IMPROVEMENT OF ACCURACY USING MFCC SPEECH RECOGNITION.

Ms. Dani Prachi P[1], Mrs. Deole M. S[2].

[1] *Student, Department of Electronics & Telecommunication Engineering, GES' RHSCOEMS &R Nashik, Maharashtra, India*

[2] *Asst. Professor, Department of Electronics & Telecommunication Engineering, GES' RHSCOEMS &R Nashik, Maharashtra, India*

ABSTRACT

This paper introduces a real time algorithm of MFCC (Mel frequency cepstral coefficients) for speech recognition. Whereas, PNCC a new feature extraction algorithm based on auditory processing is described in this paper. The features of PNCC processing include the use of a power-law nonlinearity that has been replaced by the traditional log nonlinearity used in MFCC coefficients. It also uses the medium-time power analysis, in which environmental parameters are estimated over a longer duration than is commonly used for speech, as well as frequency smoothing. PNCC is basically used for the improvement in recognition accuracy in noisy conditions. The presented results are of MFCC used for the improvement of the recognition accuracy. obtained using Matlab R2013.

Keyword: - *Speech recognition; feature extraction; Mel frequency cepstral coefficients; automatic speech recognition.*

1. INTRODUCTION

Automatic speech recognition by machine is an research area for speech recognition. There are several kinds of parametric representations for the acoustic signals. One of them is the Mel-Frequency Cepstrum Coefficients (MFCC) is the most widely used [1]. There are many kind of work done on MFCC, specially on the improvement of the recognition accuracy [3]. However, all these algorithms require large amount of calculations, which will increase the cost and reduce the performance of the hardware speech recognizer. The speech signal has a 10 dB signal-to-noise ratio and a spectrum between 0.3 kHz to 3.4 kHz at a sampling frequency of 8 kHz.Nowadays the performance of speech recognition systems in acoustical environments has drastically improved. Most speech recognition systems remain sensitive to the nature of the and their performance decreases sharply in the presence of sources of degradation such as additive noise, linear channel distortion, and reverberation. One of the most challenging problem is that recognition accuracy degrades significantly if the test environment is different from the training environment and if the acoustical environment includes disturbances such as additive noise, channel distortion, speaker differences, reverberation.

The presently developed systems for automatic speech recognition are based on two types of features mel frequency cepstral coefficients (MFCC) [2] and perceptual linear prediction (PLP) coefficients [4] .Spectro-temporal

has been observed that two-dimensional Gabor filters provide a reasonable approximation to the spectro temporal response, which has leads to various approaches to extract features for speech recognition.[17] This paper describe the introduction of an additional feature set for speech recognition referred as power normalized cepstral coefficients (PNCC) and implementation of MFCC speech recognition. Mel Frequency Cepstral Coefficients (MFCC) is a widely used feature extraction method implemented in multiple ways. Here MFCC for speech recognition system is tested using matlab software, which is also used in the recognition tests. Earlier PNCC does not makes the use of temporal masking, Nowadays the implementation of PNCC processing provides significantly superior recognition accuracy over a broad range of reverberation and conditions of noise using features that are computable in offline using algorithms that do not require extensive look-ahead, and with a computational complexity that is comparable to that of traditional MFCC and PLP features. Earlier versions of PNCC processing have been evaluated by various teams of researchers and compared to several different algorithms including zero crossing peak amplitude (ZCPA),RASTA-PLP, perceptual minimum variance distortionless response (PMVDR), invariant integration features (IIF) [5], and sub band spectral centroid histograms (SSCH) [21]. Results from initial comparisons, tend to show that PNCC processing provides better speech recognition accuracy than the other algorithms cited above. Over the years dozens if not hundreds of algorithms have been introduced to address this problem. Many of these conventional noise compensation algorithms have provided substantial improvement in accuracy for recognizing speech in the presence of quasi-stationary noise. Unfortunately these same algorithms frequently do not provide significant improvements in more difficult environments with transitory disturbances such as a single interfering speaker or background music. The improvements provided by PNCC are typically greatest when the speech recognition system is trained on clean speech and noise and reverberation is present in the testing environment. For systems that are trained and tested using large databases of speech with a mixture of environmental conditions, PNCC processing outperforms the best results than MFCC and PLP processing in terms of recognition accuracy.

2. LITERATURE REVIEW

Due to the importance of accuracy in speech recognition, Researchers have developed a number of methods and several techniques. The literature survey of few of them is as follows:

Wei HAN, Cheong-Fat CHAN "An Efficient MFCC Extraction Method in Speech Recognition" [1] states that a new algorithm of extracting MFCC for speech recognition. The new algorithm reduces the computation power by 53% compared to the conventional algorithm. Simulation results indicate the new algorithm has a recognition accuracy of 92.93%. There is only a 1.5% reduction in recognition accuracy compared to the conventional MFCC extraction algorithm, which has an better accuracy.

Lee Yoot Khuan, Ihsan Mohd Yassin, [2] they proposed that the performance Mel Frequency Cepstrum Coefficient (MFCC) in extracting significant feature is influence by several important parameter settings, namely the number of filter banks, and the number of coefficients used in the final representation. These settings affect the way the features are represented, and in turn, effect the performance of the classifier for diagnosis of the disease. Particle Swarm Optimization (PSO) algorithm is used in this work to adjust the parameters of the MFCC feature extraction method, together with the Multi-Layer Perceptron (MLP) classifier structure for diagnosis of infants with asphyxia. The extracted MFCC features were then used to train several MLP classifiers over different initialization values.

Gellert Sarosi1, Mihaly Mozsary "Comparison of Feature Extraction Methods for Speech Recognition in Noise-Free and in Traffic Noise Environment" [3] says that the investigate several novel front-end techniques and compare them to multiple baselines. Recognition tests were performed on studio quality wide band recordings on Hungarian as well as on narrow band telephone speech including real-life noises collected in six languages: English, German, French, Italian, Spanish and Hungarian.

Zhang Wanli [4] gives the concept of new approach presented for speaker recognition using the improved Mel frequency cepstral coefficients (MFCC).The experimental database consists of 30 speakers, 15 male and 15 female, collected in a sound proof room. The result of this experiment certificates that the improved Mel frequency cepstral

coefficients derived parameters perform better than traditional Mel frequency cepstral coefficients based on hidden Markov models.

Haofeng Kou, Weijia Shang, Ian Lane, Jike Chong tweis,[5] gives a research for Mel-Frequency Cepstral Coefficient (MFCC) feature extraction and describe the optimizations required for improving throughput on the Graphics Processing Units (GPU). It not only demonstrate that the feature extraction process is suitable for GPUs and a substantial reduction in computation time can be obtained by performing feature extraction on these platforms, but also discus about the optimized algorithm.

Adrian Pass Ji, Ming Philip Hanna Jianguo, Zhang Darryl Stewart,[6] gives a new approach to visual speech recognition which improves contextual modelling by combining Inter-Frame Dependent and Hidden Markov Models. This approach captures contextual information in visual speech that may be lost using a Hidden Markov Model alone. Contextual modelling is applied to a large speaker independent isolated digit recognition task, and compares the approach of two commonly adopted feature based techniques for incorporating speech dynamics.

Tianyu T.Wang and Thomas F. Quatieri,[7] develops a 2-D model of speech in local time{frequency regions of narrowband spectrograms using sinusoidal-series-based modulation. It is described to distribute vocal tract and onset content based on source information (e.g., noise and voicing) in a transformed 2-D space, thereby explicitly representing different classes of energy modulations commonly observed in spectrograms.

Niko Moritz, Jorn Anem uller and Birger Koll Meier [20] presents an amplitude modulation _lter bank (AMFB) that is used as a feature extraction scheme in ASR systems. The time-frequency resolution of the employed FIR filters, an i.e. bandwidth and modulation frequency setting that was originally proposed to describe data from human psychoacoustics. Investigations on modulation phase indicate the need for preserving such information in amplitude modulation features.

3. Proposed Method for Implementation of MFCC

The fundamental objective of noise in an speech is that it completely removes the noise from the speech signal while it preserves the original speech signal as it is. Firstly in the removal of noise from a speech signal there is use of MFCC speech processing. [1]. There will be a training set which is trained on a clean speech and then testing is done to recognize the accuracy. The important idea of MFCC is to extract the different features contained in the speech signal. Improvement of accuracy is also done by PNCC in noisy environment by performing medium time power analysis, in which environmental parameters are estimated and frequency smoothing is done. Also accuracy and computational complexity plays an important role in various speech processing applications. It is still an active area of research. It helps greatly to improve the accuracy of a speech signal. In this paper features are extracted and accuracy is obtained by using MFCC (Mel frequency cepstral coefficients)

This System consists of five modules:
1. Pre-emphasis.
2. Windowing.
3. FFT (Fast Fourier Transform)
4. Mel frequency filter bank.
5. Cepstrum.

The steps for MFCC algorithm implementation:

Step-1: Pre-emphasis
In speech processing audio signals pre-emphasis is mainly designed to increase the magnitude of some frequencies with respect to the magnitude of other frequencies in order to improve the overall signal-to-noise ratio by minimizing the adverse effects of such phenomena as saturation of recording media in subsequent parts of the system. The inverse operation is called de-emphasis, and the system as a whole is called emphasis. For pre-emphasis default value of 'a' is 0.97[2]

Step-2: Windowing
Windowing of a simple waveform like cos ωt causes its Fourier transform to develop non-zero values at frequencies other than ω. The leakage tends to be worst (highest) near ω and least at frequencies farthest from ω. To reduce the edge effect hamming window is hamming window is applied to each frame.

Step-3: FFT (Fast Fourier transform)
A fast Fourier transform (FFT) algorithm computes the discrete Fourier transform (DFT) of a sequence, or its inverse. Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. To obtain a good frequency resolution, 256-point FFT is used.

Step-3: Mel frequency filter bank
Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

Step-5: Cepstrum
The Cepstrum is the result of taking the inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal. It may be pronounced in the two ways given, the second having the advantage of avoiding confusion with 'kepstrum' which also exists. There is a complex cepstrum, a real cepstrum, a power cepstrum, and a phase cepstrum. The power cepstrum in particular finds applications in the analysis of human speech.

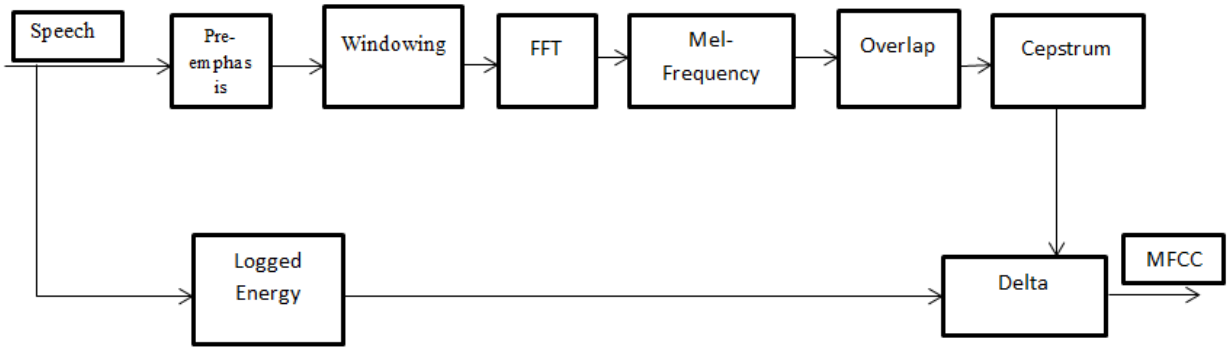

Figure1. Proposed Method for MFCC Implementation

4. IMPLEMENTATION

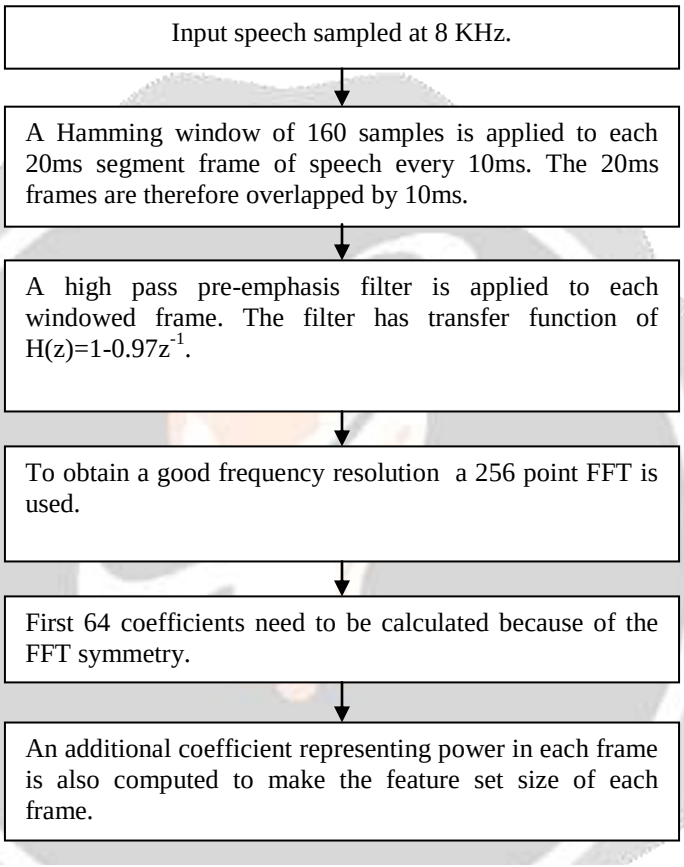

Figure 2: General Algorithm for MFCC Speech Recognition

The basic idea of MFCC is to reduce the edge effect, therefore a 160 points hamming window is applied to each frame. The mathematical formula for Hamming window is ,

$$Ham(N) = 0.54 - 0.46\cos\left(2\pi\frac{n-1}{N-1}\right)$$

Where N=160, the number of points in one frame

To obtain a good frequency resolution, a 256 point FFT is used [5]. Because of symmetry property of FFT there is need to calculate 128 coefficients .The mapping from linear frequency to Mel-frequency is given by,

$$Mel(f) = 1127\ln(1+\frac{f}{700})$$

Also, the equation to calculate the Mel-frequency cepstrum from the output power is given by,

$$c_n = \sum_{k=1}^{K} (logS_k) \cos[n(k - 0.5)\frac{\pi}{K}]$$

Where S_K is the output power of the k^{th} filter of the filter bank. The logged energy of each frame as one of the coefficients, is calculated by

$$E = log \sum_{n=1}^{160} s_n^2$$

Which is calculated without any windowing and pre- emphasis? To enhance the performance of speech recognition system, time derivatives are added to the basic static parameters. The delta coefficients are obtained from the following formula.

$$dc_1 = \frac{(c_{t+2} - c_{t-2}) + (c_{t+1} - c_{t-1})}{10}$$

5. RESULT

i. The major contribution of this paper is to extract the features from MFCC (Mel frequency cepstral coefficients) using Speech recognition technique and to improve the accuracy using MFCC.
ii. After removing noise by proposed method, accuracy of speech signal is checked. This gives good results.
iii. The accuracy after implementing the Proposed MFCC method is better than conventional algorithm. The accuracy is 86% that of conventional MFCC extraction method.

Speech signals	a (Pre-emphasis coefficient)	Window length	FFT Point	Recognition Accuracy
S1	0.97	160	256	86.43%
S2	0.97	160	256	85.32%
S3	0.97	80	256	82.84%
S4	0.97	80	128	79.56%

TABLE: RECOGNITION ACCURACY OF MFCC USING DIFFERENT SPEECH SIGNALS

SNAPSHOT OF EXPERIMENTAL RESULTS

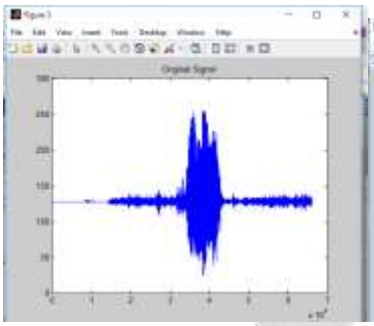

Fig.1: Original signal of speech

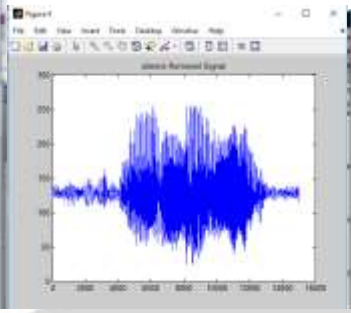

Fig.2: Silence removed signal

Fig 3: MFCC feature extraction

6. CONCLUSION

This paper presents an real time extraction algorithm called MFCC (Mel frequency cepstral coefficients). Currently, many new schemes are proposed in the field of speech recognition. So the best method among all should be found out. The proposed method is among the efficient method of all to noise removal which leads to extract features for speech recognition. Many techniques are proposed for automatic speech recognition but none of it is considered to be perfect for measurement of accuracy. Improving accuracy plays a crucial role in the field of speech processing. In this paper features are extracted using MFCC with real time database and accuracy has been checked using MFCC (Mel frequency cepstral coefficients) algorithm. From the estimated results it is found MFCC algorithm reduces the multiplication steps with improved accuracy. Logically, a bigger value of SNR is good because it means that the ratio of signal to noise is higher. Higher SNR indicate that higher removal of noise.

7. ACKNOWLEDGEMENT

First and the foremost I, take this opportunity to express gratitude to my guide, Mrs. M.S. Deole, for her constant encouragement and support throughout the project implementation. I sincerely thank Prof. S. P. Agnihotri, Head of Department of Electronics & Telecommunication Engineering for his advice and support during course of this work.

With deep sense of gratitude I thank to our Principal Dr. P. C. Kulkarni and Management of Gokhale Education Society for providing all necessary facilities and their constant encouragement and support. I also express my thanks to all teaching and non-teaching staff for their kind co-operation and guidance also.

REFERENCES

[1] Wei HAN, Cheong-Fat CHAN, Chiu-Sing CHOY, "An Efficient MFCC Extraction Method in Speech Recognition," IEEE Conference on Speech Processing,0-7803-9390-2/06/$20.00 ©2006

[2] Azlee Zabidi, Wahidah Mansor, Lee Yoot Khuan, "Three-dimensional Particle Swam Optimisation of Mel Frequency Cepstrum Coefficient Computation and Multilayer Perceptron Neural Network for classifying Asphyxiated Infant Cry", IEEE Conference on Computer applications and industrial electronics, 2011

[3] Gellert Sarosi, Mihaly Mozsary, Peter Mihajlik, "Comparison of Feature Extraction Method for Speech Recognition in Noise-Free and in Traffic Noise Environment," IEEE 978-1-4577- 0441-3/11/$26.00 ©2011

[4] Zhang Wanli," The Research of Feature Extraction Based on MFCC for Speaker Recognition,"2013 3rd International Conference on Computer Science and Network Technology.

[5] Haofeng Kou, Weijia Shang, Ian Lane, "OPTIMIZED MFCC FEATURE EXTRACTION ON GPU," ICASSP 201

[6] Adrian Pass Ji, Ming Philip Hanna Jianguo, Zhang Darryl Stewart, "Inter-Frame Contextual Modelling for visual speech recognition". IEEE 17th International conference on Image processing September 26-29,2010,Hong kong.

[7] Tianyu T. Wang and Thomas F. Quatieri,, "Two-Dimensional Speech-Signal Modeling,",IEEE transactions on audio, speech and language processing, vol. 20, no. 6,august 2012.

[8] Niko Moritz, Jorn Anem uller and Birger Kollmeier,\An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition." IEEE/ACM Transactions on audio, speech, and language processing, vol. 23, no. 11, november 2015.

[9] H. Hermansky and F. Valente, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 4165–4168.

[10] S. Y. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition," in *Proc. INTERSPEECH*, Sep. 2008,

[11] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Speech, Signal Process., Mar. 2010, pp. 4574–4577.

[12] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 2, pp. 142–152, Feb. 2008.

[13] S. Ganapathy, S. Thomas, and H. Hermansky, "Robust spectro-temporal features based on autoregressive models of hilbert envelopes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4286–4289.

[14] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Commun.*, vol. 53, no. 5, pp. 736–752, May/Jun. 2011.

[15] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Commun.*, vol. 53, no. 5, pp. 753–767, May/Jun. 2011

[16] F. Müller and A. Mertins, "Contextual invariant-integration features for improved speaker-independent speech recognition," *Speech Commun.*, vol. 53, no. 6, pp. 830–841, Jul. 2011.

[17] F. Kelly and N. Harte, "A comparison of auditory features for robust speech recognition," in *Proc. EUSIPCO*, Aug 2010, pp. 1968–1972.

[18] S. Ganapathy, S. Thomas, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," IEEE Signal Process. Letter., vol. 15, pp. 681–684, Nov. 2008.

[19] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in Proc. INTERSPEECH, Sep. 2010, pp. 2058–2061.

[20] C. Lemyre, M. Jelinek, and R. Lefebvre, "New approach to voiced onset detection in speech signal and its application for frame error concealment," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., May 2008, pp. 4757–4760.

[21] S. R. M. Prasanna and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 556–565, May 2009.