to see thirty of you there.	9
13	Freeze! Don't enter. You break the rule.	7
14	Hello there. Your destination is in the east, fifty-eight kilometres from here.	13
15	The temperature is at zero degree.	6
16	Histogram is a type of bar chart.	7
17	The car park is wide and open.	7

p-ISSN 1675-7939; e-ISSN 2289-4934

2.2 Pre-processing

The pre-processing consists of the process such as normalization, pre-emphasis, framing and overlapping, and windowing. Normalization was done to reduce the mismatch between signals and making the normalized signal more comparable regardless of the amplitude [13]. The normalization equation as in Equation (1).

$$\text{sigN}(n) = \frac{[sig(n) - \mu]}{max|sig(n) - \mu|} \quad (1)$$

where sigN(n) and sig(n) are the normalized and original signals with μ is the mean while *max* is the maximum values of speech amplitudes.

Next is the pre-emphasis procedure which is a filter to balance the frequency spectrum [14]. The speech signal was routed through a high-pass filter (FIR) to compensate for the attenuation from lip radiation [15] with the FIR equation as in Equation (2).

$$\text{sigP(n)} = \text{sigN}(n) - \alpha \cdot \text{sigN}(n-1) \quad (2)$$

The filter coefficient, α, typically has a value between from 0.9 to 1. However, in this study, the parameter of α was set to 15/16 (0.9375) as a fixed-point implementation since it is the most common value used in past research [16]. Figure 2 shows the difference between original, normalized and pre-emphasis signals.

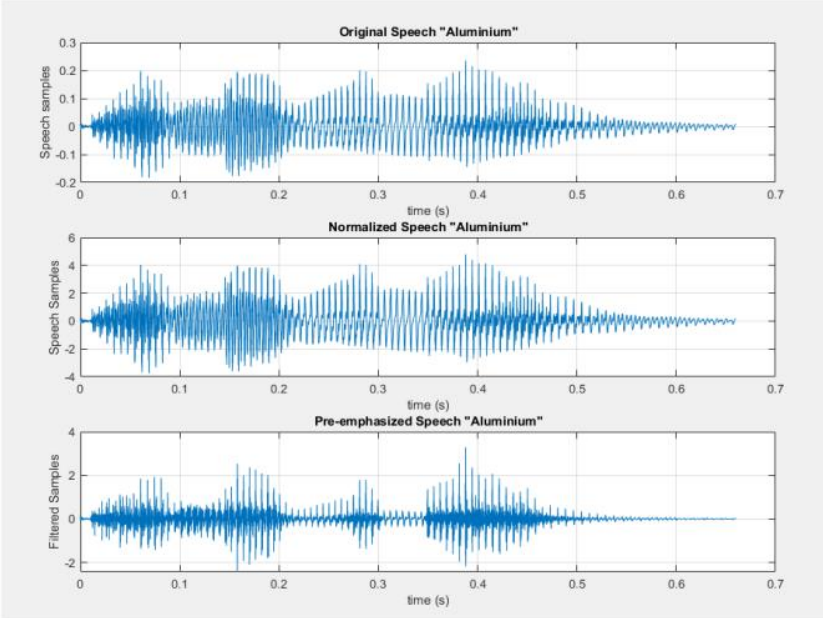

Figure 2: Difference in Normalization and Pre-Emphasis Signal

From Figure 2, the pre-emphasized speech signal shows that the pre-emphasized process had flattened the spectral to increase the signal-to-noise ratio (SNR). Besides, this process enhanced the audio signal to a more suitable signal for the feature extraction.

p-ISSN 1675-7939; e-ISSN 2289-4934

Next, to obtain the pseudo-stationary feature, the pre-emphasis signal was frame blocked into a 32 ms short-time frame since the speech signal cannot be processed in a whole non-stationary signal [17]. This short-time frame was then overlapped by 50% to prevent any critical data losses due to windowing function. The popularly used Hamming window is expressed as in Equation (3) was applied to every short-time frame to reduce signal distortion at both ends of the frames [18]. The length of window and hop sizes employed in this study were fixed to N = 512 and M = 256 respectively. Figure 3 shows the pre-emphasized signal and windowed signal.

$$\mathrm{w(n)} = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{3}$$

$where\ 0 \leq n \leq N-1$

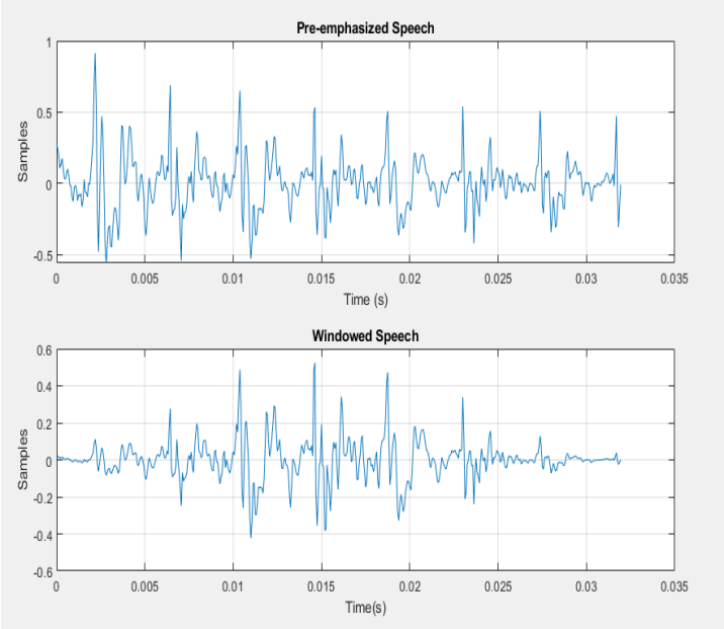

Figure 3: A snapshot of pre-emphasized and windowed signals

2.3 Feature Extraction

The method of extracting feature vectors or numerical features from a voice stream is known as feature extraction. The recorded voice signal is unsuitable to use as an input to a speech recognition system, necessitating the use of feature extraction methods [19]. With careful feature selection, useful information can be extracted from the input to accomplish the intention to identify the discriminative characteristics from the speech signal. This procedure could make the classification of the system to be done more efficiently.

2.3.1 Linear Prediction Coding

In a voice recognition system, LPC provides a suitable acoustic model. Based on the linear combination of earlier speech samples, this approach estimates the subsequent speech signal sample. LPC generates a coefficient of prediction that is comparable to the original signal but with a lower bit value. As a result, LPC is commonly employed to compress voice signals [8].

p-ISSN 1675-7939; e-ISSN 2289-4934

The estimated speech is calculated as in Equation (4) [16]. The LPC procedure's block diagram is summarized in Figure 4.

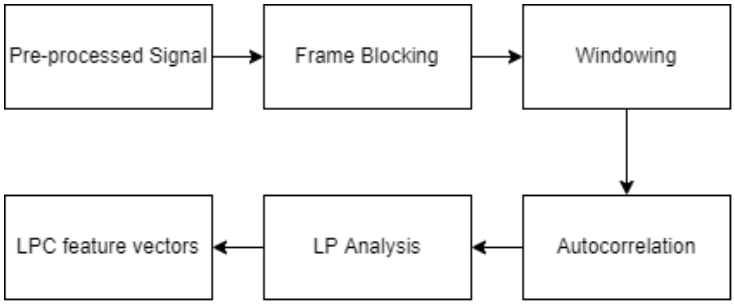

Figure 4: LPC Block Diagram

$$\tilde{x} = \sum_{k=1}^{p} a(k)x(n-k) \tag{4}$$

where $x(n)$ and $\tilde{x}$ are speech samples and their estimates while a(k) is the LPC parameters and p is the filter order.

Additionally, the spectra from the LPC order p = 10, 16 and 22 for the FFT and LPC methods are illustrated in Figure 5. It reveals that if the LPC filter is increased, it could produce more reasonable poles that are able to reflect the FFT spectrum excellently.

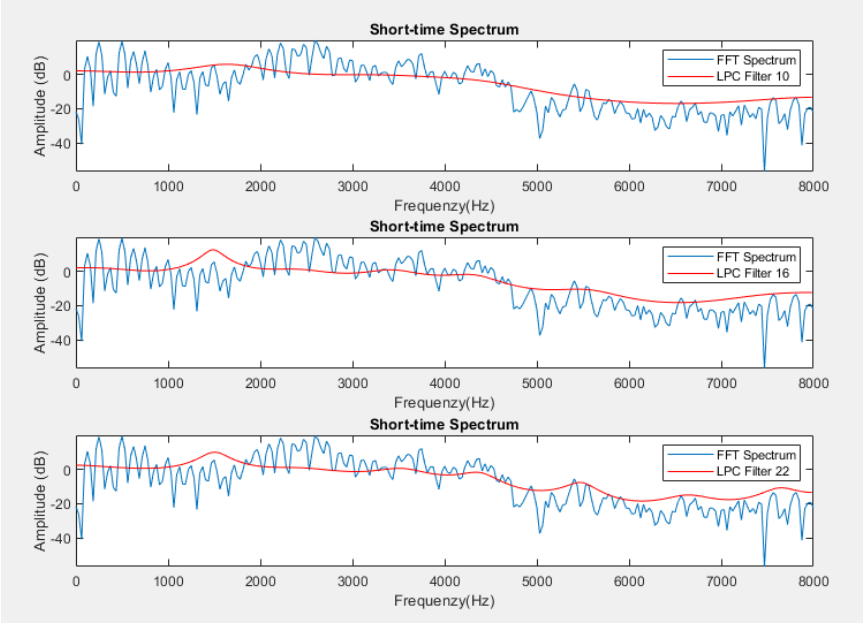

Figure 5: Spectra of FFT and LPC filter for the isolated word "aluminum"

p-ISSN 1675-7939; e-ISSN 2289-4934

2.3.2 Mel-frequency Cepstral Coefficients

In speech recognition systems, MFCC is a popular feature extraction approach. MFCC was built on a collection of filter banks consisting of multiple band pass filters in the form of triangle shape window functions that used mel-scale warped frequency to decode speech sounds [20]. Equation (5) states the mapping of acoustic linear frequency, f into perceptual mel frequency.

$$\text{mel freq} = 2595 \log_{10} \frac{f}{700} \tag{5}$$

The same pre-processing steps were applied to MFCC as with LPC extraction. A set of Mel filterbank was applied to the power spectra and the logarithm of all filterbank energies were calculated. Then, discrete cosine transform (DCT) procedure was used in the final stage of MFCC. In essence, DCT converts the cepstral frequency domain into a coefficient called the quefrency domain. Mel-scale cepstral coefficients were generated by the cepstrum coefficients produced by the DCT transform. Equation (6) was used to represent the procedure using DCT to obtain the MFCC coefficients.

$$c_m = \sum_{k=1}^{N} E_k \cos\left[\frac{m(k-0.5)\pi}{N}\right] \tag{6}$$

where variables $c(.)$ and $E(.)$ represent the m^{th} cepstral coefficient (cepstrum) and k^{th} log-energy respectively. N is the number of filters in the filter banks and the number of cepstrum takes in this order, m = 1, 2, … M. The mel filter bank and block diagram of MFCC can been seen in Figure 6 and Figure 7 respectively.

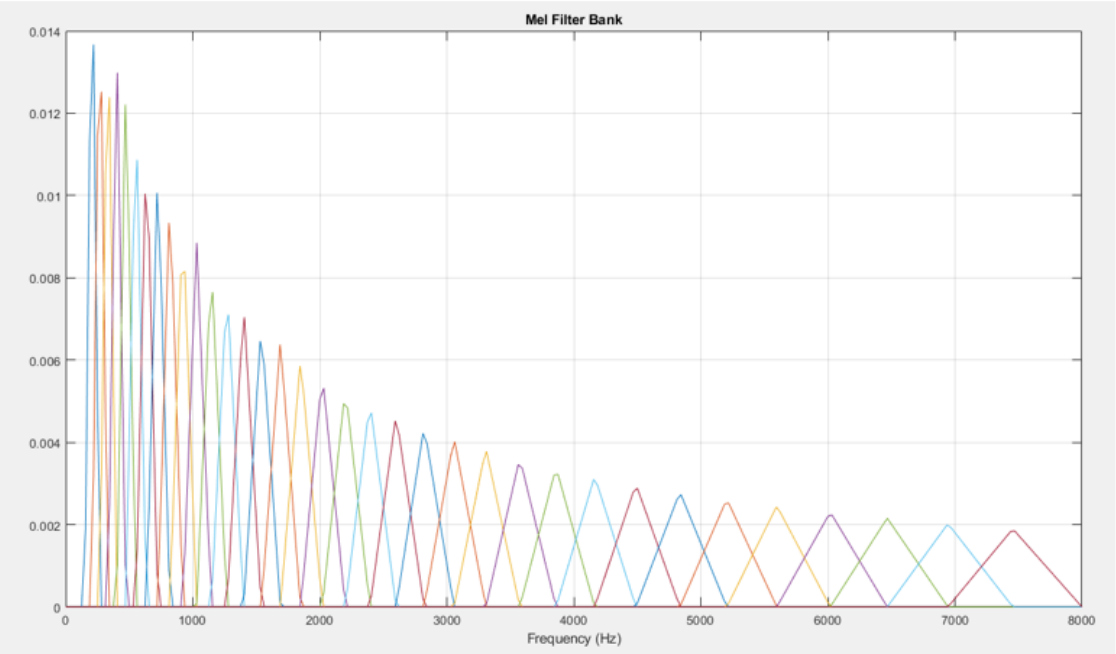

Figure 6: Mel Filter Bank

p-ISSN 1675-7939; e-ISSN 2289-4934

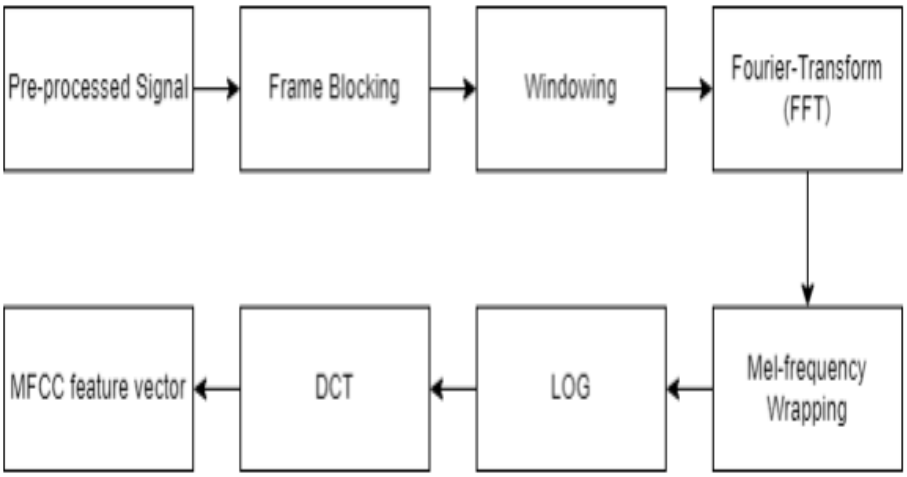

Figure 7: MFCC Block Diagram

2.4 Classification

The classification or speaker modelling was used to generate speaker recognition algorithms for voice feature matching [21]. The discriminant analysis (DA) approach was used in this study for the classifier. The DA classifier can be varied based on five functions: *mahalanobis*, *quadratic*, *diagonal quadratic*, *linear*, and *diagonal linear*. All these functions were used to test the performance. In the testing and validation method using DA, cross-validation utilising the *k*-fold approach was applied. For example, feature extraction datasets were randomly divided into 10 subsets, with one of the ten subsets assigned as a testing dataset and the remaining subsets were intended as training datasets. The process was repeated until the final subset was designated as the testing dataset. The results were evaluated using a confusion matrix [16, 20].

3. RESULTS AND DISCUSSION

In this study, the performance of ASpkR using LPC and MFCC features were compared after experimenting with the optimal parameters of the features and classifier.

3.1 Selection of filterbank for MFCC

Firstly, this study conducted analysis for MFCC to determine the number of filters, N_F in the filterbank would produce the best outcome for speaker recognition performance. The MFCC order for this analysis was arbitrarily fixed at c = 10 and the function of the classifier was set to *mahalanobis*. The frequency range of the filterbank was set from 150 Hz to 8 kHz as there is little information below 150 Hz for clean speech [22]. Then, starting from 15, the number of bands were increased in steps of 5 up to 40 filters. Figure 8 shows the results of varying the numbers of filters to the performance of speaker identification (Speaker ID).

The results concluded that N_F = 35 number of filter bank was the optimal number of bands with the highest accuracy rate of 86.53% for c = 10. Therefore, the number of bands in the filterbank of this research was fixed at 35 for the following analysis involving MFCC.

p-ISSN 1675-7939; e-ISSN 2289-4934

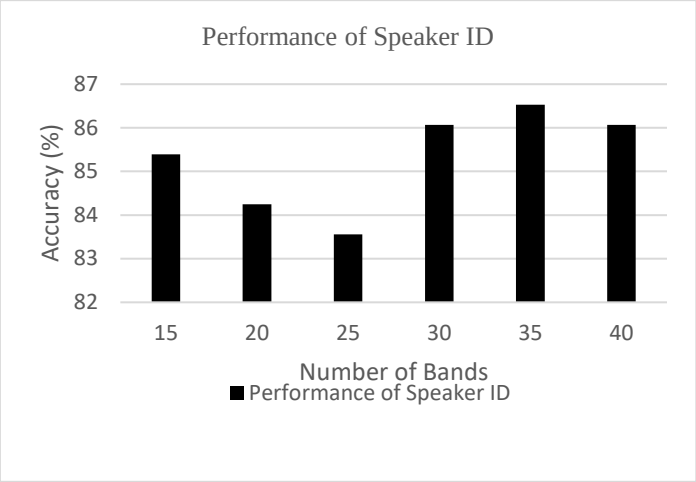

Figure 8: Performance of Speaker ID with different number of filer bank

3.2 Selection of no of coefficients for LPC and MFCC

Next, the best coefficients for the two feature extraction methods were determined by fixing the previous settings. The experiment was carried out by increasing the number of coefficients by 2 at a time. Figure 9 shows the performance using LPC and MFCC features for the coefficients, *p* and *c* of LPC and MFCC varied between 10 and 30 respectively.

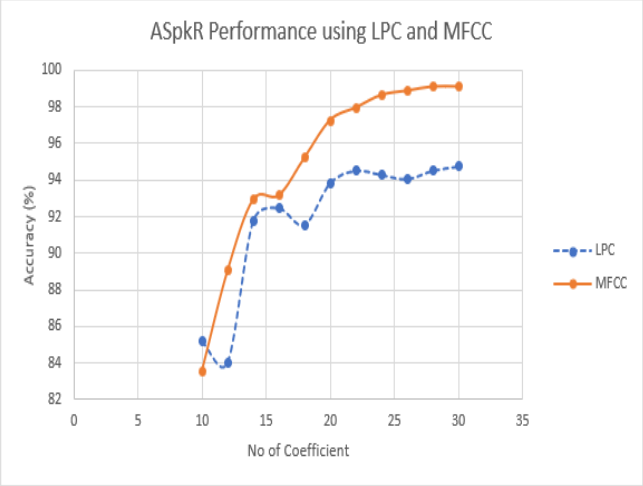

Figure 9: Comparison of LPC and MFCC for varying no of coefficients

In general for both features, there was a hike in performance between 10 and 14 coefficients. The trends continued to increase between 16 and 20 at slower rates and started to stagnate after 28 coefficients. MFCC surpassed LPC with its best result of 97.94 % at $c = 28$ as compared to LPC yielded the optimal result of 94.52 % at $p = 22$.

3.3 Selection of distance functions for Discriminant Analysis

The selection of functions in the classifier was also an important parameter setting. In this experiment, it is found that *mahalanobis* and *quadratic* functions achieved better results than the other two functions as shown in Figure 10. The overall accuracy for LPC using *mahalanobis*

p-ISSN 1675-7939; e-ISSN 2289-4934

and *quadratic* were 94.75 % and 94.29 % respectively while that of MFCC were 99.09 % and 98.17 % respectively. DA with *diagonal quadratic* resulted in the worse performance of 59.13 % and 86.99 % for LPC and MFCC respectively.

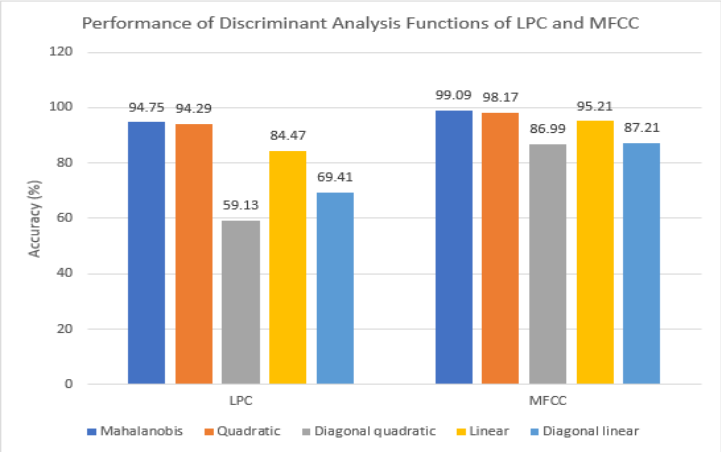

Figure 10: Performance of Discriminant Analysis Functions of LPC and MFCC

3.4 Individual classes performance

From the confusion matrix, apart from overall class performance, the individual class performance can also be obtained. The analysis maintained the best parameter setting for each feature set namely, p = 22, c = 28 and N_F = 35. The results were illustrated in Figure 11. The individual performance for both feature sets were both dominated by *Speaker 2* with the highest average accuracy rate of 95.89% for LPC and 100% for MFCC respectively. Hence, it may be said that *Speaker 2* surpassed the other speakers due to his unique speaking identity.

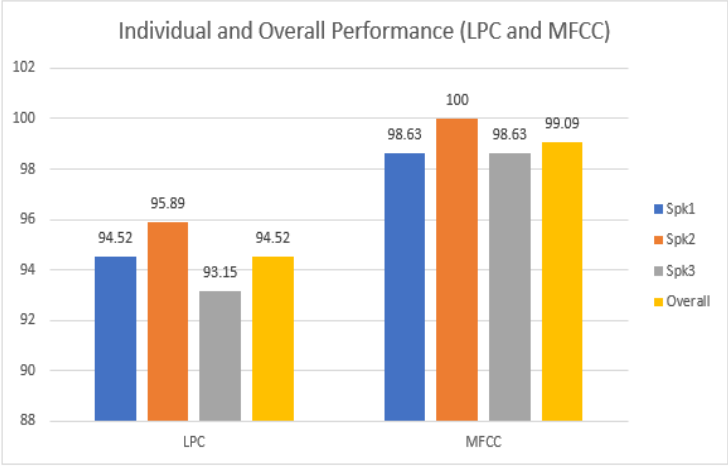

Figure 11: Individual and Overall Performance for LPC and MFCC

4. CONCLUSION

In conclusion, speaker recognition is one of the complex systems in processing human information which plays a vital role in human-machine interaction. Many crucial signal processing methods such as pre-emphasis, frame blocking and windowing, must be performed

p-ISSN 1675-7939; e-ISSN 2289-4934

to provide good signal conditions before extracting the speech features. Then, the system's performance is a direct function of feature extraction and classification techniques. In this paper, a comparative analysis of text-independent speaker recognition using LPC and MFCC on Malaysian English databases for male speakers was investigated. The outcomes were promising with the overall accuracy rates of 94.75 % and 99.09 % for LPC and MFCC respectively. Comparing the two features, MFCC outperformed LPC by 4.34 % overall accuracy rate using the *mahalanobis* DA function as the classifier. Other than that, when the three speakers were examined individually, *Speaker 2* exhibited the most unique speech characteristics with the highest overall accuracy rate of 100 % using MFCC. In future, other relevant methods in ASpkR could be explored to improve the performance of the system.

ACKNOWLEDGMENT

The authors would like to express our deepest gratitude to Universiti Teknologi MARA, Cawangan Pulau Pinang for the research support.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] H. Jiang and H. Yu, "Research on Speaker Recognition Technology Based on Feature Model," *Proceedings of the 3rd Asia-Pacific Conference on Image Processing, Electronics and Computers, Dalian, China*, 2022.

[2] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114591, 2021.

[3] A. Kusuma and D. P. Lestari, "Atom Aligned Sparse Representation Approach for Indonesian Emotional Speaker Recognition System," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, IEEE, pp. 1-4. 2020.

[4] N. Washani and S. Sharma, "Speech recognition system: A review," *International Journal of Computer Applications*, vol. 115, no. 18, pp. 7-10, 2015.

[5] S. Shaikh Naziya and R. Deshmukh, "Speech recognition system—a review," *IOSR J. Comput. Eng*, vol. 8, no. 4, pp. 3-8, 2016.

[6] M. Jakubec, E. Lieskovska, and R. Jarina, "An Overview of Automatic Speaker Recognition in Adverse Acoustic Environment," in *2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA)*: IEEE, pp. 211-218, 2020.

[7] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*,: IEEE, pp. 130-133, 2019.

[8] E. R. Swedia, A. B. Mutiara, and M. Subali, "Deep learning long-short term memory (LSTM) for Indonesian speech digit recognition using LPC and MFCC Feature," in *2018 Third International Conference on Informatics and Computing (ICIC)*, : IEEE, pp. 1-5. 2018

[9] D. Salvati, C. Drioli, and G. L. Foresti, "A late fusion deep neural network for robust speaker identification using raw waveforms and gammatone cepstral coefficients," *Expert Systems with Applications*, vol. 222, p. 119750, 2023.

[10] S. R. Hasibuan, R. Hidayat, and A. Bejo, "Speaker Recognition Using Mel Frequency Cepstral Coefficient and Self-Organising Fuzzy Logic," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*: IEEE, pp. 52-55, 2020.

p-ISSN 1675-7939; e-ISSN 2289-4934

[11] Y. Tu, W. Lin, and M. W. Mak, "A Survey on Text-Dependent and Text-Independent Speaker Verification," *IEEE Access*, vol. 10, pp. 99038-99049, 2022, doi: 10.1109/ACCESS.2022.3206541.

[12] M. A. Yusnita, M. P. Paulraj, S. Yaacob, M. N. Fadzilah, and A. B. Shahriman, "Acoustic Analysis of Formants Across Genders and Ethnical Accents in Malaysian English Using ANOVA," *Procedia Engineering*, vol. 64, pp. 385-394, 2013.

[13] B. M. Nema and A. A. Abdul-Kareem, "Preprocessing signal for speech emotion recognition," *Al-Mustansiriyah Journal of Science*, vol. 28 (3), pp. 157-165, 2018.

[14] M. M. Hasan, H. Ali, M. F. Hossain, and S. Abujar, "Preprocessing of Continuous Bengali Speech for Feature Extraction," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*: IEEE, pp. 1-4, 2020.

[15] M. Yusnita, M. Paulraj, S. Yaacob, R. Yusuf, and M. N. Fadzilah, "Robust Accent Recognition in Malaysian English using PCA-Transformed Mel-Bands Spectral Energy Statistical Descriptors," *Indian Journal of Science and Technology*, vol. 8, p. 20, 2015.

[16] M. A. Yusnita, M. P. Paulraj, S. Yaacob, S. A. Bakar, and A. Saidatul, "Malaysian English accents identification using LPC and formant analysis," in *2011 IEEE International Conference on Control System, Computing and Engineering*: IEEE, pp. 472-476, 2011.

[17] O. K. Hamid, "Frame blocking and windowing speech signal," *Journal of Information, Communication, and Intelligence Systems (JICIS)*, vol. 4 (5), pp. 87-94, 2018.

[18] M. Manjutha, J. Gracy, P. Subashini, and M. Krishnaveni, "Automated speech recognition system—A literature review," *Computational Methods, Communication Techniques And Informatics*, vol. 205, pp. 740-741, 2017.

[19] Y. Astuti, R. Hidayat, and A. Bejo, "Comparison of Feature Extraction for Speaker Identification System," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*: IEEE, pp. 642-645, 2020.

[20] M. Yusnita, E. Noorsal, N. F. Mokhtar, S. Z. M. Saad, M. H. Abdullah, and L. C. Chin, "Speech-based gender recognition using linear prediction and mel-frequency cepstral coefficients," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, pp. 753-761, 2022.

[21] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79236-79263, 2021.

[22] M. Yusnita, M. Paulraj, R. Y. Sazali Yaacob, and A. Shahriman, "Analysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in Malaysian English," *International Journal of Automotive and Mechanical Engineering*, vol. 7, pp. 1053-1073, 2013.

p-ISSN 1675-7939; e-ISSN 2289-4934