RESEARCH ARTICLE OPEN ACCESS

# A Comparative Study on Feature Extraction Technique for Isolated Word Speech Recognition

Easwari.N[1], Ponmuthuramalingam.P[2]
[1,2](PG & Research Department of Computer Science, Government Arts College, and Coimbatore)

**Abstract:**

One of the common and easier techniques of feature extraction is Mel Frequency Cestrum Coefficient (MFCC) which allows the signals to extract the feature vector. It is used by Dynamic Feature Extraction and provide high performance rate when compared to previous technique like LPC. But one of the major drawbacks in this technique is robustness. Another feature extraction technique is Relative Spectral (RASTA). In effect the RASTA filter band passes each feature coefficient and in both the log spectral and the Spectral domains appear linear channel distortions as an additive constant. The high-pass portions of the equivalent band pass filter effect the convolution noise introduced in the channel. The low-pass filtering helps in smoothing frame to frame spectral changes. Compared to MFCC feature extraction technique, RASTA filtering reduces the impact of the noise in signals and provides high robustness.

*Keywords* — **Automatic Speech Recognition, MFCC, RASTA, Isolated Word, Speech Chain, Signal Noise Ratio.**

## I. INTRODUCTION

Digital Speech Signal Processing is the process of converting one type of speech signal representation to another type of representation so as to uncover various mathematical or practical properties of the speech signal and do appropriate processing to support in solving both fundamental and deep troubles of interest. Digital Speech Processing chain has two different main models. They are Speech Production Model/Generation Model which deals with acoustic waveform and Speech Perception Model/Recognition Model deals with spectral representation for recognition process. Digital Speech Processing used to achieve reliability, flexibility, accuracy, real-time implementations on low-cost digital speech processing chip, facility to integrate with multimedia and data, encryptability/security of the data and the data representations via suitable techniques. The overall process of production and recognition of speech is to convert the speech signal from the device or human, and to understand the message is speech chain. In other word, the process of converting the speech signals into acoustic waveform is speech processing. Speech Production is the process of converting the text from the messenger to acoustic waveform. Speech Perception is the process of analyzing the acoustic waveform into the understandable message.

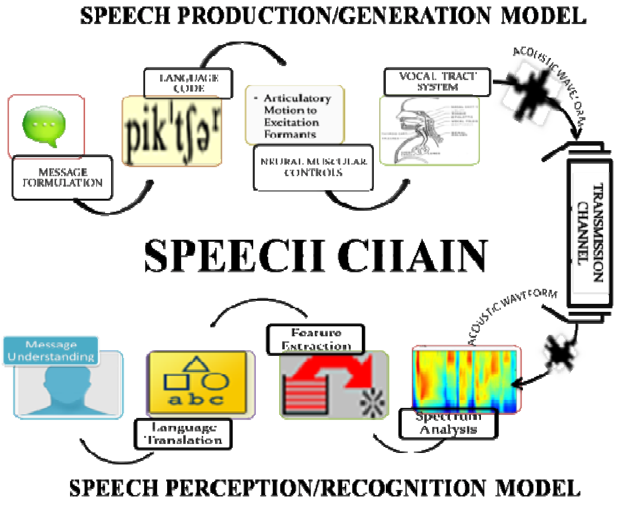


Fig. 1 Cyclic Representation of Speech Chain

The speech chain comprises of speech production, auditory feedback to the speaker, speech transmission and speech perception and understanding by the listener. The message from a speaker to a listener in speech chain is represented in different levels. The Speech Chain which produce continuous and discrete output in the Information rate of 30 kbps to 50 bps from the discrete input and continuous input of the signals. In practical applications, using of real world speech will leads to an issue of noise and distortions [1].

Automatic speech recognition is a method by which a computer maps an acoustic speech signals into text. Automatic speech recognition processed under two phases. They are Training phase, where the speech signals are recorded, represented as parameters and stored in database and the next one is Recognition phase where the features from speech signals are extracted and referred to the templates which are already existed. Speech recognition is also known as automatic speech recognition or computer speech recognition which means understanding voice of the computer and performing any required task or the ability to match a voice against a provided or acquired vocabulary. speech recognition system consists of a microphone, for the person to speak into; speech recognition software; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation [1].An efficient Speech Recognition system has the major considerations for developing higher recognition accuracy, achieving low word error rate and addressing the issues of variability in the source. Speech Recognition Methodologies have four different stages. Each stage of methodologies deals with various analyses of speech signals, extracting algorithms, identification of signal and related word matching.

## II. FEATURE EXTRACTION

Feature Extraction is the process of removing unwanted and redundant information and find out the set of properties called as parameter of utterances by processing of the signal waveform of the utterances. These parameters are the features. After preprocessing the feature extraction is performed. It produces a meaningful representation of speech signal. The most important part of the speech recognition system which distinguishes one speech from another. First of all, recording of various speech samples of each word of the vocabulary is done by different speakers. After the speech samples are collected; they are converted from analog to digital form by sampling at a frequency of 16 kHz. Sampling means recording the speech signals at a regular interval. The collected data is now quantized if required to eliminate noise in speech samples. The collected speech samples are then passed through the feature extraction, feature training & feature testing stages. Feature extraction transforms the incoming sound into an internal representation such that it is possible to reconstruct the original signal from it. There are various techniques to extract features like MFCC (Mel Frequency Cepstral Coefficient), PLP (Perceptual Linear Prediction), RASTA(RelAtive SpecTrAl Filtering), LPC (Linear Predictive Coding), PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), ICA (Independent Component Analysis), Wavelet etc but mostly used is MFCC [2].

### *A. Process Of Feature Extraction*

The steps for feature extraction are based on the speech signals. After filtering the cepstral coefficient value will be the features. The most common feature extraction technique was MFCC which provides more accurate than the others.

- Speech signals are sampled and quantized.
- Pre-emphasis - The speech samples are sent through a high-pass filter to amplify the frequencies above 1 KHz in the spectrum because hearing is more perceptive in this region.
- Frame blocking: The speech samples are blocked into frames of N samples (amounting to a time period of 10-30 ms) with an overlap of some samples between frames.
- Fast Fourier Transform (FFT): FFT is performed on each of the frames to obtain the magnitude values of the frequency response.

- Triangular Band-pass Filtering: The magnitude frequency response is multiplied by a set of triangular band-pass filters to get the log energy value of each filter. The positions of these filters are evenly spaced along the Mel frequency scale.
- Discrete cosine transform or DCT: The DCT is applied on the logarithm of the energy obtained from the triangular band pass filters [3].

In speech recognition, feature extraction requires much attention only because of main recognize process depends heavily on this phase. Among the different techniques of feature extraction the common technique MFCC and the other one RASTA are discussed briefly.

### *B. Mel Frequency Cepstral Coefficents*

Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) and were the main feature type for automatic speech recognition (ASR).The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance.

The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which extract can not perceive frequencies over 1 Khz. In other words, in MFCC is based on known variation of the human ears critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. MFCC consists of seven computational steps. Each step has its function and mathematical approaches as discussed briefly in the following:

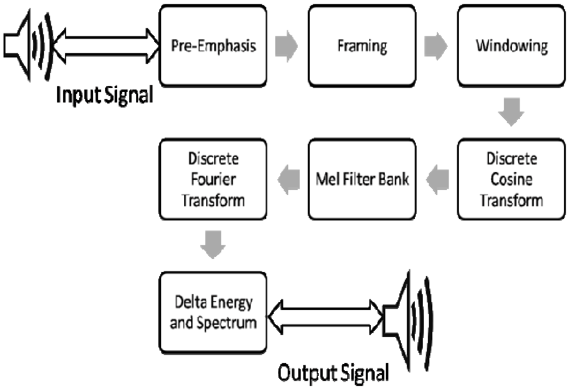

Input Signal
Pre-Emphasis
Framing
Windowing
Discrete Cosine Transform
Mel Filter Bank
Discrete Fourier Transform
Delta Energy and Spectrum
Output Signal

Fig. 2 Block Diagram of MFCC

**Step 1:** Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$\mathbf{Y\,[n] = X\,[n] - 0.95\,X\,[n-1]} \qquad \mathbf{(1)}$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

**Step 2:** Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

**Step 3:** Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

If the window is defined as W (n), $0 \leq n \leq N-1$ where

N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W(n) = Hamming window, then the result of windowing signal is shown below:

$$\mathbf{Y\ (n) = X\ (n) * W\ (n)} \qquad (2)$$

$$\mathbf{W\ (n)\ -\ 0.54\ -\ 0.46\ \cos\left[\frac{2\pi n}{N-1}\right]\ 0\ \leq n\ \leq N-1} \quad (3)$$

**Step 4:** Fast Fourier Transform

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$\boldsymbol{Y\ (w) = FFT\ [h\ (t) * X\ (t)] = H\ (w) * X\ (w)} \quad (4)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

**Step 5:** Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Mel scale filter bank, from a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

$$\mathbf{F\ (Mel) = [2595 * \log 10\ [1 + f\ ]\ 700\ ]} \quad (5)$$

**Step 6:** Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

**Step 7:** Delta Energy and Delta Spectrum

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time . 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented at the equation below:

$$\mathbf{Energy = \sum X^{2}\ [t]} \qquad (6)$$

Each of the 13 delta features represents the change between frames in the equation 8 corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

$$\mathbf{d(t) = \frac{c(t+1) - c(t-1)}{2}} \qquad (7)$$

The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficients is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vectors. The speech waveform is cropped to remove silence or acoustical interference that may be present in the beginning or end of the sound file. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The FFT block converts each frame from the time domain to the frequency domain. In the Mel-frequency wrapping block, the signal is plotted against the Mel spectrum to mimic human hearing. In the final step, the Cepstrum, the Mel - spectrum scale is converted back to standard frequency scale. This spectrum provides a good representation of the spectral properties of the signal which is a key for representing and recognizing characteristics of the speaker [4].

### *C. Features of MFCC*

Commonly allow remote person authentication. Reduces the frequency information of the speech signals into a small number of coefficients. It is easy and relatively fast to compute. It reduces the influence of low-energy components. Most efficient and better anti-noise ability than other vocal tract parameters, such as LPC. The reason for MFCC being most frequently used for extracting features is that it is most nearest to the actual human auditory speech perception. MFCC is worn to recognize information automatically verbal into a telephone, airline reservation, voice recognition system for security purpose etc.

*D. Limitations of MFCC*

MFCC is Noise Sensitive where it is common to normalize their values in speech recognition system to reduce the influence of noise. And also need some concentration to reduce the influence of low energy components. it is conceivably more invariant to background noise and could capture characteristics in the signal where MFCCs tend to fail. The feature space of a MFCC obtained using DCT is not directly dependent on speech data, the observed signal with noise does not show good performance without utilizing noise sup pression methods. The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by the number of filters, the shape of filters, the way that filters are spaced and the way that the power spectrum is warped. MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to reduce the influence of noise [9].

*E. Relative Spectral Processing (RASTA)*

In speech recognition the process of decoding the linguistic message in speech. The speech signal reflects the movement of vocal tract. The rate of change of nonlinguistic components in speech often lies outside the typical rate of change of the vocal tract shape. The **RelAtive SpecTrAl (RASTA)** suppress the spectral components that change more slowly or quickly than the typical range of change of speech. RASTA technique implements to improve the performance of recognition among the convolution and additive noise.

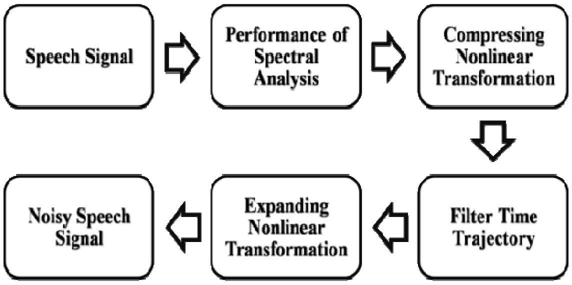

Speech Signal
Performance of Spectral Analysis
Compressing Nonlinear Transformation
Filter Time Trajectory
Expanding Nonlinear Transformation
Noisy Speech Signal

Fig. 3 Block Diagram of RASTA Filtering

To compensate for linear channel distortions the analysis library provides the power to perform rasta filtering. The rasta filter is used either within the log spectral or cepstral domains. In result the rasta filter band passes every feature coefficient. Linear channel distortions seem as an additive constant in each the log spectral and therefore the cepstral domains. The high-pass portion of the equivalent bandpass filter alleviates the result of convolution noise introduced in the channel. The low-pass filtering helps in smoothing frame to border spectral changes [5].

*F. Advantage of RASTA*

It is a band pass filtering technique. Designed to reduce impact of noise as well as enhance speech. It is a technique which is generally used for the speech signals that have background noise or simply noisy speech. Removes the slow varying environmental variations as well as the fast variations in artifacts. This technique does not depend on the choice of microphone or the position of the microphone to the mouth, and so it is robust. Capture frequencies with low modulations that correspond to speech. RASTA gives a better performance ratio.

*G. Limitations of RASTA*

This technique causes a minor deprivation in performance for the clean information but it also slashes the error in half for the filtered case.

## III. EVALUATION OF FEATURE EXTRACTION TECHNIQUE

Isolated word recognition, require a quiet gap between each utterance on both side of sample windows. It accepts single words or single utterances at a time .This is having “Listen and Non Listen state”. Isolated utterance might be better name of this class. The pre-recorded dataset are used for extracting the feature from speech signal.

### *H. MFCC for Isolated Word*

The feature extraction is usually a non-invertible (lossy) transformation. Making an analogy with filter banks, such transformation does not lead to perfect reconstruction, i.e., given only the features it is not possible to reconstruct the original speech used to generate those features. Computational complexity and robustness are two primary reasons to allow loosing information. Increasing the accuracy of the parametric representation by increasing the number of parameters leads to an increase of complexity and eventually does not lead to a better result due to robustness issues. The greater the number of parameters in a model, the greater should be the training sequence.

Speech is usually segmented in frames of 20 to 30 ms, and the window analysis is shifted by 10 ms. Each frame is converted to 12 MFCCs plus a normalized energy parameter. The first and second derivatives (D’s and DD’s) of MFCCs and energy are estimated, resulting in 39 numbers representing each frame. Assuming a sample rate of 8 kHz, for each 10 ms the feature extraction module delivers 39 numbers to the modeling stage. This operation with overlap among frames is equivalent to taking 80 speech samples without overlap and representing them by 39 numbers. In fact, assuming each speech sample is represented by one byte and each feature is represented by four bytes (float number), one can see that the parametric representation increases the number of bytes to represent 80 bytes of speech (to 136 bytes). If a sample rate of 16 kHz is assumed, the 39 parameters would represent 160 samples. For higher sample rates, it is intuitive that 39 parameters do not allow reconstructing the speech samples back. Anyway, one should notice that the goal here is not speech compression but using features suitable for speech recognition [6].

The construction of technique is implemented by loading the signal in the MFCC code and after filtering the decoded signals are allowed to calculate the feature. The calculation provides the range of mean based on the mean calculation, peak by the calculation, MFCC filter by MFCC coding , pitch and finally their vector that are quantatized using the delta energy.

### *I. RASTA for Isolated Word*

The same pre-recorded speech dataset is implemented in spectral filtering feature extraction technique. To obtain better noise suppression for communication systems the fixed RASTA filters were replaced by a bank of non causal FIR Wiener-like filters. The output of each filter is given as

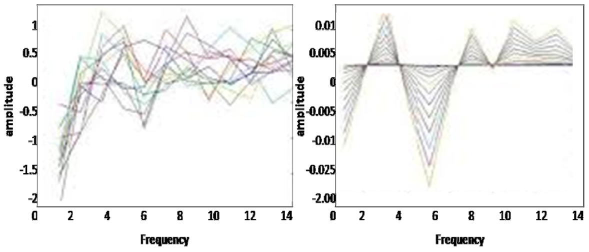
$$\mathbf{S_i\,(k)} \;= \sum_j^M w_i\,(j)\,\mathbf{Y_i\,(k\text{-}j)}$$

Here, $S(k)$ is estimate of clean speech in frequency bin “i” and frame-index “k”, $Y_i(k)$ is noisy speech spectrum, $j$ are the weights of the filter and $M$ is order of the filter. In this method the weights $w(j)$ are obtained .Such that $S_i(k)$ is least square estimate of clean speech $S_i(k)$ or each frequency bin $i$. The order $M$ = 10 corresponds to 21tap non-causal filters. The filters were designed based on optimization on 2 minutes of speech of a male speaker recorded at 8 kHz sampling over public analog cellular line from a relatively quiet library. The published response of the filter corresponding to bins in the frequency range 300Hz to 2300 Hz is a band-pass filter, emphasizing modulation frequency around 6-8 Hz. Filters corresponding to the 150-250 Hz and 2700-4000 Hz regions are low gain, low-pass filters with cut off frequency of 6 Hz. For very low frequency bins (0-100 Hz) the filters have flat frequency response with 0 dB gain [8].

When speech signal are implemented for analysis of spectral form those analyzed signals are compressed as static nonlinearities. The banks of compressed signals are next allowed for the linear bandpass filters. After that the band pass are expanded as static nonlinear spectrals and optional processing like decompression are implemented if needed. The RASTA method differs from the other feature calculation by using a filter with a broader pass-band. The RASTA processing does filtering between two static nonlinearities that are not necessarily the inverse of one another. The features can be viewed as a spectral case of temporal RASTA processing trajectories of cepstral coefficient.

## IV. RESULTS AND DISCUSSIONS

The speech signals are derived for isolated words by recording the words. An isolated utterance HMM recognizes was used for the experiment. One same speech signal are implemented for both the extraction technique and compared each other. As a first step the isolated word speech signal are filtered by MFCC coefficients and the results are noted. As the second step, same speech signals are filtered using RASTA filters and results are noted. The performances are analyzed and implemented by using an efficient tool MATLAB. The performance metrics namely signal noise Ratio, accuracy and time factors are considered. By using the technique RASTA, efficient results were produced. The left side of the fig.4 shows the result of Speech Signals which are extracted from MFCC. The Right side denotes the result of extracted, filtered, noiseless speech signals from RASTA Filtering.



MFCC Coefficients

After Applying RASTA

Fig.4 Filtered Signal of MFCC and RASTA

Compared with the MFCC filters it cut off the frequency of all the shortest cepstral mean subtraction. The main mean difference between the MFCC and RASTA processing is log spectral domain that merely removes the component of short term log spectrum and enhances the spectral transitions. A speech signal which is filtered by both extraction filters are shown in the fig.4. applying MFCC the signals are sampled along with the noises are compared with RASTA filters . In RASTA Technique, the feature vector is extracted better than the MFCC technique. The following figure shows the clean data which is compared with MFCC and RASTA. The two speech signals are implemented in both MFCC and RASTA technique. When first speech signal are filtered their results are plotted in a graph on the basis of signal noise ratio and their error rate in the form of percentage are declared. After that the second speech signal are filtered using spectral filter and the error rate are also plotted in the graph. In this graph o denote signal without peak and + denote signal with peak.

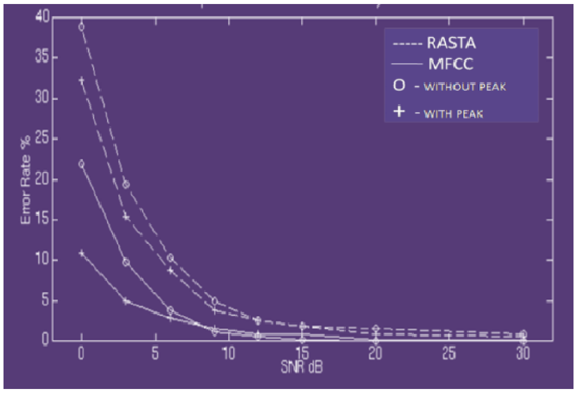


Fig.5 Comparative Results of MFCC and RASTA

## V. CONCLUSIONS

In automatic speech recognition system, Feature extraction includes the process of converting speech signals to the digital form and measures important characteristics of signal i.e. energy or frequency and augment these measurement with meaningful derived uttered speech signal for utilizing in recognition. Short sections of the speech signal are isolated and given for processing. Speech processing is the ability of converting the machine language into speech signals. For extracting speech signal, techniques like Linear Predictive Code, Perceptual Linear Prediction, Mel Frequency Cepstrum Coefficient, Wavelet and RelAtive SpecTrAl are used. Mel-scale Frequency Cepstrum Coefficients (MFCC) is the most frequently used for speech recognition.

This is because MFCCs considers observation sensitivity of human ear at different frequencies, and hence, is appropriate for speech recognition. RelAtive SpecTrAl filtering (RASTA) is an improved feature extraction technique which is used to enhance the speech when recorded in a noisy environment. The time trajectories of the representations of the speech signals are band pass filtered in RASTA. Initially, it was just used to reduce the impact of noise in speech signal but know it is also used to directly enhance the signal. From the Comparative study the extraction of speech signals works more effective in the two techniques MFCC and RASTA than other extracting techniques. As a result the speech signal was extracted on the basis of their Signal Noise Ratio, frequency and Error Rate. The Results shows that the feature extraction technique provides best outcome in RASTA filtering technique than MFCC. This research work can be further extended in the following directions: The two techniques are combined to achieve high level accuracy and robustness which is the fundamental problem of background noise in the speech signals. It is also extended in the direction to develop the complete accurate applications and will focus on the hearing impaired people.

## REFERENCES

1. Lawrence R. Rabiner And Ronald W, “Introduction To Digital Speech Processing”, Foundations And Trends In Signal Processing Vol. 1, Nos. 1–2 (2007) 1–194.
2. Kishori R.Ghule, R. Deshmukh,“Feature Extraction Techniques For Speech Recognition: A Review”, International Journal Of Scientific & Engineering Research, Vol. 6, Issue 5, May(2015).
3. Vibha Tiwari, “Mfcc And Its Applications In Speaker Recognition”, International Journal On Emerging Technologies 1(1): 19-22(2010).
4. Lindasalwa Muda, Mumtaj Begam And I. Elamvazuthi, “Voice Recognition Algorithms Using Mel Frequency Cepstrum Coefficient (MFCC) And Dynamic Time Warping (DTW) Techniques” Journal Of Computing, Vol. 2, Issue 3, March (2010).
5. Hynek Hermansky, “Rasta Processing Of Speech”, Ieee Transactions On Speech And Audio Processing, Vol.2, No.4 Oct(1994).
6. Neha P.Dhole, Ajay A.Gurjar, “Detection Of Speech Under Stress Using Spectral Analysis”,International Journal of Research In Engineering And Technology Issn: 2319-1163 Vol.2 Issue: 04, Apr(2013).
7. BhupinderSingh,Rupinder Kaur,Nidhi Devgun,Ramandeep Kaur, “The Process Of Feature Extraction In Automatic Speech Recognition System For Computer Machine Interaction With Humans: A Review”, International Journal Of Advanced Research In Computer Science and Software Engineering, Vol. 2, Issue 2, Feb(2012).
8. Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare , Pukhraj P. Shrishrimal, “A Comparative Study Of Feature Extraction Techniques For Speech Recognition System”, International Journal Of Innovative Research In Science, Engineering And Technology Issn: 2319-8753 Vol. 3, Issue 12, Dec(2014).
9. Anchalkatyal, Amanpreet Kaur, Jasmeen Gil, “Punjabi Speech Recognition Of Isolated Words Using Compound Eemd & Neural Network” International Journal Of Soft Computing And Engineering, Issn: 2231-2307, Vol.4, Issue-1, March (2014).