ISSN:0975-9646

Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words

Vimala.C, Radha.V

Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamil Nadu, India

***Abstract*— Speech feature extraction which attempts to obtain a parametric representation of an input speech signal plays a crucial role in the overall performance of an Automatic Speech Recognition (ASR) system. A good feature extraction technique must capture the important characteristics of the signal also should discard some irrelevant attributes. The main motivation behind this paper is to provide a suitable feature extraction method for the speech recognition system based on various analyses. Among the various feature extraction methods available today the recent attempt on Gammatone Filtering and Cochleagram Coefficients (GFCC) which purely represents auditory features provide promising results and also improves the robustness of an ASR system. Hence, the main objective of this paper is to evaluate the performance of gammatone filter bank features with the conventional feature extraction techniques. Also, the metrics of these features are investigated with the most popular speech recognition techniques namely Dynamic Time Warping (DTW), Hidden Markov Model (HMM), Multi Layer Perceptron (MLP), Support Vector Machine (SVM) and Decision Trees. For the experiments, the speaker independent isolated speech recognition system for Tamil language using various feature extraction and pattern matching techniques has been designed and developed. The most suitable feature vectors for Tamil isolated speech recognition are discovered based on various analyses and experiments. Based on the study it is observed that the GFCC features outperformed the conventional features and achieved better results. For this work, highest word recognition accuracy is achieved with GFCC features for both training and testing data.**

***Keywords*— Tamil Speech Recognition, Feature Extraction, Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Gammatone Filter banks.**

I. INTRODUCTION

Automatic Speech Recognition (ASR) is an important task in digital signal processing related applications. It is the process of automatically converting the spoken words into written text by the computer system. Over the past few decades speech recognition has made widespread technological advances in many fields such as call routing, automatic transcriptions, information searching, data entry etc. Speech recognition has been accomplished by combining various algorithms drawn from different disciplines such as statistical pattern recognition, signal processing and linguistics etc [1]. Among them, feature extraction also called signal processing front-end, which converts the speech signal into some useful parametric representation has a greatest importance. It extracts a small amount of data from the speech signal which are used to build a separate model for each speech utterance or each speaker. This parametric representation is then used for further analysis to represent the specific speech utterance or speaker. These feature characteristics can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for various task. The most successful methods also include the attributes of the psychological processes of human hearing into analysis. Many feature extraction methods use cepstral analysis to extract the vocal tract component from the speech signal [2]. The Fig.1 explains the methodology of the developed speech recognition system.

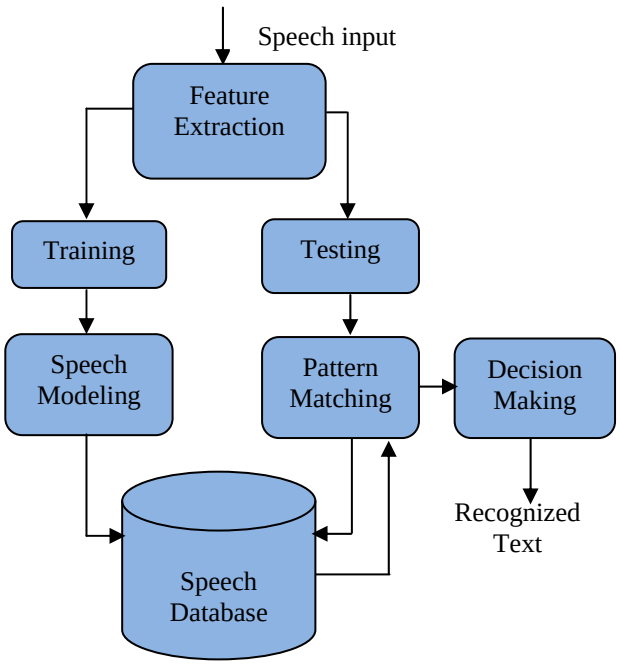

Fig.1 Methodology of the Developed Speech Recognition System

Initially, the speech signal is given as an input for the system. Next, the signal is divided into smaller frames where the useful feature vectors are extracted. Then, these feature vectors are divided into training and testing features. From the training features the speech model has been developed. During testing process, the pattern matching analysis will be done against the reference

features which are stored in the speech database. Finally, decision will be taken based on the best match.

The paper is organized as follows. The section 2 explains about the overview of the speech feature extraction process. The section 3 broadly discusses the feature extraction techniques adopted for this study. The section 4 explains the speech recognition techniques adopted for this work. The experimental results are given in the section 5 and the findings and discussions are deeply presented in the section 6. The conclusion and future works are discussed in the section 7.

II. Overview Of Speech Feature Extraction Techniques

Speech feature extraction is the signal processing front-end which converts the speech waveform into some useful parametric representation. These parameters are then used for further analysis in various speech related applications such as speech recognition, speaker recognition, speech synthesis and speech coding. It plays an important role to separate speech patterns from one another. Because every speech and speaker has different individual characteristics embedded in their speech utterances [3]. But extracted feature should meet some criteria while dealing with the speech signal such as:

- Easy to measure extracted speech features
- Not be susceptible to mimicry
- Perfect in showing environment variation
- Stability over time

In general, the speech signals are slow varying time signals that are also called as quasi-stationary. Because of this variability in a speech signal, it is better to perform feature extraction in short term interval that would reduce these variability. Hence, these signals are examined over a short period of time (10-30 msec), where the characteristics of speech signal becomes stationary. In general, a speech signal contains some acoustic information which can be represented by short term amplitude spectrum. The motivation behind this computation is the cochlea of the human ear performs a quasi-frequency analysis. The resultant analysis in the cochlea on a nonlinear frequency scale becomes the bark scale or the mel scale [2]. The feature vectors can be extracted from these analyses. The following section explains the most popular feature extraction techniques and the adopted methods for this work.

III. Speech Feature Extraction Techniques

The widely used feature extraction techniques are listed below:

- Linear Predictive Coding (LPC)
- Linear Predictive Cepstral Coefficients (LPCC)
- Perceptual Linear Predictive (PLP) Coefficients
- Power spectral analysis
- Mel-Frequency Cepstral Coefficients (MFCC)
- Relative spectra filtering of log domain coefficients (RASTA)
- Wavelet features
- Auditory features

The two types of features which are considered for any ASR system are static and dynamic features. These feature vectors are used to classify or recognize the similar patterns of speech utterance. In this research work, various static and dynamic features are extracted for implementations which are explained below.

A. *Mel Frequency Cepstral Coefficients (MFCC)*

The MFCC is the most evident cepstral analysis based feature extraction technique for speech and speaker recognition tasks. It is popularly used because it approximates the human system response more closely than any other system as the frequency bands are positioned logarithmically [3]. Computing MFCC is based on the short-term analysis, and thus from each frame a MFCC feature vector is computed. In order to extract the coefficients, the speech sample is taken as the input and it is divided into number of frames. After that, the hamming window is applied to minimize the discontinuities between the frames where Discrete Fourier Transform (DFT) is used to generate the Mel filter bank.

According to Mel frequency wrapping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included [4].

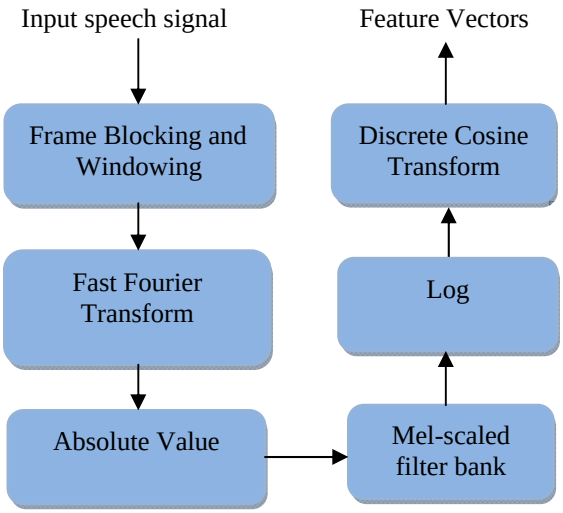

Fig. 2 Steps involved in MFCC Feature Extraction

After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transform (IDFT) is used for the cepstral coefficient calculation. It transforms the log of the quefrench domain coefficients to the frequency domain [5]. MFCC can be computed by using the formula **"Eq. 1".**

$$Mel(f)= 2595*\log10(1+f/700) \quad \text{------(1)}$$

In this work, 13 coefficients are extracted and used for the experiments. The Fig.2 shows the steps involved in the MFCC feature extraction.

B. *Linear Predictive Coding (LPC)*

LPC feature extraction has become the predominant technique because it provides an accurate estimate of the speech parameters. It is also an efficient computational

method for modeling speech. The basic idea behind the LPC analysis is that a speech sample can be approximated as a linear combination of past speech samples. Like MFCC, LPC is a frame based analysis of the speech signal which is performed to provide observation vectors of speech. To compute LPC features, initially the speech signal is blocked into frames of N samples. Each frame is multiplied by an N-sample Hamming window, and this windowed frame is passed to perform short term auto correlation. Then, LP analysis is performed based on Levinson-Durbin recursion algorithm [4]. It provides 2Q-by-T matrix of observation features, where *T* is the number of frames. The LPC coefficients are then converted to Q cepstral coefficients, which are weighted by a raised sine window.

The first half of an observation vector is the weighted cepstral sequence for frame *t*, the second half is the time differenced weighted cepstral coefficients which is used to add dynamic information [6]. In this work, 24 feature vectors are extracted using LPC analysis. The Fig. 3 shows the steps involved in LPC feature extraction.

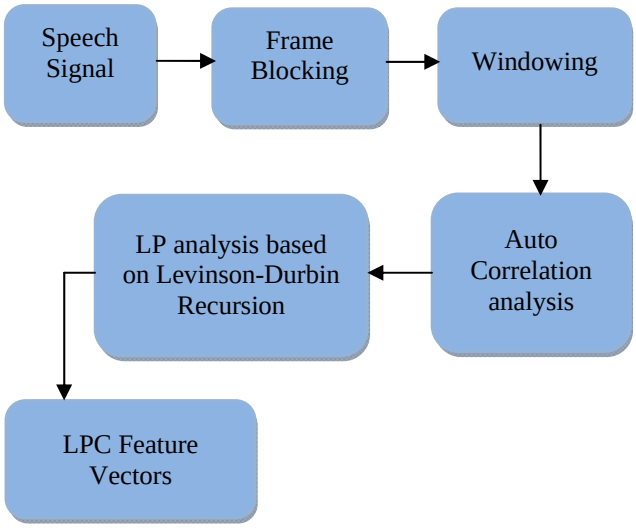

Fig. 3 Steps involved in LPC Feature Extraction

C. *Perceptual Linear Predictive (PLP) Coefficients*

PLP feature extraction is similar to LPC analysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, LP modifies the short-term spectrum of the speech by several psychophysically based transformations [2]. PLP performs spectral analysis on speech vector with frames of N samples with Nband filters. Here, 256 window size and 24 filter bank are used for the experiments. After that the PLP filters are created with preemphasis and bark scale. Then the power spectrum is estimated with power law. This PLP power spectrum along with the nearest frequencies is then passed for LP analysis. Finally, LP analysis is done with FFT and the final observation vectors are extracted by taking the real values of inverse FFT.

D. *Gammatone Filtering and Cochleagram Coefficients (GFCC)*

The auditory features always have a great range of improvement in speech related applications. Among them, the Gammatone Filters (GF) is designed with more mathematical forms of frequency response and is successfully applied in speech signal processing applications. In this paper, a Gammatone filtering and a Cochleagram generation in time domain is used, which is in a form of cascade of four identical filters [7]. Initially it converts the ERB (Equivalent Rectangular Bandwidth) rate scale to normal frequency scale. The upper and lower bound of ERB are estimated for ERB segment where the center frequency arrays are indexed by channel [8]. In this work, the center frequency ranging from 80Hz to 5000Hz are chosen. Then the signal is rearranged using zero-padding based on the frame number. After that, the Gammatone filtering is done in time domain for which the pre-emphasis is used with low-pass filter. It is represented as

H(z) = 1 + 4m*Z^(-1) + m*Z^(-2).

The standard form of Gammatone filter can be represented as

G(z) = (1 - 4m*Z^(-1) + 6m^2*Z^(-2) - 4m^(3)*Z^(-3) + m^(4)*Z^(-4))^(-1).

Here, Basilar membrane response is calculated using frequency shift by taking the real part of a signal at time *t* on channel *m*. Finally, the Cochleagram is generated by using smoothing and average-framing method [7]. It computes the dynamic features using DCT [8]. Dynamic features are generally helpful in capturing temporal information. Totally 58 features are extracted where the first 29 dimension represents the static features and the second 29 dimension represents its dynamic features. Since both spectral and temporal information are extracted it effectively helps in pattern matching process. In this research work, very good results were achieved using GFCC features and the experimental results are given in the following sections.

IV. SPEECH RECOGNITION TECHNIQUES

For more than 50 years, researchers are putting lots of effort to make a machine to understand the fluently spoken speech. Many techniques were proposed and successfully applied for this task. Firstly, the dynamic programming techniques have been proposed for spoken word recognition based on template matching approach. Succeeding researches were done on developing statistical pattern matching approach such as HMM and GMM. These methods have offered great improvement in ASR by using probability distribution density. Based on these probabilities, the models are created with the entire data for each speech patterns. These models will have the complete knowledge and description of the actual problems hence the improved accuracy is possible with these methods [9]. Next to these approaches, the machine learning techniques like Artificial Neural

Networks(ANNs), Support Vector Machines (SVMs), are proposed to replace the conventional HMM/GMM systems [10]. Recent research works focusing on building hybrid techniques to combine the metrics of these techniques for increased performance. In this work, totally five speech recognition techniques are used namely DTW, HMM, MLP, SVM and Decision Trees. The results achieved from these techniques are briefly explained below.

V. EXPERIMENTAL RESULTS

For this work, the analyses are done with speaker independent isolated speech recognition for Tamil language under Matlab environment. The experiments are done with 10 Tamil spoken digits (0-9) and 5 spoken names from 4 different speakers. The utterances consist of 10 repetitions from one male and three females within the age group of 20-35. The total size of the dataset is 15*4*10=600. To make the utterance variation, the speakers uttered the same word at different interval of time. The utterances were recorded at 16 KHz using high quality microphone using audacity software at a silence environment. Furthermore, preprocessing steps were done before extracting features using preemphasis, framing and windowing and silence removal. The same speech samples were used for different experiments with different feature extraction and pattern matching algorithms. In this experiment, the dataset is divided into training and testing data where 60% data is given for training and the remaining 40% data are given for testing. The same speaker's data are used for the experiments. The performances are measured based on the recognition accuracy and the processing time taken for given technique.

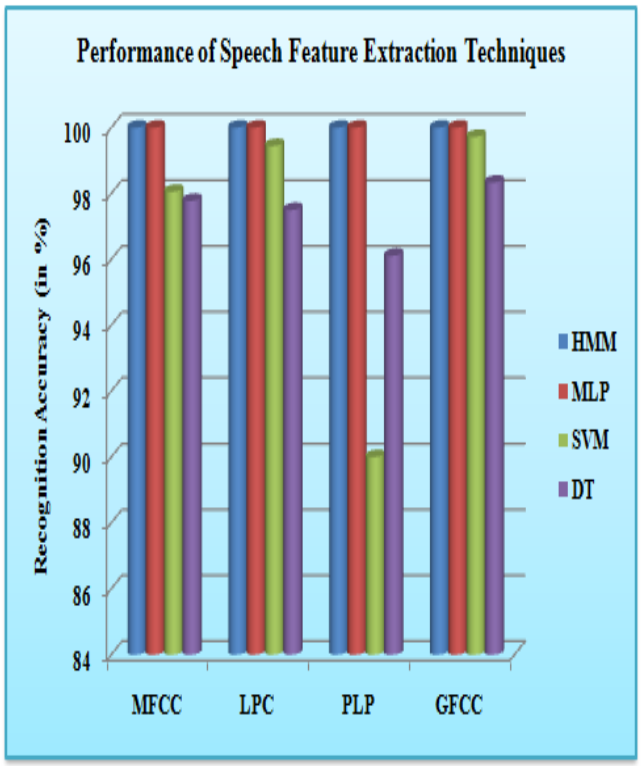

Fig. 4 Training Accuracy

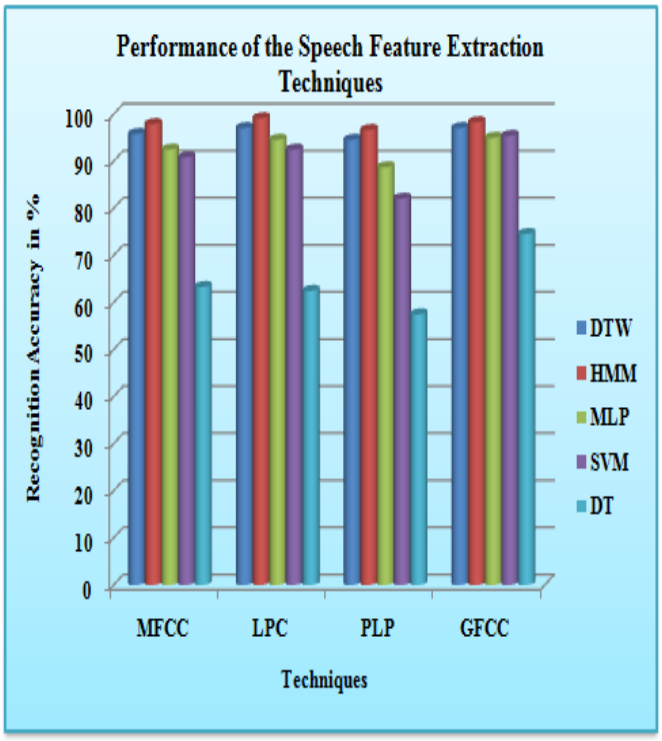

Fig. 5 Testing Accuracy

The Fig.4 and Fig.5 shows the recognition accuracy achieved for the training and testing data respectively. Table I shows the average word recognition accuracy achieved for each technique during training and testing process for all the speakers. The Table II illustrates the average time taken for training and testing the above system for all the speakers. The Table III illustrates the average word level accuracy achieved during training and testing process. The highest word recognition accuracy achieved for the system is highlighted.

Since DTW does not require training its performance are measured only for testing data. From the above two figures it is clearly observed that the GFCC feature extraction technique has offered better results for all the speech recognition techniques involved here

VI. FINDINGS AND DISCUSSIONS

The feature extraction and analysis is an important component of an ASR system. It plays a vital role in speech recognition process as the decision logic is completely depends on the features that are given as an input for modeling the speech. The main goal of this research work is to provide a comparative study of the most popular speech feature extraction techniques namely MFCC, LPC, PLP and GFCC. The performances of these feature extraction techniques are clearly observed by applying with the speech recognition techniques. It is clearly observed from the experiments and analysis that the GFCC feature extraction technique outperformed all the other feature extraction techniques adopted for this work. It offered high recognition accuracy for all the speech techniques and speakers involved here. By examining the recognition results, better accuracy is obtained with HMM and DTW followed by MLP and SVM techniques for both training and test data. The figure 3 and 4 has shown the training and

testing accuracy. According to the results, these three techniques are found to be the good speech recognition techniques for the above developed system. Based on the investigations, it is proved that better results were achieved for all the speakers enrolled in this study.

By considering the processing time factor, it is noticed that the GFCC was found to be time consuming when compared with other techniques. This is a drawback of this technique but when it is applied with machine learning techniques the processing time was reduced. It is also observed from the table 3 illustration that GFCC method has offered better results for word level accuracy. In this work, GFCC features provided high recognition accuracy for more than ten words out of fifteen words enrolled in this work. Next to GFCC, the LPC features were found to be a better feature extraction method for the above developed system. The conclusion and summary are given in the following section.

.

TABLE I
AVERAGE WORD RECOGNITION ACCURACY FOR TRAINING AND TESTING PROCESS

Speech Recognition Techniques	Training Accuracy				Testing Accuracy			
	MFCC	LPC	PLP	GFCC	MFCC	LPC	PLP	GFCC
DTW	-	-	-	-	95.83	**97.08**	94.58	**97.08**
HMM	**100**	**100**	**100**	**100**	97.92	**99.17**	96.67	98.33
MLP	**100**	**100**	**100**	**100**	92.50	94.58	88.75	**95.00**
SVM	98.06	99.44	90.00	**99.72**	90.83	92.50	82.08	**95.42**
DT	97.78	97.50	96.11	**98.33**	63.33	62.50	57.50	**74.58**

TABLE II
AVERAGE TIME TAKEN FOR TRAINING AND TESTING THE SYSTEM

Speech Recognition Techniques	Training Time(in seconds)				Testing Time(in seconds)			
	MFCC	LPC	PLP	GFCC	MFCC	LPC	PLP	GFCC
DTW	-	-	-	-	28.93072	251.066	28.63291	471.0272
HMM	90.72533	22.62792	10.60725	31.41766	61.59758	22.01567	9.924852	30.64007
MLP	2.895	5.01	4.6525	10.8525	2.7075	4.6325	3.885	9.34
SVM	3.035	2.1375	1.3975	1.69	2.3675	2.095	2.9525	3.4925
DT	0.0275	0.025	0.0275	0.0525	0.0275	0.0275	0.0425	0.0725

TABLE III
AVERAGE WORD LEVEL ACCURACY ACHIEVED DURING TRAINING AND TESTING PROCESS

Words (W-words D-digits)	Training Accuracy				Testing Accuracy			
	MFCC	LPC	PLP	GFCC	MFCC	LPC	PLP	GFCC
D0	98.96	**100**	98.96	**100**	**90**	85.00	71.13	86.25
D1	100	**100**	98.96	97.92	88.75	**91.25**	73.88	80
D2	97.92	96.88	93.75	**98.96**	80	**83.75**	60.38	**83.75**
D3	97.92	98.96	96.88	**100**	83.75	80	63.50	**92.50**
D4	**100**	**100**	**100**	**100**	81.25	88.75	70.38	**90**
D5	97.92	98.96	92.71	98.96	**87.50**	83.75	68.75	**87.50**
D6	**100**	**100**	96.88	98.96	90	**97.50**	78.00	92.50
D7	**100**	**100**	96.88	**100**	**100**	90	80.38	93.75
D8	**100**	**100**	95.83	**100**	88.75	92.50	77.13	**96.25**
D9	**98.96**	98.96	94.79	**98.96**	86.25	92.50	70.63	**100**
W1	97.92	98.96	93.75	**100**	80	88.75	76.13	**95.00**
W2	96.88	98.96	96.88	**100**	86.25	86.25	72.00	**95.00**
W3	**98.96**	97.92	97.92	**98.96**	87.50	91.25	79.00	**97.50**
W4	98.96	**100**	97.92	**100**	**95.00**	92.50	81.00	91.25
W5	**100**	**100**	95.83	**100**	96.25	93.75	81.25	**100**

VII. CONCLUSION AND FUTURE WORK

The main objective of this research work is to provide a detailed comparative analysis and implementation of the most popular speech feature extraction techniques for speaker independent Tamil isolated speech recognition system. The most popular speech feature extraction and pattern matching techniques were implemented and analyzed. Totally, four feature extraction algorithms namely MFCC, LPC, PLP and GFCC are implemented and its performances were deeply observed. The potential pattern matching algorithms that are widely used for speech recognition such as DTW, HMM, MLP, SVM and Decision Tree were implemented with the above feature extraction techniques. By investigating these feature vectors along with the recognition techniques it was found that the GFCC features gave better results and outperformed the other algorithms for all the speech recognition techniques. The HMM and DTW followed by MLP and SVM techniques was found to be best speech recognition methods for this research work. Highest word recognition accuracy is achieved with GFCC techniques for boh training and tesing data. Based on the satisfactory results and metrics of this technique the feature combination method will be proposed in future.

REFERENCES

1. MarutiLimkara., RamaRaob., and VidyaSagvekarc, “Isolated Digit Recognition Using MFCC AND DTW”, International Journal on Advanced Electrical and Electronics Engineering (IJAEEE), ISSN (Print): 2278-8948, Volume-1, Issue-1, 2012.
2. Urmila Shrawankar, “Techniques for Feature Extraction in Speech Recognition System: A Comparative Study”, SGB Amravati University.
3. Vimala.C., Radha.V, “A Review on Speech Rec ognition Challenges and Approaches”, World of Computer Science and Information Technology Journal (WCSIT), ISSN: 2221-0741, Vol. 2, No. 1, pp. 1-7,2012.
4. Shivanker Dev Dhingra., Geeta Nijhawan., and Poonam Pandit, “Isolated Speech Recognition using MFCC and DTW”, International Journal of Advanced Research in Electrical Electronics and Instrumentation, 2013.
5. 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), 2013.
6. J.R Deller., J.G. Proakis and F.H.L. Hansen. 2000. *Discrete-Time Processing of Speech Signals.* IEEE Press, chapter 12.
7. L. Bezrukov., H. Wagner., and H. Ney, “Gamma tone features and feature combination for large vocabulary speech recognition”, ICASSP 2007, vol. 4, pp 649–654.Engineering. (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 8, August 2013.
8. Jun Qi., Dong Wang., Yi Jiang., and Runsheng Liu, “Auditory Features Based On Gammatone Filters for Robust Speech Recognition”, IEEE International Symposium on Circuits and Systems (ISCAS) 2013, Page(s):305 – 308. ISSN: 0271-4302, Print ISBN:978-1-4673-5760-9,DOI:10.1109/ISCAS.2013.6571843.
9. Ibrahim M., M. El-emary., Mohamed Fezari and Hamza Attoui, “Hidden Markov model/Gaussian mixture models (HMM/GMM) based voice command system: A way to improve the control of remotely operated robot arm TR45”, Scientific Research and Essays, Academic Journals. Vol. 6(2), pp. 341-350, 2007.
10. DO VAN HAI, “Hybrid architectures for speech Recognition”, Conformation Report Submitted to the School of Computer Engineering. Nanyang Technological University, 2011