Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Automatic speech recognition using Mel- frequency cepstrum coefficient (MFCC) and vector quantization (VQ) techniques for continuous speech

Amit Verma *, Amit Kumar, Iqbaldeep Kaur

Chandigarh Engineering College, Landran (Mohali), India

ARTICLE INFO

Article history:
Received 6 November 2017
Received in revised form
7 February 2018
Accepted 15 February 2018

Keywords:
Automatic speech recognition
Mel frequency cepstrum coefficient
Vector-quantization

ABSTRACT

Automatic speech recognition is a field related to the interaction between user and machine using effective techniques. ASR is one of the very hot concepts in these days. A lot of researchers worked on different techniques to achieve the best accuracy for speech recognition. In previous research techniques used provides accuracy for a single utterance. Due to which for continuous utterance combination of the technique used in this research work which provides best accurate performance with less noisy interaction. For this research work, Mel Frequency Cepstrum Coefficient (MFCC) and Vector Quantization (VQ) techniques are used. These techniques provide easy speech processing with Mel-frequency scale which includes spacing of linear frequency less than 1000 Hz. Due to which MFCC provides high accuracy, less complexity and high performance with capturing main characteristics of speech. This approach provides efficient and more accurate results than other techniques for a continuous speech by minimizing the distortion created by noise. In this research work algorithms for each technique are represented. This research work presents best possible accuracy for continuous speech signal as compared to other feature extraction techniques.

1. Introduction

Automatic speech recognition (ASR) is the process of translation of independent spoken language into text in real time. It is the process by which a machine identifies spoken words which mean talking to a machine and having it correctly understand what speaker is saying. By understand means, the application to react appropriately or to convert the input speech to another mode of conversation which is further processed by another application that can process it properly and provide the user the required result. Speech recognition is done by various algorithms from various fields like signal processing, pattern recognition, and linguistic. From these feature extraction is called signal processing which is used to convert speech signal to different useful parameters. In this data is extracted from speech signals to build a model for each utterance differently. The methodology for speech recognition system (Plannerer, 2005) is described in Fig. 1.

Automatic Speech recognition consists of two parts. The first part is training part where the whole speaker database is created and another part is testing part where speaker recognition occurs. Different phases of speech recognition are:

- Feature Extraction (Speech analyzer)
- Matching process (Speech Classifier)

1.1. Feature extraction process

Feature Extraction (Vimala and Radha, 2014) is needed because there is large variability in the digital waveform, thus it reduces the variability. Feature extraction extracts the features that help in analyzing the speaker. There are different feature Extraction techniques presents like Linear Predictive Coding (LPC), Perceptual Linear Coding (PLC) and Mel- Frequency Cepstrum Coefficient (MFCC) etc. But In this thesis work using MFCC technique because its complexity is less due to which it give better recognition accuracy than other techniques.

1.2. Mel frequency cepstrum coefficient (MFCC) technique

MFCC (Gaikwad et al., 2010) represents the power spectrum for speech signal on basis of

* Corresponding Author.
Email Address: dramitverma.cu@gmail.com (A. Verma)
https://doi.org/10.21833/ijaas.2018.04.009
2313-626X/

transformation of the speech signal. MFCC generate mimics of the human auditory system. In Mel frequency scale, linear frequency spacing is less than 1000Hz and for a log is higher than 1000Hz. The whole MFCC process is described in Fig. 2.

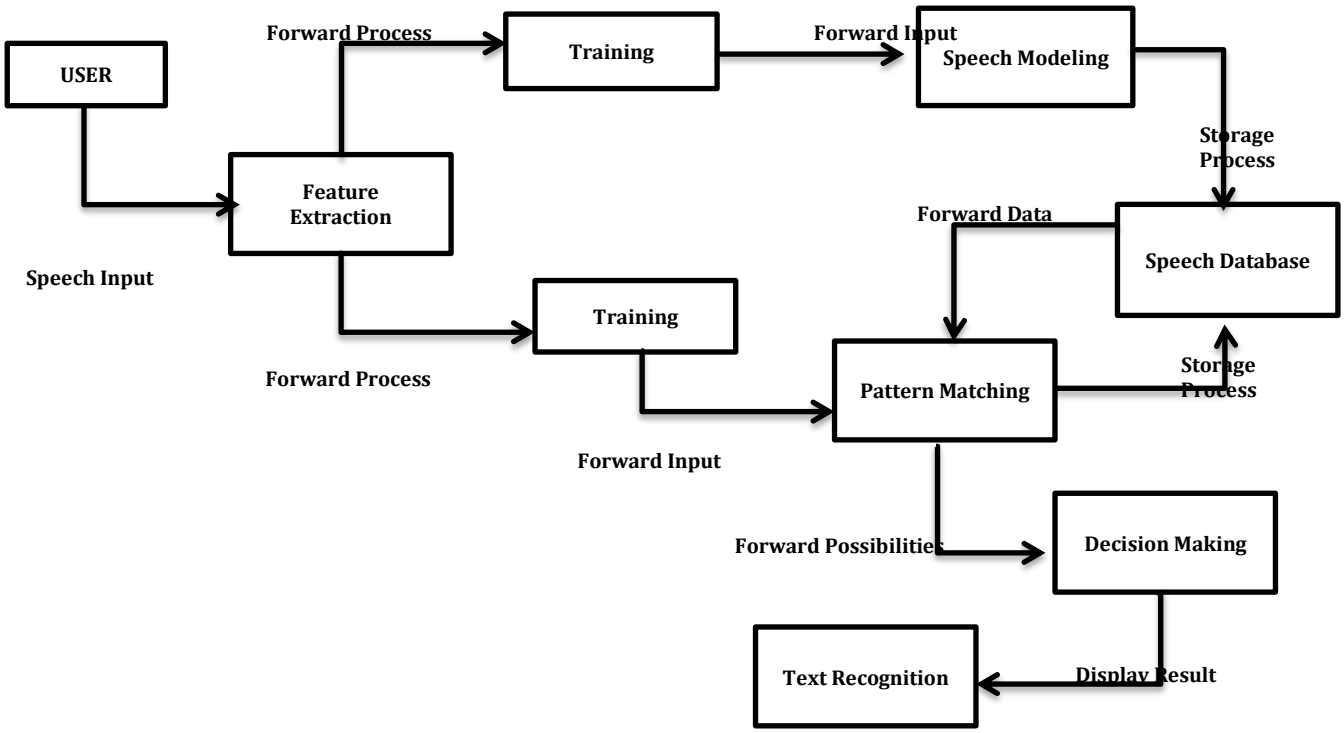


Fig. 1: Methodology of speech recognition system

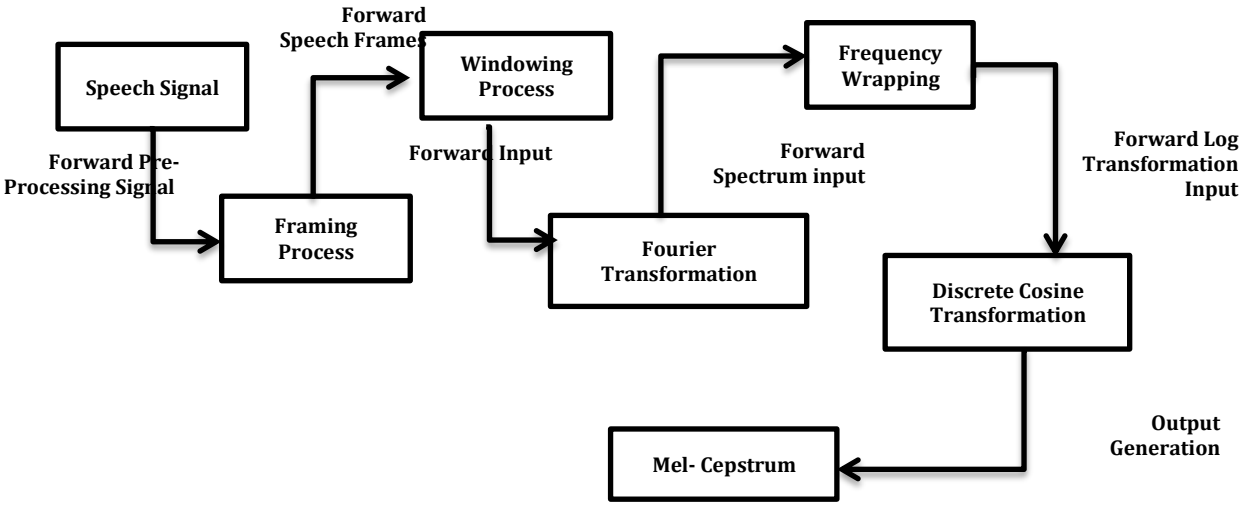


Fig. 2: Working of MFCC feature extraction technique

Fig. 2 provides an overview of feature extraction process (Sledevic et al., 2013). In feature extraction process, continuous speech is entered as input for windowing. Before transformation stage windowing reduces the disruption process. After that, a speech signal which is in a continuous form converted in frames of the window. Then these frames are passed on to Fourier transformation process which transforms frames of the window into a spectrum. After that spectrum is analysed and Mel-spectrum is obtained at Mel- frequency scale with fixed resolution. Then speech signal passes on to log transformation and then to the inverse process of transformation that is inverse Discrete Fourier Transformation. Then the final result of Mel-spectrum is generated (Singh et al., 2012).

1.3. Speech matching process

In speech matching, process speaker identification is matched with the database created in the first phase of speech recognition that is using MFCC (Ittichaichareon et al., 2012). For feature matching, different techniques are present like Dynamic Time wrapping (DTW) and Vector Quantization (VQ) (Hasan et al., 2004). But In thesis work, VQ is used for matching process.

1.4. VQ matching technique

VQ (Singh et al., 2012) is the process of vector mapping from large space to small regions of vector. In these regions are named as the cluster and

represented by its centre which is named as a codeword and collection of code words named as a codebook. Vector quantization (Nijhawan and Soni, 2014) codebook formulation is described in Fig. 3.

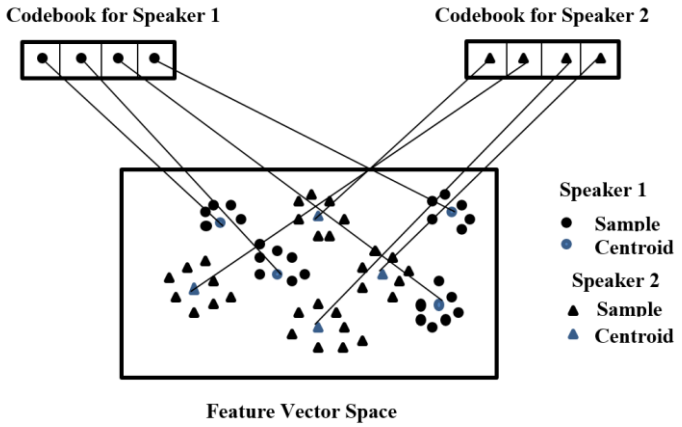


Fig. 3: Vector quantization codebook formulation (Singh et al., 2012)

Fig. 3 shows the VQ (Gray, 1984) codebook formulation for speech recognition. In this case, only two speakers are considered and the sample of each speaker is shown. In training part, VQ codebook is generated specific to each speaker. Centroid is also described which describes the resulting codeword and vector to closest codeword distance is called VQ distortion. In Speech recognition, VQ distortion is checked and speaker from codebook with minimum distortion is identified.

This research paper represents the Mel-Frequency Cepstrum Coefficient (MFCC) for the speech recognition process. Organization of this paper is as follows in section 2 Proposed Method and Implementation has been presented. Result and discussions have been given in section 3. In section 4 and section 5, Conclusion and references have been presented respectively.

2. Proposed method and implementation

In Automatic Speech Recognition, speech inputs have been given with microphone and stored in a database for training purpose. MFCC and VQ algorithms are implemented for speech recognition. Framework for Automatic Speech Recognition for continuous speech is represented in Table 1.

Table 1: Framework of automatic speech recognition for continuous speech (Algorithm 1)

Input: Speech signal (S_I)
Output: feature set extracted and representation of image (I_O)
Begin
For each S_I do
P_i and Apply MFCC and VQ for training
Compute G_M and M_D for MFCC
Extract feature set F_i for Testing
Return F_i
End

Table 1 describes the framework for Automatic Speech Recognition for continuous speech. In this S_I denotes Speech Input Signal, P_i denotes Pre-processing of Speech Signal input, F_i denotes Feature set , G_M denotes Gaussian Modelling and M_D denotes Mahanalobis distance. In training phase samples of audio and images are processed. In this case, speech signal is processed by MFCC and VQ techniques which extract features. In testing phase, duration is given and then processed by MFCC matching speech is presented. Algorithms for MFCC have been represented in Table 2. In this O_S denotes overlap-Size, F_S denotes Hamming frequency, N_F denotes Number of filters, FL_T denotes letters array, S_S denotes Sample Size, F_N denotes File Name, S_R denotes Speech Regions. In Mel frequency Cepstrum Coefficient speech signals are converted into frames of twenty to thirty milliseconds with overlapped size of 0.5 or 0.33. Then windowing is done which help in reducing the disorientation around the corners of speech signals. Then Fourier transformation occur which transforms domain from time to frequency. In Mel frequency wrapping linear transformation of speech signals is formed which is easily processed by machine. Approximation computation formula of the Mel coefficient for a frequency f in Hz is given by Eq. 1.

$$Mel(f) = 2595 * log10\left(1 + \frac{f}{700}\right) \quad (1)$$

Table 2: Working of Mel frequency cepstrum coefficient (MFCC) technique (Algorithm 2)

Input: Speech samples Data set (D_S)
Output: Mel coefficient for speech Samples
Begin
Initialize O_S, F_S, L_T, S_S
Load D_S;
Compute N_F;
For i=1 to size(L_T) do
For j=1: S_S do
F_N=strcat (L_T(i),int2str(j));
Evaluate(char(F_N));
Compute Threshold S_R;
Compute Windowing for S_R;
Calculate FFT(S_R);
Mel 2595*log10(1+ FFT(S_R)./700);
Compute DCT;
Return Mel.
End

Table 3 shows the Vector quantization algorithm for codebook generation. In this C_I denotes Initial

Codebook, C_F Final Codebook, C_E denotes centroid and D denotes distortion.

Table 3: Working VQ for codebook formation (Algorithm 3)

Input: Initial Codebook (C_I)
Output: Final optimized Codebook (C_F)
Begin
Load C_I;
S1: Locate C_E;
Split C_E;
S2: Compute m ←2*m;
Locate C_V and C_E;
Compute D;
Compute F_R← (D'-D)/D;
if (F_R<ϵ)
if (m<M)
goto S1;
else
Return C_F;
else
D'=D;
goto S2;
End

3. Result and discussion

Speech recognition can be evaluated on the parameter of accuracy. To check the credibility of the presented method, the present work used total 10 speaker speech samples for isolated words for same word and alphabets and continuous speech. Accuracy table for MFCC speech recognition techniques is shown in Table 4.

In Fig. 4, frequency of different samples of speech signal is represented. In this time vs Amplitude graph is plotted for different voice signal from different speakers for isolated word "Zero".

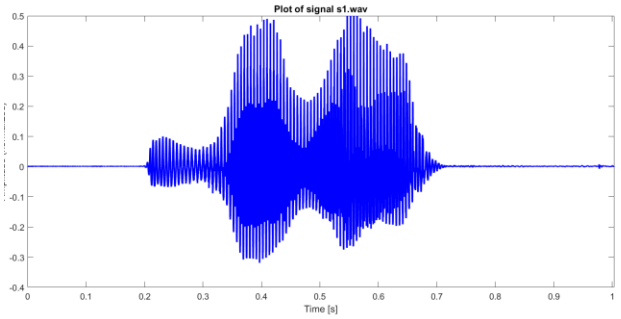
In Fig. 5, frequency of different samples of speech signal is represented. In this time vs Amplitude graph is plotted for different voice signal from different speakers for continuous speech.

In Fig. 6, plot of Vector Quantization codebook for different speaker is represented.

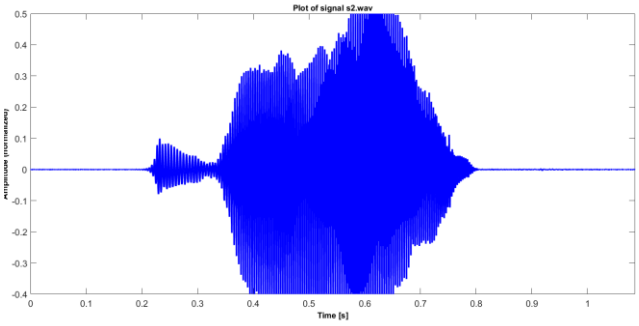
Table 4: Average accuracy for isolated words and continuous speech using MFCC

Word(W), Alphabets and C_S	Average Accuracy (%)	Overall Accuracy (for W,A_B, C_S)
W-Zero	95.3	60.63
W-Five	45	
W-point	41.6	
A	52.5	51.7
B	49.4	
C	53.3	
V	51.6	
C_S- "Hello, how are you"	85.5	

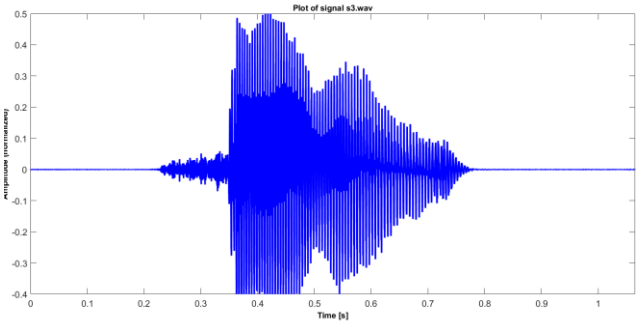
MFCC-Mel frequency Cepstrum coefficient, C_S- continuous Speech, A_B- Alphabets



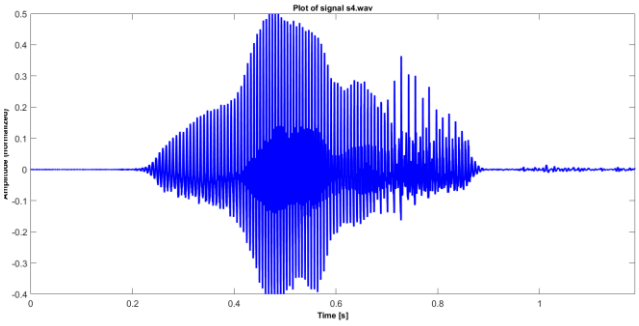
Speech signal frequency plot for speaker 1



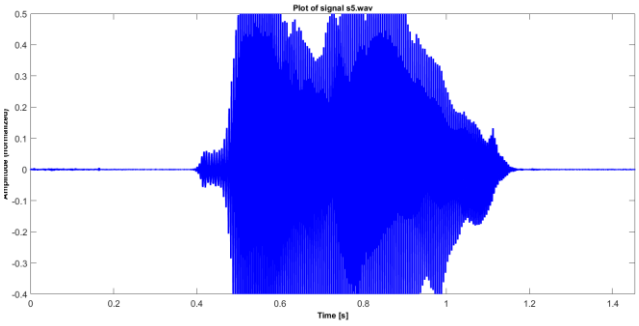
Speech signal frequency plot for speaker 2



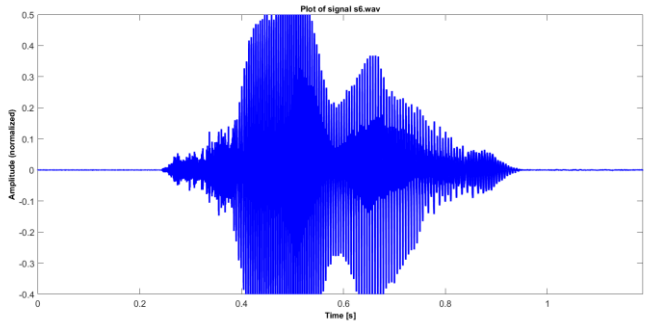
Speech signal frequency plot for speaker 3



Speech signal frequency plot for speaker 4



Speech signal frequency plot for speaker 5



Speech signal frequency plot for speaker 6

Fig. 4: Plot of different frequency from different speaker for same isolated word "Zero"

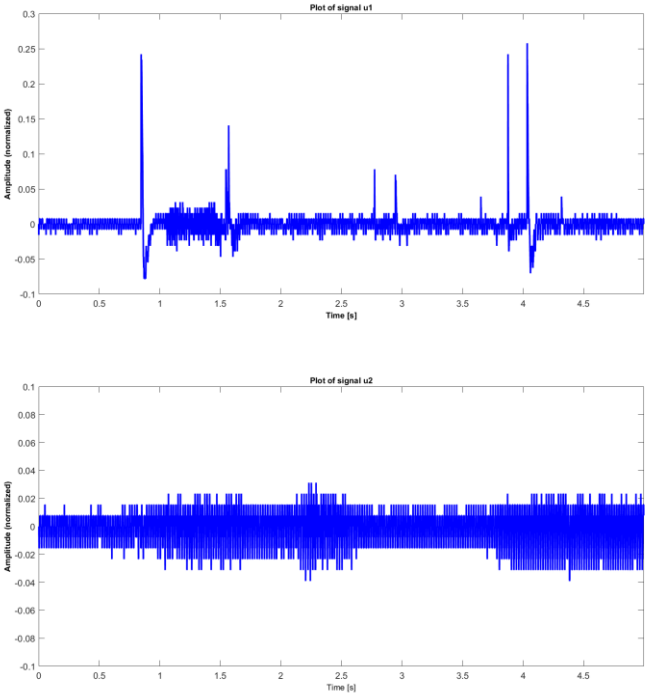
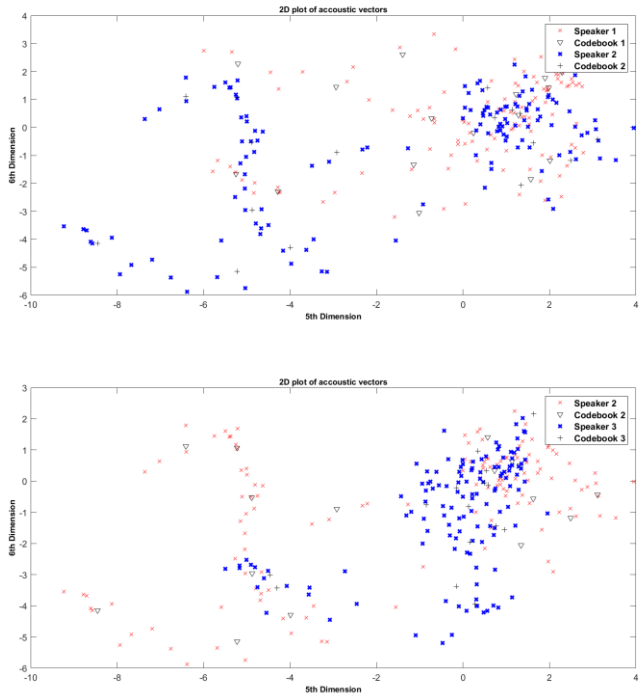


Fig. 5: plot of frequency for continuous speech signal



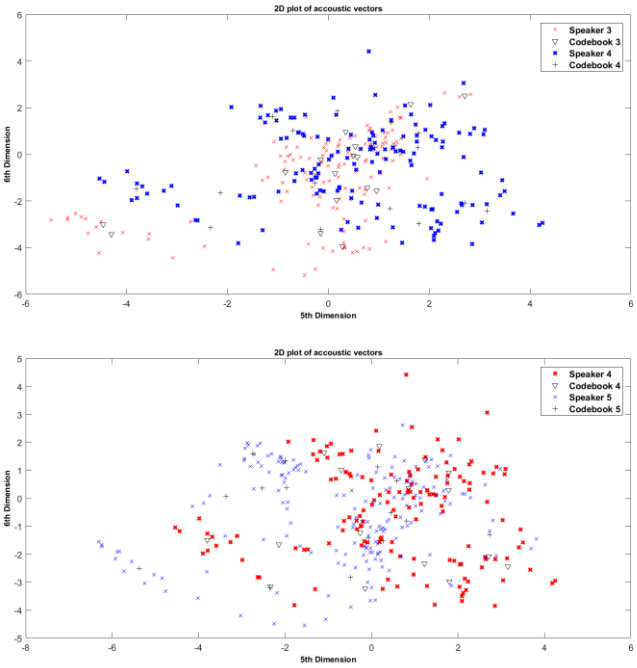


Fig. 6: plot of VQ codebook for different Speakers

4. Conclusion

The proposed approach performs better for continuous speech in comparisons to existing approaches in terms of accuracy. Speech accuracy is increases for continuous speech because distortion created by noise is reduce. The average performance of MFCC and VQ for isolated words is 60.63 and for alphabets 51.7 and for continuous words is 85.5.

In future, the processes of the speech recognition can also be increased with the help of HMM and Neural network. Neural network gives the self-learning mechanism. So the proposed approach if attached with neural network then it might enhance the accuracy of speech recognition for continuous speech.

References

Gaikwad SK, Gawali BW, and Yannawar P (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3): 16-24.

Hasan MR, Jamil M, Rabbani MG, and Rahman MS (2004). Speaker identification using mel frequency cepstral coefficients. In the 3rd International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 1(4): 565-568.

Ittichaichareon C, Suksri S, and Yingthawornsuk T (2012). Speech recognition using MFCC. In the International Conference on Computer Graphics, Simulation and Modeling, Pattaya, Thailand: 135-138.

Nijhawan G and Soni MK (2014). Speaker recognition using MFCC and vector quantisation. International Journal on Recent Trends in Engineering and Technology, 11(2): 211-218.

Plannerer B (2005). An introduction to speech recognition. Tech. rep., University of Munich Munich, Germany.

Singh N, Khan RA, and Shree R (2012). Mfcc and prosodic feature extraction techniques: A comparative study. International Journal of Computer Applications, 54(1): 9-13.

Sledevic T, Serackis A, Tamulevic˘ius G, and Navakauskas D (2013). Evaluation of features extraction algorithms for a real-

time isolated word recognition system. International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, 7(12): 303-307.

Vimala C and Radha V (2014). Suitable feature extraction and speech recognition technique for isolated tamil spoken words. International Journal of Computer Science and Information Technologies (IJCSIT), 5(1): 378-383.

Gray R (1984). Vector quantization. IEEE Assp Magazine, 1(2): 4-29.