Feature Extraction Using MFCC for Speech Recognition

Sherry Vijh*, Parminder Singh** and Manjot Kaur Gill***
Guru Nanak Dev Engineering College, Ludhiana (India)
*vijh_sherry@yahoo.in, **parminder2u@gmail.com, ***gill.manjot@gmail.com

Abstract: Speech is one of the stellar means of verbal expression used by humans to interact with the machines. Speech is produced when air is rushed from lungs through vocal cords causing trembling of the vocal cords. With the introduction of digital techniques and methods researchers could transform the speech signals in analog form to digital form at minimal cost and with much higher efficiency. Speech recognition is a mechanism to convert the digital speech signal into textual form. Feature extraction is the most significant phase of the mechanism which is concerned with extraction of features from given signal. For efficacy of recognition mechanism robust features are required. In this paper focus is on mechanism of feature extraction from isolated words taking MFCC as the standard technique to extract the features. The paper highlights the comparative advantage of MFCC and elaborates its complete procedure.

Keywords: Speech Recognition, Feature extraction, Acoustic Approach, Pattern Recognition, Artificial Intelligence, Mel Frequency Cepstrum Coefficients.

Introduction

Speech is the widely preferred form of communication opted by the humans for interaction. But this mode of communication is a complicated and difficult to be handled. Based on the requirements and technology available automatic speech recognition systems are being developed for translation of speech signals. Speech Recognition (ASR) is a field that involves conversion of human speech into text form. Speech recognition is an integration of concepts of linguistics, natural language and computer science [1]. Applications of speech recognition are increasing in number due to interactive template that makes the access to devices easier and faster. These applications find more utility for disabled people or for people more use to their respective native languages. Speech recognition is a difficult task and it requires efficient techniques to produce efficacy in results. The main focus is to build software for transforming the speech into text using various techniques and methods. Speech recognition is useful for extracting the message in the speech and directing the machine to act accordingly. Further it is useful for security purpose by restricting the use of a device by differentiating and identifying the different speakers. There are different approaches for speech recognition [2]. These are:

- Acoustic Phonetic Approach: An Acoustic Phonetic approach is based on browsing sounds and providing apt labels to each type. In this approach the spoken speech is differentiated on the basis of acoustic properties of the finite phonetic units.
- Pattern Recognition Approach: The pattern recognition approach involves pattern training followed by testing which is performed by matching the patterns. The significant characteristic of this is that it is performed using proper mathematical framework for computation to establish steady pattern representation for effective pattern matching by training and testing the system.
- Artificial Intelligence Approach: The Artificial Intelligence approach is a combination of both above stated approaches. In this approach neural networks are used to compute the results by modifying the set of input parameters to obtain the desired outputs. This approach is more robust and fault tolerant. It is also called as the knowledge based approach.

Mechanism of Speech Recognition

The procedure of the speech recognition process involves three phases. Initially the system translates the input speech into the text format. The first phase of speech recognition is the Pre-processing of the speech signal which is done to obtain a clear speech signal. This intensifies the grade of the speech signal quality and thus increases the perfection of the recognition system [3]. The second phase of recognition system is the Feature extraction phase which is used to extract the relevant, robust and useful features from the speech signal to extract useful data from it. Finally the pattern matching is performed to recognize the input speech and provide its textual description. This step involves the comparison of extracted parameters with the parameters present in the database of the recognition system [1]. The acoustic models depict the phonetic properties, gender variation, variations according to environment and other acoustic characteristics. The language model accommodates the syntax and semantic knowledge required for the recognition mechanism. The mechanism of speech recognition is as follows in “Fig. 1”.

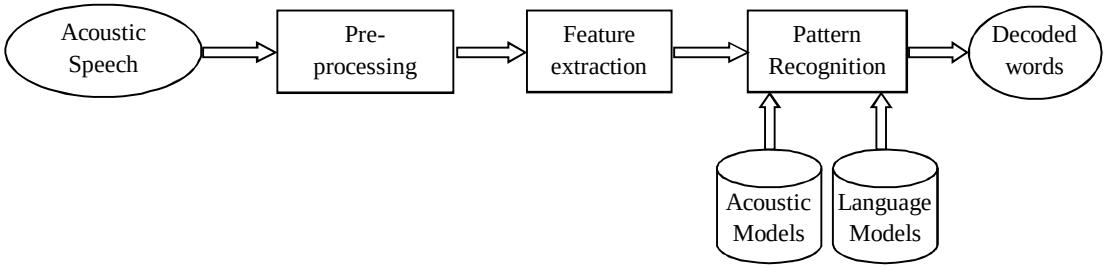

Figure 1. Speech recognition mechanism

The performance of recognition system is computed through the various parameters likerecognition accuracy, system robustness and computation complexity. The data of the system is split into two parts, namely training data and test data. The word error rate during system working is calculated for both types of data which represents the accuracy of the system. Robustness of system is highly dependent on the pre-processing and feature extraction phase of the system. Training of the system is done under certain conditions. During testing with live input the conditions might change which affect the recognition accuracy. Computation complexity is to observe the complexity in terms of model or the computation. There is a trade-off between accuracy of the system and complexity. This parameter is important for real time environment response.

Feature Extraction

Feature extraction captures the relevant and important information regarding the speech. These extracted features should be sturdy to the noisy background. The speech signal is converted into a vector sequence. The windowing of the signal provides a vector form each window. However with removal of irrelevant part of speech signal there is a risk to lose the relevant information [5]. Feature extraction helps to model the features of speech signal for further classification task. This phase is followed after the pre-processing of speech signal and is basically aimed at parametric representation of the speech signal. Feature extraction provides the features of signal in a form that can be easily used for analysis or any other processing. The speech analysis phase obtains the raw features for description of short term speech power spectrum. Then the static and dynamic features are combined to produce extension vector. Feature extraction captures the relevant and important information regarding the speech. These extracted features should be sturdy to the noisy background. The basic procedure followed in the feature extraction:

a) Sampling
b) Short term analysis
c) Transformation to frequency domain
d) Filtering
e) Optimization of class separability

Feature extraction is used not only for recognition process but is also applicable and an utmost important mechanism for the speech coding, speech processing or the analysis process. The major focus of the feature extraction is to obtain the acoustic set of features that are robust and have lower change rate and are largely beneficial for the processing of the speech signal. The extended vectors are then converted to compact vectors. These vectors are obtained from enhanced and pre-processed speech signal and are used for classification of types of sound [6].There arevarious kinds of feature extraction techniques are as stated below:

- *Principal Component Analysis (PCA):* This technique is a non-linear feature extracting which uses Eigen values and performs mapping.
- *Linear Discriminant Analysis (LDA):* This technique is based on supervised type of learning which is similar in working, but better than thePCA technique.
- *Linear Predictive Coding(LPC):* This is a static technique which produces coefficients of lower order.
- *Mel-Frequency Cepstral Coefficients (MFCC):* This technique is a frequency domain technique which performs with closer approximation to human hearing.

Mel Frequency Cepstral Coefficients

There are many feature extraction techniques available for the feature extraction like PLP, LPC, MFCC, etc. The most widely used features are the acoustic features namely Mel Frequency Cepstral Coefficients (MFCC). This technique takes frequency domain as its standard base and thus sounds more appropriate than time domain techniques. This provides closer proximity to human hearing [7]. MFCC computation is dependent upon the number and type of filters, shape of filters and way the spectrum is warped. The working of MFCC technique is as described in Fig. 2.

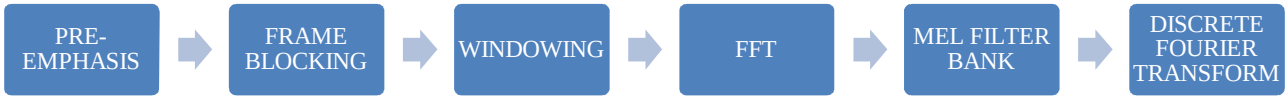

Figure 2. Working of MFCC

Pre-emphasis
The captured signal y(n) is passed through high-pass filter:

$$X(n) = y(n)-a*y(n-1) \quad (1)$$

Where, x(n) is the output and the value of a lies between 0.85 to 1.0. The major focus of pre-emphasis is compensation of the high order frequencies suppressed and intensify higher-frequency formants.

Frame Blocking
The input signal is segmented into a number of short term frames which supports 50% overlapping. Frame size chosen is usually taken between 20-30ms.Generally the frame size chosen is of power of two. The short span frames provide the static information about the speech signal [3]. Overlapping equal to 50% of size of frame is done to achieve continuity in the signal.

Windowing
All the frames are multiplied with the window function to ensure continuity at the end points of the frames. Most appropriate window function used is the hamming window function. This window function provides with the sum of a rectangle and a Hanning window. The graph of the Hamming Window is as follows in Fig. 3. This window provide slow onset and minimize the generation of side lobes in the spectrum. Hamming window:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) \\ 0, \quad otherwise \end{cases} \quad (2)$$

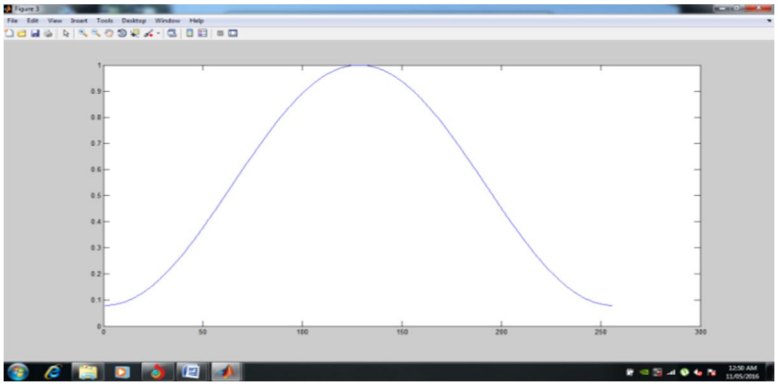

Figure 3. Hamming window

Fast Fourier Transform (FFT)
This is performed to compute the magnitude frequency response for every frame. While application of FFT we consider static and continuous nature of signal in each frame. Before FFT the signal should be continuous since the discontinuity can result in the wide and inefficient peak in the frequency response. Sharp peaks are required for best results. The result of the FFT is as shown in Fig. 4.

Filter Banks
The magnitude frequency response of each frame is multiplied by the chosen set of number of triangular bandpass filters to obtain the corresponding log energy of each triangular filter. These filters are uniformly spaced. Mel frequency is proportionate to log value of linear frequency. The triangular filters are placed with overlapping to deal with the non-linear nature of the human speech. These are commonly named as Mel filters. The mid-point of each filter represents the Mel frequency. Output of each frame is obtained by summing up these mid points of the filters [8]. The reason behind choosing

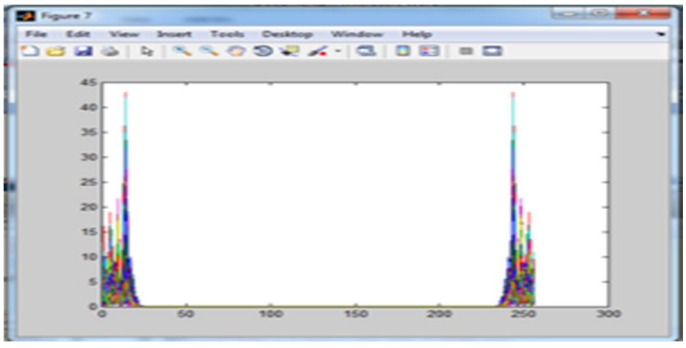

Figure 4. Fast Fourier Transform

the triangular filter is to smoothen the spectrum by flattening the harmonics to obtain the spectrum envelope. The other reason is minimize the size of the respective features. Frequency is computed as stated below:

$$Mel(f) = 2595 \times log10(1 + f/700) \quad (3)$$

Discrete Cosine Transform

Discrete cosine transform is applied to the log values of the filters to represent signal data in thetime domain.DCT is performed on the 20 log energy (S_k) computed from the triangular filters to calculate the Mel scale coefficients. The formula to compute DCT is as follows:

$$C_n = \sum_{k-1}^{k}(\log(sk))\left[\frac{(k-0.5)\pi}{K}\right] \quad (4)$$

Where, K denotes Number of coefficients being extracted and S_k= energy obtained from filter banks. This converts the frequency domain into time-domain. The features computed are alike to the cepstrum and thus called as Mel Frequency Cepstral Coefficients. MFFC can singly be used to compare the features and execute recognition process. Other add-ons can be the log energy, delta function, etc. The result of the MFCC computation is as follows in Fig. 5.

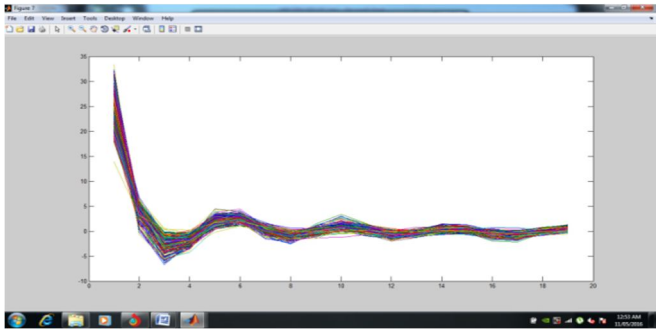

Figure 5. MFCC coefficients

Conclusion

In this paper the major area under discussion is the feature extraction mechanism along with the detailed procedure for the MFCC coefficients. There are a large number of techniques for the feature extraction but the advantage of using MFCC technique as a method for extracting features is of producing robust and concise features that comes out with high accuracy in the recognition process and provide with effective results during pattern recognition mechanism. The accuracy computed with MFCC technique comes out to be 95% which is best result obtained in comparison with other feature extraction techniques. MFCC also proves to be beneficial for the pattern recognition using artificial neural network since it reduces the complexity while comparing the features and thus helps to simplify the computation. The paper also gives complete overview of the mechanism of feature extraction specifying the benefits derived from this phase of speech recognition.

References

[1] B.S. Atal and L.R. Rabiner, "A pattern recognition approach to voiced-unvoiced silence classification with applications to Speech Recognition", IEEE transactions on Acoustic, Speech, and Signal Processing, vol. 24, no. 3, 2005, pp. 201-212.

[2] S.K. Katti, “Speech Recognition by Machine: A Review”, International Journal of Computer Science and Information Security, vol. 6, no. 3, 2009.
[3] C.S Kumar, “Design of an Automatic Speaker Recognition System using MFCC, Vector Quantization and LBG Algorithm”, International Journal on Computer Science and Engineering, vol. 3, pp. 2942-2954, 2011.
[4] R.K. Ghule, “Automatic Speech Recognition System Using MFCC and DTW for Marathi Isolated Words”, International Journal of Technology Enhancements and Emerging Engineering Research, vol. 3, no. 9, 2006.
[5] L.R. Rabiner. and B.H Juang , “Fundamentals of Speech Recognition”, AT&T, Prentice-Hall Inc, 2010.
[6] A.N. Mishra, “Isolated Hindi Digits Recognition: A Comparative Study”, International Journal of Electronics and Communication Engineering, vol. 3, pp. 229-238, 2010.
[7] R. Jain and S.K. Saxena, “Advanced Feature Extraction &its Implementation in Speech Recognition System”, International Journal of Services Technology and Management, vol. 2,no. 3, 2011.
[8] U. Shrawankar and V.M. Thakare, “Techniques for Feature Extraction in Speech Recognition System: A Comparative Study”, International Journal of Computer Applications in Engineering, Technology and Sciences, vol. 2, no. 1, pp. 412-418, 2010.