Gender Prediction by Speech Analysis

Nur Adlin Nazifa 1, Chong Yen Fook 1*, Lim Chee Chin 1, Vikneswaran Vijean 1, Eng Swee Kheng 2

1 Biosignal Processing Research Group (BioSIM), School of Mechatronic Engineering, Universiti Malaysia Perlis, Arau, Perlis, Malaysia
2 School of Mechatronic Engineering, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia.

*yenfook@unimap.edu.my

Abstract. Speech is one of the important methods of communication for humans. The speech signal itself contain linguistic information that can be used to identify the speaker information such as gender, emotions and many more. There are some problems that involve in detection gender of the speaker. In forensic analysis, the police need to detect criminal profile from any evidence such as voice from any calls and while in healthcare aspect, some vocal fold pathologies can be bias to a particular gender such as vocal fold cyst can be seen particularly in female patients and the patient will have problem with their voices. Three features are extract from the speech signal which are Mel Frequency Cepstrum Coefficient (MFCC), Linear Prediction Coding (LPC) and Linear Prediction Coding Coefficient (LPCC). While for the classification, two classifier are used which are Support Vector Machine (SVM) and k-Nearest Neighbour (KNN). The recognition rate is higher for the combination of MFCC and LPCC compared to other features. SVM classifier had outperformed KNN classifier and obtained highest recognition rate of 97.45%. Lastly a graphical user interface system is develop that will record the voice of the speakers, pre-process the signal, extract MFCC and LPCC and classify it using SVM.

1. Introduction
Humans can deliver message and information between each other through speech. Human voice is particularly a piece of human sound creation in which the human vocal chords is the essential sound source, which plays an important role in the conversation [1]. Linguistic information such as gender, emotions and others can be recognize through speech signal. Speech processing is the study of speech signals, and various of methods used to process them. Many applications such as speech coding, speech synthesis, speech recognition and speaker recognition technologies is employed using speech processing. One of important step in speech and speaker recognition is gender recognition [2]. Physiological differences such as vocal fold thickness or vocal tract length and differences in speaking style of humans can be identify as partly the reason gender based differences in human speech [3]. Since these differences are reflected in the speech signal, acoustic measures related to those properties can be helpful for gender classification. Female voice usually has higher resonance frequencies and higher pitch than male voice [4]. The main objective of this paper is to investigate the relationship between the speech signals towards the gender of the speakers. In order to predict the gender of the speaker, this research has proposed a method to extract several features from the speaker voice and class it according to its
gender. This study involved the methods of pre-processing of speech signal, extracting features from the filtered signal and classified it to two classes which are male or female.

2. Experiment
2.1. Data Acquisition
There are twenty four subjects which included twelve male and twelve female had volunteered to participate in this experiment. Microphone used for data acquisition in this experiment is Philips Speechmike Premium LFH3500. It is USB dictation microphone with push-button operation that can connected with the laptop. A Graphical User Interface (GUI) in MATLAB is developed to record the voice for each subject and stored the data. Sampling frequency was set to 16 000 Hz and 16 bit. User can record the voice and also play back the recorded voice.

2.2. Data Acquisition
First, a brief is given to the subjects on how the workflow of the experiment and what they need to do. Then, demonstration is given to them on how to handle the microphone and what button they need to push. The subjects need to sit in front of the laptop and sit with relax. The device is placed 3 to 4 cm from the mouth. The recorder will start and the subjects will pronounce a sequence of number from ‘0’ to ‘9’. The sequence of number is repeat for ten trials with random sequence. Lastly, the recorder will stop and end of data acquisition process.

3. Speech Signal Processing
The recorded speech signal was analysed in signal processing which comprises of three stages. It started with pre-processing of the signal, feature extraction from the signal and lastly classification into two classes which are male and female.

3.1. Pre-processing
The signal needs to be processed before undergoes feature extraction. If there are some noises, it needs to be eliminate and filter to enhance the speech signal features on the waveform. Most of the gender information in the speech signal lies in the voiced region of the speech signal. The signal needs to be segmented for each number from ‘0’ to ‘9’. The silence and unvoiced region are removed by set up a threshold value of energy for the voiced region. The energy value of signal above the threshold value is considered as voiced region and the energy value below the threshold value will considered as unvoiced region. Next, pre-emphasis is used to neutralize pronunciation effect of speech signal from the lip. It boosts input frequency range that most susceptible to noise and makes the position of dominant energy in frequency more crucial. In this case, the value of pre-emphasis used is 0.98.

3.2. Feature Extraction
Some useful information needs to extract from the signal to distinguish whether the speaker is male or female. The feature extraction that will be use in this project is Mel Frequency Cepstrum Coefficient (MFCC), Linear Prediction Coefficient (LPC) and Linear Predictive Cepstral Coefficient (LPCC).

3.2.1. Mel Frequency Cepstrum Coefficient (MFCC)
Mel Frequency Cepstrum Coefficient (MFCC) are based on the known variation of the human ear’s critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech [5]. The signals were divided into small frame size with the length 30 ms. Hamming window is used in this research to eliminate discontinuities at the edges and integrates all the closest frequency lines. The third step is to apply Fast Fourier Transform (FFT) which convert each frame of N samples from time domain to frequency domain and to extract the frequency components of the signal. After that, the signal needs to go through
Mel Filter Bank to filter the signal so it approximates to Mel scale. Each filter output is the sum of its filtered spectral components. Lastly, Discrete Cosine Transform (DCT) is computed by converting log Mel spectrum into time domain. Final result of the conversion is what it called Mel Frequency Cepstrum Coefficient.

3.2.2. Linear Predictive Coding (LPC)
Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model [6]. Pre-processing techniques of this method is including pre-emphasis, frame blocking and Hamming window. The autocorrelation analysis is used to identify fundamental frequency or pitch from the speech signal. Lastly, the signals need to undergo LP analysis.

3.2.3. Linear Predictive Cepstral Coefficient (LPCC)
Linear Predictive Cepstral Coefficient (LPCC) can generates vocal tract coefficients and it can represent the spectral magnitude of speech signal. It behaves in the Cepstrum domain which is also used in Linear Predictive Coefficient (LPC). In this method, speech signal also need to undergo pre-processing technique such as pre-emphasis, frame blocking and Hamming window and autocorrelation analysis. Auto correlation is used as the method to obtain the parameters by assuming that the output of the system is predicted as linear combination of past output samples while input is unknown [7]. Next, in LPC analysis, it will convert auto correlation coefficient into LPC parameters. The Levinson-Durbin recursive algorithm can be used to identify the coefficients [8].

3.3. Classifier
Classifier that had been chosen is Support Vector Machine (SVM) and K-Nearest Neighbour (KNN).

3.3.1. Support Vector Machine (SVM)
Firstly, kernel function is applied that performs a low to high dimensional feature transformation and in this case linear kernel function, Gaussian kernel function and polynomial kernel function is used. Next, this classifier constructs a maximum margin hyperplane to draw the decision boundary between classes. SVM search for an optimal hyperplane $H$ that separates this data space into two distinct subspaces and each corresponding to a class. This hyperplane can be described as the following equation [9-10]:

$$H: x^T w + b = 0$$

Where $w$ is the perpendicular vector to the hyperplane $H$ and $b$ is the bias of this function.

3.3.2. K-Nearest Neighbour (KNN)
K-Nearest Neighbour is base of an unknown sample on the “votes” of K of its nearest neighbour rather than only it’s single neighbour. The input of KNN classifier in the feature space consists of the k closest training example [11-13]. However, the output of the KNN are depends on the usage of the KNN either it is used for classification or regression. For the KNN classification, the class membership is the output of the classifiers. The data is classified using majority vote of the neighbours and the data is assigned to its classes according to its most common its k nearest neighbour [14-15]. The nearest neighbour are found by minimising a distance function, usually the Euclidean distance. K-value used is varies from 1 to 10 folds with 10-fold cross validation.
4. Result and Discussion

4.1. Comparison of SVM and KNN
Table 1 shows the recognition rate of SVM based on features extracted using several kernel functions.

| Kernel function | MFCC  | LPC   | LPCC  | MFCC+LPCC |
|-----------------|-------|-------|-------|------------|
| Linear          | 89.43%| 89.70%| 89.08%| 92.82%     |
| Gaussian        | 92.90%| 91.25%| 93.31%| 95.06%     |
| Polynomial      | 95.54%| 95.59%| 95.96%| 97.45%     |

From Table 1, it can be seen that polynomial kernel function has highest recognition rate of gender prediction compared to the linear and Gaussian kernel functions. The accuracy achieved is 95.54% for MFCC, 95.59% for LPC, 95.96% for LPCC and the highest is 97.45% for MFCC+LPCC. Table 2 shows the recognition rate of KNN based on features extracted using different k-values.

| k-value | MFCC  | LPC   | LPCC  | MFCC+LPCC |
|---------|-------|-------|-------|------------|
| 1       | 88.87 | 87.45 | 89.76 | 90.89      |
| 2       | 87.63 | 85.77 | 88.67 | 89.37      |
| 3       | 86.31 | 84.48 | 87.25 | 87.98      |
| 4       | 85.69 | 83.96 | 86.69 | 87.44      |
| 5       | 85.55 | 83.83 | 86.51 | 87.40      |
| 6       | 85.50 | 83.65 | 86.36 | 87.36      |
| 7       | 85.36 | 83.36 | 86.17 | 87.26      |
| 8       | 85.25 | 83.13 | 85.99 | 87.19      |
| 9       | 85.15 | 83.00 | 85.89 | 87.11      |
| 10      | 85.07 | 82.91 | 85.81 | 87.05      |

From Table 2, combination of MFCC and LPCC has the highest recognition rate using KNN in all k-value compared to the others feature which the highest value is 90.89%. Followed by MFCC and LPCC that have higher accuracy than LPC. Then, the lowest rate is achieved by LPC which most of its recognition rate is below than 90%. Furthermore, the result shows that combination of MFCC and LPCC is the most suitable features in gender prediction and able to distinguish the gender since MFCC and LPCC is more sensitive and nearest feature to human voice in both classifiers.

4.2. Graphical User Interface (GUI) Development
GUI was developed to build a system for prediction of gender of the speaker by using MATLAB software. This GUI can be used to record the voice and display the speech signal, segment word by word and display the signal, extract MFCC and LPCC features and classify the gender of the speaker. The layout of the GUI for gender prediction system is shown in Figure 1.
5. Conclusion
From the analysis of results, SVM classifier had outperformed KNN classifier and obtained highest recognition rate of 97.45%. This could be because KNN is a lazy algorithm that depends barely on statistics and comparison, and it must keep track of large amount of features whereas SVM uses offline learning to find the optimal hyperplane so it can distinguish gender well and very effective.

6. References
[1] Kumar P, Baheti P, Jha R K, Sarmah P and Sathish K 2018 Voice Gender Detection Using Gaussian Mixture Model Journal of Network Communications and Emerging Technologies (JNCET) 8(4) 132-136.
[2] Deiv D S 2011 Automatic Gender Identification for Hindi Speech Recognition International Journal of Computer Applications 31(5) 1–8.
[3] Upadhyaya P, Farooq O, Abidi M R and Varshney Y V 2018 Continuous Hindi speech recognition model based on Kaldi ASR toolkit Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2018(2) 786–789.
[4] Rashmi C R 2014 Review of Algorithms and Applications in Speech Recognition System International Journal of Computer Science and Information Technologies 5(4) 5258–5262.
[5] Jayasankar T, Vinothkumar K and Vijayaselvi A 2017 Automatic gender identification in speech recognition by genetic algorithm Appl. Math. Inf. Sci. 11(3) 907–913.
[6] Carol Y Espy-Wilson 1994 A feature-based semivowel recognition system The Journal of the Acoustical Society of America 96(1) 65-72.
[7] Kaur G, Srivastava M and Kumar A 2018 Genetic Algorithm for Combined Speaker and Speech Recognition using Deep Neural Networks Journal of Telecommunications and Information Technology 2018(2) 23–31.
[8] Yusnita M A, Hafiz A M, Fadzilah M N, Zulhanip A Z and Idris M 2017 Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal 2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang 372-377.
[9] Ahmad J, Muhammad K and Kwon S 2016 Dempster-Shafer Fusion based Gender Recognition for Speech Analysis Applications 2016 International Conference on Platform Technology and Service (PlatCon), Jeju, 1-4.

[10] Fook C Y, Hariharan M, Yaacob S, and Adom A H 2012 Malay speech recognition in normal and noise condition IEEE 8th International Colloquium on Signal Processing and its Applications (CSPA 2012) 409–412.

[11] Qawaqneh Z, Mallouh A A and Barkana B D 2017 Age and gender classification from speech and face images by jointly fine-tuned deep neural networks Expert Systems with Applications 85 76–86.

[12] Fook C Y, Hariharan M, Yaacob S, Adom A H 2016 Subtractive clustering based feature enhancement for isolated malay speech recognition ARPN Journal of Engineering and Applied Sciences 11(13) 8517-8524.

[13] Jonsson P and Wohlin C 2004 An evaluation of k-nearest neighbour imputation using Likert data 10th International Symposium on Software Metrics, 2004. Proceedings., Chicago, Illinois, USA 108-118.

[14] Chin L C, Fook C Y, Vikneswaran V, Abdul Rahim M A H, Hariharan M 2018 Body Mass Index (BMI) Of Normal and Overweight/Obese Individuals Based on Speech Signals Journal of Telecommunication, Electronic and Computer Engineering 10 1-16.

[15] Chin L C, Basah S N, Yaacob S, Juan Y E and Kadir A K A 2015 Camera Systems in Human Motion Analysis for Biomedical Applications AIP Conference Proceedings 1660(1) 090006.