INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue4)
Available online at www.ijariit.com

Speech Recognition in Noisy Environment-an implementation on MATLAB

Nishitha Danthi
Dayananda Sagar College of Engineering
nishithadanthi@gmail.com

Dr. A. R Aswatha
DSCE, Bangalore-78
aswath.ar@gmail.com

Abstract: Speech is one of the ways to express ourselves naturally. So, speech can be used as a means to communicate with machines. In this work, using MATLAB as a platform isolated word recognizer is achieved.
Speech signals get distorted by many kinds of noises. Hence, it is necessary to reduce the noise contained in the speech signal. This is called speech enhancement. Speech enhancement aims at improving the intelligibility of the speech. Noise has been removed using Spectral Subtraction with Over Subtraction technique. The feature extraction is carried out using MFCC and feature matching is achieved using HMM.

Keywords: Speech Enhancement, Windowing, VAD, MFCC, HMM.

I. INTRODUCTION

Speech is one of the approaches to communicating naturally. In this way, speech can be utilized as a way to speak with machines. In this venture, utilizing MATLAB as a platform isolated word recognizer is accomplished.
Speech signals get contorted by numerous sorts of noises. Hence, speech enhancement is very much important. It is done through Spectral Subtraction with Over Subtraction method. The feature extraction is done MFCC and feature matching is achieved by HMM.
Features must be extracted from the speech signals. This is achieved by parameterizing the speech signal into appropriate feature vectors. These vectors are utilised for the recognition purpose, i.e. for models training and testing as well.
A statistical approach to speech processing, the advantage of using hidden Markov modeling is the very noticeable advantage of training cost and computational simplicity. Although a neural network is a huge application of approach, its training process becomes highly expensive with an increase in the number of states, which can result in large numbers of intermediate neurons and weight as well. For simple and more basic levels of speech processing, statistical approaches are more viable. Therefore, even today, very efficient recognition systems employ a combination of these two approaches.

II. METHODOLOGY

Speech signals are subject to processing by following the steps of (HMM) Hidden Markov Model, which is the most effective automatic speech recognition (ASR) system. Noise is reduced by spectral subtraction with over subtraction techniques. The signal is then converted to the first parametric form, which is the Mel-Frequency Cepstrum Coefficients (MFCC), and then it is subjected to training and recognition. Fig1pictorially represents the methodology. The models are designed and tested for all the words that are recognized in MATLAB.

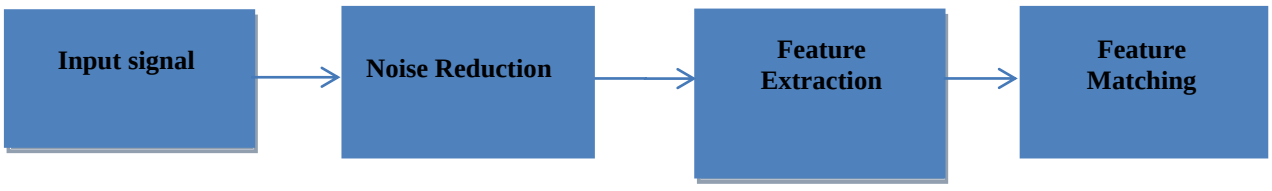


Fig1: Block Diagram of Speech Recognition System.

A. Voice Sample Database

The voice sample database is created by recording speech samples from both male and female speakers. It is recorded using Audacity software and the sampling frequency is kept at 8 KHz. Within one part of the audio signal, there is one voice frequency or one of the voice band frequencies. The frequency band of voice approximately ranges from 300Hz to 3400Hz. An adult male has 85Hz to 180Hz as fundamental frequency and for an adult female, it is 165Hz to 255Hz.

B. Noise Addition

To test various techniques proposed in this section, the noise signal has been added to a clear speech signal. The algorithm should be tested with AWGN noise. The purpose of this is to find out which of the three algorithms is best for removing any particular type of noise.

C. Frame Blocking and Windowing

Dividing the signal into a matrix form having an appropriate length of the time for every frame is the main intention of frame blocking. It is assumed that a signal is stationary in a frame of 32ms. A frame having 256 samples is obtained by making the sampling frequency as 8 KHz.

After frame blocking, for every frame Hamming window is applied. The discontinuities between the frames are reduced by applying the window and thus preventing unnecessary high frequencies if the Fourier transform is applied to it.

The equation that defines the length K of a humming window:

$$w(k) = 0.54 - 0.46\cos\left(\frac{2\pi k}{K-1}\right) \quad (1)$$

D. FFT on Each Block

In the matrix, 256 points FFT is applied on the frame with each window. This is a spectrum of noise-speech because it is a representation of variations of frequencies over time.

E. Noise Reduction Using Spectral Subtraction With Over Subtraction

This technique is mainly concentrated to lessen the "musical noise" effect. In this method, the modification done to the spectral subtraction method is by introducing an over-estimate factor, i.e. from the power spectrum of an over-estimate is subtracted. By doing so, it can be prevented that the spectrum of the resultant going below a minimum pre-set floor value of the spectrum. The effect of the musical noise is reduced by this modification and it also reduces the perception of the excursions of the narrow spectrum [7].

$$\left|\hat{X}_k\right|^2 = \begin{cases} |Y_k|^2 - \left|\hat{D}_k\right|^2, if\ |Y_k|^2 \geq (\alpha+\beta)\left|\hat{D}_k\right|^2 \\ \beta\left|\hat{D}_k\right|^2, otherwise \end{cases} \quad (2)$$

$\left|X_k\right|$ is the magnitude spectrum of clean speech, $\left|\hat{D}_k\right|$ estimated noise magnitude spectrum, $\left|Y_k\right|$ is the noisy signal magnitude spectrum, α and β is an over-subtraction factor and spectral floor parameter respectively, with α >1 and 0 < β ≤ 1. The parameter α is the function of signal to noise ratio (SNR) given by the equation.

$$\alpha = \alpha_0 - \frac{3}{20}SNR, -5dB < SNR < 20dB \quad (3)$$

$\left|X_k\right|$ is the magnitude spectrum of clean speech, $\left|\hat{D}_k\right|$ represents the estimated magnitude spectrum of the noise, $\left|Y_k\right|$ represents the magnitude spectrum of the noisy speech signal. α is an over-subtraction factor and β is spectral floor parameter. Here the values must be α >1, 0 < β ≤ 1. α is a parameter which is a function of (SNR) signal to noise ratio. It is given by the following equation.

F. Inverse FFT

To convert the enhanced speech signal from frequency domain to time domain we do inverse FFT and take the real part of it [2].

G. De-Framing

Overlap And Add Method:

Speech is divided into overlapping frames before it was further processed. The enhanced speech is reframed using the overlap-add method.

III.FEATURE EXTRACTION USING MEL-FREQUENCY CEPSTRAL COEFFICIENTS

The information which will be analysed must be adjusted while speech recognition system for recognising isolated words is created. Information present in the analog speech signal which is in a discrete and parametric shape is very helpful in recognising speech using HMM. This is the reason why the analog speech signal is converted into the parametric MFCC.

A. Pre-Emphasis

The signal needs to be leveled clearly. The pre-emphasizer, often to emphasize high-frequency components, the first order represents the high pass FIR filter. In the time domain, the composition is as described in the equation

$$h(n) = (1, -0.95) \quad (4)$$

B. Voice Activation Detection(VAD) & Silence Removal

When we have an approach to the sampled as well as discrete signal, the samples representing the signal values is of interest and not that of noise. Hence, it is important to reduce the size of the data. This means there is a requirement for a good Voice

Activation Detection. It can be achieved through many ways. In MATLAB's Voicebox Toolbox, VAD is a function defined for the *VAD Sohn*. This intermediate eliminates silent frames with no significant signal content, in which the signal adds the data to a fully useful point. There is a silent-speech-silent format in a specific speech sample where the silence has to be removed. The entire pronunciation is divided into frames of approximately 32 milliseconds (i.e. 256 sampling points for 8 kHz sampling) and the silent frame has been removed based on the maximum energy value of each frame. The threshold value must be set carefully based on the background noise level and the signal values in the sample, x_vad (n).

C. Normalization

The *x_vad(n)* has to be normalized to the range -1 to 1, so as to prevent signal amplitude from being a parameter.

D. Frame Blocking and Framing

The signal after subjected to pre-processing is split into units called frames. Each and every frame is subjected to extraction features. The frame length is 256 samples for sample rates of 32ms i.e., 8000Hz. The frame is overlapped by an overlapping factor of 75%, so each frame is different from 25%, which are 128 samples

After frame blocking, for each frame Hamming window is applied. The discontinuities between frames are lessened after applying windowing and thus preventing unnecessary high frequencies when Fourier transform is applied to it.

$$w(k) = 0.54 - 0.46\cos\left(\frac{2\pi k}{K-1}\right) \tag{5}$$

Where, w(k) – window function.

E. FFT on Each Block

256 point FFT is applied on each windowed frame. The resulting matrix is having the same dimensions as the *x_windowed*.

This *x_fft(m,n)* is a spectrogram since it is a representation of the variation of frequencies with time.

F. Mel-Frequency Cepstral Coefficients

The Mel -scale is related to perceived frequency of a real tone, or pitch. It is the measured frequency in actual. Human beings are more sensible at discerning little changes in the pitch at frequencies at the lower band than the frequencies at the higher band. After this scale is incorporated, more people meet people who listen to us than our characteristics.

$$F_{mel} = 2595\, log_{10}\left(1 + \frac{F_{Hz}}{700}\right) \tag{6}$$

$$F_{Hz} = 700\left(10^{\frac{F_{mel}}{2595}} - 1\right) \tag{7}$$

Using triangular Mel scale filter banks, practical wrapping is achieved. Frequency from Hz to frequency in Mel scale is wrapped and it is as shown in Fig 2.

These filters are applied on each FFT frame to get x_mel (m, n), where m = 26 and n = the number of frames to obtain x_dct (m, n), a Discrete Cosine Transform (DCT) has been applied to x_mel.

The pitch contribution is removed when the Discrete Cosine Transform is performed according to the following equation [18]. So we get the spectral envelope, which describes the noise filter which defines the nature of the speech.

$$cep_s(n,m) = \sum_{p=0}^{N-1} \alpha_k \log(fmel_k)\cos\left(\frac{\pi(2n+1)p}{2N}\right) \tag{8}$$

n=0, 1, 2 ….N-1

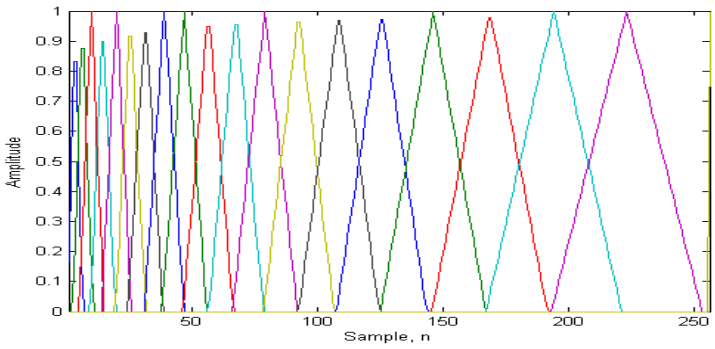


Fig 2: Mel Filter Bank

G. Delta & Acceleration Coefficients

In order to increase the content of information of the perception of the human, delta and acceleration coefficients are calculated. The delta coefficients are all about the time difference and the acceleration coefficients are all about the second time derivative.

The delta coefficients are calculated according to the following equation:

$$\delta^{[1]} = \frac{\sum_{k=-K}^{K}\left(C_h(n,m+k) - C_h(n.,m)\right)k}{\sum_{k=-K}^{K} k^2} \tag{9}$$

The acceleration coefficients are calculated according to the following equation:

$$\delta^{[2]} = 2\left(\frac{\sum_{k=-K}^{K} k^2 C_h(n, m+k) - (2k+1)\sum_{k=-K}^{K} C_h(n, m+k)k^2}{\left(\sum_{k=-K}^{K} k^2\right)^2 - (2k+1)\sum_{k=-K}^{K} k^4}\right) \quad (10)$$

H. Normalization

Normalization refers to enhancement. The feature vectors are normalized under time in order to obtain mean as zero and variance as unity. The generalization facility forces the vectors to the same numerical range [14]. The mean vector, called $\bar{f}_{\mu}(n)$, can be calculated according to the equation:

$$\bar{f}_{\mu}(n) = \frac{1}{M}\sum_{m=0}^{M-1} x\_mfcc(n, m) \quad (11)$$

To normalize the feature vectors, the following operation is applied:

$$\bar{f}_{\mu}(n; m) = x\_mfcc(n, m) - \bar{f}_{\mu}(n) \quad (12)$$

IV. HIDDEN MARKOV MODEL

The method used for recognition of speech as mentioned in the introduction part is (HMM) Hidden Markov Model. Training of models is achieved through this method, which is used to represent an utterance of the spoken word. To test the utterance, this model is only used later. This model is later used to test an utterance and probability of the model having created the vector sequences.

The Hidden Markov Model is represented by **λ = (π, A, B)**, where:

π = Initial state distribution vector.

A = State transition probability matrix.

B = Continuous observation probability density function matrix.

When MFCC is achieved, all the given training negotiations are required to be generalized. The number of matrix States is divided into several coefficients. Then all these metrics are used to calculate the mean and variance.

Amid the experimentation with the quantity of entrance inside the re-estimation of A the last assessed estimations of A where seen to stray a considerable amount from the earliest starting point estimation. The last introduction estimations of A are instated with the accompanying esteems rather, which will probably the reassessed values (the re-estimation issue is managed later on in this segment.

Changes in the initial values are not an important event, so according to the estimated process, the estimation again adjusts the value to the right people.

A. Initial State Distribution Vector

The initial state distribution vector is initialized with the probability to be in state one at the beginning, which is assumed in speech recognition theory [19]. It is also assumed that *i* is equal to five states in this case.

$\pi_i = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}$, $1 \le i \le$ number of states, in this case $i = 5$

B. Continuous Observation Probability Density Function Matrix

As specified the Hidden Markov Model, the intricacy of the immediate perception of the condition of the discourse procedure is unrealistic there is a requirement for some measurement computation. This is finished by presenting the persistent perception likelihood thickness work grid, B. The thought is that there is a likelihood of mentioning a specific objective fact in the express, the likelihood that the model has created the watched Mel Frequency Cepstrum Coefficients. There is a discrete perception likelihood contrasting option to utilize. This is less entangled in counts yet it utilizes a vector quantization which produces a quantization blunder. A greater amount of this option is to peruse in [17].

Because of that, the MFCC are regularly not recurrence disseminated a weight coefficient is important to utilize when the blend of the pdf is connected. This weighting coefficient, progressively the quantity of these weights are utilized to demonstrate the recurrence capacities which prompt a blend of the pdf.

C. Forward Algorithm

While hunting down the likelihood of a perception grouping O_N = (o1, o2, o3… oT) a model is given λ = (π, A, B) you have to discover an answer for the issue, likelihood evaluation [19]. Its answer is about the model (expecting they exist), which is well on the way to create the perception arrangement. The normal method of doing this is to assess each conceivable arrangement of length T states and after that to include these two together.

$$P(O \mid \lambda) = \sum_{q_1, q_2, \ldots, q_T} \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \quad (13)$$

D. Backward Algorithm

In the event that the recursion portrayed for figuring the forward factor is invert, at that point, you will get βt(i), in reverse factors. This variable is characterized with the accompanying definition:

$$\beta_t(i) = P(o_{t+1} o_{t+2} \ldots o_T \mid q_t = i, \lambda)$$

The meaning of $\beta_t(i)$ is that $\beta_t(i)$ there is a probability of time t and in the state i provided for the model, made fractional perception arrangement up to t + 1 supervision up to the perception number.

E. Log (P(O|λ))

The log(P(O| λ)) is spared in a network to see the re-appraisal arrangement modification. For each redundancy, there is a summation of the entirety (log (scale)), add up to likelihood. This summation is contrasted with the past summation in past cycle. In the event that the distinction between measured esteems is not as much as a limit, at that point one ideal can become too. On the off chance that fundamental, quantity of settled emphases can be resolved to diminish math.

F. Re-Estimation of The Model Parameters.

The prescribed calculation utilized for this object is a recursive Baum-Welch calculation which amplifies the likelihood capacity of a given model λ = (π, A, B) [12], [17], [18]. For every emphasis, the calculation appraises the most extreme estimation of the HMM parameter nearer to "Worldwide" (in numerous areas). The significance is that it is discovered that the neighborhood most extreme is first worldwide; generally, a wrong mix is found.

The Baum-Welch algorithm is based on a combination of the forward algorithm and the backward algorithm.

G. Testing of an Observation

Looking at the perception succession O_N = (o1, o2, o3… OT) with a model, λ = (π, A, B) you have to discover an answer for two issues [17]. The arrangement is about finding an inexact grouping and the most extreme succession of states for the model q = (q1, q2, q3… qT) is the diverse arrangement on what the ideal arrangement implies. On account of the in all probability state grouping in its consummation, the calculation Viterbi calculation [12], [17], [18] has been taken to augment P (q| O, λ), the state move potential outcomes have been taken by this calculation, Which is not done when you ascertain the most noteworthy conceivable state way. Because of issues with the Viterbi calculation, duplication with the conceivable outcomes, Alternative Viterbi calculation is utilized. Tests are done in such cases that the correlation which is contrasted and each model is analyzed, and after this, there is a score characterized for every examination.

Log(B):

Steady perception likelihood thickness work framework is computed in the past area. The distinction is that logarithm is utilized on the grid because of obstructions to the Alternative Viterbi calculation.

Delta:

To have the capacity to look the greatest measure of a state course, the prerequisite of the accompanying sum δ t (i) is important

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1 q_2 .. q_{t-1}, q_t = i, o_1 o_2 .. o_t \mid \lambda) \tag{14}$$

The amount δ t (i) is the likelihood of watching o1 o2 o3 . . . . ot utilizing the best way that finishes in state i at time t, given the model. Along these lines by acceptance, the incentive for δ t+1 (i) can be recovered.

$$\delta_{t+1}(j) = b_j(o_{t+1}) \max_{1 \le i \le N} \delta_t(i) a_{ij} q_2 \ldots q_{t-1}, q_t = i, o_1 o_2 \ldots o_t \mid \lambda) \tag{15}$$

Psi:

The ideal state arrangement is recovered by sparing the contention which augments δt+1(j), this is spared in a vector ψt(j) [17]. Note that while ascertaining bj(o t) the μ, Σ is accumulated from the distinctive models in correlation. The calculation is handled for all models that the perception arrangement ought to be contrasted.

Log(P*):

Likelihood computation is for the in all probability state succession. The greatest contention is on the last state.

qT:

Computing the state which gave the biggest Log(P*) at time T. Utilized as a part of backtracking later on.

Path:

State arrangement backtracking. Discover the ideal state succession utilizing the ψt ascertained in the acceptance part.

Score:

The score is as per the Alternative Viterbi calculation, like the ascertained esteem log (P *), the expansion of the likelihood of a comparative way is spared thus for every examination. The most astounding score is in all probability the examination demonstrates given by the similar model. The test explanation has been delivered.

V. RESULTS AND DISCUSSION

A. Noise Reduction Using Spectral Subtraction with Over Subtraction Technique

The analog speech signal is preferably recorded at ≥ 2fs in a relatively noiseless environment. The frequency band of Speech is 3 Hz to 4 KHz [10]. Speech signal “Bangalore” in waveform sampled is given in Fig 3 The speech signal was added with AWGN Noise and is as shown in Fig 4.

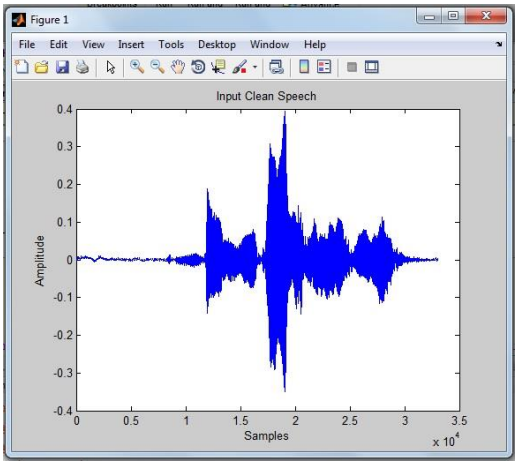


Fig 3: Recorded Speech Signal with Utterance "Bangalore"

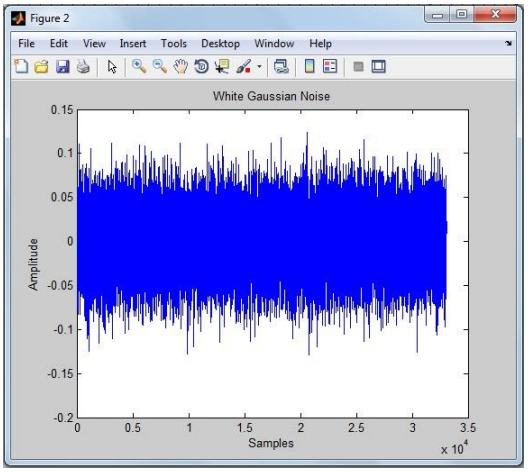


Fig 4: AWGN Noise

The noisy speech signal is as shown in Fig 5. The outcome from the Spectral Subtraction with Over Subtraction Technique is as shown in Fig 6.

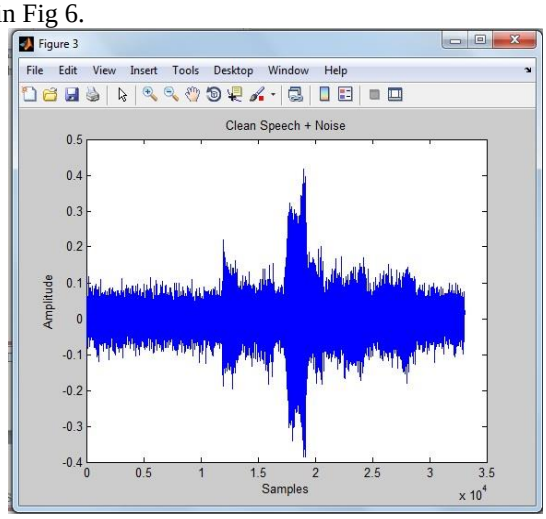


Fig 5: Speech with AWGN Noise

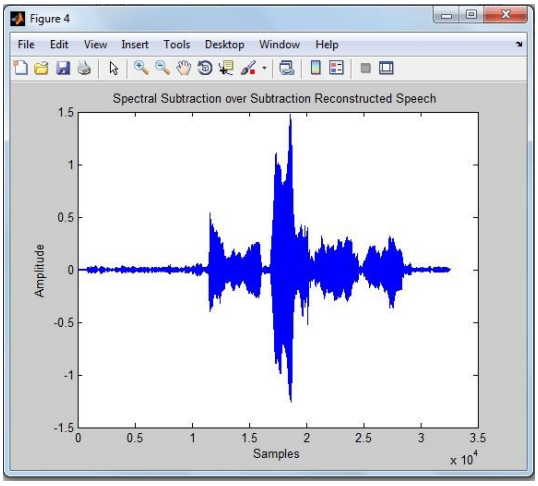


Fig 6: Spectral Subtraction with Over Subtraction Technique Output.

B. Extraction of Commands

The different utterances of the same word are recorded and are stored in the memory. The program accesses the.WAV format files in the memory and lists them under an array of different Words.

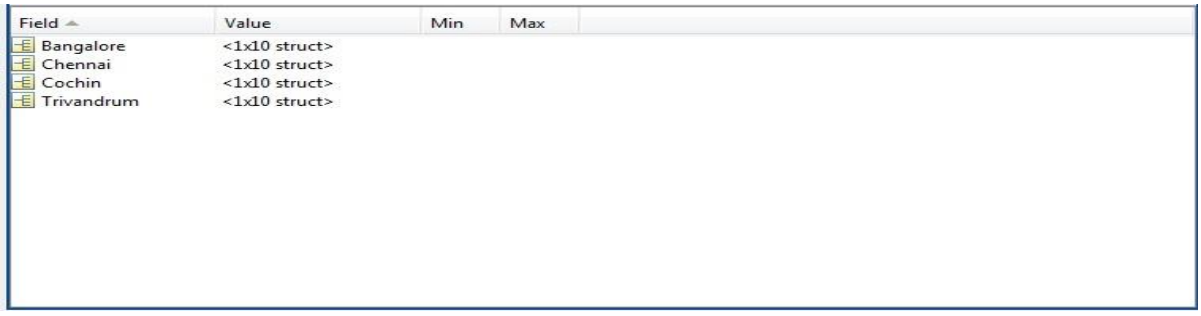


Fig 7: Arrays of words.

C. MFCC Features in MATLAB

MFCC facilities have Mel Frequency Cepstral coefficients MFCC is extracted from time signals. The steps for extraction of facilities have been discussed in the previous chapter. Fig 8 shows the conspiracy of MFCC facilities for Bangalore. MFCC features are 36 for each sample. In this example, features are drawn in 122-time instants, so the features of MFCC are 36 X 122. The changes in colors in Fig. 8 are due to the dimensions of the data present in that situation. The bar near the plot indicates the range of values that are proportional to any particular colour, the darkest red is the highest value of 3 and the darkest blue colour has the lowest value of -3. Other colours come in between these colours. The values of all the colours mentioned at that time are

those 36 coefficients; there are 12 cepstrum coefficients, 12-24 delta coefficients and 24-36 acceleration coefficients are for the next part of the modeling. We use these MFCC coefficients instead of the original signal.

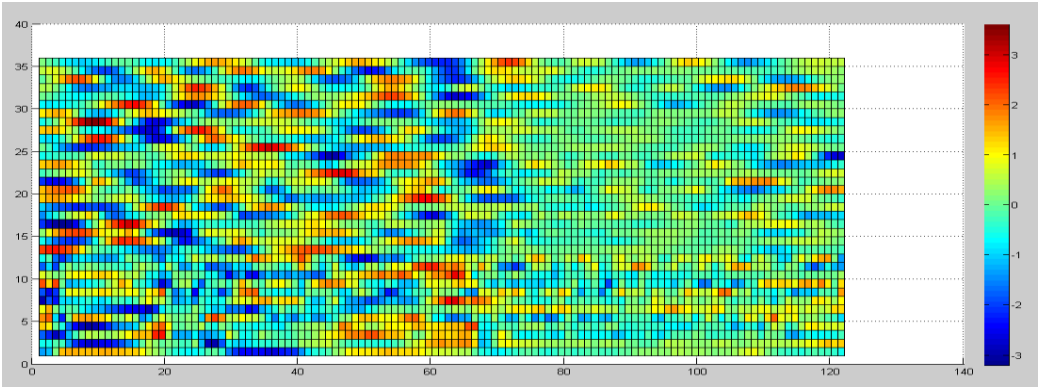
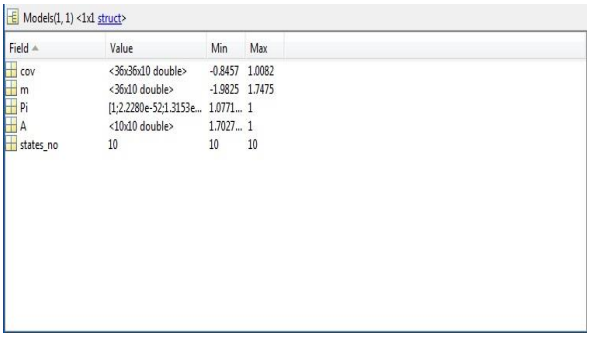


Fig8: Plot of MFCC features.

D. Preparation of Models

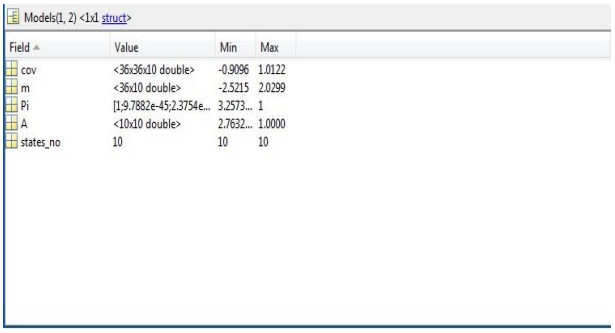
After preparing the commands, each element of the command is accessed as the input of the models. As we have seen, the first characteristics of modeling are extracted for all the words and then the states, mean, covariance, the initial state probability matrix and state transition probability metrics are received for each word and these parameters together are called Models. The Figures 9, 10, 11 and 12 show the models which are prepared for different words.



Models(1, 1) <1x1 struct>

Field	Value	Min	Max
cov	<36x36x10 double>	-0.8457	1.0082
m	<36x10 double>	-1.9825	1.7475
Pi	[1;2.2280e-52;1.3153e...	1.0771...	1
A	<10x10 double>	1.7027...	1
states_no	10	10	10

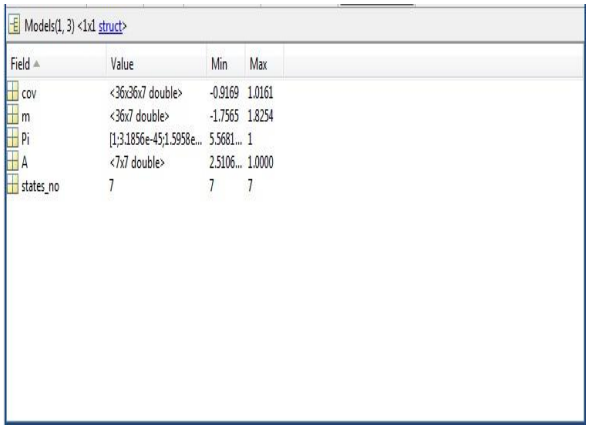
Fig 9: Model of the word BANGALORE



Models(1, 2) <1x1 struct>

Field	Value	Min	Max
cov	<36x36x10 double>	-0.9096	1.0122
m	<36x10 double>	-2.5215	2.0299
Pi	[1;9.7882e-45;2.3754e...	3.2573...	1
A	<10x10 double>	2.7632...	1.0000
states_no	10	10	10

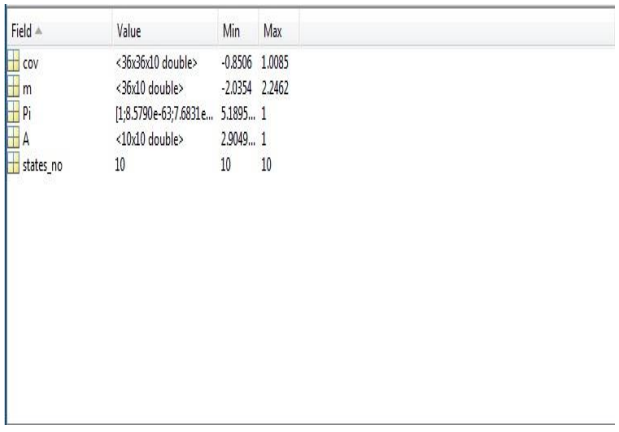
Figure 10: Model of the word CHENNAI



Models(1, 3) <1x1 struct>

Field	Value	Min	Max
cov	<36x36x7 double>	-0.9169	1.0161
m	<36x7 double>	-1.7565	1.8254
Pi	[1;3.1856e-45;1.5958e...	5.5681...	1
A	<7x7 double>	2.5106...	1.0000
states_no	7	7	7

Fig11: Model of the word COCHIN



Field	Value	Min	Max
cov	<36x36x10 double>	-0.8506	1.0085
m	<36x10 double>	-2.0354	2.2462
Pi	[1;8.5790e-63;7.6831e...	5.1895...	1
A	<10x10 double>	2.9049...	1
states_no	10	10	10

Fig 12: Model of the word TRIVANDRUM

E. Testing

Prepared models and testing utterances are given in the form of input. The first step in the test is to prepare the model for the testing utterances. Models of the test utterance are prepared with the same algorithm, with which the models are ready. The mean and covariance of the testing utterance model are compared with all the models. The result provides some probability of the event for each word, while the test statement is one which is most likely.

While tracing along the row the column which encounters highest (least negative) value will be the command contained in the test utterance.

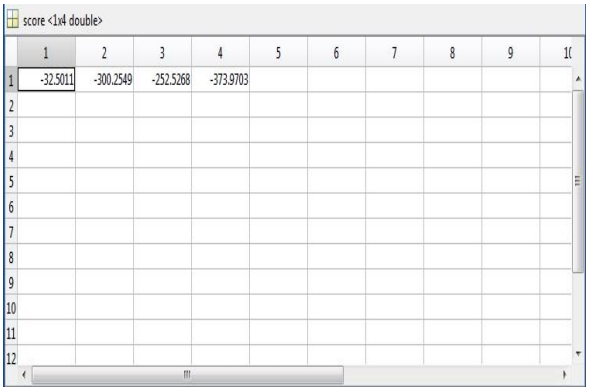


Fig 13: Scores of testing BANGALORE in MATLAB

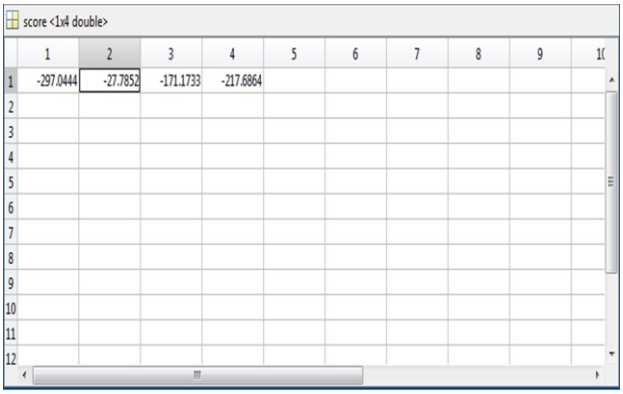


Fig 14: Scores of testing CHENNAI in MATLAB

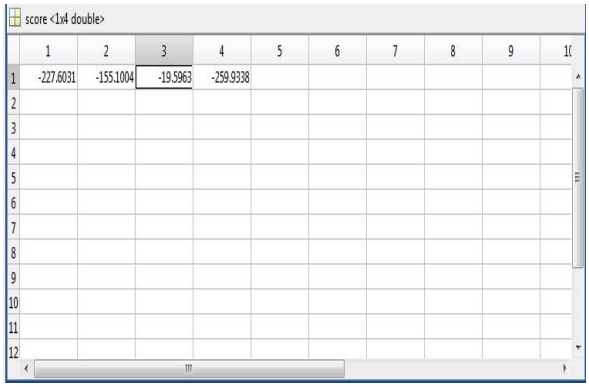


Fig 15: Scores of testing COCHIN in MATLAB

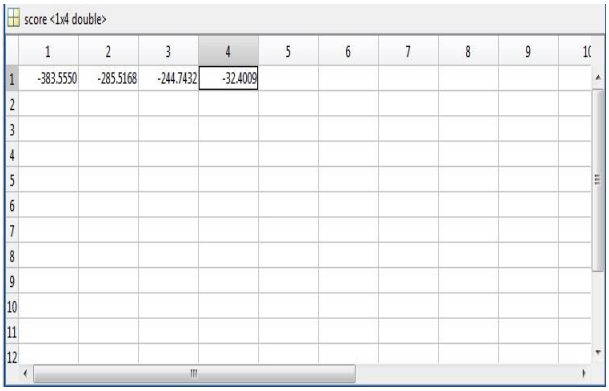


Fig 16: Scores of testing TRIVANDRUM in MATLAB

VI. CONCLUSION AND FUTURE WORK

The Noise is reduced by Spectral Subtraction with Over Subtraction Technique. Hidden Markov Modelling of speech signals proved to be quite accurate and efficient to identify and test small databases of simple words. Unlike neural networks, which include computational costs for large databases, HMM is a better option. Today, the use of HMM is generally used for all cases and interpretation of speech processing as well as manuscript analysis.

In this line for this project, words based on terminology based on the implementation of HMM for sentence modeling, synonyms for one word and word recognition of HMM in this line will apply in states instead of an arbitrary division of state MFCC data.

REFERENCES

[1] Javier Ortega-García and Joaquín González-Rodríguez, Overview of speech enhancement techniques for Automatic Speaker Recognition

[2] Lalchhandami and Rajat Gupta, Different Approaches of Spectral Subtraction Method for Speech Enhancement

[3] Vishv Mohan, "Analysis And Synthesis Of Speech Using Matlab" International Journal of Advancements in Research & Technology, Volume 2, Issue 5, M ay-2013 373 ISSN 2278-7763

[4] Weinstein, Clifford J. "Opportunities for advanced speech processing in military computer-based systems." Proceedings of the IEEE 79.11 (1991): 1626-1641.

[5] Englund, Christine. "Speech recognition in the JAS 39 Gripen aircraft-adaptation to speech at different G-loads." Master degree thesis 57 (2004).

[6] Kumar, P. Sathish, Suraj, S. Subramanian, R. Venkata, Raghavan and Vinay V, "Voice Operated Micro Air Vehicle", International Journal of Micro Air Vehicles, Jun 2014, Vol. 6 Issue 2, p129.

[7] Robust Automatic Transcription of Speech (RATS), for Information Processing Techniques Office (IPTO), Defense Advanced Research Projects Agency (DARPA), DARPA-BAA-10-34, 2012.

[8] Tolba, Hesham, and Douglas O'Shaughnessy. "Speech recognition by intelligent machines." IEEE Canadian Review 38 (2001): 20-23.

[9] Javier Ferreiros, Rubén San-Segundo, Roberto Barra, Víctor Pérez, Increasing robustness, reliability and ergonomics in speech interfaces for aerial control systems, Aerospace Science and Technology, 13(2009), 423–430.

[10] Cary R. Spitzer, Chapter-8, Digital Avionics Handbook, CRC Press, 2001, 978-1- 4200-3687-9

[11] Sledevic, Tomyslav, et al. "Evaluation of features extraction algorithms for a real- time isolated word recognition system." International Journal of Electronics 7.12 (2013).

[12] Mikael Nilsson and Marcus Ejnarsson, "Speech Recognition using Hidden Markov Model", Degree of Master of Science in Electrical Engineering, Department of Telecommunications and Signal Processing, Blekinge Institute of Technology Ronneby, March 2002.
[13] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm", Proceeding of European signal processing conference, 2009.
[14] Dimitrios S. Koliousis, "Real-time speech recognition system for robotic control applications using an ear microphone", thesis submitted to the naval postgraduate school, Monterey, California, 2007.
[15] Kumar, Kshitiz, Chanwoo Kim, and Richard M. Stern. "Delta-spectral Cepstral coefficients for robust speech recognition." Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011.
[16] Acero, Alex, and Xuedong Huang. "Augmented Cepstral normalization for robust speech recognition." Proc. of IEEE Automatic Speech Recognition Workshop. pp. 146-147, 1995.
[17] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol.77, No.2, pp.257-286, February 1989.
[18] Chris Karlof, David Wagner "Hiddn Markov Modelling Cryptanalysis", Computer Science Division (EECS), University of California, Berkeley, California

