International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 38741-38745

Speaker Identification and Spoken word Recognition in Noisy Environment

Shaik Shafee
Dept. of. Electronics and Communications Engineering S.V.University College of Engineering Tirupathi, Andhrapradesh, INDIA shafee.shaik@yahoo.co.in

Prof. B.Anuradha
Dept. of. Electronics and Communications Engineering S.V.University College of Engineering Tirupathi, Andhrapradesh, INDIA anubhuma@yahoo.com

Abstract
Speaker identification and spoken word recognition becomes difficult in noisy environment. The performance of ASR (Automatic Speech Recognition) systems would decrease if the noise levels are different from training to testing environment. In this paper an attempt has been made to improve the recognition success rate using Gammatone frequency cepstral coefficients (GFCC) [2][3][4] in noisy environment (Home/Office environment). Different set of HMMs (Hidden Markov Models) [9] have been designed using different types of speech features such as MFCC (Mel Frequency Cepstral Coefficents) [1], PLP (Perceptual Linear Prediction) coefficients [1] and Gammatone Frequency Cepstral coefficients (GFCC) [1] in MATLAB environment and the results are analyzed. Experiments have been carried out on 8 different Telugu spoken words among 10 different Telugu Speakers. Speech wave forms for training set are recorded in almost clean environment and the speech wave forms for testing recorded in noisy environment (Home/Office environment where noise levels are up to 50dB). With the above experiments it is clear that the Speaker Identification success rate as well as the Spoken word recognition success rate have been improved using GFCC features in Noisy environment over MFCC and PLP coefficients.

Index Terms— Gammatone Frequency Cepstral Coefficients (GFCC); Mel Frequency Cepstral Coefficients (MFCC);Perceptual Linear Prediction(PLP) coefficients; Hidden Markov Models(HMM),Speaker Identification, Telugu word recognition

I. INTRODUCTION

Generally Speech recognition means converting speech into text or automatic speech recognition (ASR). Speech Recognition Systems can be of different types such as speaker dependent or speaker independent, isolated words or continuous speech, large vocabulary or limited vocabulary. ASR systems performs two major tasks: Signal/Speech processing in which features like LPC, PLP, MFCC etc.. are extracted which can be called as front-end task and the other task is modeling like Hidden Markov Models, Artificial Neural Networks, Self Organizing Maps, Learning Vector Quantization Neural networks etc… which can be called as back-end task. Feature extraction process mainly consists of framing, windowing. Modeling consists of vector quantization of features and trains the quantized feature vectors.

In Recent years, many researchers are working on ASR systems to improve the performance in Noisy environment. The ASR Systems which are designed using MFCC speech features giving good results in clean environment or in very low noise levels but the performance decreased in noisy environment. In this paper an attempt has been made to improve the results of speaker Identification and spoken word recognition in a common noisy environment (Home/Office environment: Whisper noise (up to 30 dB), Office noise (up to 40 dB), Window Air Conditioner noise (up to 50 dB).The accuracy of ASR mainly depends on the type of feature extraction process, still research is going on to identify the best suitable features for ASR. Three different concepts have been experimented and the results are analyzed in this paper: Speaker Identification, Spoken word recognition and the combination of speaker identification and the spoken word recognition. All the three experiments were conducted using HMM models for 3 different types of speech features such as PLP cepstral coefficients, MFCC and GFCC features for Telugu Speakers/Words in MATLAB 2013(a) environment.

II. FEATURE EXTRACTION TECHNIQUES

A. Perceptual Linear Prediction (PLP) Cepstral coefficients

Bark scale filter bank is used in PLP analysis. The relationship between Bark scale frequency and the linear frequency is as shown in Equation (1)

$$f_{bark} = 6\ln\left(\frac{f}{600} + \sqrt{\left(\left(\frac{f}{600}\right)^2 + 1\right)}\right) \quad (1)$$

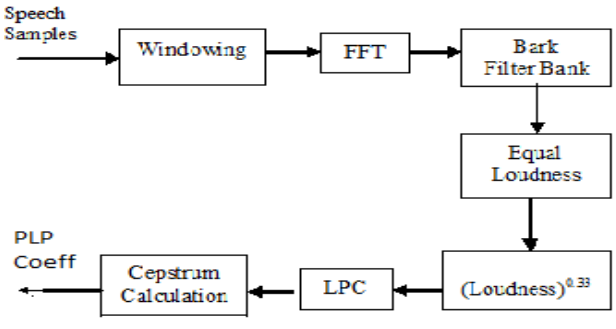

Fig 1. Flow diagram for PLP Cepstral Coefficients extraction process

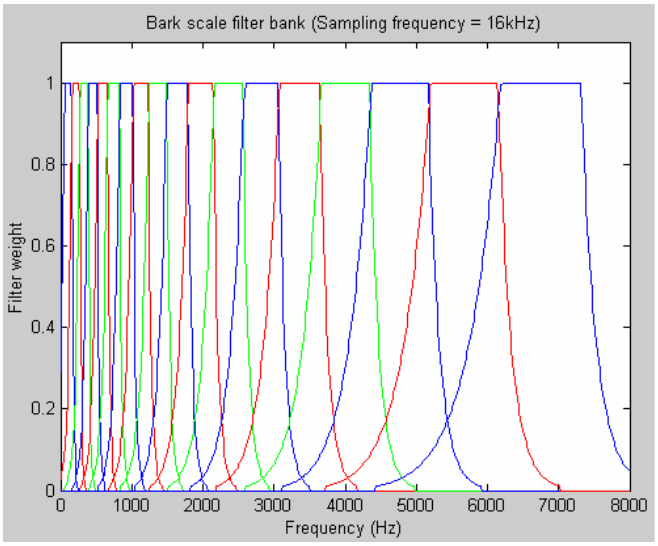

Fig2 : Bark scale filter bank (Sampling frequency = 16kHz) applying for PLP Cepstral Coefficients

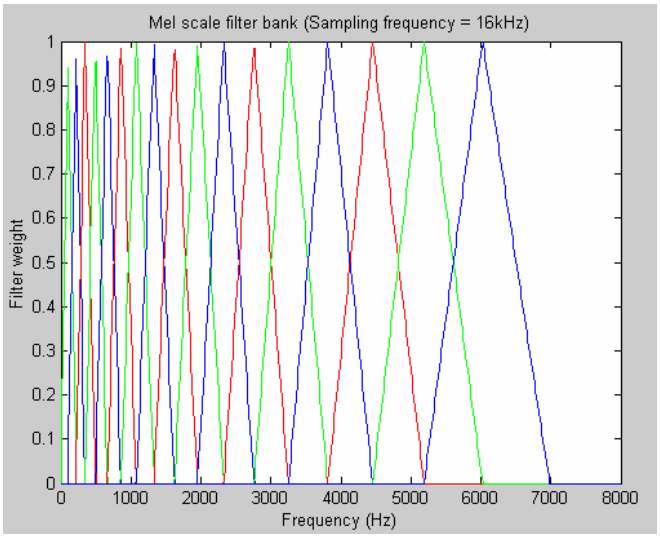

Fig.4.: Mel scale filter bank (Sampling frequency = 16kHz) applying for MFCC Cepstral Coefficients

B. Mel Frequency Cepstral coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) is one of the most common feature extraction method used in speech recognition systems.

$$f_{Mel} = 2595 \times \log\left(\frac{f}{700} + 1\right) \quad (2)$$

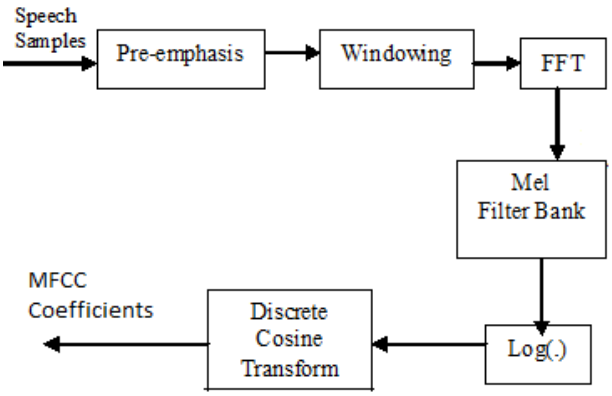

Fig.3. : Flow diagram for MFCC features extraction process

C. Gammatone Cepstral coefficients (GFCC)

Gamma tone function is defined in time domain by its impulse response as in Equation (3)

$$G(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft + \phi) \quad (3)$$

Where
n is the order of the filter;
b is the bandwidth of the filter;
a is the amplitude;
f is the filter centre frequency;
φ is the phase;

Patterson and Moore proved that the gammatone function is the best fit to the human auditory system. The Equivalent Rectangular bandwidth (ERB) of the auditory filter with the function has been proposed as in Equation (4)

$$ERB = 24.7 \times \left(\frac{4.37}{1000} f + 1\right) \quad (4)$$

It was also suggested that a 4th order filter (n=4) would be a good model for the human auditory filter.

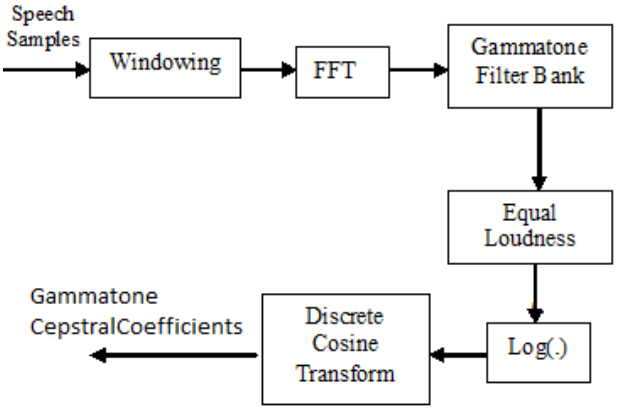

Fig.5: Flow diagram for GFCC features extraction process

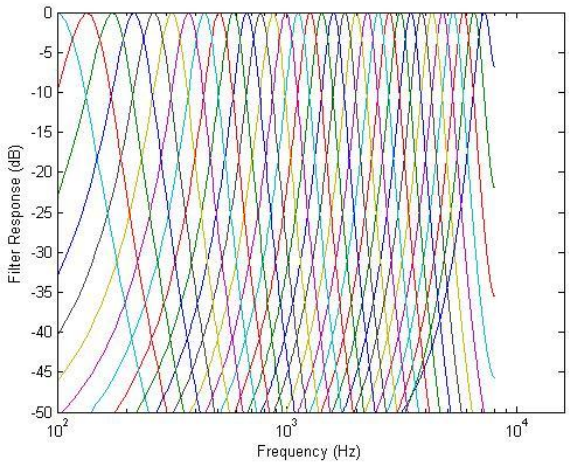

Fig.6:Gammatone filter bank (Sampling frequency = 16kHz,32 Channel) applying for GFCC features

III. K-MEANS ALGORITHM

The recorded speech samples may not be having of same length though the same words are collected from the same speaker. So there is a need to process the speech waveforms to a fixed set of feature vectors (same number of feature vectors to all the speech waves) to apply to the training system. By using End point detection algorithm the unvoiced samples will be removed at both the ends of speech waveforms. Then speech wave will be segmented to overlapped frames and then compute the feature vectors (MFCC/PLP/GFCC coefficients) frame wise. By using k-means algorithm with 'K' centroids, all the speech wave forms feature vectors are processed to a fixed set of k-feature vectors to each of the speech wave form.

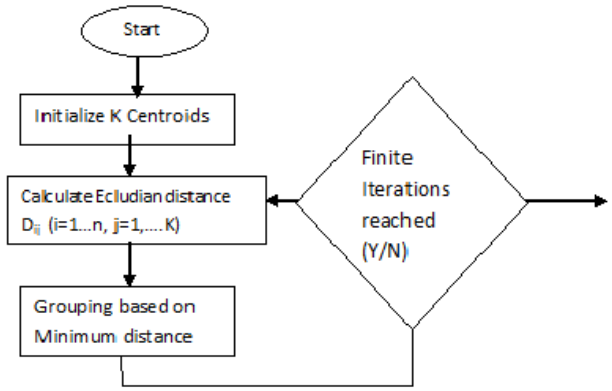

Fig.7.Flow chart of K-Means Algorithm

IV. HIDDEN MARKOV MODELS

HMMs are very popular statistical stochastic approach which will be used as back-end task since many years in speech recognition systems. There are three basic problems arise in applying HMM models to Speech recognition task. Problem1 can be treated as speech recognition problem: How efficiently the Probability $P(O/\lambda)$ be computed for the given Observation sequence $O=(O_1,O_2,O_3,\ldots,O_T)$ and the HMM model ($\lambda=\{A,B,\pi\}$).Problem2 can be treated as hidden part: To find the Optimal state sequence for the given Observation sequence $O=(O_1,O_2,O_3,\ldots,O_T)$ and the HMM model ($\lambda=\{A,B,\pi\}$). And problem 3 can be treated as the training problem: How the Model ($\lambda=\{A,B,\pi\}$) be adjusted to maximize the probability $P(O/\lambda)$.

Feature vectors of each spoken word are applied to K-means algorithm and obtain a uniform training sequence for each spoken word.Then HMMs are built to each spoken word using quantized feature vectors. The same process is applied for designing the HMMs for Speaker Identification. In Recognition task the unknown word is applied to all the designed HMMs and calculate the $P(O/\lambda)$. The HMM for which the maximum value is calculated will be chosen as the recognized word.

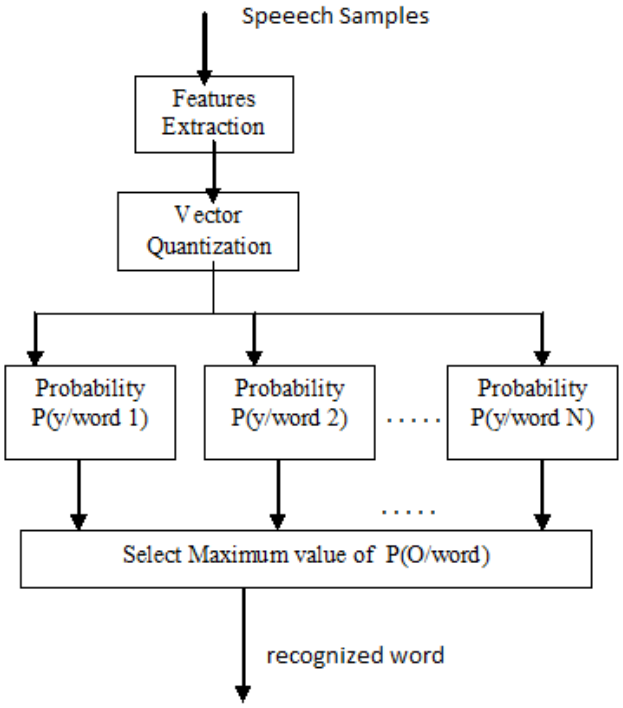

Fig 8: Isolated word recognizer for N words

The same flow process is applied in speaker Identification where word HMMs are replaced by Speaker HMM models for the given spoken word.

V. EXPERIMENTAL SETUP AND RESULTS

8 different Telugu words (aagu (□□□)-STOP, edama (ఎడమ)-LEFT, kadulu (□□□□□)-START, kudi (□□□□)-RIGHT, kinda (□□□□)-DOWNWARDS, paina (□□□)-UPWARDS, venakki (□□□□□□□)-BACKWARDS, munduki (□□□□□□□)-FORWARD) from 10 different Telugu speakers (5 Male and 5 Female Speakers) have been recorded with 16 KHz sampling frequency, each word recorded 10 times in clean environment for training the HMMs and 5 times in noisy environment (Home/Office environment where noise levels up to 50 dB) for testing. Total of 800 samples (10 speakers*8 words*10 times) for training) and 400 samples (10 speakers*8 words*5 times) for testing are recorded in.wav form.

HMM models designed based on the spoken words as well as on speaker identification, three different experiments are analyzed using the above trained HMM models. In case1: Recognizing the spoken word success rate. In case2: Identification of Speaker success rate. And In case3: Success rate for the combination of spoken word and the speaker identification.

The experiments conducted for three different Speech features (PLP Cepstral coefficients, Mel Frequency Cepstral Coefficients and Gammatone Frequency Cepstral Coefficients. All the three different cases experimented many times and their average results are tabulated as below.

TABLE I Case1: Spoken word Recognition

Speech feature extraction type	No. of samples tested in each experiment	No.of Avg. samples recognized correctly	Success rate (%)
PLP	400	271	67.8
MFCC	400	279	69.7
GFCC	400	288	72.0

TABLE II Case2: Speaker Identification

Speech feature extraction type	No. of samples tested in each experiment	No.of Avg. samples recognized correctly	Success rate (%)
PLP	400	292	73.0
MFCC	400	297	74.2
GFCC	400	308	77.0

TABLE III Case3: Combination of both word Recognition and Speaker Identification

Speech feature extraction type	No. of samples tested in each experiment	No.of Avg. samples recognized correctly	Success rate (%)
PLP	400	208	52.0
MFCC	400	213	53.2
GFCC	400	221	55.2

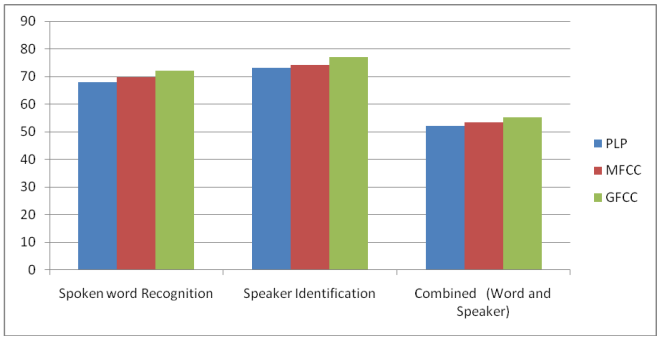

Fig.8.Comparison diagam of recognition success rates (%) in three different cases.

VI. CONCLUSIONS

From the results, it is clear that the word recognition and speaker Identification success rate using GFCC features giving better results over MFCC and PLP cepstral coefficients in noisy conditions (Office/Home environment) for the HMM as backend training models. The work may be extended by examining the different noise conditions for different feature extraction procedures. Other models such as ANNs (Artificial Neural Networks), Self Organizing Maps (SOM) and Learning Vector Quantization neural networks may be experimented and examined for the improvement of recognition success rate in noisy conditions.

REFERENCES

[1] A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios,Jitong Chen, Yuxuan Wang, and DeLiang Wang, Fellow, IEEE IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, December 2014

[2] Auditory processing-based features for improving speech recognition in adverse acoustic conditions-Hari Krishna Maganti and Marco Matassoni-Magantiand Matassoni EURASIP Journalon Audio, Speech, and Music Processing May-2014

[3] Gammatone Wavelet Cepstral Coefficients for Robust Speech Recognition-Anirudha adiga and Chandrasekhar seemanthuls IISC Bangalore-TENCON-2013

[4] Auditory features based on gammatone filters for robust speech recognition-Jun Qi, Dong Wang, Yi Jiang, Runsheng Liu-Tsinghua University, Beijing, China-2013

[5] Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum-Interspeech-2012-Md Jahangir Alam, Patrick Kenny, Douglas O'Shaughnessy-University of Quebec, Montreal, Quebec, Canada

[6] Signal Processing for Robust Speech Recognition Motivated by Auditory Processing-Thesis by Chanwoo Kim, Language Technologies Institute School of Computer Science Carnegie Mellon University-2010

[7] Xiaojia Zhao, Yang Shao and DeLiang Wang, "CASA-based Robust Speaker Identification," IEEE Trans. on Audio, Speech and Language Processing, vol.20, no.5, pp.1608-1616, 2012.

[8] Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition-International Journal of Information and Communication Technology Research-Volume 2 No. 12, December 2012

[9] Speech Recognition using Artificial Neural Networks and Hidden Markov Models-IEEE multidisciplinary engineering education magazine, vol. 3, no. 3, september 2008

[10] isolated speech recognition using artificial neural networks: Prasad D Polur, Ruobing Zhou, Jun Yang, Fedra Adnani, Rosalyn S. Hobson-23rd Annual International Conference of the IEEE ENgineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey.

[11] Neural Networks used for Speech Recognition-Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, Senior Member, IEEE-journal of automatic control, university of belgrade, vol. 20:1-7, 2010

[12] Automatic Noise Recognition Based on Neural Network Using LPC and MFCC Feature Parameters-Proceedings of the Federated Conference on Computer Science and Information Systems pp. 69–73 ISBN 978-83-60810-51-4

[13] the optimal performance of multi-layer neural network for speaker-independent isolated spoken malay parliamentary speech-issn 2231-7473 2010 Faculty of Computer and Mathematical Sciences, Universiti Teknolgi MARA, (UiTM), Malaysia

[14] Role of neural network models for developing speech systems-S¯adhan¯ a Vol. 36, Part 5, October 2011, pp. 783–836. c Indian Academy of Sciences

[15] recent advances in deep learning for speech research at microsoft-Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA-ICASSP 2013

[16] Neural Network Toolbox For Use with MATLAB® Howard Demuth Mark Beale User's Guide Version 4 copyright 1992-2004 by The MathWorks, Inc.

[17] voice recognition using neural networks-Ganesh K Venayagamoorthy, Viresh Moonasar and Kumbes Sandrasegaran* Electronics Engineering Department, M L Sultan Technikon, Durban, South Africa gkumar@saiee. org. co.za *Institute for Information Sciences and Technologv (IIST), Massey Universiv, New Zealand K. Sandrasegaran@massey. ac. nz 0-7803-5054-5.0029 1998 IEEE

[18] Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques-journal of computing, volume 2, issue 3, march 2010, issn 2151-9617

[19] Development of Isolated Word Speech Recognition System-INFORMATICA, 2002, Vol. 13, No. 1, 37–46 2002 Institute of Mathematics and Informatics, Vilnius

[20] Recognition of Spoken Arabic Digits Using Neural Predictive Hidden Markov Models-Rafik Djemili, Mouldi Bedda, and Hocine ourouba-The International Arab Journal of Information Technology, Vol. 1, No. 2, July 2004