Classification of Language Speech Recognition System

Khin May Yee[1], Moh Moh Khaing[2], Thu Zar Aung[3]

[1]Faculty of Computer Systems and Technologies Computer University, Myitkyina, Myanmar
[2]Information Technology Department, Technological University, Taunggyi, Myanmar
[3]Information Technology Department, Government Technological High School, Meiktila, Myanmar

How to cite this paper: Khin May Yee | Moh Moh Khaing | Thu Zar Aung "Classification of Language Speech Recognition System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.946-951, **https://doi.org/10.31142/ijtsrd26546**

IJTSRD26546

ABSTRACT

This paper is aimed to implement Classification of Language Speech Recognition System by using feature extraction and classification. It is an Automatic language Speech Recognition system. This system is a software architecture which outputs digits from the input speech signals. The system is emphasized on Speaker-Dependent Isolated Word Recognition System. To implement this system, a good quality microphone is required to record the speech signals. This system contains two main modules: feature extraction and feature matching. Feature extraction is the process of extracting a small amount of data from the voice signal that can later be used to represent each speech signal. Feature matching involves the actual procedure to identify the unknown speech signal by comparing extracted features from the voice input of a set of known speech signals and the decision-making process. In this system, the Mel-frequency Cepstrum Coefficient (MFCC) is used for feature extraction and Vector Quantization (VQ) which uses the LBG algorithm is used for feature matching.

KEYWORDS: *Text-dependent, speech identification*

1. INTRODUCTION

Speech recognition is useful as a multimedia browsing tool. It allows us to easily search and index recorded audio and video data. Speech recognition is also useful as a form of input. Speech contains information about the identity of the speaker.

A speech signal also includes the language that is spoken, the presence and type of speech pathologies, the physical and emotional state of the speaker. Often, humans are able to extract the identity information when the speech comes from a speaker acquainted with.

The recording of the human voice for speaker recognition requires a human to say something. In other words, human has to show some of the speaker speaking behavior. Therefore, voice recognition fits within the category of behavioral biometrics. A speech signal is a very complex function of the speaker and his environment that can be captured easily with a standard microphone. In contradiction to a physical biometric technology such as fingerprint, speaker recognition is not fixed and static and does not have physical characteristics. In speaker recognition, there is only information depending on an act. The state of the art approach to automatic speaker verification (denoted as ASV) is to build a stochastic model of a speaker based on speaker characteristics extracted from the available amount of training speech. In speaker recognition, there are differences between low-level and high-level information. High level-information is valued like a dialect, an accent, the talking style and subject manner for context. These features are currently only recognized and analyzed by humans.

The low-level information is denoted like pitch period, rhythm, tone, spectral magnitude, frequencies, and bandwidths of an individual's voice. These features are used by speaker recognition systems. Voice verification works with a microphone or with regular telephone handset although the performance increase with higher quality captures devices. The hardware costs are very low because today nearly every PC includes a microphone or it can be easily connected. However, voice recognition has got its problems with persons who are husky or mimic with another voice. If this happens, the user may not be recognized by the system. Additionally, the likelihood of recognition decrease with poor-quality microphones if there is background noise. Voice verification will be a complementary technique for example, finger-scan technology as many people see finger recognition technology as a higher authentication form. In general, voice authentication has got a high EER; therefore it is generally not used for identification. The speech is variant in time, therefore adaptive templates or methods are necessary.

This paper is organized as follows: related words are an investigation in section 2. Speaker identification algorithms are described in section 3. In section 4, the proposed system design is presented. In section 5, test and result for text-dependent speaker identification are mentioned. Finally, in section 6, the paper has been concluded.

RELATED WORKS

Speaker Identification is the process of finding the identity of an unknown speaker by comparing his or her voice with voices of registered speakers in the database, was presented by Markowitz 2003 [1]. In this system, pitch provides very important and useful information for identifying speakers. In the current speech recognition systems, it is very rarely used as it cannot be reliably extracted, and is not always present in the speech signal. An attempt is made to utilize this pitch and voicing information for speaker identification. Ran D.Zilca [2] studied Text-Independent. F. K. Soong, A. E. Rosenberg and B. H. Juang [3] were described as a vector quantization approach to speaker recognition. A. E. Rosenberg and F. K. Soong [4] expressed recent research in automatic speaker recognition.

Text-Independent Speaker Verification was described by Gintaras Barisevicius. In this system, the applications for text-independent verification system are vast: starting with telephone service and ending up with handling bank account.

SPEAKER IDENTIFICATION STEPS

The most important parts of a speaker recognition system are the feature extraction and classification methods. The aim of the feature extraction step is to strip unnecessary information from the sensor data and convert the properties of the signal, which are important for the pattern recognition task to a format that simples the distinction of the classes.

Usually, the feature extraction process reduces the dimension of the data in order to avoid the curse of dimensionality. The goal of the classification step is to estimate the general extension of the classes within feature space from the training set.

A. SPEECH PARAMETERIZATION (FEATURE EXTRACTION)

Feature extraction: Before identifying any voices or training person to be identified by the system, the voice signal must be processed to extract important speech characteristics, the amount of data used for comparisons is greatly reduced and thus, less computation and less time is needed for comparisons. The steps used in feature extraction are frame blocking, windowing, Fast Fourier transform, Mel-frequency wrapping and cepstrum Figure 1. *Frame blocking and windowing:* In this step, the signal is put into frames (each 256 samples long). This corresponds to about sound per frame. Each frame is then put through a hamming window. Windowing is done to avoid problems due to truncation of the signal. The hamming window has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \qquad (1)$$

Wher N=256 is the length of the frame.

The system of home meter reading is composed of a control terminal in distance, GPRS module and user metering module is shown in Figure 2.

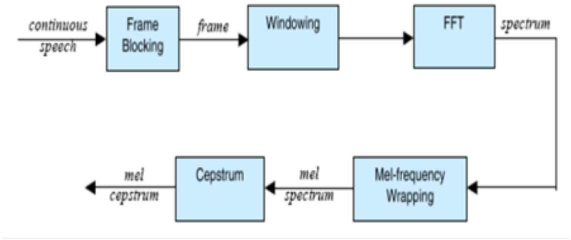


Fig.1 Block diagram of the Mel-frequency Cepstrum Coefficient (MFCC) processor

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}, n = 0,1,2,..., N-1 \qquad (2)$$

Where, j is the imaginary unit, , i.e. j = $\sqrt{-1}$.

Mel-frequency wrapping: In the fourth step, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The meal-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 meals. Therefore it can be used the following approximate formula to compute the models for a given frequency, f, in Hz:

$$\text{Mel}(f) = 2595 * \log 10(1 + f / 700) \qquad (3)$$

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel-frequency component. That filter bank has a triangular bandpass frequency response, and the spacing, as well as the bandwidth, is determined by a constant meal-frequency interval. The modified spectrum of S(w) thus consists of the output power of these filters when S(w) is the input. The number of mel spectrum coefficients, K, is typically chosen as 20. This filter bank is applied in the frequency domain, therefore it simply amounts to taking those triangle-shaped windows on the spectrum. A useful way of thinking about this mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

Cepstrum: In this final step, it converts the log mel spectrum back to time. The result is called the Mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, it can be converted to the time domain using the Discrete Cosine Transform (DCT). Therefore if it denotes those mel power spectrum coefficients that are the result of the last step are $\tilde{S}_k, k = 1,2,..,K,$ the MFCC's, $\tilde{c}_n$, can be calculated as follow:

$$\tilde{C}_n = \sum (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right], \qquad (4)$$

n=1,2,...,K

Note that it excludes the first component, $\tilde{c}_0$, from the DCT since it represents the mean value of the input signal which carried little speaker-specific information.

PATTERN MATCHING AND CLASSIFICATION

Speaker identification is basically a pattern classification problem preceded by a feature extraction stage [5]. Given a sequence of feature vectors representing the given test utterance, it is the job of the classifier to find out which speaker has produced this utterance [6]. In order to carry out this task, the acoustic models are constructed for each of the speakers from its training data.

In the classification stage, the sequence of feature vectors representing the test utterance is compared with each acoustic model to produce a similarity measure that relates the test utterance with each speaker. Using this measure, the speaker identification system recognizes the identity of the speaker.

There exists a lot of model for classification: Template models and statistical models. Templates models used Dynamic Time Wrapping (DTW) and Vector Quantization (VQ) models. Statistical models include a Gaussian Mixture Model (GMM), Hidden Markov Models (HMM) and Artificial Neural Network (ANN) [7, 8]. Vector quantization methods were outlined and an example of such classification was displayed.

1. Dynamic Time Wrapping (DTW)

In this paper, Vector Quantization (VQ) method was used for classification so Dynamic Time Wrapping (DTW) method didn't discuss in detail. The main idea of this approach is that training template T consisting of N_T frames and test utterance R consisting of NR frames, the Dynamic Time Wrapping model is able to find the function m=w(n) , which maps the time axis n of T to time axis m of R.

Thus, the system makes the comparison between the test and training data of the speaker evaluating the distance between them and makes the decision whether in favor of the user identify or not identify [8].

2. Vector Quantization (VQ)

Vector Quantization (VQ) method could be to use all the feature vectors of a given speaker occurring in the training data to form this speaker's model. However, this is not practical as there are too many feature vectors in the training data for each speaker. Therefore, a method of reducing the number of training vectors is required.

It is possible to reduce the training data by using a Vector Quantization (VQ) codebook consisting of a small number of highly representative vectors that efficiently represent the speaker-specific characteristics [2, 9]. Note that the VQ-based classifiers were popular in earlier days for text-dependent speaker recognition.

There is a well-known algorithm, namely, Linde, Buzo and Gray (LBG) algorithm, for clustering a set of L training vectors onto a set of M codebook vectors. Each feature vector is the sequence X is compared with all the stored codeword in the codebook and the codeword with the minimum distance from the feature vectors is selected as a proposed command.

For each codebook, a distance measure is computed, and the command with the lowest distance is chosen.

$$d(x, y) = \sqrt{\sum_{j=1}^{k}(x_i - y_{ij})^2} \qquad (5)$$

The search of the nearest vector is done exhaustively, by finding the distance between the input vector X and each of the codewords C1-CM from the codebook. The one with the smallest distance is coded as the output command.

PROPOSED SYSTEM DESIGN

Figure 2 shows the basic structure of the speaker identification system. It is two-phase in this system. The first phase is the training phase. In this phase, the voices are recorded.

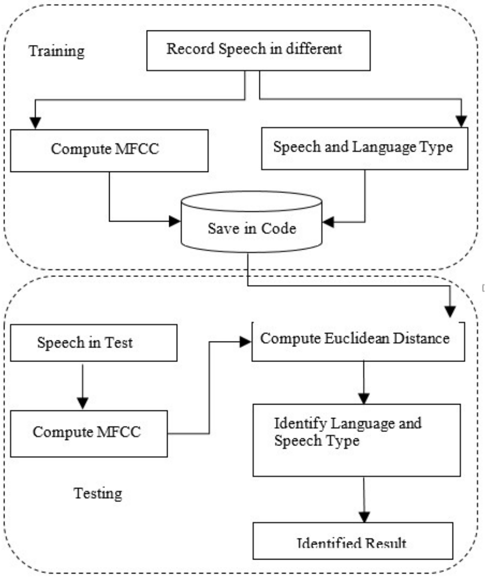


Figure2. The basic structure of the speaker identification system

The recorded voices are then extracted. The features extracted from the recorded voices are used to develop models of the languages. The second phase in the system is testing. In this phase, the entered voice is processed and compared with the languages model to identify the language.

In the training part, there are Pre-processing, Feature Extraction and Speaker Modeling. To get the feature vectors incoming voice, pre-processing will be performed. For feature extraction, Mel-frequency Cepstral Coefficient (MFCC) algorithm is used. For codebook construction, Vector Quantization (VQ) algorithm is used. After extracting the features, speaker models are built and then save to the database. In the testing part, features of speech signals are extracted and matched it to the speaker model in the database. And then make a decision based on the minimum distance between the input pattern and speaker model. To decide, a threshold is used with each speaker. If the speaker minimum distance if lower than this threshold, then the language is identified. However, if the speaker minimum distance exceeds this threshold, the language is not identified. Figure 3 shows the flow chart of the testing process.

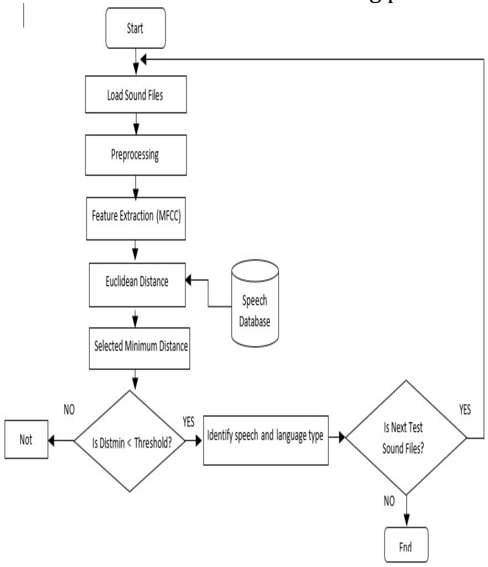
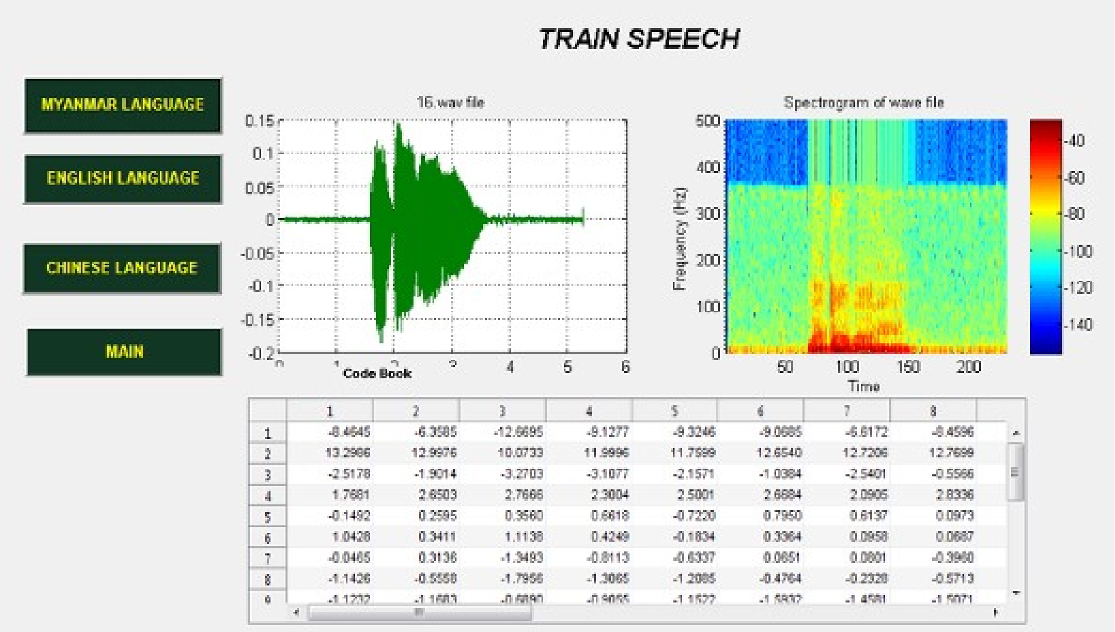

Start
Load Sound Files
Preprocessing
Feature Extraction (MFCC)
Euclidean Distance
Speech Database
Selected Minimum Distance
NO
Not
Is Distmin < Threshold?
YES
Identify speech and language type
Is Next Test Sound Files?
YES
NO
End

Figure3. Flowchart of the speech recognition system for the testing phase

TEST AND RESULTS

Firstly, the speech signal is recorded and saved as wave files for the person.

In this step, voice signal is saved as wave files in the trained data set to match the input voice files from the user. So, the user can see the wave files which are saved in the train data set by clicking any language push button. If the user chooses the Myanmar language button, the feature database display appears as shown in Figure 4.



	1	2	3	4	5	6	7	8
1	-8.4645	-6.3585	-12.6695	-9.1277	-9.3246	-9.0685	-6.6172	-8.4596
2	13.2986	12.9976	10.0733	11.9996	11.7599	12.6540	12.7206	12.7699
3	-2.5178	-1.9014	-3.2703	-3.1077	-2.1571	-1.0384	-2.5401	-0.5566
4	1.7681	2.6503	2.7666	2.3004	2.5001	2.6684	2.0905	2.8336
5	-0.1492	0.2595	0.3560	0.6618	-0.7220	0.7950	0.6137	0.0973
6	1.0428	0.3411	1.1138	0.4249	-0.1834	0.3364	0.0958	0.0687
7	-0.0465	0.3136	-1.3493	-0.8113	-0.6337	0.0651	0.0801	-0.3960
8	-1.1426	-0.5558	-1.7956	-1.3065	-1.2085	-0.4764	-0.2328	-0.5713
9	-1.1232	-1.1683	-0.6890	-0.9055	-1.1522	-1.5932	-1.4581	-1.5071

Figure4. Feature database display of Myanmar language

Table 1 shows the parameters of code minimum and maximum values for each language. The data are calculated as average for each type of languages. The recording duration also depends on the test person ascend. The threshold value for distance vector is defined by checking and testing many times according to the recording database.

TABLE I COMPARISON OF PARAMETERS

Language Type	Min Codebook value	Max Codebook value	Frequency	Average Duration	Euclidean distance (threshold)
English language	-14.5934	13.2986	80-130 Hz	4 sec	2
Myanmar language	-29.2414	7.8993	85-150 Hz	5 sec	2
Chinese language	-18.1614	12.5227	100–190 Hz	4.5 sec	2

Recognition is to test the unknown speaker; the steps are open speech file, recognize the language, play wave file and main as shown in Figure 5.

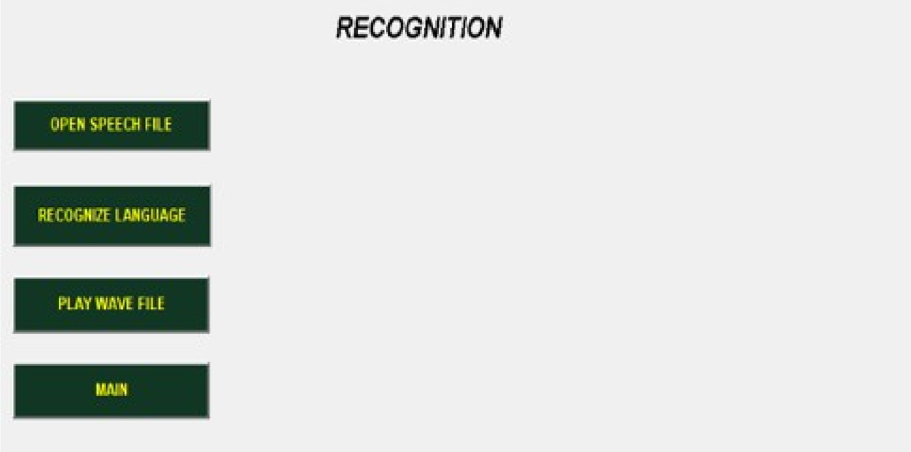
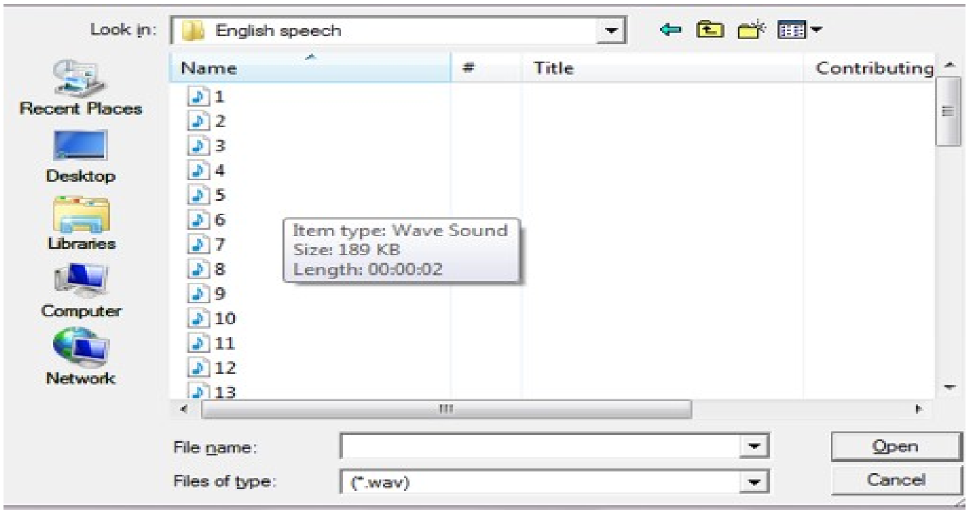


Figure5. Recognition dialog box

The open speech file button shows the wav files that are saved. If the user selects a wav file want to recognize, the speech file is loaded as shown in Figure 6 and 7.



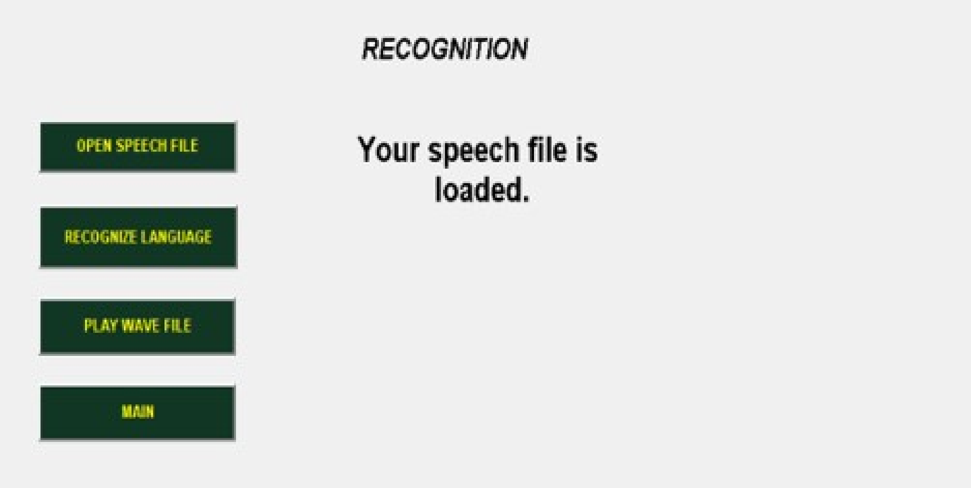
Look in: English speech

Name	#	Title	Contributing
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			

Item type: Wave Sound
Size: 189 KB
Length: 00:00:02

File name:
Files of type: (*.wav)

Figure6. Open voice dialog box

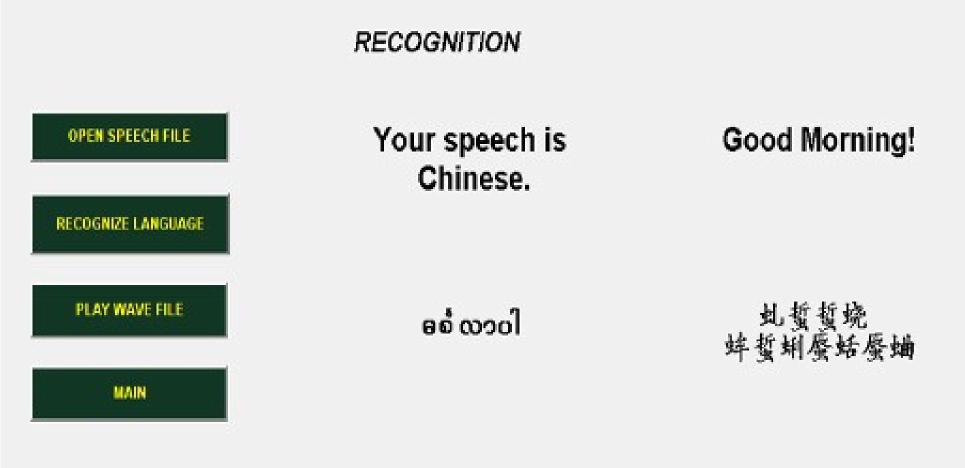


RECOGNITION

Your speech file is loaded.

Figure7. Loaded speech file dialog box

The recognize language button identifies the result as a text of language such as Myanmar, English and Chinese as shown in Figure 8. The result was identified as a voice if the user has to click the play wav file button. The main button is going back to Main Window of the Language Speech Recognition System based on Feature Extraction and Classification.



RECOGNITION

Your speech is Chinese.

Good Morning!

မင်္ဂလာပါ

虬蜇蜇蛲
蚌蜇蜊蜃蛞蜃蛐

Figure8. Result screen of recognized languages as a text

The accuracy of the system is calculated with 25 train files for each language per person. Five persons are recorded to get precise accuracy in each language. So, totally 125 files are recorded for each language. For testing, the untrained files are also checked 4 files for each language per person. Different persons are also tested and the result is shown in table II. The known trained persons have achieved 100% accuracy. But the untrained persons have got low accuracy due to the ascent of their voice and the level of voice.

The Chinese language is most difficult to recognize for untrained persons. The style of speaking and the wood frames are different from other languages. The accuracy can also change according to the test person is whether the native speaker or not

TABLE II IDENTIFICATION ACCURACY OF THE SYSTEM

LANGUAGE TYPE	No of train file	No of the Test file	% Correct for train file	% Correct for Test file
English language	125	100	100	60
Myanmar language	125	100	100	54
Chinese language	125	100	100	40

CONCLUSION

The goal of this paper was to create a speaker recognition system and apply it to the speech of an unknown speaker. By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify the unknown speaker. To evaluate the performance of the proposed system, the system trained the 125 wave files of the Myanmar, English and Chinese language and tested 100 wave file by untrained persons. The result is 100% accuracy in the trained speaker verification system by using MFCC (Mel Frequency Cepstral Coefficients). But the untrained speaker could not get the good accuracy, it got around 50 % accuracy depends on the speaker ascend. The function 'melcepst' is used to calculate the mel cepstrum of a signal. The speaker was modeled using Vector Quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. In this method, Linde, Buzo and Gray (LBG) algorithm, for clustering a set of L training vectors onto a set of M codebook vectors. In the recognition stage, a distortion measure which based on the minimizing the distance was used when matching an unknown speaker with the speaker database. During this paper, we have found out that the VQ based clustering approach provides us with the faster speaker identification process.

ACKNOWLEDGMENT

The author would like to thank all teachers, from the Department of Information Technology, Technological (Taunggyi), who gives suggestions and advice for her submission of the paper.

REFERENCES

[1] Markowitz, J. A and colleagues: *"J. Markowitz, Consultants",* (2003).

[2] Ran Zilca, D: *"Text-Independent Speaker Verification Using Utterance Level Scoring and Covariance Modeling",* (2002).

[3] F. K. Soong, A. E. Rosenberg, and B.H. Juang, *"A vector quantization approach to speaker recognition,"* AT & T Journal, vol, 66, no.2, pp. 14-26, 1987.

[4] A. E Rosenberg, and F. K. Soong, *"Recent research in automatic speaker recognition,"* in Advances in Speech Signal Processing, S. Furui, M. Sondhi, Eds. New York: Marcel Dekker Inc., pp. 701-737, 1992.

[5] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, New Jersey; Prentice-Hall, pp. 141-161.pp. 314-322, pp. 476-485, 1978.