I.J. Image, Graphics and Signal Processing, **2016, 7, 22-31**
Published Online July 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2016.07.03

Efficient Acoustic Front-End Processing for Tamil Speech Recognition using Modified GFCC Features

Vimala. C, V. Radha
Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women,
Coimbatore – 641043, Tamil Nadu, India.
E-mail: vimalac.au@gmail.com, radhasrimail@gmail.com

Abstract**—**Giving suitable input and features are always essential to obtain better accuracy in Automatic Speech Recognition (ASR). The type of signal and feature vectors given as an input is highly essential as the pattern matching algorithms strongly depends on these two components. The primary goal of this paper is to propose a suitable Pre-processing and feature extraction techniques for speaker independent speech recognition for Tamil language. The five pass Pre-processing and three types of modified feature extraction techniques are introduced using Gammatone Filtering and Cochleagram Coefficients (GFCC) to achieve better recognition performance. The modified GFCC features using multi taper Yule walker AR power spectrum, combinational features using Formant Frequencies (FF), combined frequency warping and feature normalization techniques using Linear Predictive Coding (LPC) and Cepstral Mean Normalization (CMN) are investigated. The experimental results prove that the proposed techniques have produced high recognition accuracy when compared with the conventional GFCC feature extraction technique.

Index Terms**—**Gammatone Filter banks, Multi Taper window, Yule Walker AR, Formant Feature extraction, Cepstral Mean Normalization and Tamil Speech Recognition.

I. INTRODUCTION

Speech recognition became an active topic of research in recent years. It has been applied in many research areas like dictation, dialog system and voice based information search etc. Developing an ASR system comprises of two significant components, namely, speech front-end and back-end processing. The front-end processing is used to build unique models for each speech patterns and back–end processing is used to perform pattern matching [1] [2]. Figure 1 shows the basic components of an ASR system.

Speech front-end processing includes various tasks, namely, speech acquisition, Pre-processing and feature extraction. Back-end processing is used to create an unique model for each speech patterns by extracting the most essential properties of a speech signal based on Pre-processing and feature extraction techniques [3].

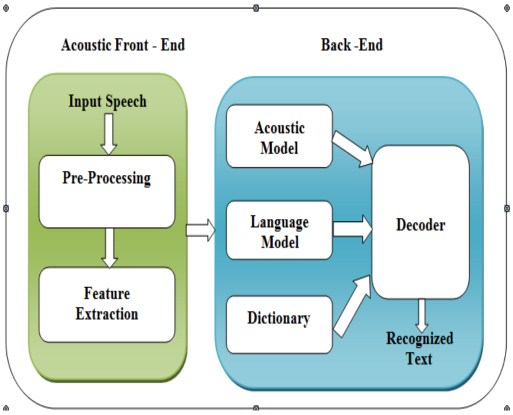

Fig.1. Basic Components of an ASR System

Pre-processing is performed in signal processing applications, in order to make useful spectral analyzes. The general Pre-processing techniques applied for signal processing are, Pre-emphasis, framing and windowing.

Feature extraction technique is used to transform the speech signal into useful parametric representations. These parameters are used to group the similar patterns and to recognize them [4]. The most popular feature extraction techniques used for speech related applications are Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Perceptual Linear Predictive (PLP) coefficients, wavelet features and auditory features. Even though many feature extraction techniques are available for speech recognition, selecting suitable features is always important, as the entire process depends on the feature vectors given as an input. Speech recognition systems never give good performance, if the selected features are not suitable [4].

To avoid this, psychological studies of the human auditory and articulatory systems are always necessary [1] [4]. During the recent years, GFCC features that are work based on an inner Ear model is widely used because of their significant improvements in the speech related applications [5].

However, for Tamil Speech Recognition most of the experiments are based on conventional MFCC and LPC features with HMM and Neural Networks (NN). Hence, *the particular contribution of our work is involving GFCC features with other machine learning techniques, like MLP and SVM.* This research work is mainly focused to propose efficient front-end processing techniques for Tamil Speech Recognition by implementing Five Pass Pre-processing and modified GFCC features.

The paper is organized as follows. Section 2 presents the related works on GFCC technique. Section 3 broadly discusses the proposed Pre-processing and feature extraction techniques introduced for Tamil ASR. The experimental results of the existing and proposed techniques are briefly presented in section 4. The findings and discussions are given in section 5. Finally, conclusion and future works are discussed in section 6.

II. Related Works

This section presents the related research work carried out particularly for GFCC features used for various speech related applications. The research findings from our previous experiments by using GFCC features are also presented at the end of this section.

R. Schluter, L. Bezrukov, H. Wagner and H. Ney (2007) say that the Gammatone features lead to competitive results for large vocabulary speech recognition. Further, different methods to combine Gammatone features with a number of standard acoustic features such as MFCC, PLP, MF-PLP and Vocal Tract Length Normalization (VTLN) were investigated. Best results were obtained when combining all features using weighted ROVER, resulting in a relative improvement of about 12% in word error rate compared to the best single feature system [5].

Hui Yin, Volker Hohmann and Climent Nadeu (2011) have proposed a variety of features based on Gammatone filter banks. The phase modulation is represented by the sub band Instantaneous Frequency (IF) and it is explicitly used by concatenating envelope-based and IF-based features [6]. Experiments are done with Chinese mandarin digits corpus under both clean and multi-condition using HMM. Their results prove that the proposed features can improve recognition rates in both conditions compared to the MFCC-based recognizer.

Shaveta Sharma and Parminder Singh (2014) have done Speech Emotion Recognition using GFCC and BPNN for English speech data. The authors have considered two emotions SAD and HAPPY [7]. The experiments were done under matlab and the results prove that the BPNN with GFCC feature extraction method performs better. The authors have also discussed about extracting GFCC features for three emotions namely, sad, happy and angry (2015) [8].

P.K. Sahu, Astik Biswas, Anirban Bhowmick and Mahesh Chandra (2014) have discussed about auditory Equivalent Rectangular Bandwidth (ERB) based admissible wavelet packet features for TIMIT phoneme recognition [9]. The authors has introduced a new filter structure using admissible wavelet packet for English phoneme recognition. In the proposed filters, the Central frequencies of ERB scale are equally distributed along with the frequency response of human cochlea. The new sets of features derived from wavelet packet transform with multi resolution capabilities were found to be better than conventional features as the frequency bands spacing is similar to the auditory ERB scale. In order to test the robustness of wavelet based features, various noise conditions are involved using NOISEX-92 database. The experimental outcome proved that the WERBC is better when compared to the WMFCC especially in case of noisy condition.

Shruti and Bharti Chhabra (2016) have implemented a singer identification technique using Artificial Neural Network (ANN). The authors have focused on the basic concepts of the feature extraction and classification techniques in speech identification system [10]. The total dataset of 45 songs are used by involving 9 singers and 5 songs of each singer. In their work, DCT is applied to derive cepstral features, GFCC is used for feature extraction and ANN is applied to classify. For the experiments, 88.9 % of singers were correctly identified using the above mentioned techniques.

Hari Krishna Maganti and Marco Matassoni (2014) have discussed about Auditory processing-based features for improving speech recognition in adverse acoustic conditions [11]. The proposed features incorporates a combination of gammatone filtering, modulation spectrum, non-linearity emulating the cochlear and the middle ear to improve robustness. The experimental results revealed that the proposed features provide reliable and considerable improvement in terms of robustness in different noise conditions by using standard Aurora-4 large vocabulary database. Their future work is to focus on evaluating the performance of the proposed features for reverberant environments and large vocabulary tasks.

Shaik Shafee and B.Anuradha (2016) have done an experiment for speaker identification and spoken word recognition in noisy background using ANN [12]. The main objective of the work is to find the better combination of speech feature extraction and ANN for speaker identification combined with spoken word recognition in general noisy environment. Experiments are done by using Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) Cepstral Coefficients and Gammatone Frequency Cepstral Coefficients (GFCC) in combination with Radial Basis Neural Networks and Learning Vector Quantization. Three different test categories such as Spoken word recognition, Speaker Identification, and the combination of both speaker and spoken word recognition have been experimented for the above mentioned combinations. From the above results it is suggested that Radial basis neural networks with GFCC can be chosen for the combination of both spoken word recognition and speaker identification in general noisy conditions. It was also observed that the Radial basis network models were found to be less time consuming compared to Learning Vector Quantization neural networks.

In our previous experiment (2014), four types of feature extraction techniques namely, MFCC, LPC, PLP and GFCC are implemented with DTW, HMM, MLP, SVM and Decision Tree techniques [13]. Experiments are done with 10 Tamil spoken digits (0-9) and 5 Tamil spoken names collected from 4 different speakers with 10 repetitions. The total dataset size of 600 (15*4*10) were used and the performances of the above algorithms were deeply observed. Based on the outcome, it was found that the GFCC features provided better results and outperformed the other features for all the above mentioned speech recognition techniques. Particularly, the HMM, MLP and SVM techniques were found to be better for Tamil speech recognition. In another experiment (2015), the utterance level and speaker level performance for the adopted speech recognition techniques were clearly presented [14].

It is clearly observed from the prior experiments and analyzes that, the GFCC technique has outperformed all the other feature extraction techniques for Tamil speech recognition. Since, GFCC technique was found to be better for speaker independent Tamil speech recognition, next attempt is made for further improvements using *modified GFCC feature extraction techniques*. The subsequent section briefly explains about the proposed front-end processing.

III. Efficient Acoustic Front-End Processing

Presenting good speech input and suitable feature vectors are very essential to achieve better results in ASR. Because, any speech pattern matching technique strongly depends on the type of input signal and feature vectors given as an input. The proposed technique addresses these two factors by producing better speech input signal and suitable features as an effective speech front-end processor. For this purpose, five pass pre-processing and modified GFCC feature extraction techniques are proposed and they are explained below.

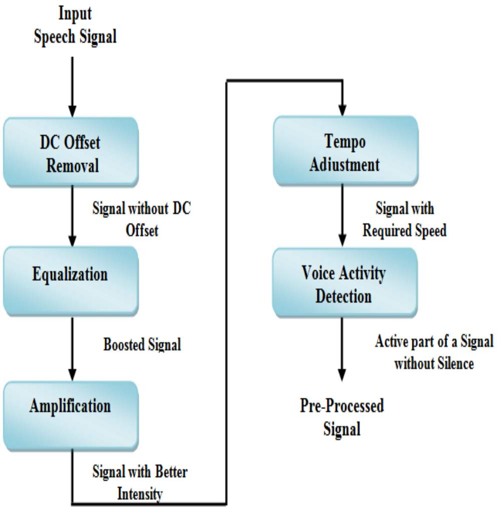

Fig.2. Five Pass Pre-processing

A. *Five Pass Pre-processing*

Pre-processing is always required to improve the quality and clarity of an input. The speech signals need to be pre-processed before extracting the feature vectors, in order to increase the accuracy and efficiency of the speech recognition task. In this paper, *five pass pre-processing has been proposed using a series of five significant pre-processing techniques, namely, DC offset removal, amplification, equalization, tempo adjustment and silence removal* using Voice Activity Detection (VAD). Figure 2 shows the steps involved in five pass Pre-processing technique.

1) DC offset Removal

Removing the DC offset from the speech signal is an important step in speech processing. DC offset generally refers to the mean amplitude, which is added to the signal during speech acquisition. It is normally undesirable in any signal processing related task, because it causes inaudible low level distortion to the input signal. Hence, it needs to be removed before performing Analog-to-Digital (A-D) conversion. Removal of DC offset performs a calculation to make the average positive and negative sample value. Therefore, the *mean amplitude value will become zero once the DC offset is being removed from the signal* [15].

2) Equalization

Equalization is an important technique which is mainly applied to manipulate the signals by boosting or reducing some specific frequencies. It removes any signal distortions caused by low and high frequency components. In this work, boosting has been applied between 80Hz and 5000Hz and the frequencies below this specified range are reduced. *This technique helps to understand the spoken words much better*. However, equalization may reduce the loudness of a signal. Thus, it is used before performing amplification.

3) Amplification

Speech recognition applications are generally affected by many types of signal variations. Among them, the most common variation occurs due to the change in intensity also called amplitude. These variations are assumed to be slowly changing, which allows the system to update the amplitude scaling factor at relatively long intervals. The sensitivity of a human Ear varies according to the frequency and quality of the sound, although signal with a greater intensity level usually sounds louder.

Signal with better intensity can help to distinguish speech with one another. However, if the amplitude is not properly normalized, it provides limited performance for the ASR. Therefore, if the loudness of a signal is good, then the system can achieve better recognition performance.

Based on the above discussions, the amplitude adjustment is employed to modify the loudness of an input signal. The loudness provides the information about the intensity or energy of a signal. Better intensity information can be extracted by strengthening the loudness of an input signal. It also helps to

- Improve the speech signal to maintain the original speech intensity.
- Reduce variations during speech recordings.

- Reduce common energy loss in transmission.

In order to achieve the above quality of an input signal, the amplitude adjustment technique has been executed. The amplification is applied to set the maximum volume required for the given signal. It is calculated and adjusted automatically without clipping. In this research work, an amplification level is set to 2.7dB. *Both amplification and equalization techniques help to reduce within word variation*.

4) Tempo Adjustment

Speaking rate is a significant factor in speech recognition applications. Therefore, speaking in a faster or slower manner also has influence on the speech signal. However, even normal speakers will have a tendency to speak faster when using a speech recognition system. But, speaking rate affects both temporal and spectral characteristics of the signal. Therefore, the performance of the acoustic model will also degrade. Benzeghiba, M (2007) says that, the faster speaking rates may also result in more frequent and stronger pronunciation changes [16].

Likewise, Matthew Richardson *et al*. (1999) say that, the state-of-the-art of an ASR system perform significantly worse on fast speech [17]. Therefore, in this research work, *tempo adjustment is employed, to change the speed of an input signal, without modifying the pitch value*. As the proposed work involves different speakers who speak in a slow and fast manner, this step helps to maintain the required speed of a signal and also assist to reduce the speaking rate variation.

5) Silence Removal using Voice Activity Detection (VAD)

Voice Activity Detection algorithm is widely used to identify the voiced and unvoiced region of a speech signal [18]. In general, a speech or speaker specific attributes are located in the voiced part, whereas the other undesirable components like silence or the background noise are located in the unvoiced part. Therefore, making useful discrimination between the voiced and unvoiced part can help to remove the irrelevant segment of a speech signal. The user/speaker usually takes a few seconds before and after saying a word while recording a speech utterance. So, the first and last 200 msec of a recorded speech signal might contain silence or irrelevant speech information. Therefore, *silence in the beginning and at the end of speech file is removed to reduce the end point detection error*.

The original input speech signal is passed through the above five steps. As a result, an effective pre-processed signal is obtained with better quality and intelligibility. Also, the *proposed pre-processed signals are found to be louder and clear when compared with original signals*. Figure 3 and Figure 4 show the spectrogram of the original and Pre-processed input speech signal for the word “poojiam” uttered by four different speakers respectively. These Pre-processed signals are then used as an input for the modified feature extraction using GFCC technique.

B. Modified Feature Extraction using GFCC

Speech recognition performance can vary according to the type of feature extraction technique adopted for the particular application. The proposed methodology focuses on providing suitable features as an effective speech front-end using three different modified GFCC features. The subsequent sections explain the same in detail.

1) Feature Extraction using Multi Taper Yule Walker AR - GFCC (MTYW-GFCC)

An efficient feature extraction technique should extract the most useful spectral information from a signal, which can improve the recognition performance. This can be accomplished by using Power Spectral Density (PSD) estimation that extracts the frequency response of a signal. A power spectrum shows the amplitudes of all frequency components present in a segment of signal that can provide better discrimination between the speech segments. It helps to understand where the average power is distributed as a function of frequency.

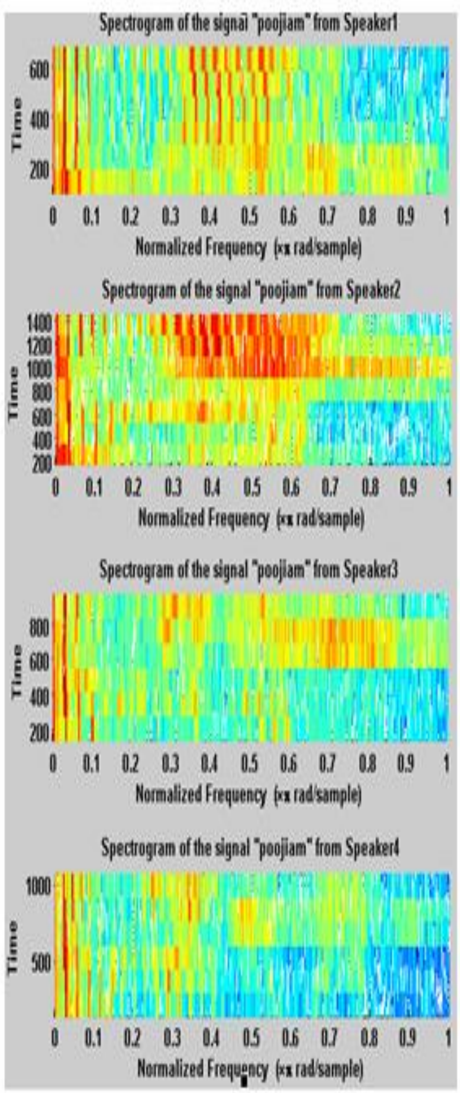

Fig.3. Spectrogram of the Original Input Speech Signal for the Word “Poojiam” Uttered by Four Different Speakers.

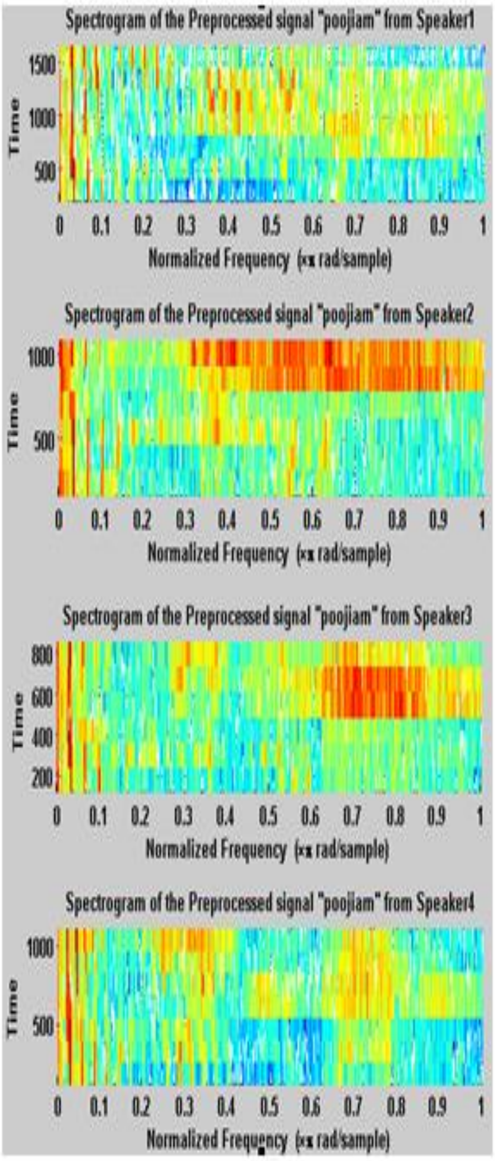

Fig.4. Spectrogram of the Pre-Processed Input Speech Signal for the Word “Poojiam” Uttered by Four Different Speakers.

One of the efficient techniques used for this purpose is windowing method. The most commonly used method is hamming window which has some limitations. Using hamming window, only one sample can be extracted in each segment. In general, taking a single sample from a signal does not provide an efficient estimate of its spectral properties and it does not help to achieve robust performance. The above problem can be solved by using the multi taper windowing method, where it is *possible to take multiple independent spectral components from a same sample* [19].

Multi Taper Windowing

In multi taper windowing, each data taper is multiplied element-wise to estimate the signal power at each frequency component. By optimizing a filter function, multiple windows can be derived which can effectively control the sidelobe and can prevent leakage of frequencies outside the bandwidth resolution. The resultant windowed signal provides the statistically independent estimate of the underlying spectrum as each taper is pair-wise orthogonal to all the other tapers. The final spectrum is obtained by taking an average of all the tapered spectra. The experimental results have proved that the multi taper method provides small bias and low variance, as long as the PSD spectrum is flat and the frequency resolution is predetermined. The resultant signal obtained from multi taper windowing is then passed to the PSD for performing spectral analyzes.

Yule-Walker AR Power Spectrum

Parametric techniques are better than non-parametric techniques in performing PSD estimation. In the previous experiments, the performance evaluation of windowing methods and PSD estimation is done [20]. The performance comparison of Yule Walker AR and Welch methods are evaluated based on both subjective and objective measures. It is noticed from the experimental outcomes that, the Yule walker AR method has produced better results. Based on the outcome, the Yule-Walker method is implemented to estimate the PSD of an input speech signal. It fits an Auto Regressive (AR) model to the windowed input data so as to minimize the forward prediction error in the least squares sense. This is done by Levinson-Durbin recursion algorithm. The output column vector contains the estimate of the PSD with equally spaced frequency points.

Multi Taper Windowing for Yule Walker AR Power Spectrum

The *particular novelty of the proposed method* is, instead of performing a regular feature extraction from an input signal, the metrics of multi taper windowing, power spectrum and weight estimation are additionally given as an input to the feature extraction. It is implemented as follows:-

- The pre-processed signal which is divided into short frames are multiplied by the multi taper windowing method,
- The spectrum and the multi taper weight have been estimated,
- Subsequently, Yule walker AR power spectrum is estimated using Levinson-Durbin Recursion algorithm, and
- In the final stage, mean value of a spectrum, multi taper weight and the estimated power spectrum from the Yule Walker AR method has been added to the signal.

The detailed descriptions of the steps involved in MTYW-GFCC feature extraction are presented in the following algorithm. In this research work, better performance has been achieved when the frame size is assigned to 240, frame shift is assigned to 120 and the number of tapers is assigned to 8. By using 4th order Yule Walker AR spectrographic analyzes, an important region of signal which is related to the speech signal is determined.

The experimental results confirm that the MTYW-GFCC features have increased the recognition accuracy and it is presented in section 4. Based on the significant improvements achieved by the above feature extraction techniques, the combinational features are implemented subsequently.

Proposed Algorithm

Step 1: Take an input signal,
Step 2: Divide the signal into N frames,
Step 3: Perform multi taper windowing, where, matrix of size (frame size X number of windows) containing the tapers (Window Functions) as column vectors,
Step 4: Estimate the weights based on the vector length, number of windows and number of FFT bins,
Step 5: Perform the PSD estimation for each frame (one frame per column),
Step 6: Take the mean value of power spectrum,
Step 7: Multiply the signal using estimated multi tapered window,
Step 8: Perform Yule Walker AR PSD estimation,
Step 9: Add Multi peak weights and the values obtained from ***step 6,7 and 8***, to the multi taper windowed signal and generate a new speech signal,
Step 10: Pass the resultant signal obtained from ***step 9*** to the gammatone filtering, and
Step 11: Extract the MTYW-GFCC feature vectors.

2) Combinational Features using MTYW-GFCC with Formant Frequencies (MTYW-GFCC-FF)

In order to make discrimination between each word, the combinational features are proposed by combining the MTYW-GFCC features with Formant Frequencies. Formant frequencies are vocal tract resonances, the frequency of which depends on the length of the tongue or tongue advancement. It can effectively distinguish the meaningful frequency components of human speech patterns. The first two formants are particularly important in speech recognition. At least three formants are generally required and up to five formants are needed for achieving high performance. In this work, the first five formant frequencies are obtained from the polynomial roots generated by the LPC that represent the vocal tract filter.

Using the normalized FFT, the amplitude of the formant frequencies is obtained. Figure 5 shows the first five formant frequencies extracted for the word "poojiam". Since these features are specifically used to distinguish vowels, it helps to recognize Tamil spoken words depend on the vowels present in an input signal. Better results are achieved with these combinational features, since the formant frequencies have the ability to reduce the mismatch between training and deployment environment.

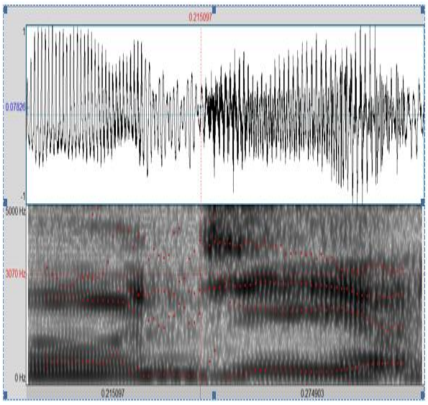

Fig.5. First Five Formants Extracted for the Word "poojiam"

3) Frequency Warping and Feature Normalization using LPC and CMN (FWCMN-MTYW-GFCC-FF)

Next, an attempt is made to reduce the speaker and channel variations which affect the ASR performance. For this purpose, frequency warping and feature normalization techniques are implemented after applying the above proposed combinational features. In this research work, initially, the frequency warping is applied to the five pass pre-processed input signal and then the modified features and combinational features are extracted. Finally, the combinational feature vectors are normalized by applying the CMN technique. Figure 6 shows the frequency warping and feature normalization using LPC and CMN.

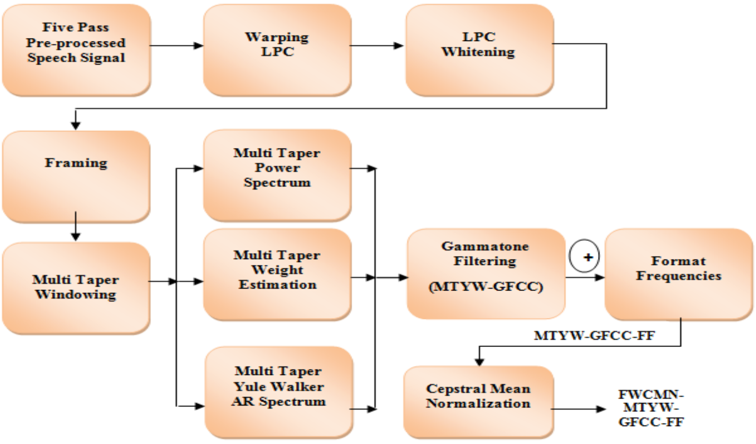
Five Pass Pre-processed Speech Signal
Warping LPC
LPC Whitening
Framing
Multi Taper Windowing
Multi Taper Power Spectrum
Multi Taper Weight Estimation
Multi Taper Yule Walker AR Spectrum
Gammatone Filtering (MTYW-GFCC)
+
Format Frequencies
MTYW-GFCC-FF
Cepstral Mean Normalization
FWCMN-MTYW-GFCC-FF

Fig.6. Frequency Warping and Feature Normalization using LPC and CMN

Variations between words will increase when the number of speakers try to utter the same word in different ways. Moreover, the speaker characteristics may vary due to the difference in the glottal waveforms, vocal tract lengths and speaking rates. All these variations can be eliminated by using the frequency warping technique which can assist to normalize the Vocal Tract Length (VTL).

The length of the VTL is an important acoustic variation among speakers, where, the length of the vocal tract is different for male and female speakers (i.e. female VTL is shorter than male VTL). Based on this characteristic, the formant frequencies produced by different speakers can vary up to 25% [21]. It causes a serious mismatch between different speakers due to the change in vocal tract shape [22]. This issue can be reduced by normalizing the VTL.

Frequency Warping using LPC

Frequency warping has a high level of significance in improving the performance of the speaker independent speech recognition systems [23]. As speaker independent system involves different speakers, the frequency warping is performed by applying the time varying all pole filter using LPC and it is explained below.

Warped Linear Predictive Coding (WLPC)

The standard form of LPC analyzes is employed to divide the signal into a smoothed spectral format. In this research work, warping LPC is applied to change the frequency resolution of an input speech signal. WLPC is an alternative technique of LPC, where the spectral representation of the system is customized using the first order all pass filter. For this purpose, all the unit delay elements are replaced by the 1st order all-pass filters.

Applying WLPC has the following benefits.

- It facilitates the signal by warping the spectra,
- Minimizes the speaker variations and help to preserve the important speech information,
- Reduces the bit rate required for a given speech signal, and
- Improves the intelligibility and the naturalness of the speech without changing pitch.

Applying WLPC is used to shift the resonant frequencies of the LPC model to the Infinite Impulse Response (IIR) filter, by substituting an all-pole system for each delay element. The LPC warping technique is used to warp an all-pole filter defined by numerator coefficients using a first order all pass substitution with alpha value. It generates a new filter with poles and zeros defined by polynomials *B* and *A*. In this research work, a 12th order LPC autoregressive model is used for warping an input signal and the resultant signal contains an all pole filter coefficients. Next, it resynthesizes the resultant LPC parameters using the noise excitation. The warping parameter α is selected between 0.1 and 0.3. Subsequently, LPC whitening method is adopted to whiten the signal and to preserve the formant frequency components of a signal. It is also be useful for whitening the noisy components present in the signal and maintains the frequency attribute of the signal.

Thus, the WLPC can effectively normalize the vocal tract length variations and moreover the frequency warped signal has also improved the perceptual quality and the intelligibility of the signal. The frequency warped signal is then used as an input for the above proposed techniques where the modified features are extracted. Finally, the resultant feature vectors are normalized using Cepstral Mean Normalization (CMN) as described below.

Feature Normalization using Cepstral Mean Normalization

The characteristics of a signal can change according to the microphone distance and the type of microphone used for speech acquisition. CMN is the most widely used to eliminate the speech signal components suffer from channel distortions. It effectively works in removing additive noise as well as channel noise. It uses the simplest method for feature normalization, by forcing the mean value of each element of the Cepstral feature vector, to be zero for all utterances [24].

Generally, the mean value of the signal conveys the spectral characteristics of the microphone and room acoustics, and it is not often reliable in signal processing. The mean value of the Cepstral coefficients is calculated across the whole utterance and then subtracted from each frame. The CMN is performed as follows:

Let X={X_0,X_1,X_{T-1}} be the Cepstral vectors computed using short term analyzes, then the sample mean is represented by the following expression (1).

$$\overline{X} = \frac{1}{T}\sum_{t=0}^{T-1} X_t \tag{1}$$

and, the CMN is represented as follows (2).

$$\hat{X}_t = X_t - \overline{X} \tag{2}$$

where, signal corresponding to X_t is processed by a linear filter. ***CMN can help to reduce three types of distortions, namely, environment distortions, channel distortions, and intra speaker effects***. Therefore, the robustness can be achieved once the mean vector is subtracted from the feature vectors. The performances of the above proposed techniques are discussed in the next section.

IV. Experimental Results

The performance evaluation of the existing and proposed techniques are presented in this section. Experiments are done with 10 Tamil spoken digits (0-9) and 5 spoken names from 30 different speakers. To make utterance variation, the speakers uttered the same word at different interval of time. The utterances consist of 10 repetitions from 15 male and 15 females. The total size of the dataset is 15*30*10=4500. The utterances were recorded at 16 KHz sampling rate using audacity software at a silence environment.

In this experiment, 60% of dataset is given for training and the remaining 40% of dataset are used for testing. The performances of these techniques are analyzed based on two metrics, namely, Word Recognition Rate (WRR) and Real Time Factor (RTF).The results obtained by applying the proposed five pass pre-processing and three modified GFCC techniques are shown in the following Table 1.

Table 1. Performance Evaluation of Proposed Techniques

Speech Recognition Techniques	GFCC	GFCC with Five Pass Pre-processing	MTYW-GFCC	MTYW-GFCC-FF	FWCMN-MTYW-GFCC- FF
HMM	**96**	**97.18**	**97.81**	**98.54**	**99.06**
MLP	**93.53**	**95.83**	**96.15**	**96.15**	**96.15**
SVM	**93.52**	**95.93**	**96.14**	**96.14**	**96.14**

From the above results, it is clear that the proposed front-end processing techniques are lead to competitive results for Tamil speech recognition. The usage of the above techniques has substantially improved the recognition rate for the dataset used. As per the experimental outcomes, it is shown that the five pass pre-processing technique has demonstrated the significant performance improvements for all the three speech recognition techniques adopted for this study. The improvements in WRR of 1.19%, 2.98% and 2.99% has been achieved for the HMM, MLP and SVM techniques respectively. In same way, the developed MTYW-GFCC features have increased the recognition accuracy for all the three speech recognition techniques involved in this research work.

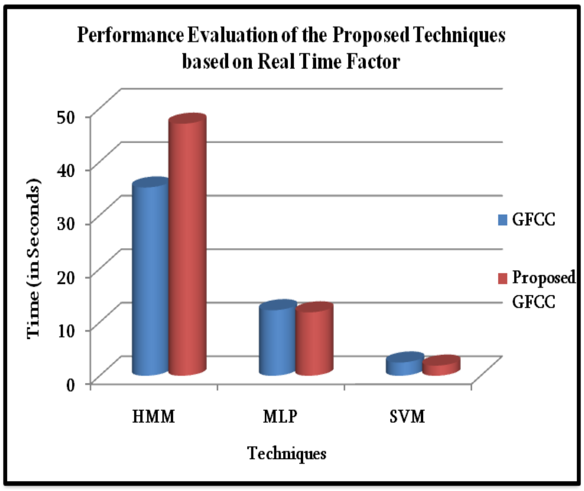

Fig.7. Performance Evaluation of Proposed Technique based on Real Time Factor

Likewise, very good results are achieved by using MTYW-GFCC-FF, where the WRR of 98.54% has been achieved for the HMM technique. Similar to the above proposed techniques, it is evident from the experiments that the FWCMN-MTYW-GFCC-FF has increased the WRR up to 99.06% for the HMM technique. Also, the techniques have some significant improvement in reducing the processing time. The average testing time taken for MLP and SVM techniques has been reduced with the proposed techniques which is shown in Figure 7.

V. Findings and Discussions

The main objective of this paper is, to propose a suitable pre-processing and feature extraction techniques for speaker independent isolated speech recognition for Tamil language. Accordingly, the merits of GFCC technique have been considered and improved using five-pass Pre-processing and modified feature extraction techniques.

In the proposed work, initially the five pass preprocessing technique has been applied and the improvements are tested with existing GFCC feature extraction method. Then, the GFCC feature extraction has been modified with multi taper and Yule walker AR power spectrum method. Subsequent improvements are achieved by using the MTYW-GFCC features for all the three techniques involved. Based on the improvements gained with the MTYW-GFCC features, the combinational features using formant frequencies were implemented. The combinational features were also improved the recognition accuracy for HMM technique.

Furthermore, in order to reduce both speaker and channel variations, frequency warping and feature normalization techniques were implemented. The FWCMN-MTYW-GFCC-FF has provided very good results for HMM when compared with the MLP and SVM. The reason is that, the statistical methods can learn as much information as possible from the data to build unique model for each speech patterns. The learning based approaches works based on the fewer assumptions made on input data. It was found that continues improvements were gained with the proposed pre-processing and modified GFCC features. The highest recognition rate achieved with the existing GFCC using HMM, MLP and SVM technique are 96.5%, 92.8% and 92.9% respectively. The maximum accuracy of 99.06%, 96.15% and 96.14% was achieved with the HMM. From the outcome, it is proved that better results were achieved for all the speakers enrolled in this study.

VI. Conclusion and Future Work

The main objective of this research work is to propose a novel preprocessing and feature extraction technique for speaker independent isolated speech recognition system for Tamil language. For Tamil speech recognition, only MFCC and LPC feature extraction techniques are implemented using HMM and Back propagation techniques. The most recent feature extraction techniques like GFCC and other machine learning techniques were not carried out for Tamil ASR. Hence, in this research work the GFCC and other machine learning techniques were implemented. The results proved that the GFCC features are best suited for all the recognition techniques involved in this work. Further, the best methods have been chosen and the proposed techniques were applied and its performances are verified.

One pre-processing technique and three different feature extraction technique were proposed namely MTYW-GFCC, MTYW-GFCC-FF and FWCMN-MTYW-GFCC-FF. The proposed pre-processing and feature extraction techniques has moderately increased the performance of the Tamil speech recognition based on WRR and RTF. Better results were achieved with HMM, MLP and SVM techniques with increased accuracy. It is also observed that the proposed methods have reduced the testing time for MLP and SVM techniques. Based on the improvements achieved with the proposed methods, this research work will be further extended for Tamil ASR under different noisy conditions in future.

REFERENCES

[1] Urmila Shrawankar and Vilas Thakare, "Techniques for Feature Extraction in Speech Recognition System: A Comparative Study", *International Journal of Computer Applications in Engineering, Technology and Sciences (IJCAETS)*, pp. 412-418, 2010.

[2] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, "Speech Recognition using MFCC" *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012,)* July 28-29, Pattaya (Thailand), 2012.

[3] J.R Deller, J.G. Proakis and F.H.L. Hansen, Discrete-*Time Processing of Speech Signals*, IEEE Press, chapter 12, 2000.

[4] Y. Lee and K.W. Hwang, "Selecting Good speech Features for Recognition", *ETRI Journal*, Vol. 18(1), 1996.

[5] Hui Yin, Volker Hohmann and Climent Nadeu, "Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency", *Speech Communication* 53, pp. 707–715, 2011.

[6] R. Schluter, L. Bezrukov, H. Wagner and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition" *in ICASSP 2007*, Vol.4, pp. 649–654, 2007.

[7] Shaveta Sharma and Parminder Singh, "Speech Emotion Recognition using GFCC and BPNN", *International Journal of Engineering Trends and Technology (IJETT)*, Vol.18 (6), pp, 321-322, ISSN: 2231-5381, 2007.

[8] Shaveta Sharma and Parminder Singh, "Extracting GFCC Features for Emotion Recognition from Audio Speech Signals", *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, Vol.5(1), pp.89-91, ISSN: 2277 -128X, 2015.

[9] P.K. Sahu, Astik Biswas, Anirban Bhowmick and Mahesh Chandra, " Auditory ERB like admissible wavelet packet features for TIMIT phoneme recognition", *Engineering Science and Technology, an International Journal*, Vol. 17, pp. 145-151, 2014.

[10] Shruti and Bharti Chhabra, "An Approach For Singer Identification Technique Using Artificial Neural Network", *International Journal of Engineering Research and Modern Education (IJERME)*, Vol. 1(1), pp-16-23, ISSN (Online): 2455 - 4200, 2016.

[11] Hari Krishna Maganti and Marco Matassoni, "Auditory processing-based features for improving speech recognition in adverse acoustic conditions", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 1(21), pp- 1-9, 2014.

[12] Shaik Shafee and B.Anuradha, " Speaker Identification and Spoken word Recognition In Noisy Background using Artificial Neural Networks, *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016*, IEEE.

[13] C. Vimala and V. Radha, Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words, *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 5 (1), pp. 378-383,ISSN:0975-9646, 2014.

[14] C. Vimala and V. Radha, Isolated Speech Recognition System for Tamil Language using Statistical Pattern Matching and Machine Learning Techniques, *Journal of Engineering Science and Technology (JESTEC)*, Vol. 10 (5), pp.617-632, 2015.

[15] Abhishek Singh and Pravin Katwe, "Study of decaying dc removal techniques", *Bachelor Thesis in Electrical Engineering*, National Institute of Technology, Rourkela, 2010.

[16] M. Benzeguiba, R.D Mori, O. Deroo, S. Dupon, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic Speech Recognition and Speech Variability: a Review", *Speech Communication* 49, pp. 763–786, 2006.

[17] Matthew Richardson, Mei-Yuh Hwang, Alex Acero, and Xuedong Huang, "Improvements on Speech Recognition for Fast Talkers", *Proceedings of the Euro speech Conference*, 1999.

[18] Nitin, N. Lokhande, Navnath, S. Nehe and Pratap, S. Vikhe, "Voice Activity Detection Algorithm for Speech Recognition Applications", *IJCA Proceedings on International Conference in Computational Intelligence (ICCIA)*, Vol. 6, pp. 5-7, 2011.

[19] M. Hansson and G. Salomonsson, "A Multiple Window Method for Estimation of Peaked Spectra", *IEEE Transaction on Signal Processing*, Vol.45 (3), pp. 778-781, 1997.

[20] V. Radha, C. Vimala and M. Krishnaveni, "Power Spectral Density Estimation using Yule Walker AR method for Tamil Speech Signal", *International Conference on Information Systems for Indian Languages (ICISIL 2011)*, Springer, pp.284-288, ISBN:978-3-642-19402-3-1865-0929, 2011.

[21] L. Lee and R.C Rose, "Speaker Normalization Using Efficient Frequency Warping Procedures", *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96*, Atlanda, GA, pp. 353-356, 1996.

[22] Wakita. K, "Normalization of vowels by vocal-tract length and its application to vowel identification", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 25, pp. 183–192,1997.

[23] Hemant Misra, "Multi-stream Processing for Noise Robust Speech Recognition", *Doctoral thesis*, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, March 2006.

[24] http://www.ee.uwa.edu.au/~roberto/research/speech/local/entropic/HAPIBook/node85.html.

Authors' Profiles

Dr.Vimala.C done her Ph.D in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women. She has more than 2 years of teaching experience and 3 years of research experience. She worked as a Project Fellow for the UGC Major Research project.

Her area of specialization includes Speech Recognition, Speech Synthesis and speech signal enhancement. She has 16 publications at National and International level conferences and journals.

Dr.V.Radha, Professor in Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India. She has more than 26 years of teaching experience and 16 years of Research Experience.

Her Area of Specialization includes Image Processing, Optimization Techniques, Speech Signal Processing, Data Mining & Data Warehousing and RDBMS. She has authored more than 99 papers published in refereed International journals and Conferences. She has obtained funding projects from UGC-MRP in the field of speech signal processing. She is a Reviewer of American Journal Operational Research and American Journal of Signal Processing. She is an Editor in Chief of International Journal of Computational Science and Information Technology (IJCSITY). She is the member of many international bodies such as IAENG, AIRCC, IACSIT etc.. Visited countries such as the USA and Singapore.

How to cite this paper: Vimala. C, V. Radha,"Efficient Acoustic Front-End Processing for Tamil Speech Recognition using Modified GFCC Features", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.8, No.7, pp.22-31, 2016.DOI: 10.5815/ijigsp.2016.07.03