International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 9, September - 2013

Performance Analysis of MFCC And LPCC Techniques In Automatic Speech Recognition

Rajiv Chechi,Reetu

Deptt. Of ECE,HCTM Technical Campus Kaithal,Haryana,India

Abstract

Automatic speech recognition (ASR) System is to accurately and efficiently convert speech signal into a text message independent of device, speaker or the environment.The computing algorithms of speech features, being the main part of speech recognition system are analyzed.The determination algorithms of Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) coefficients expressing the basic speech features are developed. The training and recognition processes are realized in both subsystems separately, and recognition system gets the decision being the results of each subsystems.

Index term : Speech Recognition, Cepstral Analysis, Frequency Cepstral Coefficient, Power Spectrum, Feature of speech, Mel scale, Linear Predictive Coding

1. Introduction

Automatic speech recognition system which allow a computer to identify the word that a person speak into telephone and microphone or other transducer and convert it into text. This is the interesting mode of interaction between human and computer. Communication between person dominated by spoken language so people need a speech interface with computer which can recognize speech in native language. It is important that the system should be independent of speaker and language characteristics such as accent, speaking styles, disfluencies, syntaxed grammer, along with the capability of handling a large vocabulary.[1]

i) Structure and Working

A. Architecture of ASR system

ASR System consists of two end such as front end and back end. Front end consist signal processing component such as preprocessing and feature extraction. The back end consist acoustic model and language model and search engine as shown in fig.1[1]

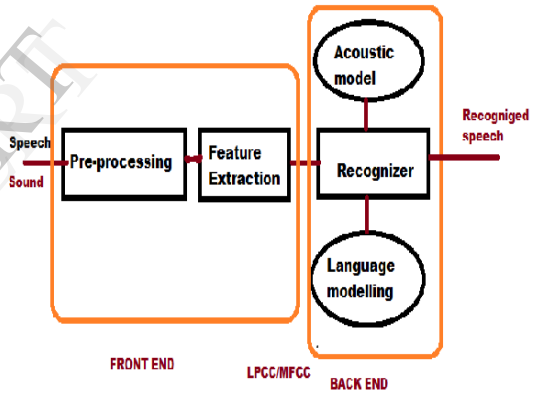

Figure 1. ASR Architecture

Main Component of Automatic speech recognition system

i. Pre-processing-translate the speech signal into a sequence.
ii. Feature Extractor-give us feature vector of the speech signal.
iii. Recognizer-tells us what a word string is likely to be pronounced
iv.

B. Working of Speech Recognition System

Automatic speech recognition system working at four stages given as pre-processing, feature extraction, modelling and testing.

The brief introduction of four stages given below

a) Pre-processing:

Preprocessing consist following three part such as pre-emphasis, framing and windowing

Pre-emphasis:
To flatten the speech spectrum ,pre-emphsizer is used to compensate the high frequency component which was suppressed during the human sound production mechanism. The mostly used filter is high pass filter whose transfer function is given as

$$H(z) = 1 - \alpha_{preem}\, z^{-1} \qquad (1)$$

Output of the filter is given as

$$S_{p(n)} = S_{in} - \alpha S_{in}\,(n-1) \qquad (2)$$

Where α is the filter coefficient ($\alpha \in (0.95; 1)$)

S_{in} -is the input signal

***Framing and Windowing*:**
The speech signal is divided into a sequence of frame where each frame can be analyzed independently and represented by a single feature vector. By applying the frame blocking to the Sp(n),we will get M vector of length N, which is correspond to Sp(k,m) where k=0,1,. . . .,N-1 and m= 0,1,. . . .,M-1 since each frame supposed to have stationary behavior ,a compromise in order to make the frame blocking is to use a 20-25ms window applied at 10ms interval. Window technique at each frame is use to reduce signal discontinuity at either end of blocking. Mostly used technique is hamming window technique.

It is calculated as

$$W(k) = 0.54 - 0.46\ \cos\frac{2\pi k}{N-1} \qquad (3)$$

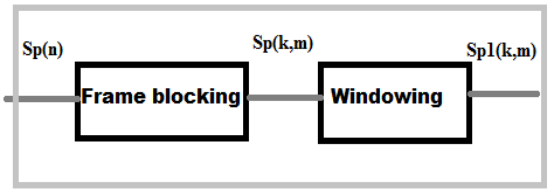

Figure 2. Framing and windowing

By applying W(k) on $S_{P\ (k,m)}$, we can calculate $S_{p1\,(k,m)}$

Feature Extraction

The next step which is most important is the feature extraction technique which extract the information from the speech frame. Common parameter used in speech recognition are Linear Predictive coding(LPC), and Mel Frequency cepstral coefficient (MFCC).these are used because of following reason

- It compute the some important characteristic of signal such as energy or frequency response.
- These technique work well in speech recognition technique application.

Recognition

The next step in speech recognition is the distance calculation between the feature vector .the two most prominent method used is the Euclidean distance and the likelihood distortion measure.

2. Popular Feature Extraction Techniques

Complete part of information in acoustic signal is too much to process and not full information is required for specific task .In ASR system, the aim of feature extraction has to find a representation which is stable for different example of same speech sound. These techniques maintain the part that represents the message in the speaker or the environmental characteristic. Several techniques have been used to extract feature from the speech. MFCC and LPC are two of most commonly used method.

A. Mel Frequency Cepstral Coefficient

Mel frequency Cepstrum Cofficient(MFCC)

i. BLOCK DIAGRAM OF CALCULATION OF MFCC

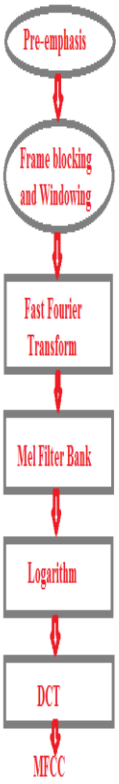

Figure 3. Block diagram of MFCC

ii. Mel Frequency Cepstral cofficient

MFCC is given by Davis and Mermelstein as a beneficial approach for speech recognition.

After applying the pre-emphasis and the frame blocking and windowing to the signal, the MFCC vectors will be obtained from each speech frame. MFCC is commonly derived as following the different step by considering that all the stages are applied over speech frames.

First take the Fast Fourier transform of each frame and obtain its magnitude. Next step is to map the power of spectrum which obtain in first step, onto the mel scale using mel filter bank. This step is used to adapt the frequency resolution to properties of the human ear means to obtain the perceptual frequency, which known as perceptual mel frequency scale. The filter bank is the set of triangular filter whose bandwidth and spacing are equal to critical band whose centre frequency covers the most important frequencies for speech perception. The input of mel filter bank is the Xframe[k] and the output of filter bank is log spectral energy vector , Eframe[m] for each frame. Filter bank sample the spectrum of speech fame at its centre frequencies that known as mel-frquency scale. Suppose that there is M(m=1,2,……M) number of mel filter bank channel and consider Hm(k) is the transfer function of the filter m ,then log spectral energy is given as

$$E[m] = \sum_{k=1}^{K-1} \ln\ X[k]^2 H_m[k] \qquad (4)$$

The main reason of using mel-filter bank is given as

- This remove the pitch of speech signal and smooth the magnitude spectrum.
- Reduce the feature size involved.

The final step of feature extraction process is applying discrete cosine transform (DCT) on log spectral energy vector and output of this stage is called as mel frequency cepstral coefficient. DCT-II is used at this stage because of following reason

- It is good for compressing information.
- Uncorrelated the vector.

Output of this stage means mel frequency coefficient is given as

$$C_I = \sqrt{\frac{2}{M}} \sqrt{\frac{2}{M}} \sum_{m=1}^{M} E_m \cos\left(\frac{\pi i}{m}\left(m - \frac{1}{2}\right)\right) \qquad (5)$$

As the coefficient increases then the variance and average value is decreases. We studied that zero cepstral coefficient is directly proportional to the log spectral energy vector. As explained above ,by using discrete cosine transform higher cepstral coefficient is discarded so M channel filter bank become only L MFCCs means L<M. By using MFCC technique for extraction of feature vector from speech signal there is some loss of information because of following reason

- In this magnitude spectrum is used so phase information is removed
- After the DCT-II stage there is trunctation from M to L so because of this more spectral detail is lost.

B. Linear predictive coefficient(LPC)

As in MFCC extraction technique filter bank analysis is used for computation of mel frequency cepstral coefficient ,linear predictive coding (LPC) is used to compute linear prediction coefficient. A difference equation for s(n) is given as

$$S[n] = \sum_{i=1}^{p} a_i \; S(n-1) + G \; e[n] \qquad (6)$$

In this equation a_i is the linear prediction coefficient and e[n] represent an error in the model(the difference between the predicted value and the actual measured value).

From the equation observed that the speech signal is the combination of previous p sample, in this we discuss the computation of feature of speech signal based on least mean square error theory. This technique is known as linear prediction .the LPC coefficients are obtained for each frame independently one of each other.

Prediction error Em for one frame according to equation (6) is given by equation (7)

$$E_m = \sum_n e_m^2 \; [n] = \sum_n X_m \; [n] - \sum_{j=1}^{p} a_j \; X_m[\text{n-j}] \qquad (7)$$

Where $X_m[n]$ is the frame of speech signal and p is the order of LPC analysis.

i. Block diagram

The block diagram for computing LPCC is given below

Figure 4. Block diagram of LPCC

ii. Linear Predictive Cepstral Coefficient Algorithm

In LPCC before converting into cepstral coefficient ,first compute a LPC specral envelope.Linear prediction coefficient used in speech recognition as well as in system identification problem in modern control system ,maximum entropy technique ,speech coding and synthesis.This process is also known as autoregressive process and applied on the spectrum which capture the vocal tract properties of vowel like sound. First three step of the LPCC is same as the MFCC technique. As in the MFCC pre-emphasis is applied to speech waveform in LPCC pre-emphasis is applied to the spectrum of input signal. LPCC smoothing the spectrum by autoregressive filter. This provide the stable parameter. In the next step inverse discrete fourier transform is applied to the power spectral density(PSD) to obtain the autocorrelation function. Then further Durbin recursion is applied to solve autoregressive coefficient then cepstral coefficient is compute from LPC coefficient. After LPC analysis input signal passing through a filter and obtain power spectrum of signal as an output. When input signal is speech signal then inverse LPC filter A(z) is used ,and power spectrum is given as

$$S(\omega) = \frac{\sigma^2}{1-\sum_{I=1}^{P} a_i e^{-ij\omega}} \qquad (9)$$

This features technique have been used by many recognition systems, being its performance comparable with MFCC technique ,another LPC based feature technique to obtain features vector called Perceptual Linear Prediction (PLP) coefficients PLP incorporating a non-linear frequency scale and other psychophysics properties of the human perception system. PLP analysis is same as to MFCC analysis, but the incorporation of more perceptual properties makes it more related to psychophysical results.

Conclusion

In this paper we have studied the implementation of ASR system using LPCC and MFCC for feature extraction. In this paper we conclude that MFCC algorithm is better and more efficient than LPCC algorithm .The use of ASR solve the problem of technology acceptance in India by playing the interesting mode of interaction between human and computer.

REFERENCE

[1] H .Fletcher," Auditory pattern Review of Modern Physics," Jan 1940.

[2] Gray Jr., A.H. and Markel, J.D. ,1976,"Distance Measure for Speech Processing", IEEE Transaction on Acoustic Speech and Signal Processing, issue 5,pp.380-391,oct 1976

[3] Molau, S., Pitz, M., Schluter, R. & Ney, H. (2001),"Computing Mel Frequency Cepstral Coefficient on the power Spectrum", IEEE international Conference on ACOUSTIC, Speech and Signal Processing Germany, pp. 73-76,2001.

[4] M.Nilsson, M. Ejnarsson,"Speech Recognition using Hidden Markove Model-Performance evaluation in Noisy Environment", Blekinge Institute of Technology. Sweden 2002.

[5] J. Darch, Ben Milner, Xu Shao, Saeed Vaseghi and Qin Yan, "Predicting Formant Frequencies from MFCC Vectors", IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.

[6] R. K. Aggarwal and M. Dave, "Implementing a speech recognition system interface for Indian Languages", Proc. IJCNLP, Workshop on NLP for Less privileged Language ,IIIT Hyderabad, pp. 105-112, 2008.

[7] R.K. Aggarwal and M. Dave, "Using Gaussian Mixtures for Hindi Speech Recognition System", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol4, No.4, Dec 2010.

[8] J. Picone ,"Signal Modeling Technique in Speech Recognition",Processing of the IEEE, Vol.81,NO.9,pp.121-1247,1993.