# Voice Command Interpretation for Robot Control

Amal Punchihewa
School of Engineering & Advanced Technology
Massey University
Palmerston North, New Zealand
g.a.punchihewa@massey.ac.nz

Zuriawati Mohd Arshad
IBM Systems & Technology Group
IBM Malaysia Sdn. Bhd.
Selangor, Malaysia
zuria@my.ibm.com

*Abstract*—**This paper presents some initial results from an analysis of performance of a voice command interpretation and authorisation system using voiceprint to identify the human-commander. Two approaches based on human voice related algorithms are proposed. Mel-frequency cepstral coefficient (MFCC) and perceptual linear predictive (PLP) are two feature extraction methods that are closely mimic the human auditory system. The two methods were applied to the proposed system to determine their suitability for use in a commander recognition system. Vector Quantization (VQ) with Linde-Buzo-Gray (LBG) iterative algorithm was used for clustering for the classification of commanders. The performance of the algorithms was evaluated to compare between two methods in MATLAB simulation environment based on, false rejection rate (rejecting an authorised commander), false acceptance rate (accepting unauthorised commander) and the execution time. Based on the initial results, both methods achieved accurate classification and PLP method has shown better execution time and lower false-acceptance rate compared to the MFCC. The combined approach (MFCC-PLP) did not show considerable improved performances to the individual feature models PLP and MFCC without incurring high computational costs that will compromise the performance of the speaker recognition tasks. Therefore, PLP method is the best candidate for command-recognition system to be developed in the second phase of this research.**

***Keywords-robot control; voice command; interpretation; identification; recognition***

## I. INTRODUCTION

In recent times, biometric systems have been used in many applications for security protection systems such as access control, computer and network security, and also personal and public safety. Since voice is a unique characteristic of human beings and it is a natural form of communication, voice biometric is one of the most popular biometric systems that have drawn the attention of most law enforcement agencies and investigators [5]. Moreover, the ability of human perception system to recognise and identify a person based on her or his voice has motivated the researchers to develop robust and effective automated speaker recognition systems [9]. In disaster situation, there are locations where human cannot reach. In aging societies around the world, service robots for eldercare are a possibility. Robots are also capable of navigating and surviving in dangerous environments such as chemical spill or nuclear radiation. A wireless robot can be manipulated by voice commands like in real human environment situation. However, it is important to assure line of command authority and identifying individual commanders. In this research, voice is used as a means to interact with the robot control system to provide a hierarchical command system using commander identification. In the proposed system, other than authenticating a person and allowing the person to gain access to the give command to robot, other commanders can be overridden based on a hierarchical command system such as in millitary. There are four basic steps in commander recognition system; digital acquisition of speech data, feature selection and extraction, feature clustering and pattern matching. Each stage will be further explained in the later sections.

## II. BACKGROUND

This section presents the background information and literature review of speaker recognition research. Firstly, the section will attempt to provide basic information on human speech production and perception systems and also human auditory systems briefly as these are some important aspects in developing a robust and effective speaker recognition system. Next, the section will give some reviews on the development and previous works in speaker recognition. Since there are a lot of researches have been done in speaker recognition area, this section will cover only the crucial parts that are mostly related to this particular research

### *A. Human Speech Production & Perception*

Speech signals can be described as a sequence of sounds that contain information that is useful for communication purposes [9]. Voice is produced when air is expelled from lungs to set up a vibration in vocal cords via the trachea. Air pressure is then set up vibration at the larynx and produces quasi-periodic pulses (frequency spectrum) due to the act of opening and closing the glottis. The frequency spectrum is then shaped by nasal and vocal tract. The end result of all processes is a longitudinal pressure waves from mouth which produce a range of sounds determined by the positions of jaw, tongue, lips, and mouth [9].

From the receivers' point of view, human hearing and perception systems are involved. Some key findings of the research in human speech and sound perceptions are highlighted in [9] as follows:

(i). Frequency is perceived as pitch on a non-linear frequency scale. (ii). Then secondly loudness is perceived on a compressive amplitude scale that rapidly becomes logarithmic above 1000 Hz. (iii). Thirdly auditory masking is a key

component in sound perception that provides robustness against noise and other interfering signals.

The acoustic waves produced by a speaker are first converted to a neural representation by the outer, middle and inner ear of the listener. Next, the neural representation of the waveform is then transmitted to the brain through the auditory nerve which is the resultant of the nerve firings at the hair cells of the inner ear. Finally, the brain will process the neural input and produce the perceived sound. The speech communication between a speaker and a listener is comprised of the processes of speech production, auditory feedback to the speaker, speech transmission to the listener either through the open air or over an electronic communication medium, and lastly the speech perception by the listener through the brain functions [9]. The processes are very efficient in understanding the speech and recognizing the speaker as well. Therefore, by achieving a good understanding of how human produce and perceive sound and speech, the more robust and efficient speaker recognition system can be designed and implemented.

### *B. Speaker Recognition System*

Voice or speaker recognition is a type of speech processing technologies. Speech processing can be classified into few categories where the most three distinct categories are speech recognition, speaker recognition and speech synthesis [4]. The difference between speech recognition and speaker recognition is that speech recognition recognises the spoken words while speaker recognition recognises the speaker identity or the person.

Speaker recognition can be further classified into two categories: speaker identification and speaker verification [1]. Speaker identification is a one-to-many process where the system identifies the speaker as one of the registered speakers. On the other hand, speaker verification is a one-to-one process where the system verifies the speaker as the one that he or she claimed to be [10]. The recognition process can be implemented to the known speakers (closed-set) or unknown speakers (open-set) where some impostors could come and trick the system. Some speaker recognition systems require cooperation from the speakers to speak a certain text which the speaker has prior knowledge on what to say, while some systems allow the users to utter any words or phrases. The former is called text-dependent system while the latter is called text-independent system [10].

In the past, researches on speaker recognition technology focused on distinguishing voices using manual human intervention. The first type of experiment in automatic speaker recognition (ASR) was called voice-print analysis in the 1960s [4]. The term 'voiceprint' is similar to the fingerprints where both methods use unique characteristics of human beings. Voice-print analysis was a semiautomatic process whereby human experts were used to visually examine the graphical representation of voice signal called speech histogram to compare the patterns. Since this method has many flaws especially it used subjective human judgment, a number of experiments were implemented in the subsequent years to address some of the discrepancies of the voiceprint analysis. The researches in speaker recognition became more successful in the mid 1980s due to the advancement of technology and computer hardware [4]. Since mid 1980s, speaker recognition has been widely used in commercial applications such as biometric solutions to control access to information, restricted areas, computer system and private services.

Generally, there are five main stages in the automatic speaker recognition (ASR) system. They are signal data acquisition, feature extraction, pattern matching, decision, and enrolment [1]. In the first step, speech sample is obtained using a microphone, and then the analogue signals are converted to digital signals so that the signals can be processed more efficiently. The unwanted signal or noise need to be filtered out from the voice signal and it needs to be amplified properly. The next step is to distil the important characteristics of voice signal so that only important characteristics are used for the speaker modelling and enrolment processes. Fig. 1 illustrates the overall structure of an Automatic command interpretation system.

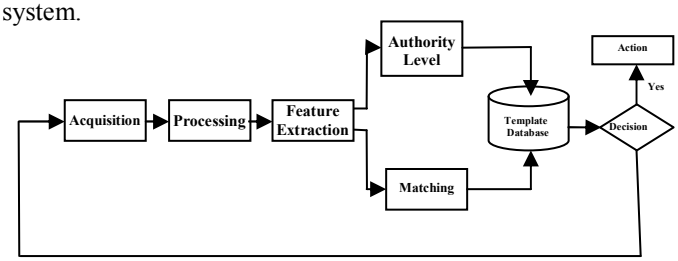


Figure 1. Automatic command interpretations and authentication system

ASR system can be further divided into two main stages: training and identification. In training stage, the important characteristics obtained from the distilling process are processed and common characteristics are obtained to produce reference templates for the system. On the other hand, in the identification stage, the characteristics obtained from the distilling process are compared with reference templates to find the best match [1]. The binary decision is then made by the decision block either to accept or reject the user. If there is not enough confidence level, the system might request additional input by providing a prompt message to the user. More explanations on the implementation of an ASR system are provided in the following sections.

### *C. Design Tradeoffs*

In order to determine the effectiveness and reliability of a recognition system, the most commonly adopted metrics are used: false acceptance rate and false rejection rate.

1. False acceptance rate (FAR) can be defined as the percentage of the system rejecting authorised individuals. Therefore,

FAR = (Number of false acceptance/ Number of trial) x100%

2. False rejection rate (FRR) can be defined as the percentage of the system accepting unauthorised individuals. Therefore,

FRR = (Number of false rejection/ Number of trial) x 100%

The main goal of an effective recognition system is to have both low FRR and low FAR. The trade-off between FAR and FRR can be adjusted by a certain threshold so that both rates are acceptable [5].

Other tradeoffs in developing a speaker recognition system are the processing speed and the data storage requirements [5]. Both aspects are very important characteristics of such system. No matter how accurate the system is, it must be prompt otherwise it will be useless. Plus, huge data storage requirement is undesirable since it can incur a huge cost for the whole system.

Therefore, in order for the system to be effective and reliable, all these aspects must not be overlooked otherwise the system will not work as desired.

### D. Speech Parameters

From the speech production system point of view, it is observed that speech signals are generated from quasi-stationary processes. Based on this knowledge, short-term spectral analysis can be applied to the short speech segments, which results in a sequence of short-time spectra. The spectral envelopes of speech signals are characterized by the vocal tract resonance frequencies, vocal tract length, and spatially varied cross-section areas [5]. This short time spectrum can be further classified into feature vectors which differ among individuals using few analysis techniques.

One of the most popular techniques is called linear prediction (LP) analysis [5]. LP analysis is performed to extract the feature vectors which are known as LP coefficients (LPCs) from the speech waveforms. LPCs contain the information of the formant frequencies and their bandwidth. Other sets of feature vectors can also be derived from LP analysis such as impulse response, autocorrelation coefficients, and cepstral coefficients [5]. Among all these features, cepstral based coefficients (LPCCs) are the most commonly used and these are the effective features for speaker recognition. The characteristics of LPCCs are the distance between LPCCs can be easily computed using simple Euclidean distance measure, the coefficients are almost orthogonal and they are independent to each other [4, 5].

Other than LP analysis, speech features can also be derived from the analysis that is based on the auditory-based principles. According to Davies and Mermelstein (1980) , mel-scale filter banks are used to extract spectral features from speech signals [2]. The resulting coefficients are called Mel-frequency cepstral coefficients (MFCCs). Other than MFCC, research has found that based on the psychoacoustics or the study of the way human perception and physiology effects sound, the human auditory system can be approximated by a set of overlapped band-pass filters whose frequencies follow a critical band scale which is more adapted to human hearing [3].

### E. Speaker Modeling

There are many speaker modelling techniques have been proposed in recent years [1]. Speaker modelling is used to generate a model or pattern for each speaker and to produce a database of information for all trained speakers for the recognition purposes. The speaker modelling techniques are chosen based on the type of speech to be used, the expected performance, the ease of training and updating, and storage and computation considerations [10]. This section describes four most popular speaker modelling techniques for comparison purposes.

#### 1) Template Matching

Template or pattern matching is used mainly for text-dependent system. It requires templates as reference for the speaker models. These templates are composed of a sequence of feature vectors extracted from a fixed phrase uttered by the speaker. During verification, a match score is produces by dynamic time warping (DTW) to align the test phrase with the reference templates [10]. The disadvantage of this technique is that the template cannot represent the extensive variability of the speech signals.

#### 2) Vector Quantization

Vector Quantization (VQ) is introduced by Soong et al. [11] to convey signals at low bit rate. VQ produces set of codebooks to the large vectors by mapping the large vector space obtained from the feature extraction methods into a smaller region called cluster. The centroid or codeword of each cluster will produce a set of codebook for each speaker. Recently, VQ has been a standard in speaker recognition systems due to some advantages such as low computational cost and it is more efficient than template matching technique [5].

#### 3) Hidden Markov Model

Hidden Markov Model (HMM) is a modelling technique that encodes both the temporal evolution of feature sequences and the statistical variation of the features [5]. The parameters for HMM are estimated from the speech signal using established automatic algorithms [5]. This technique can be used for both text-dependent and text-independent applications. For text-dependent applications, whole phrases or phonemes may be modelled using multi-state left-to-right HMMs while for text-independent applications, single state HMMs, also known as Gaussian Mixture Models (GMMs) are used [10].

#### 4) Neural Networks

Neural Networks (NN) is a technique that learns the complex mappings between data in the input and output space and it can be considered as supervised classifiers [5]. Neural network is commonly used in the system that the statistical distributions are not known in advance. There are some forms that this particular technique can have such as multi-layer perceptions, radial basis functions, the combination of those two forms and etc [10]. The advantage of this method is that the discriminative information can be found easily by using the supervised learning. However, the disadvantage of this method is that it has longer training time and hence can be computationally expensive technique.

## III. METHODOLOGY

This section presents the approach and exposition of this research. The section is divided into three sub-sections namely front-end processing, speaker modelling and speaker recognition. Each sub-section explains research analysis and algorithms to find the best solutions to simulate a reliable and robust voice control access system.

### A. Front-end Processing

The ability of human perceptions system to successfully identify someone's identity by hearing his/her voice has motivated researchers to develop robust and reliable speaker

recognition systems. The aim for the researches in this field is to produce a system that capable of recognising and identifying people as accurate as human auditory system does. Many algorithms have been developed to extract only speaker-related information from speech waveforms. Acoustic features can be obtained from two groups of features [4]:

1. Learned behavioural features such as pitch, accent, speaking rates and etc.

2. Anatomy such as shape of vocal tract and mouth.

From sub-section 2.4, it is known that there are many techniques based on the vocal tract information available for feature extraction analysis. This research is attempted to use two most popular techniques that are based on human auditory and perception systems which are the Mel-frequency cepstral coefficient (MFCC) and the Perceptual Linear Predictive (PLP) methods. These two methods are useful to extract the unique features of voice from a number of speakers and hence, allow the system to differentiate different speakers.

For speaker modelling and vector classification, the vector quantization with Linde Buzo-Gray (LBG) technique is used to map the large vector space obtained from the feature extraction methods into a smaller region by using codebooks [11]. The performance for each individual method are examined and compared. The overall implementation in the training stage is illustrated in Fig. 2.

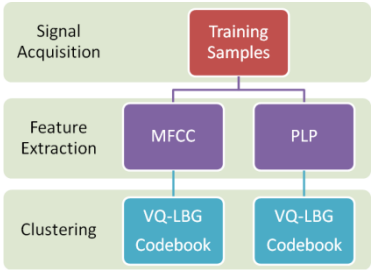


Figure 2. Implementation

## B. *Mel-Frequency Cepstral Coefficients (MFCCs)*

MFCC is one of the most popular feature extraction method used in automatic speech and speaker recognition systems. The aim of MFCC is to mimic the behaviour of the human auditory system based on the known variation of human ear's critical bandwidth with frequencies in its filters.

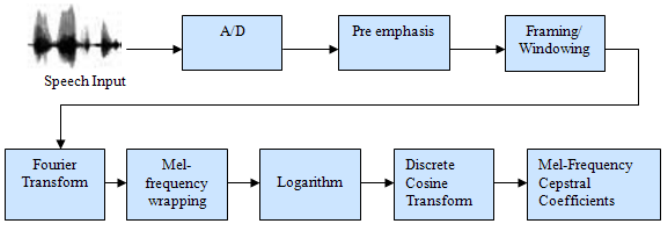


Figure 3. Mel-frequency Cepstral Coefficient Processor

Studies have shown that human auditory system does not follow a linear system [8]. Therefore, in order to mimic human auditory system, the frequency needs to be mapped according to 'Mel-scale' which uses filters that are spaced linearly at low frequencies (below 1000 Hz) and logarithmically at high frequencies (above 1000 Hz) [2]. Fig. 3 illustrates the overall process of obtaining Mel-frequency coefficients.

Figure 4 illustrates the triangular filter banks that are uniformly spaced on the Mel-scale. To convert the frequency in HZ to Mel scale, the following formula (1) is used [7]:

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

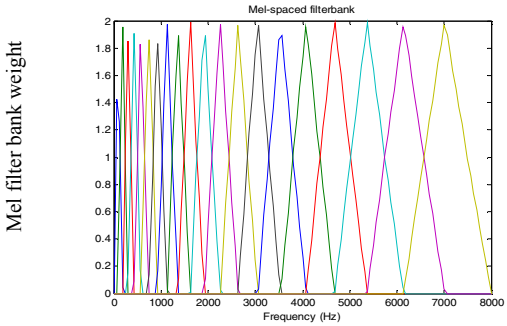


Figure 4. Mel-spaced triangular shape filterbank

From Fig. 4, it can be seen that the filter bank has a triangular band-pass frequency response. The number of Mel spectrum coefficients, $K$, is typically chosen as 20. At this stage, each filter acts as a histogram bin where all bins are overlap in the frequency domain.

## C. *Perceptual Linear Predictive (PLP)*

The second type of feature extraction technique which is used in this research is called perceptual linear predictive (PLP). This method is proposed by Hermansky and this is the most popular and successful method of auditory model [3]. The aim of PLP is to implement a speech spectrum analysis that is similar to the human auditory system [9]. A block diagram of PLP is shown in Fig. 5. PLP is more adapted to human hearing because it uses three psychophysically based transformations to modify the spectrum [3]. The three concepts are [3]:

(i). Critical band spectral resolution: Trapezoidal shaped filters with non-linear (Bark) frequency scale are used to perform critical band spectral analysis.

(ii). Equal loudness curve: Approximation of the unequal sensitivity of human hearing to the different frequency components of the signal.

(iii). Intensity loudness power law: Use of cubic root compression of the spectral level to approximate the non-linear relationship between sound intensity and perceived loudness

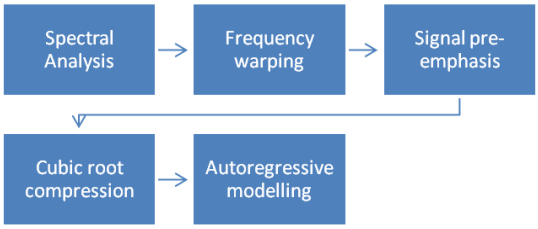


Figure 5. Perceptual Linear Predictive Processor

As shown in the Fig. 5, the first step to obtain the PLP coefficients is to perform spectral analysis to the speech signals. Firstly, the speech signal is framed and windowed using the available window function whereby in this research hamming window is used. The frame length of 25ms with 10ms time shift is used so that the useful data of the spectrum can be analysed. The next step is to transform the windowed signal into frequency domain using the Fast Fourier Transform (FFT).

Next, critical band analysis is used to warp the spectrum in radial frequency (ω) into the Bark frequency (Ω) using the following Bark-scale warping function [3]:

$$\Omega(\omega) = 6\ln\left\{\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right\} \quad (2)$$

All of the above processes are similar to the MFCC operations except the following processes which apply the three engineering approximations of human hearing process. The warped signal is then integrated within the critical-band bands. The next process is to compensate the unequal frequency sensitivity of human hearing by pre-emphasis the resulting spectrum in time-domain with first-order high pass filter.

Next step is to perform cubic root compression to the resulting spectrum. This step is performed to approximate the power law of hearing and the non-linear relationship between the intensity of sound and perceived loudness. The final step is to perform inverse FFT (IFFT) to the resulting spectrum in order to obtain a set of cepstral coefficients.

### D. Speaker Modelling

The next important step in developing a speaker recognition system is to generate a model or pattern for each speaker. At this stage, a database of information for all trained speakers is developed for the recognition purposes. Speaker modelling constitutes two important steps which are the distance measurement and the clustering process.

#### 1) Distance Measure

There are many types of distance measurement algorithm have been developed in speech processing technology. The most well known algorithms are the Euclidean and Manhattan distance to compute the distance between two vectors. These two types of distance measures can be computed using the following formulas [4]:

Euclidean distance: $d(a,b) = (\sum_{i=1}^{p}(a_i - b_i)^2)^{1/2}$ (3)

Manhattan distance: $d(a,b) = \sum_{i=1}^{p}|a_i - b_i|$ (4)

where $a_i$ and $b_i$ are the two vectors or set of vectors to be computed and p is the total number of features to compare. In this research, the Euclidean distance is used to compute distortions between vectors.

#### 2) Vector Quantization (VQ)

Computing the distortion between large numbers of vectors is not a practical approach for a speaker recognition system since the quality and reliability of a system are based on the computational time taken for the system to work. Therefore, there is a need to reduce the number of vectors without having reducing the efficiency of such system. Vector quantization is a lossy data compression method. This step is also important in the recognition stage where the template database is compared with the new provided speech data.

In the prototype system, VQ mapped the large vector space obtained from the feature extraction methods into a smaller region called cluster. The centroid or codeword of each cluster will produce a codebook for each speaker. By clustering the codebook, the difference between each speaker can be identified. The Fig. 6 shows how different speaker vectors are clustered using centroids of vectors.

In this research, a well known iterative algorithm developed by Linde, Buzo and Gray is used in VQ generation codebook [6]. The proposed algorithm requires an initial codeword and uses splitting method to produce two codewords as the initial codebook. Minimisation of Euclidean distance is used in computing the distortion between vectors and the centroid. The iterations of creating a set of codebook repeated until the average distortion is less than the threshold.

### E. speaker recognition

In order for the recognition to be more efficient, testing sample voices are also obtained using the same microphone as the training stage. This is to avoid any dissimilarity between different signal acquisition methods. In the authentication part, unknown voice will be compared with the template database to determine whether the voice belongs to the authorised person or not. The same processing was implemented to the test sample where the important characteristics from test voice signal are extracted and codebooks are generated. The generated codebooks are then compared with the set of trained codebooks in the template database. Distortion measurement is also done using Euclidean distance to find the minimum distance between test data and the database. The pseudo-code of the implementation of the recognition stage is: Extract features from voice signal; Create codebooks for test speaker; Compare test codebooks with codebooks in database; Find minimum Euclidean distance; If minimum distance less than threshold; Then give score; Else provide command prompt; Else end the system.

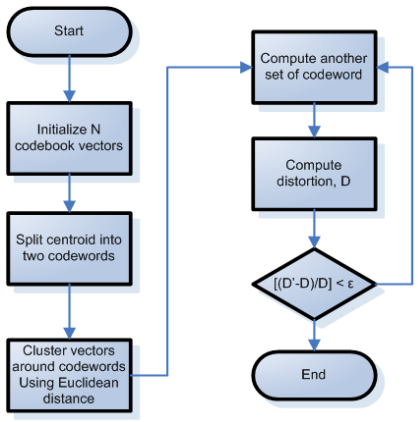


Figure 6. VQ-LBG flow chart

Implementation In training stage, 10 voice samples were recorded from 12 different speakers using Philips microphone

which is commonly used with windows audio applications. A personal laptop of 2.00 GHz Intel Core 2 Duo processor is used as the platform to implement the proposed system. The sampling rate of 16 kHz is used to record the voices using MATLAB functions and toolboxes. According to the Nyquist theorem, speech is recorded using sampling rate above 10 kHz to minimise the effect of aliasing [9]. Each speaker uttered their full name for 10 for duration of 3 seconds each. Then, the feature extraction algorithms proposed for the hands-free access control system were tested on each sample voices.

The decision threshold for both methods is determined based on the equalization of the false acceptance rate and false rejection rate. This means that the threshold are computed based on the distribution of inter and intra speaker distances according to the Euclidean distance and the chosen threshold must be able to produce optimal (minimum) error rates. The computation of FRR and FAR are as follows:

FRR = (Number of false rejection/ Number of authorised person) x 100%

FAR = (Number of false acceptance/ Number of unauthorised person) x100%

## IV. Results and Discussion

The effectiveness of the proposed method during training is measured based on the execution time and the classification rate while during testing or recognition stage, the performance was measured based on the false acceptance rate (FAR) and false rejection rate (FRR). Table 1 shows the execution time and classification rates obtained.

TABLE I. Results for MFCC and PLP methods

| Parameter | MFCC | PLP |
|---|---|---|
| Execution time for training (seconds) | 84.9318 | 12.2382 |
| FAR (%) | 2.5% | 0.83% |
| FRR (%) | 8.33% | 9.17% |

Based on the experiment, both feature extraction methods produced 100% classification rates which means that the methods were able to recognise and identify the speaker based on their speech. PLP method requires lower storage capacity since it only produced 2682 vectors for one sample for each speaker. A greater data storage requirement is needed for the MFCC since it produced 9560 for one sample for each speaker. PLP also produced shorter computation time if compared to the MFCC. Therefore, PLP is seen more efficient in term of the data storage requirement and execution time for the same classification rate.

There was tradeoff between the rates of rejecting authorised speaker and accepting unauthorised speaker. The rates were basically depending on the threshold value that was set based on the average distance between codebooks. In term of FRR, PLP has higher value than MFCC while lower value for FAR. The execution time for the combination of PLP-MFCC is greater but it still worthwhile since the false rejection rate is decreased and false acceptance rate is increased.

## V. Conclusions

This research analyzed and simulate human voice based robot command system for disaster management and dangerous environment access by comparing the performance and recognition rate of different feature extraction methods namely Mel-frequency Cepstral Coefficient (MFCC) and Perceptual Linear Predictive (PLP) and also the combination of both methods. Vector Quantization (VQ) with Linde-Buzo-Gray (LBG) methods was used as the speaker modeling for classification purpose. Based on the results of first phase of research, there was design trade-off between the chosen threshold which affected the false rejection and false acceptance rates. The initial study also revealed that both feature extraction methods were successful in identifying and recognizing the speakers based on their voice. The PLP feature model has lower computational time and data storage requirements. The combined approach (MFCC-PLP) did not show considerable improved performances to the individual feature models PLP and MFCC without incurring high computational costs that will compromise the performance of the speaker recognition tasks.

In future work as the second phase PLP method will be deployed to carry out substantial experimental works with a database size of 25 commanders and the robustness against the background noise.

## References


[1] Campbell, J. P., Jr. (1997). Speaker recognition: a tutorial. Proceedings of the IEEE, 85(9), 1437-1462.

[2] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4), 357-366.

[3] Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis For Speech. J. Acoustical Society of America, pp. 1738-1752.

[4] Klevans, R. L., & Rodman, R. D. (1997). Voice Recognition. Boston: Artech House, Inc.

[5] Kung, S. Y., Mak, M. W., & Lin, S. H. (2004). Biometric authentication : a machine learning approac. Upper Saddle River, NJ: Prentice Hall Professional Technical Reference.

[6] Linde, Y., Buzo, A., & Gray, R. (1980). An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications, 28(1), 84-95.

[7] Molau, S., Pitz, M., Schluter, R., & Ney, H. (2001). Computing Mel-frequency cepstral coefficients on the power spectrum. . Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. (ICASSP '01). 2001.

[8] Rabiner, L. R., & Juang, B. H. (1993). Fundamentals of Speech Recognition. Englewood Cliffs, N.J.: Prentice-Hall.

[9] Rabiner, L. R., & Schafer, R. W. (2011). Theory and applications of digital speech processing (1st ed ed.). Upper Saddle River: Pearson.

[10] Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002, (ICASSP '02).

[11] Soong, F., Rosenberg, A., Rabiner, L., & Juang, B. (1985). A vector quantization approach to speaker recognition. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '85.