Effect of Pre-processing along with MFCC Parameters in Speech Recognition

[1]Ankita S. Chavan, [2]Mrs.S. S Munot(Bhabad)
Dept. of Electronics and Telecommunication, K.K. Wagh college of Engg. and Research, Nasik, India

Abstract **— This paper gives brief description about feature extraction technique using MFCC. MFCC is mostly used in Automatic Speech Recognition System. Feature extraction plays an important role in ASR, which provides the set of main features. This paper includes the results for effects of normalization, down-sampling and parameter changes like window size, linear spacing. Speech recognition means pattern recognition problem, so classification is done on data using minimum distance classifier. Training and testing data are different. For down-sampling i.e Fs/4 more accuracy is achieved than normalization.**

Index Terms – **Automatic Speech Recogniton(ASR), Down Sampling(Fs), Mel Frequency Cepstral Coefficients (MFCC), Discrete Cosine Transform(DCT).**

I. INTRODUCTION

Speech is nothing but an ability to express our feelings. It is a human vocalized form. Automatic speech recognition system is one of the most interesting areas which develop the techniques for understanding natural language. And now a day it is used as an input to the machine for effective reorganization and processing. Speech recognition system will be beneficial for speech disorder people. Speech recognition system consists of four stages are analysis, feature extraction, modeling and matching. Feature extraction is the process, which provides the subset of all feature vectors and further those are used for recognition purpose. In other words it is the subset of all parameters or features. By using feature extraction technique, the result of features having ability to provide more accuracy in speech recognition is achieved. There are number of feature extraction techniques like MFCC, LPC and PLP. In domain of speech processing MFCC technique is most commonly used. In this work the parameters are decided on basis of accuracy achieved by pre-processing the signal using normalization, down the sampling frequency. Compares the results of normalization i.e without down sample and with down sample the signals. In Classification the patterns are classified into different classes and data is separated into training and testing sets.

Section II gives the description on literature review of MFCC.

Section III provides the details related to the MFCC.

II. LITERATURE SURVEY

Summarization of the effect of filter bank smoothing on the recognition performance of children's speech. Filter bank smoothing of spectra is done during the computation of the Mel filter bank cepstral coefficients (MFCCs). The results from experiments of paper indicate that unlike conventional VTLN implementation, it is better not to scale the bandwidths of the filters during VTLN only the filter center frequencies need be scaled. From experiment result the formant center frequencies may approximately scale between speakers, the formant bandwidths do not change significantly. Therefore, the scaling of filter bandwidths by a warp-factor during conventional VTLN results in differences in spectral smoothing leading to degradation in recognition performance [1]. Summarization of the several feature extraction techniques for speaker recognition was proposed in this paper. MFCC is well known techniques used in speaker recognition to describe the signal tract properties. The main aim of this paper is to create a speaker recognition system, and apply it to a 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 [2]. Summarization of many vocal fold detection systems has been developed for voice pathology detection by using sustained vowel with good results was mentioned in this paper. Voice pathology detection systems based on continuous speech will be a better alternative that is easier to be used in practice. Therefore, in this paper author presented a voice disorder detection system based on continuous. The developed system had very good performance and detection rate. MFCC is a good choice for voice disorder detection with a sustained vowel as compared to other speech features. The results of the developed system show that even with continuous speech MFCC is performing better. Disorder detection with continuous speech can be further investigated by applying other speech features and compare its performance with MFCC [3].

III. MEL FREQUENCCY CEPSTRAL COEFFICIENTS

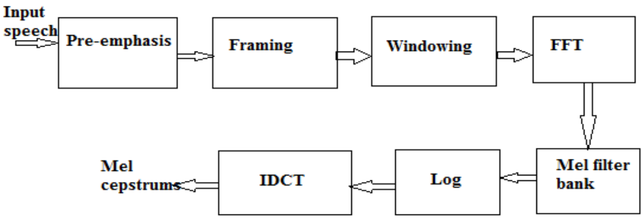

Fig.1 MFCC block diagram

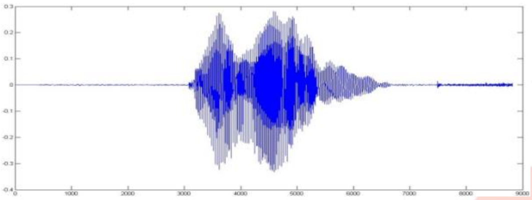

Fig.2 Input speech signal

- Pre-emphasis – In MFCC extraction process firstly the input speech signal is pre-emphasized to enhance the high frequency part of the signal at the time of speech generation. The pre-emphasized speech wave is getting divided into number of frames. The length of frames is usually of 20ms to 40ms.
 y (n) = x(n) - x(n - 1)(1) * a
 Where, y(n) is the output signal 0.95≤a≤1.
 Magnitude plots for pre-emphasis filter-For

the frequency response, substitute $exp(-j\frac{2\pi f}{fs})$for z.

$$H(f)=1\text{-}a\,exp^{(-j\frac{2\pi f}{f_s})} \qquad (2)$$

Magnitude response is given by,

$$H\,(f) = 1 + a^2 - 2a\cos(\frac{2\pi f}{f_s}) \qquad (3)$$

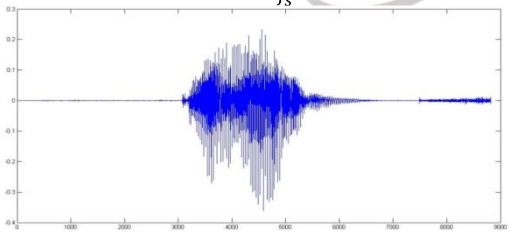

Fig.3 Pre emphasis

- Normalization –Normalization is done on signal to get the same amplitude of signal. As the speech signal has different amplitudes at different envelopes so for further processing of speech signal normalization is used. By normalizing the speech signal it becomes easier to process, also saves processing time and gives more accuracy.
- Down sampling gives more accuracy than normalization. Both normalization and down-sampling are the pre-processing steps before framing and windowing the signal to increase the accuracy.
- The frame length should not be too short and too long. If frame length is short then we don't get enough samples and for long length the signal changes. As the speech signal is non-stationary, so for analyze the speech it get divided into frames where it supposed to be stationary speech signal.
- Windowing is used to reduce the spectral effects, smooth's the signal for computation of the FFT. Overlapping is used to produce continuity within the frames. Hamming, Hanning, Rectangular, Triangular window types are used for windowing of speech signal.

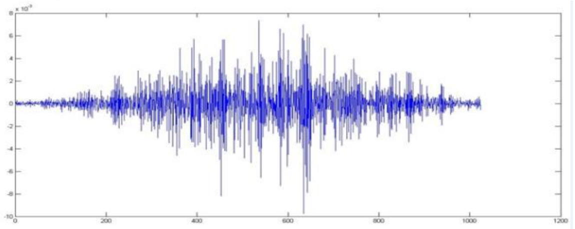

Fig.4 Framing and windowing

Mel scale -

Pitch is denoted in Mel scale. This scale approximates to the human hearing system. As, human can easily distinguish the minor variations in pitch at low frequencies than at high frequencies.

Formula from Frequency to Mel scale:

$$M(f)= 1125\ \ln(1 + f/700) \qquad (3)$$

Formula from Mel scale to Frequency:

$$M^{-1} = 700(exp^{(m/1125)}) - 1 \qquad (4)$$

- Fast Fourier transform is used to convert the signal from time to frequency domain.

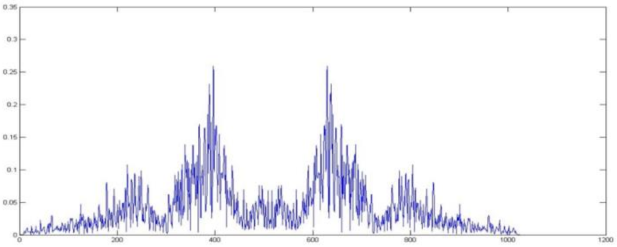

Fig.5 FFT signal

- The Mel filter bank consists of overlapping triangular band pass filters. From the Mel scale concept, the center frequencies are equally linearly separated below 1000Hz and logarithmically separated above 1000Hz.

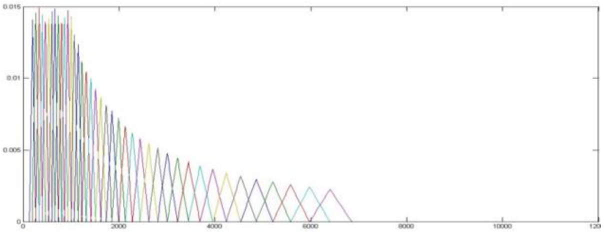

Fig.6 Mel filter bank

- Logarithm is used for wide range.

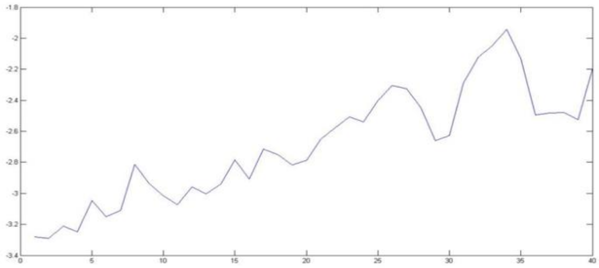

Fig.7 Log

- IDCT – It is applied to the log spectral energy vector resulting in the group of Mel frequency cepstral coefficients. DCT is mostly used due to its energy compaction, which results in its coefficients being more concentrated at lower indices than the DFT. This property allows approximating the speech signal with fewer coefficients. The Mel filter bank output is given to IDCT by Logarithm compression which results in the group of coefficients called Mel frequency cepstral coefficients. The formula for DCT is,

m=Nk=1 cos [m*(k-0.5)*pi/N] * Ek m=1,2, ..., L

where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients.

RESULTS

MFCC results for varying different parameters

1. Selection of window size by varying the sizes for normalization and downsample.

Lowest frequency = 133.333,
Linear filters = 13
Linear spacing = 66.666,
Log filters = 27
Log spacing = 1.0711703,
FFT size = 1024
Frame rate = 100,
Cepstral coefficients =24

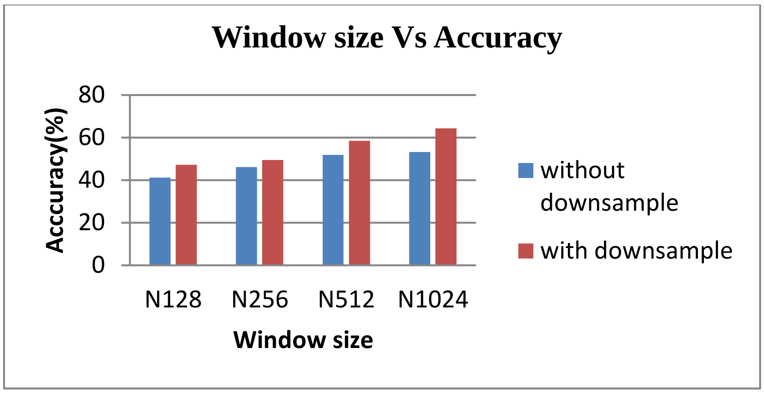

Fig.8 Graph for window size

2. Selection of linear spacing by varying the spacing for normalization and downsample.

Lowest frequency = 133.333, Linear filters = 13, Window size = 1024, Log filters = 27, Log spacing = 1.0711703, FFT size = 1024,Frame rate = 100, Cepstral coefficients =24

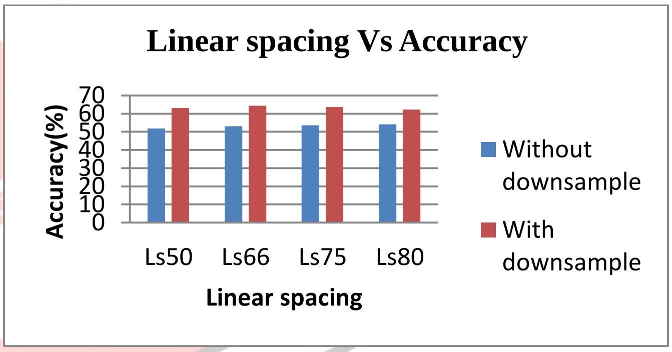

Fig.9 Graph for Linear spacing

CONCLUSION

For normalized MFCC i.e without down sampling 53% accuracy is achieved. With down the sampling frequency from Fs=44 KHz to Fs=11 KHz 64.36% accuracy is achieved for window size N=1024 and linear space Ls=66.

REFERENCES

[1] S. Umesh, Member, IEEE, and RohitSinha "A Study of Filter Bank Smoothing in MFCC Features for Recognition of Children's Speech" *IEEE Transactions on Audio, Speech, And Language Processing, Vol. 15, No. 8, November 2007.*

[2] TahiraMahboob and RuqiaBibi "Voice Recognition Using GMM with MFCC",*International Journal of Computer and Communication System Engineering (IJCCSE), Vol. 2 (2), 2015*.

[3] Zulfiqar Ali,Mansour Alsulaiman, Ghulam Muhammad, Irraivan Elamvazuthi, ”Vocal Fold Disorder Detection based on Continuous Speech by using MFCC and GMM", *IEEE GCC Conference and exhibition, November 17-20, 2013.*

[4] Vibha Tiwari, ”MFCC and its applications in speaker recognition" *International Journal on Emerging Technologies 1(1): 19-22(2010) ISSN : 0975-8364, Feb 2010.*

[5] Ben Milner “Robust Acoustic Speech Feature Prediction From Noisy Mel-Frequency Cepstral Coefficients”, *IEEE* transactions *on audio, speech, and language processing, vol. 19, no. 2, february 2011.*

[6] Xu Shao and Ben Milner ”MAP Prediction of Pitch from MFCC Vectors for Speech Reconstruction” *School of Computing Sciences, University of East Anglia, UK.*

[7] Shahidhar G. Koolagudi,DeepikaRastogi and K. SreenivasaRao“Identification Of Language Using MelFrequencyCepstralCoefficients(MFCC)”*SciVerseScienceDirect, International Conference On Modelling, Optimisation and Computing(ICMOC).*

[8] Haojun Wu, Yong Wang, and Jiwu Huang “Identification of Electronic Disguised Voices” *IEEE Transactions On Information Forensics And Security, Vol. 9, No. 3, March 2014.*

[9] S. Umesh, Member, IEEE, and RohitSinha “A Study of Filter Bank Smoothing in MFCC Features for Recognition of Children’s Speech” *IEEE Transactions on Audio, Speech, And Language Processing, Vol. 15, No. 8, November 2007.*

[10] Seiichi Nakagawa,LongbiaoWangand Shinji Ohtsuka “Speaker Identification and Verification by Combining MFCC and Phase Information” *IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 4, May 2012.*