ISOLATED WORDS RECOGNITION USING MFCC, LPC AND NEURAL NETWORK

Mayur R Gamit[1], Kinnal Dhameliya[2]

[1]*M.Tech student, E&C Department, C.G.P.I.T, Bardoli, India*
[2]*Assistant Professor, E&C Department, C.G.P.I.T, Bardoli, India*

Abstract

Automatic speech recognition is an important topic of speech processing. This paper presents the use of an Artificial Neural Network (ANN) for isolated word recognition. The Pre-processing is done and voiced speech is detected based on energy and zero crossing rates (ZCR). The proposed approach used in speech recognition is Mel Frequency Cepstral Coefficients (MFCC) and combine features of both MFCC and Linear Predictive Coding (LPC). The back-propagation is used as a classifier. The recognition accuracy is increased when combine features of both LPC and MFCC are used as compared to only MFCC approach using Neural Network as a classifier..

Keywords: *Pre-processing, Mel frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), Artificial Neural Network (ANN).*

----------***----------

1. INTRODUCTION

Automatic speech recognition (ASR) has been the most investigated topic in speech processing since early 1960s [1]. Speech recognition is a popular and active area of research, used to translate words spoken by humans so as to make them computer recognizable. It usually involves extraction of features from speech signal and representing them using an appropriate data model. ASR system involves two phases. Training phase and Testing phase. In training phase, known speech is recorded and parametric representation of the speech is extracted and stored in the speech database. In the testing phase, for the given input speech signal the features are extracted and ASR system compares it with the reference templates to recognize the utterance.

This paper proposes an approach to recognize isolated words using Mel frequency Cepstral coefficient (MFCC) and Linear Predictive Coding (LPC) as a feature extraction techniques and Neural Network as a classification technique. This paper evaluates the recognition accuracy by using only MFCC and combination features of MFCC and LPC. The MFCC gives higher recognition accuracy in speech recognition systems as compared to other techniques. The main advantage of using MFCC techniques is because of less complexity and accurate results while LPC is suitable for speaker recognition.

The Back-propagation neural network is used as a classifier. The human voice is recorded and digitized. Then it is input to the pre-processing block. The main task of this block is to separate out the unvoiced speech samples from the voiced speech samples. The detection of voiced and unvoiced speech can be done on the basis of calculating energy and zero crossing rates frame by frame basis. Framing is necessary to analyze the speech. Speech is quasi-periodic signal, so analysis can be done by fragmenting the speech into small segments called chunks or frames.

The frame work of speech recognition is shown below:

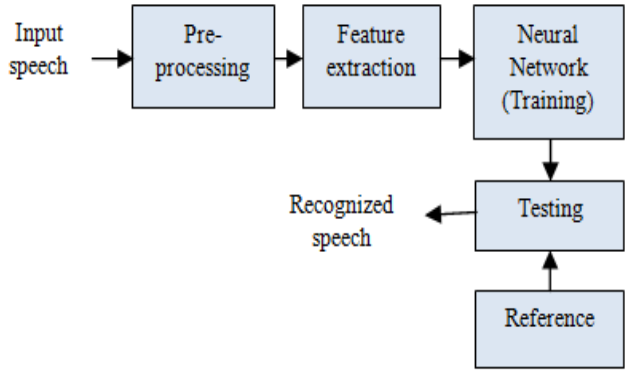

Fig. 1 Block diagram of speech recognition system.

This paper proposes a novel and efficient way to recognize isolated words by using advanced pre-processing strategy. Section 2, introduces a new pre-processing concept. Sections 3, describes the feature extraction techniques MFCC and combine features of both MFCC and LPC. Section 4, describes the neural network and section 5, describes the experimental results.

2. PRE-PROCESSING CONCEPT

The recorded speech is feed to the preprocessing block. The speech is segmented into small chunks or frames. The determination of voiced and unvoiced speech samples can be done on the basis of Energy and Zero crossing rates (ZCR) [2].

2.1 Energy

The speech signal is fragmented into small frames. Each frame is of ω samples, where ω < n as n is total number of samples. The energy of speech is calculated frame by frames. The square of each sample is done and finally summation of all squared samples is done.

The equation to calculate energy is given below:

$$Energy = \sum_{i=1}^{\omega} x_i^2 \qquad (1)$$

2.2 Zero Crossing Rate (ZCR)

The zero crossing rates means the number of times the transition of speech signal from positive to negative or vice-versa. The zero crossing rates are calculated frame by frames. If the zcr of speech samples having more zero crossing rates then it is a unvoiced otherwise voiced. According to the survey done by rabiner the ZCR of fricative speech samples are more than threshold then consider it as a voiced speech. Fricative has more zero crossing rates as compared to unvoiced speech. The equation to calculate the ZCR is as follows:

$$ZCR=\sum_{i=1}^{\omega} \frac{|sign(x_i)-sign(x_i-1)|}{2} \qquad (2)$$

2.3 Start Point and End Point Detection

The calculation of the energy and zero crossing rates can be done on the basis of frames. The start point and end point can be accurately found out on the basis of both energy and zero crossing rates. First the start and end points can be determined based on energy only. Based on zcr of the start and end points are computed. If the zcr of the frame having start point is more than the start point is shifted towards left side and same as end point shifted towards right.

- If ZCR > (threshold) then start point is shifted towards left otherwise remains same.
- If ZCR > (threshold) then end point is shifted towards right otherwise remains same.

2.4 Removal of Unvoiced Parts between Start Point and End Point

The recognition accuracy can be increased when the unvoiced part between the start and end points can be removed. There are certain samples of speech in between start and end points which have minimum energy. They do not contain any information, so by removing them makes recognition accuracy better. This advance concept has been used in this speech recognition system.

3. FEATURE EXTRACTION

In feature extraction the Mel frequency Cepstral coefficient (MFCC) and combine features of both MFCC and LPC are used. The both techniques are described below:

3.1 Mel-Frequency Cepstrum Coefficients (MFCC)

MFCC stands for Mel Frequency Cepstral Coefficient. MFCC is one of the most commonly used feature extraction method in speech recognition [3], [10].

Step1: Pre-emphasis

The signal is passed through a filter which emphasis a high frequencies [6]. This process increases the energy of signal at high frequency. High frequency also contains information. The equation used to denotes the pre-emphasis is shown below:

$$S(n) = X(n) - a * X(n-1) \qquad (3)$$

Where s(n) denotes the output sample, x(n) is present sample, x(n-1) is past sample and value of a is between 0.95 to 1.

Step 2: Framing and Overlapping

The speech signal is split into several frames such that each frame can be examined in the short time instead of the entire signal. The frame size is of the range 20-40 ms. Then overlapping is applied to frames, hamming window is applied. The equation of hamming window is as follows:

$$S(n) = X(n) * W(n) \qquad (4)$$

$$W(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \quad 0 \le n \le N-1 \qquad (5)$$

The overall process is shown in figure below:

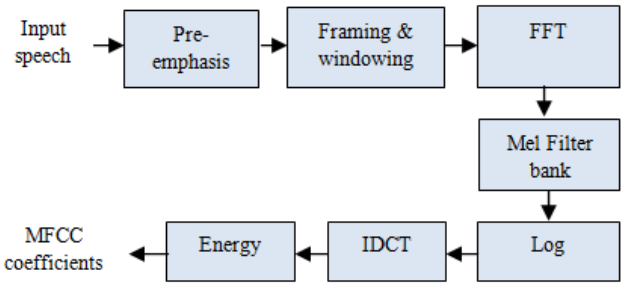

Fig. 2 Block diagram of MFCC.

Step 3: Fast Fourier Transform

The Fast Fourier Transform (FFT) converts the frames from time domain to frequency domain. The conversion is done from time to frequency domain because the information is more in frequency domain. Therefore, FFT is executed to obtain the magnitude frequency response of each frame and to prepare the signal for the next stage.

$$S(\omega) = fft\big(X(n)\big) \qquad (6)$$

Step 4: Mel Filter bank

Human ear perception of frequency contents of sounds for speech signal does not follow a linear scale. Therefore, for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the "mel scale" [9]. The Mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz [5]. To compute the Mel for a given frequency f in Hz, a following formula is used.

$$F(mel) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \qquad (7)$$

Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters.

Step 5: Log and IDCT

The output of Mel filter bank is given to log. This is the process to convert the log Mel spectrum into time domain using Inverse Discrete Cosine Transform (IDCT).

Step 6: Energy

The energy of all frames after IDCT is calculated. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

3.2 Combine Features of MFCC and Linear Predictive Coding (LPC)

The basic idea behind the Linear Predictive Coding (LPC) analysis is that a speech sample can be approximated as linear combination of past speech samples. The LPC provides a robust, reliable and accurate method for estimating the parameters that represent the vocal tract system. The autocorrelation analysis is done. Levinson Durbin's algorithm is used to analyze the LP model [11]. The LPC gives the best results for the speaker recognition rather than the speech recognition. Here, 10^{th} order LPC is taken means 11 coefficients of each frames can be obtained. The 12 coefficients of MFCC and 11 coefficients of LPC are combined frame by frame basis to yields 23 coefficients of each frames.

4. ARTIFICIAL NEURAL NETWORK

Classifier used in this speech recognition is Back Propagation Neural Network (BPNN) [8]. The back-propagation architecture takes input nodes as features based on the coefficients of MFCC and combine of both features MFCC and LPC. The hidden layer consists of 299 hidden nodes. The output layer consists of 280 nodes. The output nodes from 1-28 belongs to "zero", 29-56 belongs to "one", 57-84 belongs to "two", 85-112 belongs to "three", 113-140 belongs to "four", 141-168 belongs to "five", 169-196 belongs to "six", 197-224 belongs to "seven", 225-252 belongs to "eight", 253-280 belongs to "nine".

5. PERFORMANCE AND RESULTS

The performance of the speech recognition system is often described in terms of accuracy. Table 1 indicates the recognition accuracy by implementation of the proposed algorithm on a dataset consists of 28 speakers of which 14 are males and 14 are females.

The obtained accuracy is compared with [4]. The both features extraction techniques which are described above are used and back propagation is used as a classifiers. The results are compared with the neural network used in [4]. The recording of speech has been done in a controlled environment of laboratory using a noise removing headphone. The measuring of recognition accuracy (RA) is done based on the equation

$$RA = \left[\frac{No.\ of\ times\ word\ recognized}{Total\ no.\ of\ trials} * 100\right] \qquad (8)$$

The table 1 shows the recognition accuracy obtained during each session. The Mean square error (MSE) is 0.002482 at 10,000 epochs as compared to MSE of [4] is 0.005 at 1097 epochs for convergence.

Table 1: Recognition accuracy using MFCC and Neural Network

Number of sessions	Recognition Accuracy
1	88%
2	82%
3	77%
Average	82.33%

The table below shows the recognition accuracy by using combine features of MFCC and LPC and back-propagation neural network as classifiers. The 12 coefficients of MFCC and 11 coefficients of LPC are combine frames by frames. The accuracy at each session is shown below:

Table 2: Recognition accuracy using MFCC+LPC and Neural Network

Number of sessions	Recognition Accuracy
1	97%
2	78%
3	85%
Average	86.66%

The table below shows the comparative results of the obtained recognition accuracy with the [4]. The proposed approach gives the more recognition accuracy than described in [4].

Table 3: Comparative results [4]

Classifier	Only MFCC	MFCC+LPC +STE+ZCR	Only MFCC (Proposed)	MFCC+LPC + STE+ZCR (Proposed)
Neural Network	51.25 %	85%	82.33%	86.66%

6. CONCLUSION

The experimental results shows that by using the proposed MFCC and combination of both MFCC and LPC feature extraction techniques the results are higher as compared to [4]. The recognition accuracy is higher by using Back-propagation instead of using MLP as shown in [4]. The recognition accuracy may differ by using only MFCC, LPC and combination of both MFCC and LPC techniques as well as other classification techniques.

REFERENCES

[1] Xian Tang, "Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition, "Pacific-Asia Conference on Circuits, Communication and System, IEEE Computer society, 2009.

[2] Nidhi Desai, Prof. Kinnal Dhameliya, "Feature Extraction and Classification Techniques for Speech Recognition: A Review," International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue 12, December 2013.

[3] Shanthi Therese S, Chelpa Lingam, "Review of Feature Extraction Techniques in Automatic Speech Recognition," International Journal of Scientific Engineering and Technology, Vol. 2, Issue 6, pp.479-484, June 2013.

[4] Bishnu Prasad Das, Ranjan Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE with Neural Network Classifiers," International Journal of Modern Engineering Research, Vol. 2, Issue 3, June 2012.

[5] Revathi, Y. Venkataramani, "Speaker Independent Continuous Speech and Isolated Digit Recognition using VQ and HMM," IEEE conference, pp. 198-202, 2011.

[6] P. G. N. Priyadarshani, N. G. J. Dias, Amal Punchihewa, "Dynamic Time Warping Based Speech Recognition for Isolated Sinhala Words," IEEE Journal, pp. 892-895, 2012.

[7] Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique," International Journal of Computer Applications, Vol. 10, Issue 3, Nov. 2010.

[8] David Kriesel, "Neural Networks," research article.

[9] S. Karpagavali, P.V. Sabitha, "Isolated Tamil Words Speech Recognition using Linear Predictive Coding and Networks," International Journal of Computer Science and Management Research, Vol.1, Issue 5, December 2012.

[10] Ahmad A. M. Abushariah, Teddy S. Gunawan, Mohammad A. M. Abushariah, "English Digit Speech Recognition System Based on Hidden Markov Model," International Conference on Computer and Communication Engineering, IEEE, May 2010.

[11] Utpal Bhattacharjee, "A Comparative Study of LPCC and MFCC Features for the Recognition of Assamese Phonemes," International Journal of Engineering Research and Technology (IJERT), Vol.2, Issue 1, January 2013.