Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

Speech Feature Extraction at Different Mode with Application to Shouted Speech Recognition System used for Women Safety

Anly Paul
PG Scholar, Dept. of AEI
Rajagiri School of Engineering and Technology, Kakkanad
Kochi,India

***Abstract*— This paper presents a feature extraction algorithm for speech at different modes such ad whispered, soft, neutral, loud and shouted. Normal feature extraction methods such as LPCC, MFCC or PNCC extract cepstral components of speech and speech recognition systems build on such features provide better recognition accuracy for speech of only normal mode. But in real life, a speech recognition system that could recognize speech of only normal mode is of no use. A better method to recognize speech at whispered, normal or shouted mode is explained in this paper. Energy of the speech signal is computed for each frame and based on the energy level speech mode is categorized. Power spectrum estimated using linear prediction is used instead of power spectrum computed directly from speech signal. MFCC coefficients are calculated from the estimated power spectrum which gives the resulting features for speech recognition at different modes. GMM, GMM-UBM or DNN can be used in training and testing phase of the speech recognition system. This will result in a proper speech recognition system working in any speech mode. The effectiveness of the approach can be demonstrated using a speech recognition system for women safety which is trained and tested with shouted speech.**

Keywords—Shouted Speech Recognition, Speech mode,Linear Prediction, MFCC.

I. INTRODUCTION

This Human speech is a natural way for communication. Speech signal is non-stationary as well as one dimensional. Now a day, researchers are thinking about efficient methods for interaction between humans and machine. A best example for this interaction is systems working on Speech Recognition. Machine is trained in such a way that it should have sufficient intelligence to recognize human voice [3]. Apart from the message observed from the voice signal, much information are hidden in it which leads to the development of many other systems based on gender identification, speaker identification, emotion identification from speech etc. But the major drawback is that the system fails to recognize the speech or speaker when tested with speech at different modes other than normal mode. Background noise also reduces the efficiency of the system.

Speech signals are generally classified into five categories based on the modes of speech production. These include whispered speech, soft speech, neutral speech, loud speech and shouted speech. Whispered speech is defined as the lowest vocal mode where as shouted speech is defined as the highest vocal mode. Neutral speech is also called normal speech mode. Soft speech is an intermediate between neutral speech and shouted speech mode whereas loud speech is an intermediate between neutral speech and shouted speech mode [4].

According to studies, Sound Intensity Level (SIL) which is the power of sound transmitted along the wave is a good method for classification of speech modes. SIL is low for whispered speech and high for Shouted speech. Another method used for classification of speech mode is by measuring frame energies. Studies had proved that if the energy level for a frame is less than 10 dB, that will be the frame corresponds to whispered speech. If the frame energy is between 5 dB and 15 dB, the frame represents soft speech. For normal speech frame energy is between 15 dB and 25 dB. Frame energy level between 25 dB and 35 dB are considered as loud speech. Above 35 db, it will be of shouted speech [4].

Many feature extraction methods are introduced recently and the most widely used method for speech feature extraction are Mel Frequency Cepstral Coefficient (MFCC), Power Normalized Cepstal Coefficient (PNCC), Linear Predictive Cepstal Coefficient (LPCC) and Perceptual Linear Prediction (PLP) coefficient.

In all feature extraction methods, the speech signal is divided into frames of duration 20-25ms with a frame step of 10ms. Speech signal is considered as stationary in a frame. Sampling rate of 16 KHz is used in all cases. Short term Fourier Transform is done using Hamming window. Thus the speech signal in time domain is converted into frequency domain. Power spectrum for the signal in each frame is computed thereafter.

In MFCC, the power spectrum is applied to Mel filter bank which consist of 24-26 triangular filter banks which are overlapped. They are arranged based on the mel scale. Calculate the filter bank energies thereafter and sum them up. Take the logarithm of it.

Humans won't hear loudness in a linear scale. To incorporate this feature, log nonlinearity is used in MFCC. DCT coefficients are calculated thereafter and we get the cepstrum coefficients. 2-13 coefficients are taken out for each frame, which are the mel- frequency cepstral coefficients.

In PNCC, log nonlinearity is replaced by power law nonlinearity. The power spectrum is passed through gammatone filter bank. This is then bias corrected and then raised to the power of 0.1 before calculating the cepstal coefficient [1]. The method incorporates a noise suppression algorithm which is based on asymmetric filtering that suppresses background excitation and temporal masking. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

noise suppression algorithm for environmental compensation makes use of medium time power, by computing the running average of the power observed in a single analysis frames[2].
In LPCC, coefficients are computed from smoothed autoregressive power spectrum instead of periodogram estimate of power spectrum. The inverse Fourier transform for the logarithm of AR Power Spectrum gives Linear Predictive Cepstral Coefficients.
In this paper, Feature extraction using Power law linear Prediction is done. This incorporates power law non linearity in speech spectrum estimation by compressing /expanding the power spectrum which is estimated with autocorrelation based linear prediction [1].
Speech Recognition system is build up by training the neural network with different speech recorded in a noise free environment. These words are stored in data base. The word spoken during testing is compared with the words stored in the data base. The speech is recognized correctly, if there is a match between the spoken word and words stored in the data base [5]. Gaussian Mixture Model (GMM) is the most widely used method for training and classification. Recent studies had proved that Deep Neural Network (DNN) gives a better efficiency for Speech Recognition System. In Speech Recognition System training and testing is done using speech of normal mode. In the proposed method, speech of any mode can be trained and tested. This will improve the efficiency of Speech Recognition System.
The method can be implemented in a Women Security System based on Speech Recognition. In a difficult situation, a woman may not be able to spoke words in a normal mode. She may shout and scream. Most probable words a woman may use at that time is trained and stored in the data base. Normal mode speech as well as in Shouted mode speech are used for training. The system will be able to recognize speech when tested with words spoken at shouted mode. This improves the efficiency of the Women Security System based on Speech Recognition System. Screamed voice will be of high pitch. So If pitch is appended with the cepstral coefficient along with delta and double delta values, the system efficiency improves drastically.

II. FEATURE EXTRACTION

A. Block Diagram

The simplified block diagram for feature extraction of different speech mode is shown below. The input speech signal is sampled at 16 KHz. The speech signal is separated into voiced and unvoiced speech based on short term energy and short term average zero crossing rates. Frames with very less energy and high zero crossing rates are unvoiced speech and hence they are discarded. Voices Speech is selected and is divided into frames of size 25 ms with 15ms overlap. Windowing is performed using Hamming window to avoid Gibbs phenomenon. Short term Fourier Transform is performed on the windowed signal with a DFT of size 1024. Thus the signal in time domain is converted into frequency domain. Power Spectrum is computed thereafter with Energy Calculation. According to the energy calculated, power spectrum is compressed/ expanded. Inverse Fourier Transform is performed on the modified power spectrum. Autocorrelation of the modified signal is derived out on taking IDFT of the modified power spectrum. Initialize the prediction order. According to the prediction order, Autocorrelated signal is truncated. With this LP coefficients are calculated. Using LP coefficients, LP model is formulated which gives the LP spectrum. From it we get the LP power spectrum [1]. LP spectrum contains only the formant of the voice. Thus we get a smooth spectrum. From the LP power spectrum, Mel Frequency Cepstral Coefficients are calculated where these coefficients are the features.

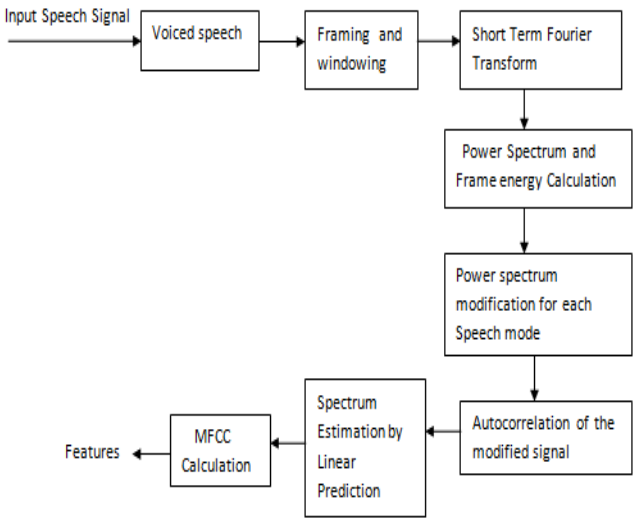

Fig 1: Block diagram for Feature Extraction at different speech mode

B. Power Spectrum Modification

The flowchart shows the Power Spectrum Modification for different speech modes is shown below. Time domain signal is converted into frequency domain and power spectrum is calculated along with frame energy. Power spectrum is modified based on frame energy. If the frame energy is above 35 dB, then the speech will be in shouted mode. Power spectrum is compressed by raising it to a very small value, very less than 1. If the frame energy is between 25 dB and 35 dB, the speech is of loud mode and the power spectrum is compressed by raising it to a value slightly less than 1. Neutral or Normal speech is between 15 dB and 25 dB. The power spectrum is keep as such. Between 5dB and 15 dB, Speech is soft. Power spectrum is therefore expanded by raising to a value slightly greater than 1. If the frame energy is less than 5 dB it corresponds to whispered speech. The power spectrum is expanded by raising it to a value very much greater than 1[1] and [4].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

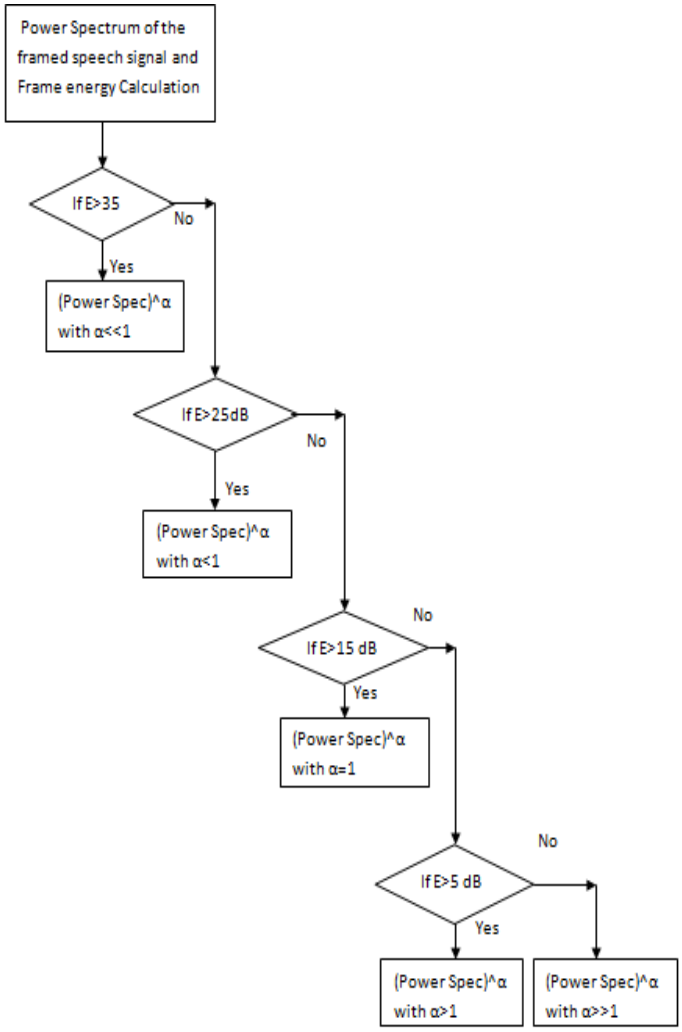

Fig 2: Power Spectrum Modification

The figure below shows the estimated spectrum for a voiced speech for different values of α. The value for α=1 gives the estmated spectrum for linear prediction in normal mode. If the speech is in shouted mode, we smooth out the power spectrum which are peaky with an α value less than 1. If the speech is in whisphered mode, peakiness of the power spectrum is inreased with the application of α value greater than 1.

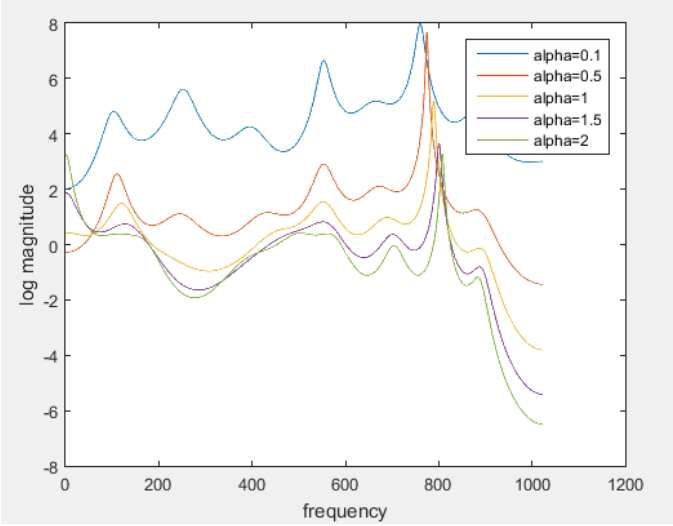

Fig 3: Modified power spectrum for different values of α

III. SHOUTED SPEECH RECOGNITION SYSTEM FOR WOMEN SAFETY

A. Shouted Speech Recognition

The simplified block diagram for shouted speech recognition is shown below. There are different approaches for speech classification and recognition. These include Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ) etc [3]. Artificial Neural Network (ANN) is widely used now a day for Speech Recognition and Speaker Identification which includes Feedforward Network, Recurrent Neural Network, and Self Organizing Maps etc. The diagram below describes a system which uses Deep Neural Network (DNN) for speech recognition. There are two phases: Training Phase and Testing Phase. In the training phase, features for different speech are extracted. The feature sequence is passed through Deep Neural Network (DNN) and based on it scores are obtained. The scores are stored in Database. In the testing phase, features of a particular speech are extracted. Score is obtained using DNN. The score obtained is compared with the already stored scores in the database. If there is a match, it means the word is already there. Thus the system recognizes the word. This is how a speech recognition system works. Feature extraction in the training and testing stage is done by the method describes above.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

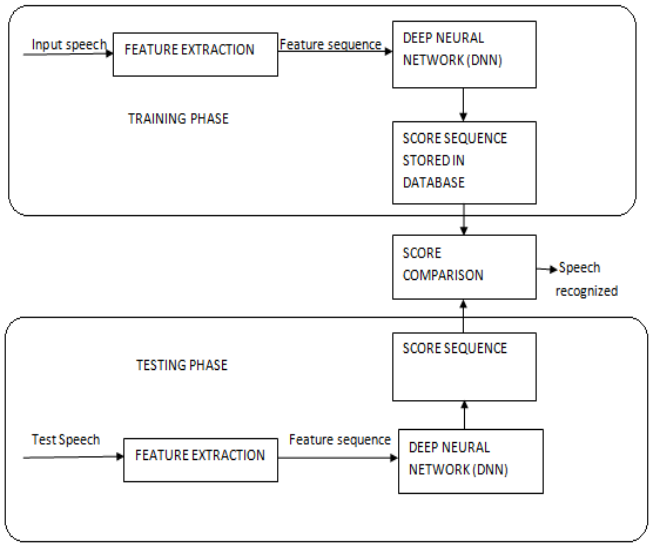

Fig 4: Shouted speech recognition system

B. Functional Block Diagram for Women Safety System

The functional block diagram for women safety system is shown below. The system is purely based on the recognition of shouted speech of a woman. MATLAB software is used for speech processing purpose. The system will be activated when a woman speaks a particular word either in normal mode or in shouted mode. The system will be trained by some set of words that can be used in emergency condition in prior. When the word is detected, the microcontroller based hardware generates the alarm. The screaming alarm makes use of real time clock to call out for help. The system tracks the location of the victim using GPS (Global Positioning System) and sends emergency message using GSM (Global System for Mobile Communication) to emergency contacts and nearby police station. The whole operation is monitored using LCD display.

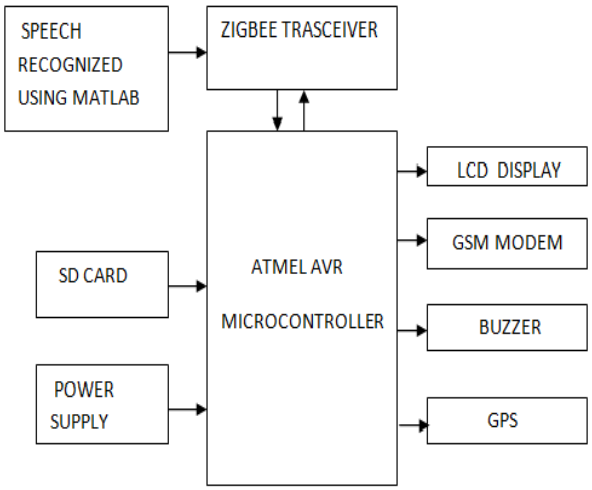

Fig 5: Block Diagram for women safety system.

IV. SUMMARY

In this paper, a different method for speech feature extraction which can be applied to any speech mode is explained. This is a different approach from the traditional feature extraction methods. The initial steps up to power spectrum calculation are same for all technique. Then power spectrum is modified using Power-Law Linear Prediction. The estimated spectrum is used for extracting Mel-frequency Cepstral Coefficients. Studies had proved that Deep Neural Network gives better recognition accuracy for speech. The application of the above described feature extraction method results an efficient speech recognition system that could work on speech other than normal speech mode. One of such application is women safety system which works on recognizing speech that is normal as well as shouted.

REFERENCES

[1] Rahim Saeidi, Paavo Alku, Tom Backstorm, "Feature Extraction Using Power-Law Adjusted Linear Prediction With Application to Speaker Recognition Under Severe Vocal Effort Mismatch," in *IEEE Trans. Audio, Speech and lang. processing, Vol No:24, No:1,January 2016*

[2] Chanwoo Kim, Richard M. Stern, "Power Normalized Cepstral Coefficient for Robust Speech Recognition," in *IEEE Trans. Audio, Speech and lang. processing, Vol No:24, No:7,July 2016*

[3] L. R. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall, 1993

[4] Chi Zhang and John H.L Hansen. "Analysis and classification of speech mode: Whispered through Shouted.

[5] Aleena Mary Paul, Anly Paul,Geethu M.M, Jishna N Gigi,Deepa Johnson, "Speech Controlled Automatic Slide Change," in *IJRCCT, Vol No.5,March 2016*