ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

SPEECH TO TEXT CONVERSION USING GAUSSIAN MIXTURE MODEL (GMM) IN NOISY ENVIRONMENT

S .SOUNDARYA , P.SRI SARANYA , R.SUMATHI , K.VINOTHINI

B.E Student, Department of Electronics and communication engineering,
Saranathan college of engineering,Panjappur, Trichy -620 012.

Abstract:
Speech to text conversion system (STT) is expedient for deaf and dump people. The input is given in the form of speech and output is text. In recent years, many feature extraction proficiency are used. They are PLP, LPC, LDA, ICA, PCA, MFCC,Kernal based feature extraction, wavelet transform and spectral subtraction. Mels Frequency Cepstral Coefficient (MFCC) is utilized to educe the characteristics feature of speech. MFCC is used for reducing the size of speech signal before recognition and also for extracting the characteristics of vocal sound. To estimate the speech parameter, various models like GMM, HMM, VQ, ANN are being used. Among those, GMM is most superior in many applications. In our project, GMM is used. GMM is a parametric probability density function which is represented as a weighted sum of Gaussian component densities. GMM model parameters are calculated by maximum likelihood estimation. Iterative expectation minimization (EM) technique is used to obtain matching and Euclidean distance to recognize the speech is acknowledged in terms of text. The proposed work is utilized by the deaf and dumb people, security systems, household appliances, mobile phones, ATM machines, computers etc.

1.INTRODUCTION

We communicate with each other in many ways such as expression, eye contact, gesture, and speech. The simple mode of communication among people is speech and also the most natural and efficient form of exchanging information amongst them in speech. Speech-to-text conversion system is extensively used in many applications. In the field of education, Speech-to-text conversion system or speech recognition system is more productive for deaf and dumb students. Speech recognition is a challenging part in speech processing systems. The significant part of the system is Feature Extraction. There are different types of feature extractions methods. Mel Frequency is based on the characteristics of the human ear's hearing, which uses a discontinuous frequency unit to reproduce the human acoustic system.Mel-Frequency scale is used to extract features of the input speech signal. These cepstral features provide the accuracy of recognition to be systematic for speech recognition and emotion recognition system At present, Speech-to-text convertor systems is mostly used in many mobile phones, computers and control systems. Accordingly, Cepstral Coefficients (MFCC) method is used. This Speech-to-text convertor systems are more useful in our day-to-day activities . In the paper, GMM and MFCC are implemented by using MATLAB.

2.LITERATURE SURVEY

MananVyas, "A gaussian mixture model based speechrecognition system using matlab", 2015.[3]

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

This paper aims at development and performance analysis of a speaker dependent speech recognition system using MATLAB. The issues that were considered are 1) Can Matlab, be effectively used to complete the aforementioned task, 2) Accuracy of the Gaussian Mixture Model used for parametric modelling, 3) Performance analysis of the system, 4) Performance of the Gaussian Mixture Model as a parametric modelling technique as compared to other modelling technique and 5) Can a Matlab based Speech recognition system be ported to a real world environment for recording and performing complex voice commands. The aforementioned system is designed to recognize isolated utterances of digits 0-9. The system is developed such that it can easily be extended to multisyllabic words as well.

G.P.S. Prasanthi, K. Sirisha, G. Ramya,B. Padma, "Speech to Text Conversion Using HMM", 2016.

"Real time speech to text" can be defined as accurate conversion of words that represents uttered word instantly after speaking." The speech-to-text conversion can provide data entry options for deaf students. The system is also used to find the disorder rate of persons affected with Parkinson‟s disease by calculating the efficiency of pronunciation. The system takes the speech at run time through a microphone and processes the sampled speech to recognize the uttered text. In the training phase, the uttered digits are recorded using the PCM modulation technique with a sampling rate of 8 KHz and saved as a wave file. MATLAB software uses the wavread command to convert the wav files to speech samples. A HMM Model is used for speech recognition, which converts the speech to text. In this *hidden* Markov model, the state is not directly visible, but output, dependent on the state, is visible. It uses baum-welch algorithm. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

3. METHODOLOGY

3.1 Silence and noise removal

Speech signal is divided into voiced and unvoiced signals. Voiced signals are periodic and one-third part of the speech signal is voiced signal which is important for intelligibility. Unvoiced signals are non-periodic, undesirable sound signals produced by the vocal tract. Unvoiced signals can also be considered as noise or the silence part of the speech signal. Silence can be removed by setting a threshold voltage which is dependent on background noise.

3.2 MFCC

The most easiest and prevalent method to extract spectral features is calculating the Mel-Frequency Cepstral Coefficients (MFCC) from human voice. It is one of the most popular methods of feature extraction used in speech recognition systems. It is based on frequency domain using the Mel scale which is based on the human ear scale. Time domain features are less accurate than the frequency domain features. The main aim of feature extraction is to reduce the size of the speech signal before the recognition of the signal. Steps involved in feature extraction are pre-emphasis, framing, windowing, fast fourier transform, Mel-

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

frequencyfiltering, Logarithmic function and Discrete Cosine Transform etc.

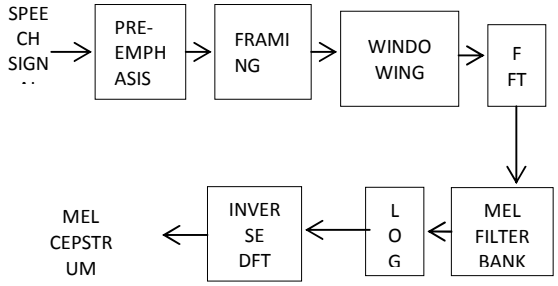

Fig 1 :Block diagram of MFCC

The first step in MFCC is pre-emphasis which is usedto boost the high frequencies of a speech signal which are lost during speech production. Pre-emphasis is needed because high frequency components of the speech signal have small amplitude with respect to low frequency components. Therefore higher frequencies are artificially boosted in order to increase the signal-to- noise ratio.

$$P(z)=1-az^{-1}$$

Next, is framing which is used to block the frames obtained by analog to digital conversion (ADC) of speech signal . The number of samples in each frame is chosen as 256 and the number of samples overlapping between adjacent frames is 128. Overlapping frames are used to acquire the information from the boundaries of the frame. Sampling rate is to be selected according to the Nyquist criteria that is

$$F_s >= 2F_{max}$$

Due to discontinuities at the start and the end of the frame causes undesirable effects in the frequency response, so windowing is used to eliminate the discontinuities at the edges. Hamming window is used which introduces least amount of distortion. Generalized hamming window equation is

$$W(n)=(1-\alpha)-\alpha \cos[2\pi n/(N-1)]$$

$$0 \leq n \leq N-1$$

After windowing, Fast Fourier Transform(FFT) is measured for every frame to extract the frequency components of the signal in time domain. Speech signal does not follow linear frequency scale used in FFT.

$$L=2\text{^}(\log(W)/\log(2))$$

Hence Mel-scale is used for feature extraction which is directly proportional to the logarithm of linear frequency. Equation is used to convert linear scale frequency into Mel-scale frequency.

$$Melf=2595*\ln(1+(f/700)) \text{ in HZ}$$

Triangular bandpass filters are used to extract the spectral envelope. 20 filters are used. Log is applied to the absolute magnitude of the coefficients of which is obtained after Mel-scale conversion. Discrete cosine transform(DCT) converts the Mel-frequency domain into time domain. The value of K ranges between 8 and 13. We choose K as 13 and hence we obtain 13 coefficients for each frame.

3.3 GAUSSIAN MIXTURE MODEL

Gaussian mixture model (GMM) is a mixture of several Gaussian distributions and can therefore represent different subclasses inside a single class . GMM combines the probability distribution of various classes and calculate the probability for a single class. GMM ‘s are considered for evaluating density and performing Clustering. The Expectation-Maximization algorithm is used for this purpose. GMMs are comprised of component function called Gausses.The number of these Gausses in the mixture model is also referred to as the number of components. The total numbr of component can be altered based on the count of

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

training data points. However,the model becomes more complex with the increase in the number of components. Mixture models are type of density models, which comprise a number of component functions, and each component is usually of Gaussian in nature. The weighted combinations of these component functions result the desired multi-modal density.

Mixture models are semi-parametric in nature and provide greater flexibility and precision in modeling the underlying statics of data.They can able to smooth-out the gaps resulting from sparse data and provide tighter constraint in assigning object membership to cluster regions. Gaussian mixture density also provides a smooth approximation to the long term sample.Distribution for the features of the particular class is decided by mean vector and covariance matrix.

4.IMPLEMENTATION

The general flow of processing and recognition of the speech signal is shown in fig(2)

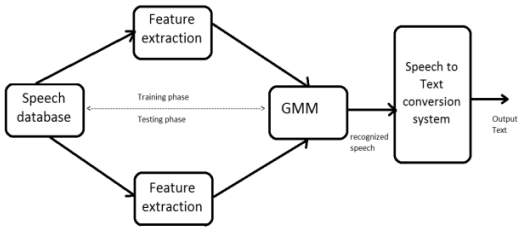

fig 2 : General block diagram of GMM model

The speech signal is recognized by using Gaussian Mixture Model. The speech signal is recorded by using 16-bit Pulse code modulation with a sampling rate of 8KHz and it is stored as a wave file by using sound recorder software in MATLAB. .wav files are converted into speech samples by using MATLAB software's wavered command. The silence part of the speech as well as background noise is removed at the initial stage of processing. A threshold voltage is set to remove the noise and the silence part which is totally dependent on environment with less ambient noise. Then the features of the speech signal is extracted by the MFCC block. The total number of samples chosen in a frame is 256 and overlapping samples with the adjacent frame will be 128. We acquire MFCC cepstral coefficients at the output of MFCC block. In GMM, K-mean algorithm is used to obtain a cluster number specific to each observation vector and sets the centriod of the observation vector. After clustering the model, it returns one centriod for each of the cluster K and refers to the cluster number closest to it. K-mean algorithm is described as the squared distances between each observation vector and its centroidscentroids. In the training section parameters of GMM model are produced iteratively by expectation-maximization(EM) algorithm. Euclidean distance is found out between observation vector and its cluster centriods to match the spoken word with the present database. The word which is recognized is displayed as text in the output.

5.RESULT

For acquiring the results the speech signal is recoded. The system is trained for multiple words such as hello,bye-bye etc. The sample speech signal which is recorded is shown in fig(3)

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

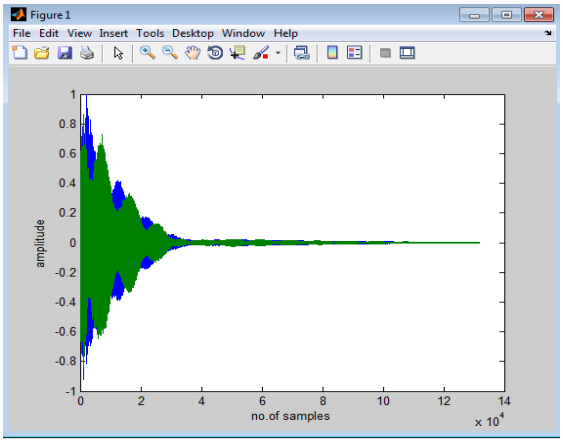

fig 3 : representation of speech signal

Framing is applied to the speech signal .framed signal is shown in fig

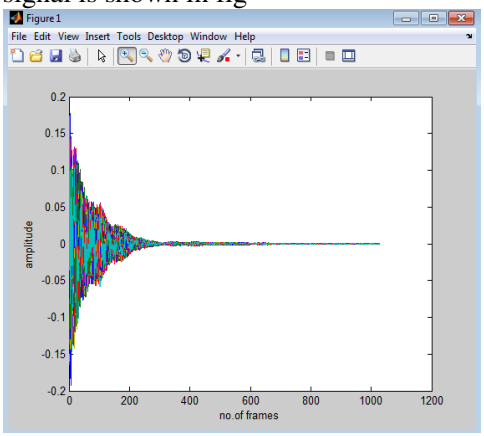

Fig 4:representation of framed signal

Words are trained. They are stored and createddatabse.the sample database is shown is the fig

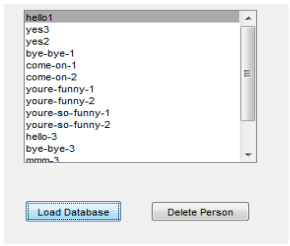

Fig 5: sample database

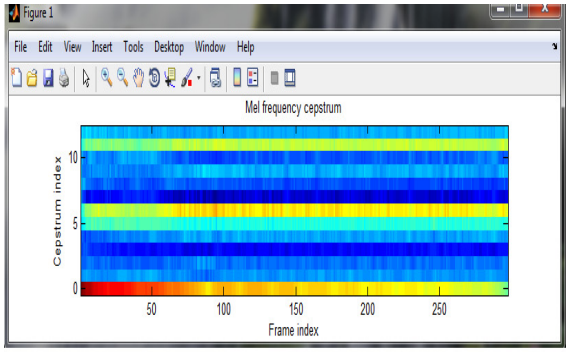

Fig 6:spectrogram of MFCC

Speech signal is loaded using speech recognition window .for an example,you're funny .wav file is loaded .screenshot of waveread operation is shown in below fig.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

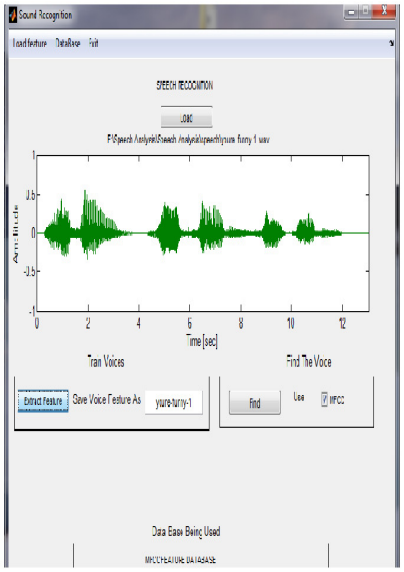

Fig 7: representation of loading the singal

A test signal is compared with already trained signal .Maximum posterior probability is calculated for all the signal which are trained. Test signal is matched with trained signal which has maximum probability .screenshot of matching is shown in fig.

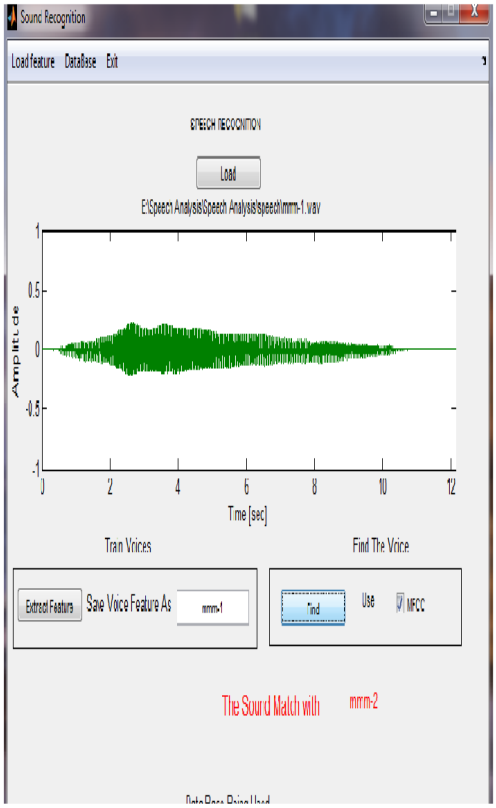

Fig 8: comparing test signal

6. Future Scope

1. Implementation of hardware can be done by using the DSP processors in real-time applications.

2. It can be used in hotels for giving the menu e.g. Samosa, Dosa, Tea etc.

7.CONCLUSION

Thus we are able to recognize multiple words and is converted into text. This system is suitable with an environment with less ambient noise. The system provides good performance

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

with respect to other systems. It can be concluded that GMM provides more accuracy.

8.REFERENCE

[1] Douglas A. Reynolds, and Richard C.Rose, "Robust TextIndependent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, pp. 74-77, 1995.

[2] Reference Book : Speech and Audio Processing(By Dr. Shaila D. Apte, WILEY INDIA Edition).

[3] MananVyas, "A Gaussian Mixture Model Based Speech Recognition System Using MATLAB", Signal and Image Processing: An International Journal(SIPU) Vol.4,No.4, August 2013.

[4] Leija, L.Santiago, S.Alvarado, C., "A System of text reading and translation to voice for blind persons," Engineering in Medicine and Biology Society, 1996.

[5] KimmuParssinen, "Multilingual Text to Speech system for Mobile Device" University of technology April 2007.

6] Ainsworth, W.,"A System for converting English text into speech, "Audio and Electroacoustics, IEEE Transactions, vol.21, no.3, pp. 288-290, Jun 1973.

[7] NiponChinathimatmongkhon, AtiwongSuchato,ProadpranPunyabukkana, "Implementing Thai text-to-speech synthesis for hand held devices", Proc Of ECTI-CON 2008.