International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 46 • 2016

MFCC based Telugu Speech Recognition using SVM Technique

Archek Praveen Kumar[a], Neeraj Kumar[b], Cheruku Sandesh Kumar[c], Ashwani Kumar Yadav[d] and Abhay Sharma[e]

[a,c-e]Assistant Professor, Amity University Rajasthan, Jaipur, India. Email: [a]archekpraveen@gmail.com
[b]Professor, Amity University Rajasthan Jaipur, India

Abstract: Language is the major path for proper communication. Telugu language is one of the most important language used in south India The speech recognition plays a major role in the data communication. This is sure that a speech has repeated symbols or characters. This paper presents recognition for Telugu language using MFCC and SV which deals with speech feature extraction through proper compression. The speech data can be recognized using various techniques but MFCC and SVM are advanced procedures to recognize the data. The speech data can be recognized by preprocessing, feature extraction, feature classification and comparison techniques. Different speech has different slang which varies the features with minute differentiation. MFCC is universal, that means that the probability of feature extraction is high. Asymptotically, the code is optimal.

Keywords: MFCC, Speech recognition, Support Vector machines, endpoint detection.

Nomenclature:

LSP: Line spectrum pairs.
PPF: Pitch prediction filter
FEC: Forward error correction.
SVM: Support Vector Machine.
MLP: Multilayer perceptron.
DWT: Discrete Wavelet Transformation
FFT: Fast Fourier Transformation
LPC: linear prediction coding.
MFCC: Mel frequency cepstral coefficients.

1. INTRODUCTION

Multimedia have many dimensions of signals, one dimension signals are the signals like A/C power supply, speech signal etc., two dimensional signals are x-ray images, sonographs, etc., three dimension signals are air pressures, longitudes etc. this paper works on one dimensional signal which is speech signal[4]. There are so many

techniques for the speech recognition with proper accuracy. Speech is exerted from air pushing through lungs flow through vocal track and comes out of mouth. Vocal track vibrates this produces sound. Speech recognition is differentiated in to two categories, firstly feature extraction and secondly feature classification. Speech is a multicomponent signal where different parameters are considered like time frequency and amplitude. Automatic speech recognition is one of the daring task in the area of multimedia speech differs with various parameters like sex, poignant state, intonation, diction, expression, adenoidal, pitch, sound, rapidity. Speech recognition deals with noisy environment, since the speech signal consists of background noise which impacts on speech recognition accuracy [8].

This paper deals with Telugu speech recognition by feature extraction technique MFCC, meaning mel frequency cepstral coefficients, and feature classification technique SVM means support vector machine. Authors developed MFC which created a huge research in sound engineering. SVM is a non-probabilistic bi linear classifier [1]. The speech society efforts to put forward a portfolio of strong recognition of various types, including pre emphasis techniques like end point detection, frame blocking, windowing and distortion techniques. [5] Many algorithms have been proposed for speech recognition in which various techniques for extraction and classification is preferably used for better recognition rate but because of versatile slang and features its will be a challenging task to get greater recognition rate [2]. The objective is to recognize Telugu speech with greater accuracy. Telugu words consists of different syllables, according to the slang the MFCC is the efficient method to extract the features. If the input size is a word, the extracted results were high. The testing data is a word, so best method is to classify the features by de-noising data first, and then extract the features. The speech data is recognized with MFCC, where the result is classified by SVM.

2. GENERAL BLOCK DIAGRAM

The input speech is recorded and speech is converted to data and that data is sequenced. At a time the whole speech signal cannot be processed so the speech signal is divided into frames. Now each frame is preprocessed by some techniques like endpoint detection, filtering i.e. removal of noise. Features are extracted by MFCC and classified by SVM as shown in Figure 1.

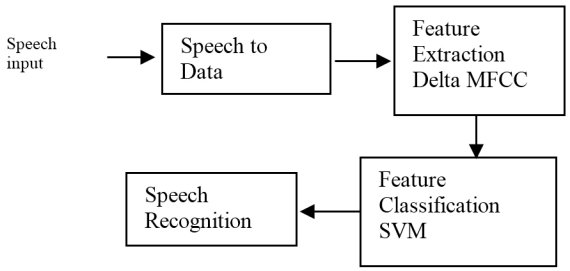


Figure 1: General Block diagram

2.1. Telugu Speech Input

Telugu is a Dravidian language. India has many regional langue's where Telugu language is originated from Sanskrit language. South india have maximum number of Telugu language speaker. Now Andhra Pradesh is divided in to Seemandhra and Telengana where all the speakers speak Telugu. Telangana, Rayalaseema and Andhra are the main 3 regions in AP with different slangs but the script is same. Telugu is a syllabic language evolved from Dravidian script with 60 symbols where 16 are vowels, 3 vowel modifiers and 41 consonants. Telugu speech recognition is the major area for advanced research since from past few years since it has wide applications in the communication world. [6] A data base has been created for Telugu language with 20 male

speakers and 20 female genders with age 20 to 40 years. Totally 40 speakers are considered, each speaker utters 10 words. The speech data is stored with 16 kHz sampling rate. International Phonetic Alphabet (IPA) is used for processing the Telugu language.

2.2. Speech to Data

Telugu speech recorded for data base is in mp3 format and this mp3 format is converted to wave files. The recorded signal for testing is also converted to wave format. The wave signal is sampled with 16 KHz and discrete pieces of signal are collected. This is called data sequence. The data sequence is quantized and converted to binary format for further processing like feature extraction and classification.

2.3. End Point Detection

End point detection is done on the speech data sequence, with zero-crossing rate. This examine the zero-crossing rate to find three occurrences of counts above the threshold level '*t*', The final endpoint is estimated If three occurrences existed and moved backward and/or forward to the location of the first of these three occurrences. The segments with spam of 100ms or less than 100ms are considered as false alarm and eliminated since Telugu words used have greater duration than 100ms. Endpoint detection is executed perfectly on average

2.4. Pre Processing

Preprocessing is the primary part in the speech recognition. Discrete wavelet algorithm [11] is used for de-noising the speech signal. Two methods are used for de-noising called soft thresholding and hard thresholding. This paper used soft thresholding the entire process is shown in Figure 2. Soft thresholding deals with absolute values which are less than the threshold are set to zero and leftover nonzero elements are reduced towards zero [9].

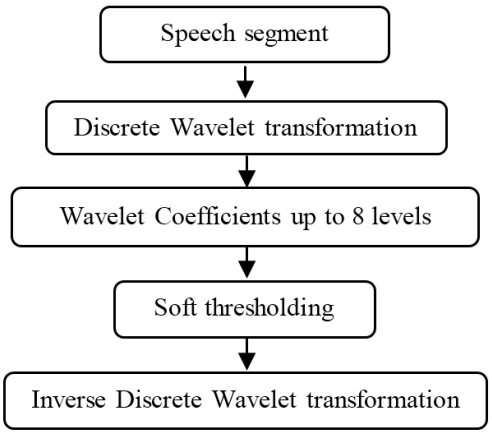


Figure 2: Proposed algorithm for Pre processing

Universal threshold given by white Gaussian noise under mean square error criterion is considered as

$$X_{soft} = \begin{cases} sign(X)\,(|X| - |\tau|) & \text{if } |X| > \tau \\ 0 & \text{if } |X| \le \tau \end{cases} \quad (1)$$

Soft threshold where X represts wavelet coefficents, *t* is threshold value is given by

$$\tau = \sigma(2\,\log(N))^{1/2} \quad (2)$$

Sigma is standard deviation, N is length of signal

2.5. Feature Extraction

Feature extraction is the main part in the speech recognition. Feature extraction is the expertise technique used for speech recognition from past decade. Many techniques are designed for extracting the features like LPCC, PLP, and RASTA-PLP are used, but MFCC Mel frequency Cepstral Coefficients was the efficient technique used by the researchers. Total speech recognition depends on Fourier transformation magnitude spectrum. The total process is shown in Figure 3 [3]

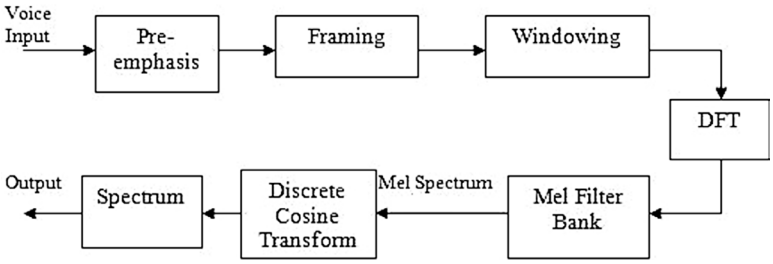


Figure 3: MFCC block diagram

Dynamic characteristics of the MFCC features are considered since; these features have information in the speech signal. The easiest way to get those dynamic features is to take difference of coefficients between successive frames. These resulting coefficients are dynamic or delta parameters reflect Cepstral changes over time. The delta-MFCC coefficients are determined as given in equation, where dk is the delta coefficient determined in terms of the corresponding coefficients $ck + \alpha$ to $ck - \alpha$. M is the configuration parameter. MFCC coefficients are passed through a linear differential filter with impulse and magnitude response which gives the plot as shown in Figure 4.

$$d_k = \frac{\sum_{\alpha=1}^{M} \alpha(c_{k+\alpha} - c_{k-\alpha})}{2\sum_{\alpha=1}^{M} \alpha^2} \tag{3}$$

First 12 MFCC features are selected which are mostly considered as features in speech recognition. Speech signal dynamic features information is considered by choosing the related MFCC coefficients calculated based on flat evaluation of the local time derivative. At last, parameters are re-scaled to the chain of the MFCC to cover gross shaped vector features [10].

2.6. Feature Classification

Classification is the second half of the speech recognition. Recent advancement in computing technology many pattern recognition techniques are evolved. Various classifiers are used like ANN, Naïve Bayes classifier, WPD etc. Basically classification consists of two phases training phase and testing phase. Training phase gives the patterns with similarity measures. Training data consists of known patterns with target points [7]. Classifier gives a model from training data which estimates the target point of testing the unknown patterns. Different applications use different classifiers, SVM, support vector machine, is most useful in speech recognition. SVM is a multi-class classifier which is a nonlinear classifier. SVM is categorized in to two parts where named as one against all and one against one. This paper uses one against one strategy.

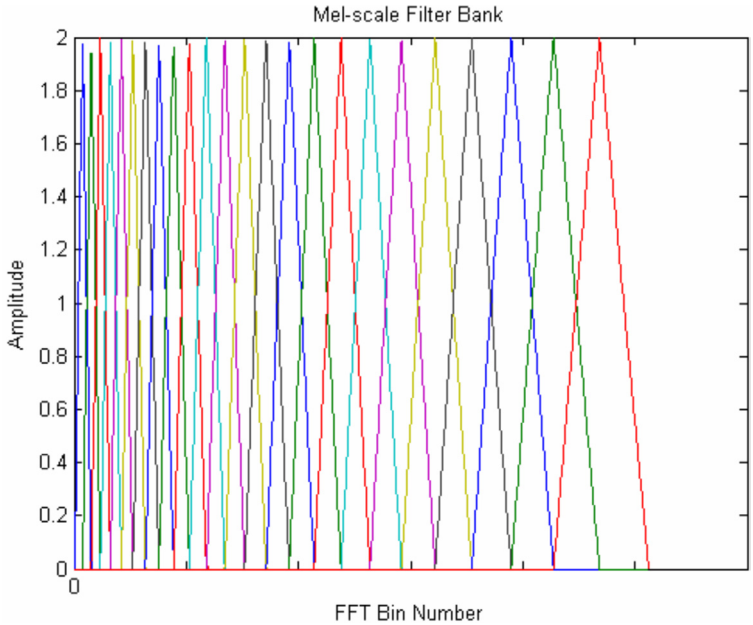


Figure 4: MFCC Plot

3. PROPOSED ALGORITHM

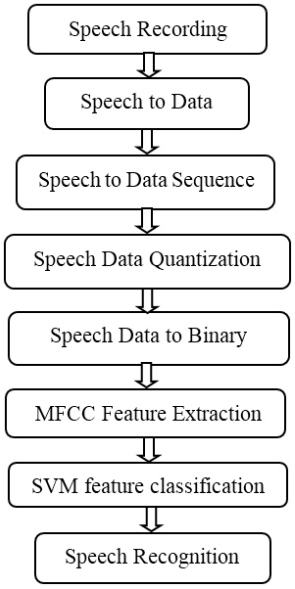


Figure 5: Proposed algorithm

4. RESULTS

Telugu words to be recognized are AMMA, NAANA, TATHA, BABU, EMITI, EVADU, ETLA, ENDUKU, AKKADA, BOMMA. Recording is the first step where Telugu speech signals are recorded randomly shown in figure 4 Audio recorder is the function used for recording the speech signal. Training phase data base is created for 20 male voices and 20 female voices with age factor of 20 to 40 and 10 words are spoken. Results plots are shown in Figure 6 for only word AMMA and NAANA.

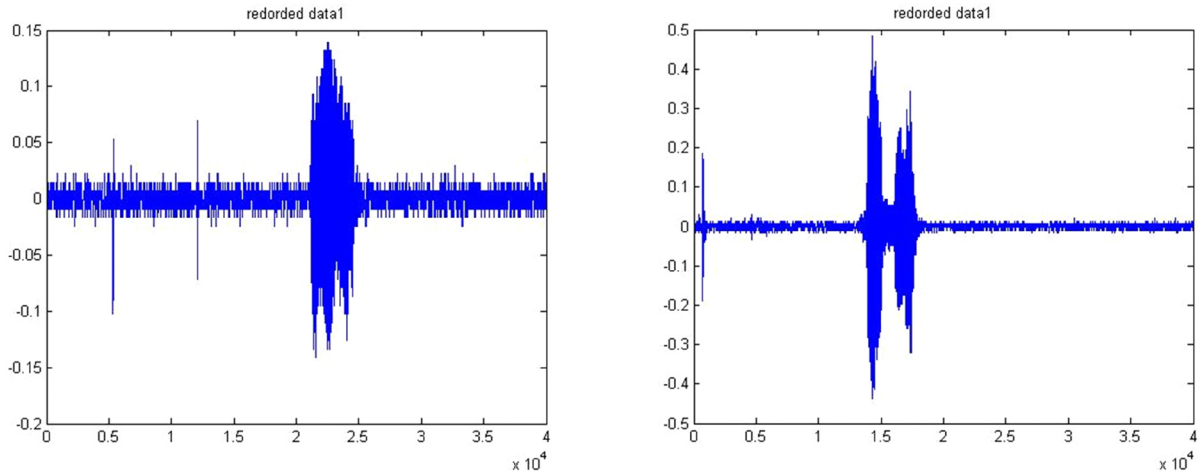


Figure 6: Recorded Speech for AMMA and NAANA

AMMA word is recorded and Sequence of the recorded data has been calculated is shown in figure 5 Absolute value is considered for plotting this sequence Hamming window is used and sampling is done at frequency of 16KHz. Spectrum of Recorded Speech Data is shown in Figure 7. Quantization is done after the sequences are generated. Now the quantized values are converted to binary if applicable is shown in Figure 8.

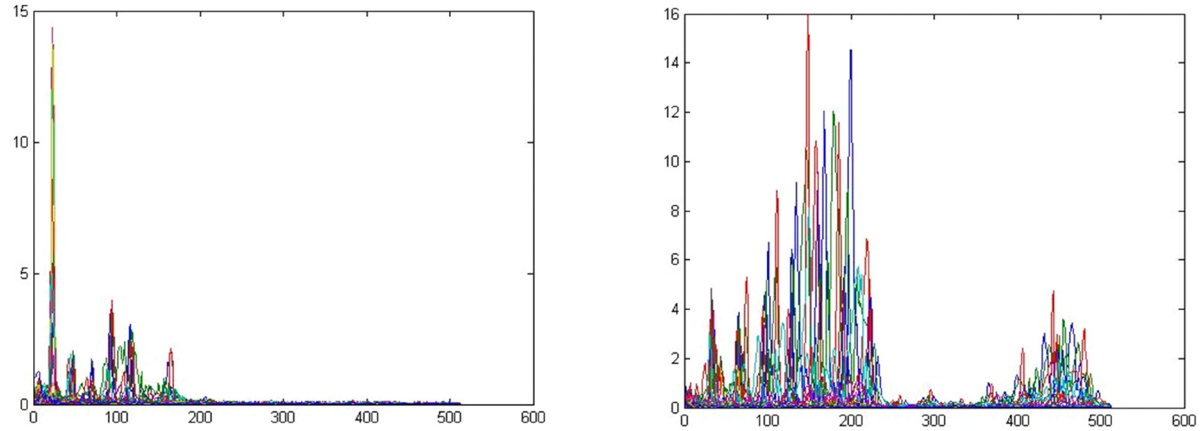


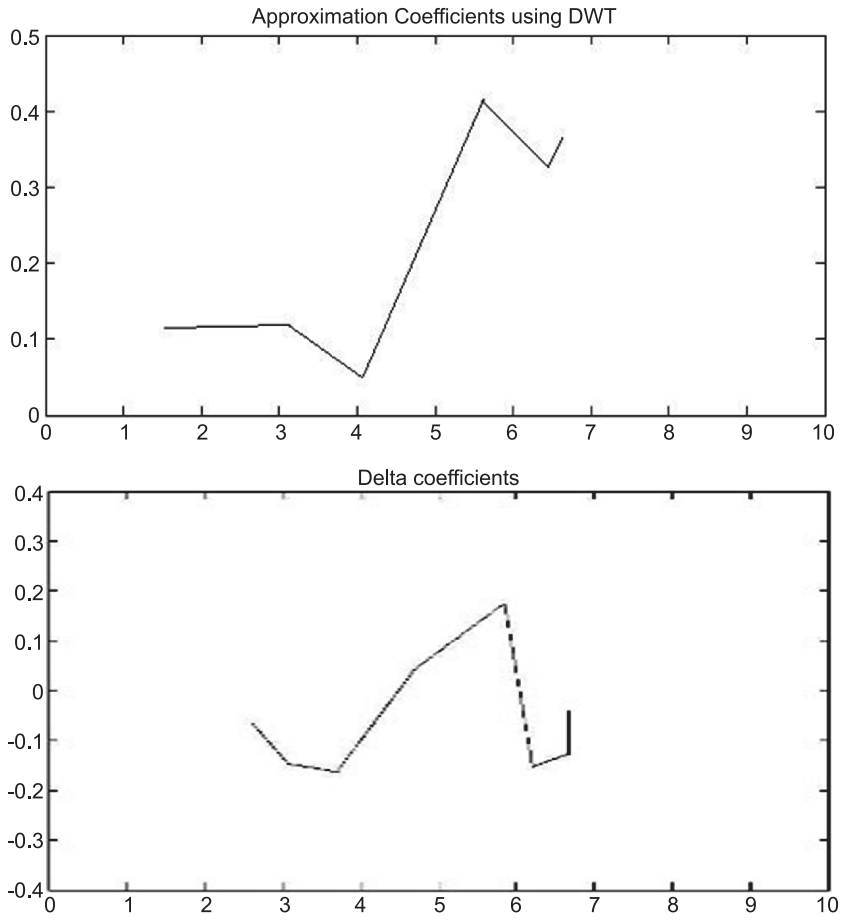
Figure 7: Spectrum of Recorded Speech Data

1	1	1	0	1	1	0	1	1	0	0	0	0	1
1	0	1	0	0	0	1	1	0	0	0	0	1	1
0	0	0	0	0	0	0	1	0	0	0	0	0	1
1	0	0	1	1	0	1	0	0	0	0	0	1	0
1	0	0	1	1	0	0	0	1	0	1	1	0	1
1	0	0	1	0	0	0	0	1	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	1	1	0
1	0	0	1	0	0	1	0	0	0	1	1	0	0
0	0	1	1	0	1	0	0	0	0	1	0	0	1
1	0	1	0	0	0	1	1	0	1	0	0	0	0
0	1	0	0	1	0	1	1	0	0	0	1	0	1
1	0	1	0	0	0	1	0	0	0	0	1	1	0

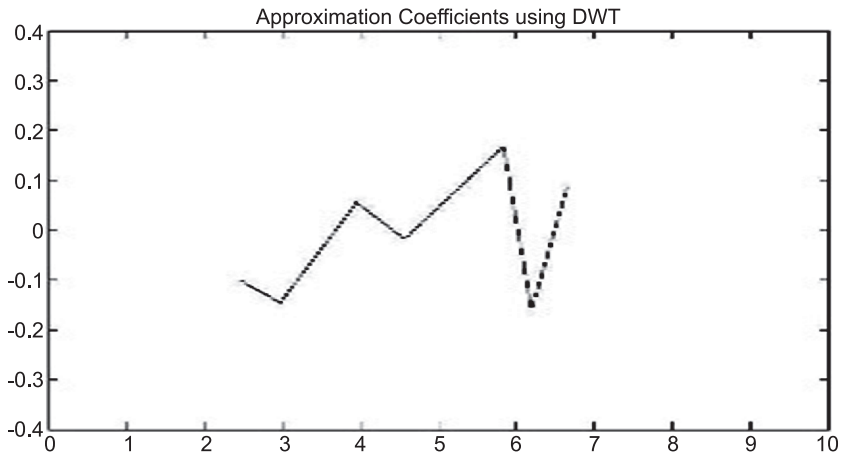
Figure 8: Binary data pattern for AMMA

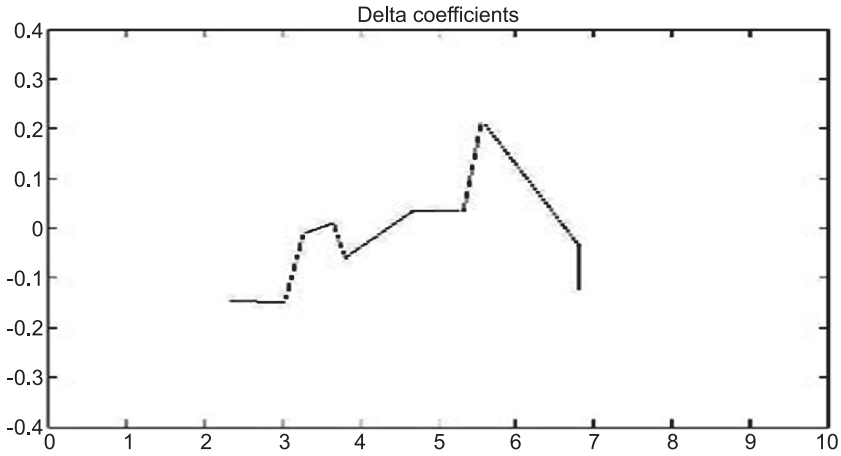
Many Feature extraction techniques were used for Recognition of speech with different frame rate approximation using DWT and Delta coefficients are shown in Figure 9 for the words AMMA and NANNA

Feature extraction using MFCC are given in TABLE I. The proposed technique shows better results. Various parameters or features extracted for 144 bits/ frames are shown in TABLE I and MFCC for 80 bits/ frame shown in TABLE II.



(a) Coefficients by DWT, Delta coefficients for AMMA





(b) Coefficients by DWT, Delta coefficients for NANNA

Figure 10: Coefficents for words AMMA and NAANA

Table 1
Features Extracted

Features	*Bits of data*
Line spectrum pairs (LSP)	31
Pitch prediction filter (PPF)	46
Code base indexes	32
Gain	30
Synchronization	1
FEC	4
Total	144

Table 2
Features extracted

Features	*Bits of data*
Line spectrum pairs (LSP)	19
Pitch prediction filter (PPF)	15
Code base indexes	31
Gain	15
Total	80

The recognition accuracy obtained by using MFCC with SVM is 90.43%.

5. CONCLUSION AND FUTURE SCOPE

Automatic Telugu speech recognition is done with losses less compression. Speech recognition is done by feature extraction with the help of MFCC. Then classification is done on extracted features by SVM. The technique MFCC provides various parameters like LSP, Pitch prediction filter, code base indexes, gain, synchronization, FEC. Two frame sizes are taken one is 144 bits/frames and other is 80 bits/frames. Recognition results are better. Future scope is instead of words, recognition can be done on large vocabulary. More number of words can be recognized. Recognition accuracy can be increased by any other feature classification techniques.

REFERENCES

[1] Singh S, Rajan E G. Vector quantization approach for speaker recognition using MFCC and inverted MFCC. International Journal of Computer Applications 2000; 17(1):1-7.

[2] Dalmiya C P, Dharun V S, Rajesh K P. An efficient method for Tamil speech recognition using MFCC and DTW for mobile applications. In Proceedings of IEEE Conference on Information and Communication Technologies; 2013 April 11-13; Thuckalay, Tamil Nadu, India2013. p.1263-1268.

[3] Hossan M, Memon S, Gregory M. A. Novel approach for MFCC feature extraction. In Proceedings of International Conference on Signal Processing and Communication Systems. 2010 Dec.; p. 1-5.

[4] Babu P R. Digital Signal Processing; Fourth edition; SciTech Publications; 2003.

[5] Kumar A P, Kumar N, Kumar C S, Yadav A K. Speech compression by adaptive Huffman coding using Vitter algorithm. International Journal of Innovative Sciences 2015; 2(5):402-405.

[6] Kalyani N, Sunitha K V N. Syllable analysis to build a dictation system in Telugu language. International journal of computer science and information technology 2009; 6(3): 171-176.

[7] Sunny S, Peter D, Jacob K. Performance of different classifiers in speech recognition. International Journal of Research and Engineering Technology 2013. 2(4):590-597.

[8] Kumar A. P, Bansal D. Digital Arithmetic Coding with AES Algorithm. International Journal of Computer Applications 2013; special Issue 1(2): 15-18.

[9] Donoho D L. Denoising by soft thresholding. IEEE Transactions on Information Theory 1995; 48:927-940.

[10] Kumar A P, Kumar N, Kumar C S, Yadav A K. Speech Recognition Using Arithmetic Coding and MFCC for Telugu Language. In Proceedings of IEEE Conference INDIACOM; 2016 March 16-18; Delhi, India 2016.

[11] Yadav A K, Roy R, Kumar A P, Kumar C S, Dhakad S.K. De-noising of ultrasound image using discrete wavelet transform by symlet wavelet and filters. Int. Conf. on Advances in Computing Comm. and Informatics (ICACCI); 2015 August 10-13; Kochi, Kerela, India 2015. p.1204-1208.