DESCRIPTIVE STATISTICS METHOD IN ISOLATED MALAY DIGITS FEATURE EXTRACTION

S. A. Majeed, H. Husain, S. A. Samad and A. Hussain
Department of Electrical, Electronic and System Engineering
National University of Malaysia, Malaysia.

ABSTRACT

Nowadays, the use of speech recognition feature extraction methods are not optimal in terms of accuracy and speed when they are applied to a specific environment and recognition task. The performance of the speech recognition system depends on the feature extraction stage and classification stage. In this paper, descriptive statistics were used after feature extraction stage to minimize the amount of feature vector elements and to maximize the peak amplitude in isolated Malay Digit speech recognition. Artificial Neural Network (ANN) was used as classifier to evaluate these new feature vectors' representations. The obtained speech recognition rate was 96.67 %. Therefore, this method shows an improvement in the recognition rate.

Keyword: Feature extraction; MFCC; Speech recognition.

INTRODUCTION

Speech recognition can be approximately divided into two stages: feature extraction and classification. Feature extraction is defined as a step to minimize the dimensionality of the input data, a minimization which certainly causes to some information loss. In typical speech recognition systems, speech signals divide into frames and extract features from each frame. Through feature extraction, speech signals are changed into a sequence of feature vectors. Then these vectors are moved to the classification stage. Information loss during the transition from speech signals to a sequence of feature vectors must be kept to a minimum. (Lee et al. 2003). Feature extraction stage plays a vital essential role in the realization of speech recognition systems. This is the result of the fact that better signal feature extraction gives rise to better recognition performance (Anusuya & Katti 2011).

Feature extraction approaches are divided into production-based and perception-based methods. Linear predictive coding (LPC) is an example of the production-based approach while Mel-frequency cepstral coefficients (MFCC) and Perceptual Linear Prediction (PLP) belong to the perception-based approach's family. In speech recognition, a premium is placed on extracting

features that are somewhat invariant to changes in the speaker. Therefore feature extraction involves analysis of speech signal. MFCC feature extraction approach gives a good discrimination and a small correlation between components. MFCCs are one of the more popular parameterization methods used by researchers in the speech technology field. It has the benefit that it is capable of capturing the phonetically important characteristics of speech. Furthermore, band limiting can easily be employed to make it suitable for telephone applications.

For classification stage several approaches have been recommended such as Hidden Markov Models (HMM), Support Vector Classifiers, Artificial Neural Networks (ANN), which among them neural networks have proven to be very efficient (Dede & Sazli 2010). Different neural network's architectures have been used in recent years for the speech recognition task (Nakagawa 1995). A contemporary study on isolated Malay digit recognition applying dynamic time warping (DTW) has 80.5 % recognition rate, and hidden Markov modeling techniques (HMM) have recognition rates of 90.7% (Al-Haddad et al. 2008). In this study, data used are isolated Malay digit from (0~9) taken from different speakers. Each speaker uttered the Malay digits sequentially.

EXPERIMENTAL PROCEDURE

In this study, a system that recognizes isolated Malay digits from zero to nine is implemented. Feature extraction of speech signals is calculated using DSP techniques, then classification of these features with an ANN.

Feature extraction process

Various methods for feature extraction are used . A famous one of those methods, the one used in this study, is the MFCC (Mel-frequency Cepstrum coefficients) algorithm. The block diagram of this algorithm is shown in Figure 1.

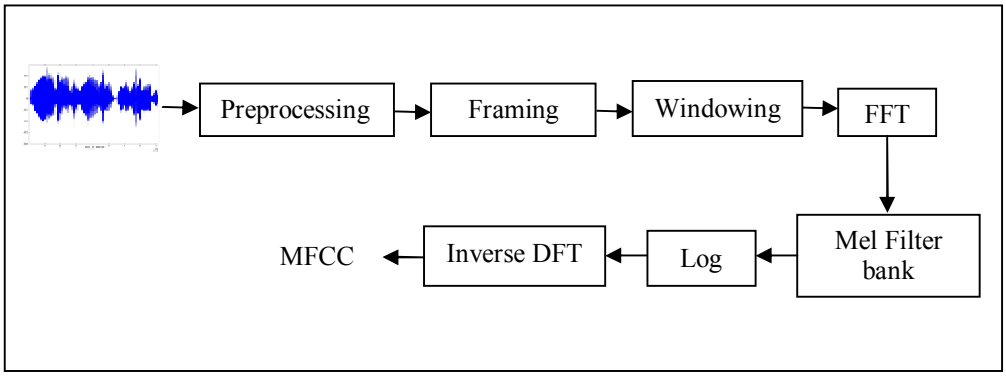

FIGURE 1 The steps involved in computing MFCCs

Speech signal is divided into overlapping frames of size 25 msec with 50% overlap, then these frames are passed through a Hamming window. Defining function of Hamming window is given in Eq. (1):

$$w\,(n)\;=\;0.54-0.46\cos(2\pi n/N-1) \tag{1}$$

The FFT is a fast algorithm to perform the discrete Fourier Transform (DFT) which is defined on the set of N samples as given in Eq. (2):

$$\mathrm{X(k)}=\sum_{\mathrm{j=1}}^{\mathrm{N}}\mathrm{x(j)W_N^{(j-1)(k-1)}} \tag{2}$$

Where $\mathrm{W_N}=\mathrm{e}^{(-2\pi\mathrm{i})/\mathrm{N}}$

The outcome after this step is often referred to as a spectrum. The mel filter bank are spaced uniformly according to the approximate formula of mel scale which is linear frequency spacing below 1000 Hz and logarithmic spacing above 1000 Hz Eq. (3).

$$\mathrm{mel(F\ )}=2595\times\log 10(1+\mathrm{f}/700) \tag{3}$$

By converting the log mel spectrum back to time, the result is called the mel frequency ceptsrum coefficients (MFFC). However, MFCC feature vector was built from first 13 MFCC coefficients (Anusuya & Katti 2011).

Rearrange MFCC feature vectors

After obtaining MFCC feature vectors, Descriptive Statistics was used to simplify them and rearrange them in order to obtain new feature vectors. The Mean, Variance, Covariance and standard deviation which are used in statistics and probability theory, some of them were applied in the output MFCC feature vectors.

The standard deviation can be calculated as:

$$\mathrm{S}=\sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})^2} \tag{4}$$

Where $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

And n is the number of column in each feature vector. Another statistical technique was the variance which is defined:

$$\mathrm{VAR\ (x)}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i}-\mathrm{x_m})^2 \tag{5}$$

Where $\mathrm{x_m}$ is the mean of feature vector.

Classification Process

This process consists of the operations in which feature vectors are modeled. In this study, ANN is used as classifier due to its profound prosperity in recognition problems. Multilayer Perceptron (MLP) was used to perform the works as classifier. MLP was used to minimize the difference between expected and real outputs of the system. The MLP topology designed for this application has the below parameters:

- Hidden layer 1: 160
- Hidden layer 2: 90

In all layer neurons, log-sigmoid transfer function is used

RESULT AND DISCUSSION

In this experiment a one hundred speech audio files were used, each file contains one utterance for Malay digit. 70 % were used as a training and the rest 30 % as testing.

After obtaining the standard MFCC feature vectors, descriptive Statistics were applied to reduce the amount of feature vector elements. As shown in Figure 2 (a) the number of feature vector elements is 215. After we applied the standard deviation and variance to the feature vectors the number of feature vector elements became 103. Figure 2 (b).

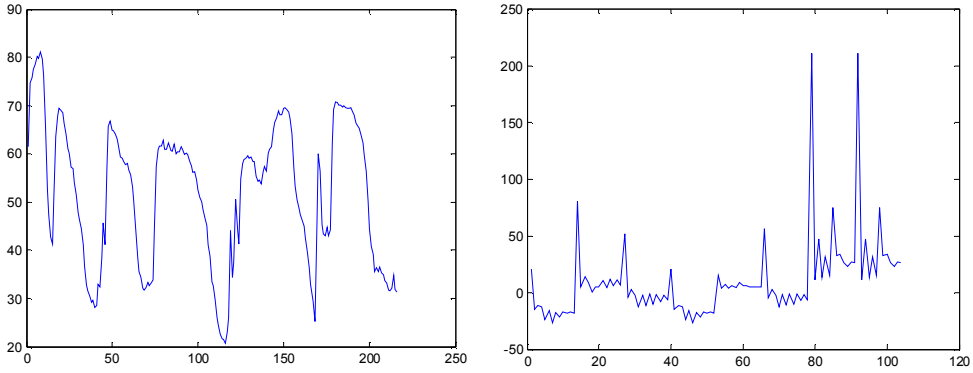

FIGURE 2 MFCCs feature vectors

These feature vectors were used in the MLP network for classification. The recognition rate for the MLP network is shown in Table 1 below:

TABLE 1 Recognition rate for Malay digits

Digit	Malay writing	Kosong	Satu	Dua	Tiga	Empat	Lima	Enam	tujuh	Lapan	sembilan	% Rec. Rate
0	Kosong	3	-	-	-	-	-	-	-	-	-	100
1	Satu	-	3	-	-	-	-	-	-	-	-	100
2	Dua	-	-	3	-	-	-	-	-	-	-	100
3	Tiga	-	-	-	3	-	-	-	-	-	-	100
4	Empat	-	-	-	-	2	-	-	-	-	-	66.67
5	Lima	-	-	-	-	-	3	-	-	-	-	100
6	Enam	-	-	-	-	-	-	3	-	-	-	100
7	Tujuh	-	-	-	-	-	-	-	3	-	-	100
8	Lapan	-	-	-	-	-	-	-	-	3	-	100
9	Sempilan	-	-	-	-	-	-	-	-	-	3	100
Total												96.67

The speech recognition system designed in this work have provided satisfactory results. With the system, a small vocabulary consisting of Malay digits have been recognized with high accuracy. Recognition accuracy resulted in this work was 96.67 %, which are high enough to show that minimizing the amount of feature vectors after applying descriptive statistics and used the MLP neural networks as classifier are both successful methods for speech recognition tasks.

In this aspect, further investigation on the number of train data set and the model of neural network topology will be studied. That means, once the network is taught with the sufficient amount of data, there is no need for increasing the train data. From this point of view, this study can be extended to an in depth analysis of Neural Network structure itself, and its optimization. Additional suggestions for further work might be the investigation of other ANN structures and new vocabularies, either in Malay or in any other language.

CONCLUSION

In this paper, a proposed method of feature extraction was used to minimize the amount of feature vector elements when transforming the speech signal to a sequence of feature vectors. By applying the descriptive statistics to the MFCC feature vectors, then MLP neural network was used as a classifier to classify the Malay digits, which shows an improvement in the recognition rate. In particular, the proposed approach provides effective, efficient and robust feature vectors.

REFERENCES

Al-Haddad, S.A.R., Samad, S.A., Hussain, A. & Ishak, K.A. 2008. Isolated Malay Digit Recognition Using Pattern Recognition Fusion of Dynamic Time Warping and Hidden Markov Models. *American Journal of Applied Sciences* 5 (6): 714-720.

Anusuya, M. A. & Katti, S.K. 2011. Front end analysis of speech recognition: a review. *International Journal of Speech Technology* - Springer.

Dede, G. & Sazli, M. H. 2010. Speech recognition with artificial neural networks. *Digital Signal Processing* 20: 763-768.

Lee, C., Hyun, D., Choi, E., Go, J. & Lee, C. 2003. Optimizing Feature Extraction for Speech Recognition. *IEEE Transactions On Speech And Audio Processing* 11 (1): 80 - 87.

Nakagawa, S. 1995. *Speech, Hearing And Neural Networks Models*. Amsterdam: IOS Press.

Rosdi, F., Ainon, R.N. 2008. Isolated malay speech recognition using Hidden Markov Models. *International Conference on Computer and Communication Engineering, ICCCE 2008*: 721-725.

Sabah, R. & Ainon, R. N. 2009. Isolated Digit Speech Recognition in Malay Language using Neuro-Fuzzy Approach. *3rd Asia International Conference on Modeling & Simulation* :336-340.

Schwarz, P., Mateˇjka, P. & Cˇernocky, J. 2006. Hierarchical structures of neural networks for phoneme Recognition. *ICASSP 2006 Proceedings. 2006 IEEE International Conference* 1: I, 14-19.