

# Speech-based gender recognition using linear prediction and mel-frequency cepstral coefficients

Yusnita Mohd Ali<sup>1</sup>, Emilia Noorsal<sup>1</sup>, Nor Fadzilah Mokhtar<sup>1</sup>, Siti Zubaidah Md Saad<sup>1</sup>,  
Mohd Hanapiah Abdullah<sup>1</sup>, Lim Chee Chin<sup>2</sup>

<sup>1</sup>Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Cawangan Pulau Pinang, Permatang Pauh, Malaysia

<sup>2</sup>Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis, Kampus Pauh Putra, Malaysia

## Article Info

### Article history:

Received Apr 8, 2022

Revised Jul 20, 2022

Accepted Aug 18, 2022

### Keywords:

Artificial neural network

Discriminant analysis

Gender recognition

Linear prediction coefficients

Mel-frequency cepstral coefficients

## ABSTRACT

Gender discrimination and awareness are essentially practiced in social, education, workplace, and economic sectors across the globe. A person manifests this attribute naturally in gait, body gesture, facial, including speech. For that reason, automatic gender recognition (AGR) has become an interesting sub-topic in speech recognition systems that can be found in many speech technology applications. However, retrieving salient gender-related information from a speech signal is a challenging problem since speech contains abundant information apart from gender. The paper intends to compare the performance of human vocal tract-based model i.e., linear prediction coefficients (LPC) and human auditory-based model i.e., Mel-frequency cepstral coefficients (MFCC) which are popularly used in other speech recognition tasks by experimentation of optimal feature parameters and classifier's parameters. The audio data used in this study was obtained from 93 speakers uttering selected words with different vowels. The two feature vectors were tested using two classification algorithms namely, discriminant analysis (DA) and artificial neural network (ANN). Although the experimental results were promising using both feature parameters, the best overall accuracy rate of 97.07% was recorded using MFCC-ANN techniques with almost equal performance for male and female classes.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Yusnita Mohd Ali

Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Cawangan Pulau Pinang

13500 Permatang Pauh, Pulau Pinang, Malaysia

Email: yusnita082@uitm.edu.my

## 1. INTRODUCTION

Gender identification or recognition is pressing matter culturally and socially as categorizing gender has become humans' second nature and remain a fundamental question in cognitive sciences. According to Whiteside [1], gender can be easily and accurately perceived through human speech alone. Voice characteristics of male and female speakers are distinguished by not only low-level pitch analysis but also contributed by timbre perception [2]. There are several acoustical differences between the two sexes resulted from anatomo-physiological nature of the speech production system unique to each gender such as the mean fundamental frequency (F0), formant frequencies, glottal function, and long-term average spectrum. In general, women speak at about an octave higher pitch than men [3] while men's voices are generally deeper and louder. The pitch average range of an adult female was reported approximately 120 to 350 Hz, while that of an adult male is around 100 to 200 Hz [2]. On the other aspect, female voices showed more aspiration noises evident by lower spectral tilt, than male voices. This is due to larger gap at the back of their vocal cords, which allows more air to pass through and gives women's voices more of a "breathy" quality than

men's voices. However, depending on pitch, formants or other speaker dependent acoustical features alone will not promise a good estimate because the intra-subject variability is very large and the range of acoustical values overlap considerably between male and female voices [1].

The urge of identifying gender can be demonstrated in many situations. Since the emergence of human-machine interaction (HMI), many fields require machines to identify gender for numerous modern applications [4]–[6]. Past research on voice pathology shows that it is gender biased. For example, vocal folds cyst [7], [8] tends to occur in female patients while dysphonia [9] is prevalently higher for boys than girls before puberty. Therefore, automatic gender recognition (AGR) plays a significant role in mobile or remote healthcare system to sort patients according to gender for proper treatment. Recently, Alhussein *et al.* [10] proposed a modified voiced contour based on time-domain voice intensity using Simpson's rule and used support vector machine (SVM) as gender detection engine to be incorporated in mobile healthcare system. The system evaluated on TIMIT and Arabic digit databases were reported to obtain accuracy of 98.27% and 96.55% respectively for clean and noisy speech. An additional AGR submodule in speech and speaker recognition systems provide a better performance for both authentication and identification for use in applications such as online banking, education, shopping, and security systems. Biometric such as voiceprint includes identification of gender could be used in forensic suspects recognition by the authorized committee for analyzing frauds using voice application systems. Forensic gender recognition developed on NIST 2003 database was proposed by Kenai *et al.* [11] using Mel-frequency cepstral coefficients (MFCC) and Gaussian mixture model.

As speech signals play an important input source in HMI, age and gender recognition was also reported in [12] using convolutional neural network (CNN) and speech spectrograms features. The proposed techniques achieved accuracy scores of 96% and 97% on common voice dataset and Korean dataset respectively. Age and gender play an important role in demographic study. Tang *et al.* [13] explored on gender and nationality information in public database called VoxCeleb1 [14] to improve speaker recognition using a mining algorithm based on those two embedding features proposed using CNN classification frameworks achieved more than 98% accuracy of gender recognition on different spectrogram features. Another field of importance that requires gender knowledge is emotion recognition for human-robot interaction since it is one of the most influential factors. Gender recognizer was incorporated in [15] and achieved accuracy closest to 98% to drive speech emotion model based on the first stage detection so that the system accuracy could be improved. Although recent research has proposed the use of deep learning in gender classification to be robust and accurate, a limitation of this type of algorithm is its resource hungry [12], [16], [17]. In other words, it requires large datasets, complex processing and time-consuming to produce a reliable model. Various statistical acoustic features [18] consisted a balance amount of male and female speech data were transformed using principal component analysis (PCA) to select 11 most salient features before feeding into SVM achieved 98.42% accuracy rate, an increase of 7.42% to using SVM alone [19]. SVM is less resource intensive and operates based on separation of data on hyperplane and proven to be good for binary classification. Until now, traditional speech extractors such as MFCC, linear prediction coefficients (LPC) and linear prediction cepstral coefficients (LPCC) are still in widely used [20]–[23] for speech operated gender detection systems known for the dedicated purpose and good performance. These general speech features are powerful over other types of speaker dependent features such as pitch, formants, chroma, energy and entropy.

Based on previous studies, gender recognition from speech signal is still undergoing interesting exploration using various speech analysis techniques and classification algorithms. This study basically aims to compare the effectiveness of two traditional and powerful speech extractors designed based on human speech production vocal tract filter model i.e., LPC and human auditory system model i.e., MFCC on a common speech database. The study utilizes the strength of vowel-based words for features extraction to discriminate gender using speaker-independent database.

## 2. RESEARCH METHODS

The proposed AGR system consists of two main blocks namely front-end processing and back-end processing as depicted in Figure 1. The front-end stage consists of pre-processing and feature extraction blocks to provide suitable speech parameters as input to the back-end stage. Front-end processing plays a crucial role in converting raw speech waveforms into a set of feature vectors contain considerably lower information rate than the original signal. These parameters are extracted in frame by frame wise to maintain quasi-stationary characteristics in the complex speech signal. This section explains about feature extraction using LPC and MFCC apart from pre-processing steps implemented in this work. As for the classifiers, this work proposes the use of simple discriminant analysis (DA) adopted in [24], [25] and

artificial neural network (ANN) adopted in [26], [27] as the classification engines to compare the performance of AGR under study.

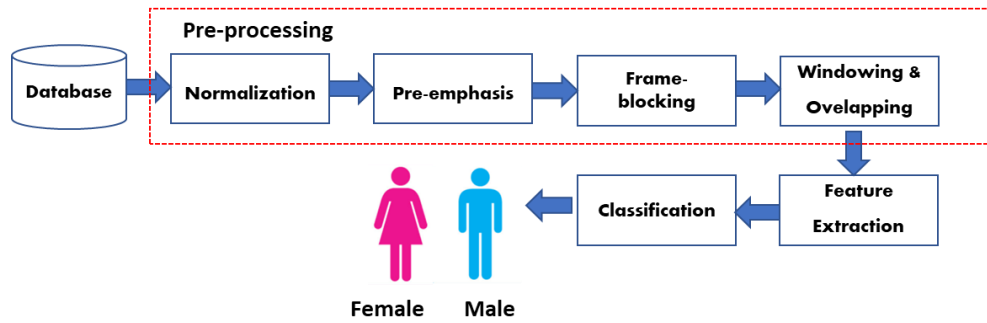


Figure 1. Overall system block diagram for automatic gender recognition system (AGR)

## 2.1. Experimental Database

Data collection is crucial and plays an important role in this project. This study adopts an open-source database consisted of 1116 speech samples recorded from male and female adult speakers. Ninety-three (93) speakers participated in this study composed of forty-five (45) males and forty-eight (48) females, uttering 12 different vowel sounds (short, long and diphthongs) accumulated to 540 male utterances and 576 female utterances. The speech corpus [28] used in this research belonged to Western Michigan University. The recorded 12 vowel sounds were pronounced in American English consisted of /ae/ in "had", /ah/ in "hod", /aw/ in "hawed", /eh/ "head", /er/ in "heard", /ei/ in "hayed", /ih/ in "hid", /iy/ in "heed", /oa/ in "hoed", /oo/ in "hood", /uh/ in "hud" and /uw/ as in "who'd". The recordings were clean speech recorded at a sampling rate of 16 kHz using linear PCM format.

## 2.2. Pre-processing

This pre-processing stages consist of normalization, pre-emphasis, frame blocking, overlapping, and windowing. The first stage is normalization performed in time-domain. The technique being applied here is mean and maximum-mean-subtraction normalization (MMS) as in (1). The dynamic range of signal amplitude is limited to  $\pm 1$  using MMS to prevent volume differences among speakers or differences due to microphone and recording settings.

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

where  $\text{sigN}(n)$  and  $\text{sig}(n)$  are respectively the normalized and original speech signals at discrete time  $n$ , with  $\mu$  and  $\max$  represents the mean and maximum values of speech amplitudes.

The second stage is applying pre-emphasis filtering. For fixed-point implementation, the filtering coefficient of  $\alpha = 15/16$  is selected to compensate the attenuation due to lip radiation using a simple first order high-pass finite impulse response (FIR) filter [29]. The filter transfer function and filtered signal are expressed in (2) and (3).

In the third and last steps of pre-processing, the speech was frame blocked into 32 msec short-time frame to ensure pseudo-stationary property for valid functions of LPC and MFCC processors. These short-time frames of the whole signal are chunk using a hop size of 50% and convolved with Hamming window function [22], [30] for minimizing spectral distortion and compensate attenuation of audio features towards both ends of an analysis frame. The length of analysis window and hop size used in this work are  $N = 512$  and  $M = 256$  points respectively.

$$H(z) = 1 - \alpha z^{-1} \quad (2)$$

$$\text{sigP}(z) = \text{sigN}(n) - \alpha \times \text{sigN}(n-1) \quad (3)$$

where  $H(z)$  is the  $z$ -transform of filter transfer function, and  $\text{sigP}(n)$  and  $\text{sigN}(n)$  are respectively the pre-emphasized and 1st order delayed normalized speech signals at discrete time  $n$ , with  $\alpha$  as the filtering coefficient. The popularly used Hamming function is represented in (4) where  $n$  and  $N$  are discrete time index and the size of the window function respectively.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \leq n \leq N-1 \quad (4)$$

### 2.3. Linear Prediction Coefficients

LPC is a speech analysis technique that estimates future speech samples from a linearly weighted summation of past  $p$ -samples using method of least squares. The speech is modeled as a  $p$ th order autoregressive system using all-pole IIR filter to represent human vocal tract as LTI system over short intervals. LPC analysis determines the coefficients of a forward linear predictor by minimizing the prediction error in the least square sense [30], [31]. The estimated speech is calculated as in (5).

$$\tilde{x} = -\sum_{k=1}^p a(k)x(n-k) \quad (5)$$

where  $x(n)$  and  $\tilde{x}$  are speech samples and their estimates, and  $a(k)$  is the feature vector of LPC coefficients where  $p$  is the linear predictive filter order. The autocorrelation function (ACF) of each frame signal can be computed [10] using (6).

$$r(i) = -\sum_{n=0}^{N-1-i} X^F(n)X^F(n+i) \quad (6)$$

where  $r(i) = [r(0), r(1), \dots, r(p)]$  is the ACF of a frame signal denoted as  $X^F(n)$  of  $N$ -points. Next, the Yule-Walker equations are solved using the Levinson-Durbin recursive algorithm to obtain the coefficients of the prediction filter as in (7).

$$\sum_{k=1}^p a(k)R(i-k) = -r(i) \quad 1 \leq i \leq p \quad (7)$$

where  $R(i-k)$  forms the ACF matrix ( $p$ -by- $p$ ) which is a symmetric Toeplitz matrix. The LP coefficients,  $a(k)$  can be solved efficiently by taking the inverse of ACF matrix multiplied by ACF vector shown in (8).

$$a(k) = -R^{-1}r \quad (8)$$

### 2.4. Mel-frequency Cepstral Coefficients

If LPC mimics the human vocal tract model, MFCC is best at representing human auditory system. MFCC speech extractor is based on a set of filter banks constructed from several band pass filter in a form of triangular-shape window functions using Mel-scale warped frequency domain to mimic the different area in the eardrum to decode the speech sounds. The transformation formula from Mel-scale to Hz is related (9). The centre frequencies of the series bandpass filters are designed according to this perceptually motivated scale, the known variation of the human ear's critical bandwidths [32].

$$Mel = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \quad (9)$$

Equation (9) reveals that both frequency in Hz and in Mel-scale are almost linearly related below 1 kHz and logarithmically related for frequency above 1 kHz. Mel-scale filtering is used to focus on the lower frequency components that are more important in speech analysis. By summing all the product of log-energy spectrum using Fourier transform from each individual bandpass filter and then applying discrete cosine transform (DCT), the cepstral coefficients of these filter banks are generated as in (10).

$$C_m = \sum_{k=1}^N E_k \cos[m(k-0.5)\pi/N] \quad (10)$$

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,  $m = 1, 2, \dots, M$ .

## 3. RESULTS AND DISCUSSION

In this section the performance of AGR system using LPC and MFCC feature datasets extracted from the selected vowel-emphasized words database were compared. Furthermore, two supervised classifiers i.e., DA and ANN were adopted and tuned for the best parameters settings. For DA classifier, class label '1' represented the male and class label '2' represented the female. As for ANN, the class label '1' and '2' were transformed into a two-bit output neuron of '10' for male and '01' for female in the output layer.

In the testing and validation phases using DA, cross-validation using  $k$ -fold method was employed using 10 folds wherein the data was randomized and stratified into 10 subsets, each time only one subset was used as testing dataset and the remaining ( $k-1$ ) subsets were utilized for training. The process was repeated until the last subset became the testing dataset. The results were accumulated and measured using confusion matrix. For ANN, independent testing method employing data partition of 60%, 15% and 25% for training, validation, and testing datasets was used respectively. The binary transformation of the simulated output using threshold and margin criteria was adopted wherein a real output of above 0.6 was considered '1' while below 0.4 as '0'. For each data block, randomization and stratification were performed to avoid bias results. The results are discussed as the following.

### 3.1. Impact of varying feature dimension

In this work, the performance of AGR using LPC and MFCC feature space was compared using DA classification method due to its simple mechanism. The optimal number of LPC input parameters was selected by varying the  $p$ -th order from 8 to 22. Similarly, for MFCC, the optimal number of cepstral coefficients was selected by varying its value from 8 to 18. For the MFCC the number of filter banks used was 20. The filter channels were densely located for low frequencies, but sparsely located for higher frequencies. This shows that filtering using the Mel scale emphasized the lower frequency components that are more important in speech analysis.

Figure 2 shows the performance measures of LPC and MFCC modelled using DA with linear function. Four performance measures were reported i.e., overall classification rate (CR), sensitivity, specificity, and precision, obtained from confusion matrix while increasing the  $p$ -th order of LPC from 8 to 22 in Figure 2(a). The results reveal that the peak performance of most types of measures occurred at  $p = 18$ . The best accuracy rates were 87.72%, 86.30%, 89.06% and 88.09% respectively for the overall CR, sensitivity, specificity, and precision. Higher accuracy rates in specificity than sensitivity shows that the female speakers were better recognized than the male speakers by approximately 3%. For MFCC, the number of cepstral coefficients in MFCC extractor was varied from 8 to 18 as in Figure 2(b). It is observed that MFCC outperformed LPC by 1.62% when the number of cepstral coefficients was set to  $c = 18$ . The results for overall CR, sensitivity, specificity, and precision were 89.34%, 89.81%, 88.89% and 88.34% respectively. The results show the male class achieved slightly higher accuracy about 1% than the other class.

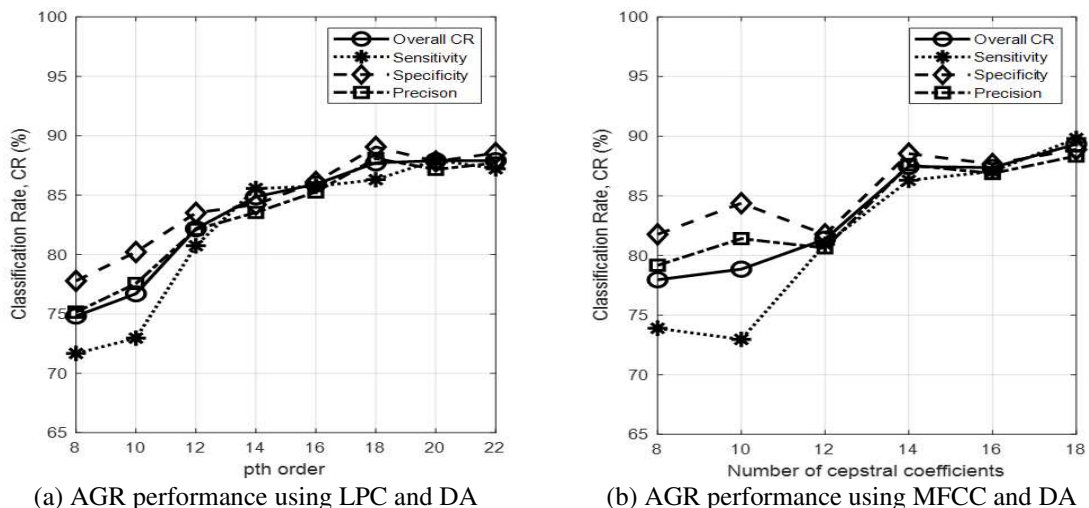


Figure 2. Varying the (a)  $p$ -th order in LPC and (b) cepstral coefficients in MFCC extraction

### 3.2. Impact of varying the classifiers' parameters

The input feature parameters were fixed at  $p = 18$  and  $c = 18$  for the subsequent analysis. The performance of AGR was tested with five different functions available in DA i.e., linear, diagonal linear, quadratic, diagonal quadratic and Mahalanobis for both LPC and MFCC inputs. Figure 3 shows the performance of AGR using different features and classifiers. Figure 3(a) depicts the results of varying the function of DA classifier and training algorithm of ANN classifier. The results suggested that both LPC and MFCC features yielded the best accuracy rates using quadratic and Mahalanobis functions. The overall CR

for LPC using quadratic and Mahalanobis were 90.14% and 90.77% respectively while that of using MFCC were 95.07% and 95.16%. MFCC surpassed LPC by approximately 5% using Mahalanobis function.

Next, the performances of 18-LPC and 18-MFCC were tested using another classifier namely ANN. The popularly used Multilayer Perceptron (MLP) was adopted as alternative classifier to DA. The activation functions used in the input layer and hidden layers were *logsig*. The stopping criterion used was mean square error,  $mse = 0.01$ . The learning rate and momentum rate were set to 0.3 and 0.9. In this experiment the number of neurons for LPC input was set to  $n_h = 30$  and that of MFCC input was set to  $n_h = 10$  to sufficiently start the training phase. Four backpropagation training algorithms used were Levenberg-Marquardt (trainlm), Bayesian regularization (trainbr), Scaled conjugate gradient (trainscg) and Resilient (trainrp). The results are shown in Figure 3(b). Both feature types show the best overall accuracy rates using Levenberg-Marquardt i.e., 90.77% and 94.89% for LPC and MFCC features. MFCC features outperformed LPC for all training algorithms. Levenberg-Marquardt is fast and so far, produce excellent performance in pattern recognition problem.

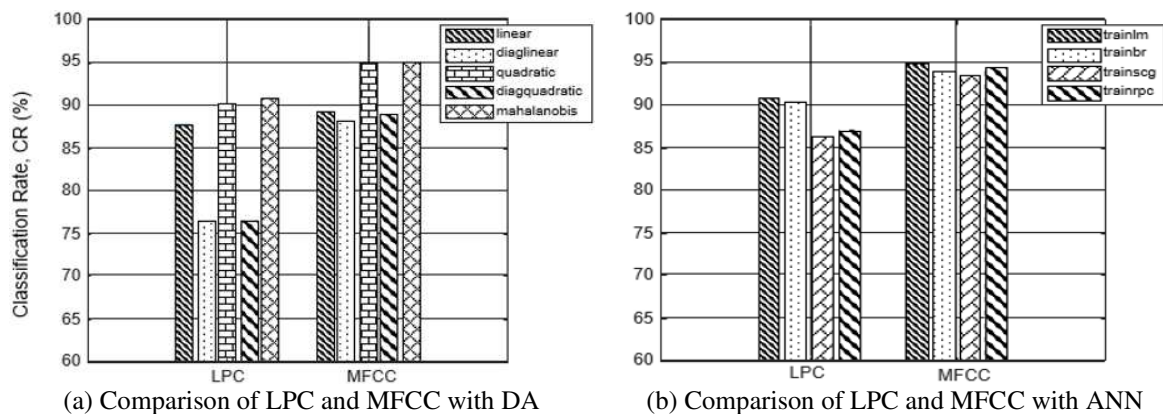


Figure 3. Varying the (a) function in DA and (b) training algorithm in ANN for LPC and MFCC features

The number of hidden neurons was varied from  $n_h = 30$  to 70 for LPC and  $n_h = 10$  to 70 for MFCC in 10 step size in subsequent step to find the optimal settings. ANN model using LPC required larger hidden neurons to be developed. The input feature vector dimension was fixed at 18 and Levenberg-Marquardt as the training algorithm. Figure 4 shows the performance of ANN for varying number of hidden neurons. From the conducted experiment, it was found that the optimized number of neurons to train LPC input was  $n_h = 50$  while MFCC required lesser number of neurons in the hidden layer,  $n_h = 20$  to avoid overfitting.

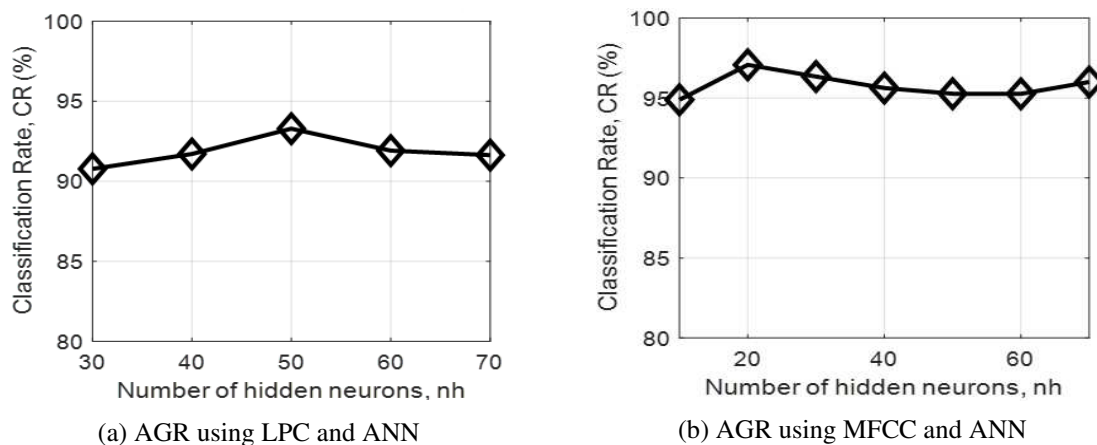


Figure 4. Varying the number of hidden neurons in ANN for LPC and MFCC inputs

### 3.3. Performance comparison of LPC features versus MFCC features

Lastly, we compare the performance of gender recognition system built using the optimal settings of both feature sets and classifiers to conclude their success in detecting gender from speech signal. Table 1 and Table 2 show the comparison results using DA and ANN classifiers. For DA, as explained at the beginning of this section, there was 1,116 samples accumulated after all 10-folds were used as test dataset using cross-validation techniques. Whilst for ANN, about 277 to 279 samples was taken as test dataset which constituted 25% of the database size. From the results, it can be concluded that MFCC consistently outperformed LPC features using both types of classifiers namely, 4.4% higher using DA and 3.8% higher using ANN. Additionally, female class outperformed male class by approximately 1.1% using LPC and DA classification algorithm. On the other hand, MFCC features resulted in a better performance of male class by about 2.6% than the other using DA classifier. Meanwhile, the capability of identifying male and female speakers using ANN was found to be quite equal i.e., less than 1% difference in accuracy.

Table 1. Performance measures using discriminant analysis

	LPC	MFCC
Accuracy	90.77	95.16
Sensitivity	90.19	96.48
Specificity	91.32	93.92
Precision	90.69	93.71

Table 2. Performance measures using artificial neural network

	LPC	MFCC
Accuracy	93.28	97.07
Sensitivity	93.02	96.99
Specificity	93.53	97.14
Precision	93.02	96.99
Classified rate	96.40	97.85
Unclassified rate	3.60	2.15

## 4. CONCLUSION

Gender recognition is one of the complex human processing information problems used in HMI systems. It has been potentially perceived through speech signal using appropriate speech analysis tools and classification algorithms. The performance of the system is closely linked to the selected parameters in feature extractors and the employed classification model. Apart from that, the choice of database affects the validity of the developed model especially using data driven machine learning such as neural network. In this study we use a limited vowel-emphasized words in American English to develop and test the efficacy of the system developed using LPC and MFCC features with DA and ANN machine learning techniques. The results were promising using both LPC and MFCC with the best accuracy rates of 93.28% and 97.07% respectively. Comparing the two features, MFCC surpassed LPC by 3.8% to 4.4% using both ANN and DA classifiers with ANN as the better classifier in this study.

## ACKNOWLEDGEMENTS

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

## REFERENCES




- [1] S. P. Whiteside, "Identification of a speaker's sex: a study of vowels," *Percept. Mot. Skills*, vol. 86, no. 2, pp. 579–584, 1998, doi: 10.2466/pms.1998.86.2.579.
- [2] C. R. Pernet and P. Belin, "The role of pitch and timbre in voice gender categorization," *Front. Psychol.*, vol. 3, p. 23, 2012, doi: 10.3389/fpsyg.2012.00023.
- [3] M. Latinus and M. J. Taylor, "Discriminating male and female voices: differentiating pitch and gender," *Brain Topogr.*, vol. 25, no. 2, pp. 194–204, Apr. 2012, doi: 10.1007/s10548-011-0207-9.
- [4] Z. S. Marzoog, A. D. Hasan, and H. H. Abbas, "Gender and race classification using geodesic distance measurement," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 27, no. 2, p. 820, Aug. 2022, doi: 10.11591/ijeecs.v27.i2.pp820-831.
- [5] J. I. Al-Nabulsi and B. E. A. Badr, "Adaptive gender-based thermal control system," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 2, p. 1200, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1200-1207.
- [6] S. Bekhet, A. M. Alghamdi, and I. F. Taj-Eddin, "Gender recognition from unconstrained selfie images: a convolutional neural network approach," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 2, p. 2066, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2066-2078.
- [7] M. Bouchayer *et al.*, "Epidermoid cysts, sulci, and mucosal bridges of the true vocal cord: a report of 157 cases," *Laryngoscope*, vol. 95, no. 9, p. 1087, Sep. 1985, doi: 10.1288/00005537-198509000-00014.



- [8] M. M. Johns, "Update on the etiology, diagnosis, and treatment of vocal fold nodules, polyps, and cysts," *Curr. Opin. Otolaryngol. Head Neck Surg.*, vol. 11, no. 6, pp. 456–461, Dec. 2003, doi: 10.1097/00020840-200312000-00009.
- [9] M. Nygren, M. Tyboni, F. Lindström, A. McAllister, and J. van Doorn, "Gender differences in children's voice use in a day care environment," *J. Voice*, vol. 26, no. 6, pp. 817.e15–817.e18, Nov. 2012, doi: 10.1016/j.jvoice.2012.05.001.
- [10] M. Alhussein, Z. Ali, M. Imran, and W. Abdul, "Automatic gender detection based on characteristics of vocal folds for mobile healthcare system," *Mob. Inf. Syst.*, vol. 2016, pp. 1–12, 2016, doi: 10.1155/2016/7805217.
- [11] O. Kenai, S. Djeghiour, N. Asbai, and M. Guerti, "Forensic gender speaker recognition under clean and noisy environments," *Procedia Comput. Sci.*, vol. 151, pp. 897–902, 2019, doi: 10.1016/j.procs.2019.04.124.
- [12] A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors*, vol. 21, no. 17, p. 5892, Sep. 2021, doi: 10.3390/s21175892.
- [13] Y. Tang *et al.*, "Attention based gender and nationality information exploration for speaker identification," *Digit. Signal Process.*, vol. 123, p. 103449, Apr. 2022, doi: 10.1016/j.dsp.2022.103449.
- [14] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Interspeech 2017*, Aug. 2017, pp. 2616–2620. doi: 10.21437/Interspeech.2017-950.
- [15] A. Guerrieri, E. Braccili, F. Sgrò, and G. N. Meldolesi, "Gender identification in a two-level hierarchical speech emotion recognition system for an Italian social robot," *Sensors*, vol. 22, no. 5, p. 1714, Feb. 2022, doi: 10.3390/s22051714.
- [16] A. A. Abdulsatar, V. V. Davydov, V. V. Yushkova, A. P. Glinushkin, and V. Y. Rud, "Age and gender recognition from speech signals," *J. Phys. Conf. Ser.*, vol. 1410, no. 1, p. 012073, Dec. 2019, doi: 10.1088/1742-6596/1410/1/012073.
- [17] A. A. Alashban and Y. A. Alotaibi, "Speaker gender classification in mono-language and cross-language using BLSTM network," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2021, pp. 66–71. doi: 10.1109/TSP52935.2021.9522623.
- [18] M. Buyukyilmaz and A. O. Cibikdiken, "Voice gender recognition using deep learning," 2016. doi: 10.2991/msota-16.2016.90.
- [19] G. Sharma and S. Mala, "Framework for gender recognition using voice," in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2020, pp. 32–37. doi: 10.1109/Confluence47617.2020.9058146.
- [20] N. A. Nazifa, C. Y. Fook, L. C. Chin, V. Vijejan, and E. S. Kheng, "Gender prediction by speech analysis," *J. Phys. Conf. Ser.*, vol. 1372, no. 1, p. 012011, Nov. 2019, doi: 10.1088/1742-6596/1372/1/012011.
- [21] S. Goyal, V. V. Patage, and S. Tiwari, "Gender and age group predictions from speech features using multi-layer perceptron model," in *2020 IEEE 17th India Council International Conference (INDICON)*, Dec. 2020, pp. 1–6. doi: 10.1109/INDICON49873.2020.9342434.
- [22] M. A. Yusnita, A. M. Hafiz, M. N. Fadzilah, A. Z. Zulhanip, and M. Idris, "Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal," in *2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, Nov. 2017, pp. 372–377. doi: 10.1109/ICCSCE.2017.8284437.
- [23] S. Chaudhary and D. K. Sharma, "Gender identification based on voice signal characteristics," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Oct. 2018, pp. 869–874. doi: 10.1109/ICACCCN.2018.8748676.
- [24] R. Gupta and G. Aggarwal, "Human speech sentiments recognition: A data mining approach for categorization of speech," in *Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016*, 2016, pp. 3987–3991.
- [25] E. Mezghani, M. Charfeddine, H. Nicolas, and C. Ben Amar, "Speaker gender identification based on majority vote classifiers," Mar. 2017, p. 103410A. doi: 10.1117/12.2268741.
- [26] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "An artificial neural network approach to automatic speech processing," *Neurocomputing*, vol. 140, pp. 326–338, Sep. 2014, doi: 10.1016/j.neucom.2014.03.005.
- [27] A. D. Dongare, R. R. Kharde, and A. D. Kachare, "Introduction to artificial neural network," *Int. J. Eng. Innov. Technol.*, vol. 2, no. 1, pp. 189–194, 2012.
- [28] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3099–3111, 1995, doi: 10.1121/1.411872.
- [29] M. A. Yusnita, M. P. Paulraj, S. Yaacob, R. Yusuf, and M. Nor Fadzilah, "Robust accent recognition in Malaysian English using PCA-transformed mel-bands spectral energy statistical descriptors," *Indian J. Sci. Technol.*, vol. 8, no. 20, 2015, doi: 10.17485/ijst/2015/v8i20/78959.
- [30] S. Furui, "Digital speech processing, Synthesis, and Recognition," in *Digital Speech Processing, Synthesis, and Recognition*, CRC Press, 2018. doi: 10.1201/9781482270648.
- [31] Y. M. Ali *et al.*, "Voice command intelligent system (VCIS) for smart home application using mel-frequency cepstral coefficients and linear prediction coefficients," *J. Phys. Conf. Ser.*, vol. 1535, no. 1, p. 012008, May 2020, doi: 10.1088/1742-6596/1535/1/012008.
- [32] M. Zakariah, R. B. Y. Ajmi Alotaibi, Y. Guo, K. Tran-Trung, and M. M. Elahi, "An analytical study of speech pathology detection based on MFCC and deep neural networks," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–15, Apr. 2022, doi: 10.1155/2022/7814952.





## BIOGRAPHIES OF AUTHORS







**Yusnita Mohd Ali**    is a senior lecturer at the Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Penang Campus, Malaysia. She received her PhD Degree in Mechatronic Engineering from Universiti Malaysia Perlis in 2014 specializing in Audio/Acoustic Engineering. She was conferred with a Master Degree in Electronics System Design Engineering from University Sains Malaysia in 2004. She completed her Bachelor Degree in Electrical & Electronics Engineering from the same university in 1998. Her field of interest includes speech processing, speech analysis, human-machine interaction, brain-machine communication and artificial intelligence. She can be contacted at email: yusnita082@uitm.edu.my.









**Emilia Noorsal**     is a senior lecturer at the Universiti Teknologi MARA Penang Campus, Malaysia. In April 2014, she obtained her PhD in biomedical engineering from Institute of Microelectronics, Ulm, Germany. Her research interests include digital design circuit in ASIC, FPGA, mixed-signal circuit design, power electronics and electronics for biomedical applications. She can be contacted at email: emilia.noorsal@uitm.edu.my.







**Nor Fadzilah Mokhtar**     is a senior lecturer attached to the Centre for Electrical Engineering Studies, Universiti Teknologi MARA Penang Campus, Malaysia. She received her MSc. degree in Electronics System Design Engineering from Universiti Sains Malaysia in 2004. Her research interests include digital design circuit in ASIC, embedded system, advanced control system, artificial intelligence, and power electronics. She can be contacted at email: norfadzilah105@uitm.edu.my.







**Siti Zubaidah Md Saad**     is a lecturer at the Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Penang Campus, Malaysia. She received her master's degree in Electrical Engineering from UiTM Shah Alam in 2015 and currently pursuing PhD degree with Universiti Sains Malaysia. She completed her bachelor's degree in Microelectronic Engineering from Universiti Kebangsaan Malaysia in 2004. Upon completed her bachelor's degree, she worked as process engineer at Infineon Technologies Kulim for five years. Her field of interest includes analogue Integrated Circuit, IC design and semiconductor process. She can be contacted at email: zubaidah7173@uitm.edu.my.



**Mohd Hanapiah Abdullah**     received a BSc. (Hons) Electrical and Electronic Engineering from MARA University of Technology, Malaysia, and M. Sc (Microelectronics) in the field of Optoelectronics from University of UKM, Malaysia in 2000 and 2005, respectively. In 2015, he finished his PhD in the field of Nanotechnology Device Fabrication for Green Technology from UiTM, Shah Alam, Malaysia. Now, he is a Senior Lecturer in Electronic Engineering at the Department of Electronics, Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Permatang Pauh Penang, Malaysia. He is a member of the NanoElectronic Center (NET), Innovation Center (IOS), UiTM Shah Alam. He is looking forward to exploring new area of research and collaborations with other parties. He can be contacted at email: hanapiah801@uitm.edu.my.



**Ts. Dr. Lim Chee Chin**     received the Bachelor of Engineering (Hon.) in Biomedical Electronic Engineering from University Malaysia Perlis, in 2012 and the PhD in Electronic Biomedical Engineering under UniMAP in 2016. Currently, she is the senior lecturer in Biomedical Electronic Engineering, Faculty of Technology Electronics Engineering University Malaysia Perlis, Malaysia. Her research interest area is medical signal and image processing, Biomechanics and Healthcare, and Bioinstrumentation design. She responsible as secretary of the Final Year Project committee (JKPTA) and given the task of coordinating the final year project of Electronic Biomedical Engineering (RK85) program. She can be contacted at email: cclim@unimap.edu.my.