International Journal of Engineering Innovation & Research
Volume 2, Issue 2, ISSN: 2277 – 5668

Automatic Speech Recognition System for Real Time Applications

S. Preethi
Department of ECE
Easwari Engineering College
skppreethi@gmail.com

B. Arivu Selvam
Department of ECE
Easwari Engineering College
arivuselvamb@yahoo.in

***Abstract* - Speech is human's most efficient communication mode. Beyond its efficiency, humans are comfortable and familiar with speech. Other modalities require more concentration, restrict movement and cause body strain due to unnatural positions. This need brings the development of a speech to text conversion. In spoken language, syllables are often considered as the phonological "building blocks" of words. Depending on the language and the sounds used, a phoneme may be written consistently with one letter; however, there are many exceptions to this rule. The range of the possible applications is wide and includes: voice-controlled appliances, fully featured speech-to-text conversion, automation of operator-assisted services, and voice recognition aids for the handicapped. This project implements a removal of additive noise and conversion of speech to text form. Spectral substraction is used to remove the noise present in speech. Next is the segmentation process done with the help of group delay algorithm. Recognition plays a major role in speech to text conversion. Letter recognition is achieved with simple Gaussian Mixture Model (GMM). Word recognition is a challenging scenario for researchers and is extracted by HMM with more accuracy main application of the framework is hand free data entry. Mobile and medical environment also use this speech to text conversion.**

***Keywords* – Automatic Speech Recognition, GMM, Real Time ASR System, HMM, System Architecture.**

I. INTRODUCTION

Automatic speech recognition (ASR) systems consist of two major parts: the speech processing and the recognition. The speech recognition process involves two phases: Training Phase and Testing Phase. In the training phase, each speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. Then the different features for the training samples are extracted. The extracted features are fused together and stored collectively. In the testing phase, the same features are extracted from the speech signal and some distance measures are calculated between the stored data and the test input. The final output is a collective analysis of the speech to text conversion using MATLAB.

Speech is a non-stationary signal and processing and is usually conducted over short-time frames where stationarity can be assumed. For example, linear prediction approaches model the speech by a set of coefficients which represent the filter coefficients of an all-zero model of the human speech production system. Cepstrum analysis, on the other hand, uses homomorphic transformations to extract speech features and apply well known spectrum linear operations. "Psychoacoustic" properties were taken into account along with cepstrum analysis to derive the Mel Frequency Cepstral Coefficients (MFCC), which are widely used in speech recognition nowadays.

II. GENERAL DESCRIPTION

Timing

Timing is the most important consideration for a real-time ASR system implementation. Speech is captured and divided into overlapping frames prior to any processing. Then, short-term processing techniques are applied in each frame for end-point detection and feature extraction. The real-time system must have the computational power to complete these steps before the next frame is captured. Thus, compromises must be made between the maximum duration of a speech frame, invoking stationary considerations and the time required for the system to complete frame processing.

Furthermore, the time needed to obtain recognition results may introduce delays between spoken words. Large delays demand a speaker to be silent for large periods of time, making the system inefficient. It is available immediately after the current word's last frame processing is finished and before the first frame of the next word is captured.

In preprocessing involves, Noise removal using spectral subtraction. In spectral subtraction is used to noise reduction. Segmentation using Group Delay Algorithm, which find the short term energy of the signal. Normalization is to regenerate the original signal from the noise removed segments. Next Feature Extraction process used Audio feature MFCC, which cover most energy of sounds that are generated by humans. Frame blocking and windowing method is used to the MFCC. Modeling audio using the GMM (Gaussian Mixture Model) and HMM (Hidden Morkov Model). In GMM are used as a parametric model of the probability distribution of continuous measurements of features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. HMM, this is used in continuous speech recognition system using the forward algorithm and viterbi algorithm. Audio Feature must be compared with the speech database and finally the output word in text format.

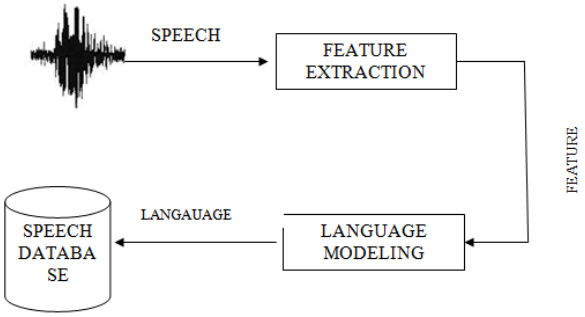

Fig 1: Training Phase

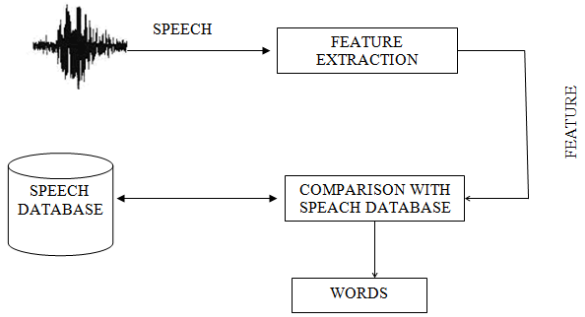

Fig 2: Testing Phase

III. GMM (GAUSSIAN MIXTURE MODEL)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(x|\_) = M\ Xi=1\ wi\ g(x|\mu i,\_i), \quad (1)$$

where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), wi, i = 1, . . . ,M, are the mixture weights, and g(x|µi,_i), i = 1, . . . ,M, are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$g(x|\mu i,\_i) = 1(2\_)D/2|\_i|1/2\ exp\_{-}12(x - \mu i)'\ \_{-}1i\ (x - \mu i)\_, \quad (2)$$

with mean vector µi and covariance matrix _i. The mixture weights satisfy the constraint that PMi=1 wi = 1.The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities.

$$= \{wi, \mu i, \_i\}\ i = 1, \ldots, M. \quad (3)$$

There are several variants on the GMM shown in Equation (3). The covariance matrices, _i, can be full rank or constrained to be diagonal. Additionally, parameters can be shared, or tied, among the Gaussian components, such as having a common covariance matrix for all components, The choice of model configuration is often determined by the amount of data available for estimating the GMM parameters and how the GMM is used in a particular biometric application. It is also important to note that because the component Gaussian are acting together to model the overall feature density, full covariance matrices are not necessary even if the features are not statistically independent. The linear combination of diagonal covariance basis Gaussians is capable of modeling the correlations between feature vector elements. The effect of using a set of M full covariance matrix Gaussians can be equally obtained by using a larger set of diagonal covariance Gaussians.

A GMM acts as a hybrid between these two models by using a discrete set of Gaussian functions, each with their own mean and covariance matrix, to allow a better modelling capability. The GMM not only provides a smooth overall distribution fit, its components also clearly detail the multi-modal nature of the density.

IV. HMM

In a typical hidden Markov model (HMM)-based ASR system three main stages are involved. The first stage is *feature extraction*. Its main purpose is to convert a speech signal into a sequence of acoustic feature vectors, **o**T1 = *{***o**1, **o**2, . . , **o***T }*, where *T* is the number of feature vectors in the sequence. The entire speech signal is segmented into a sequence of shorter speech signals known as frames. The time duration of each frame is typically 25 ms with 15 ms of overlapping between two consecutive frames. Each frame is characterized by an acoustic feature vector consisting of *D* coefficients. One of the widely used acoustic features is called mel frequency cepstral coefficient (MFCC). Feature extraction continues until the end of the speech signal is reached. The next stage is the calculation of the *emission probability* which is the likelihood of observing an acoustic feature vector. The emission probability densities are often modeled by Gaussian mixture models (GMMs). The last stage is *Viterbi search* which involves searching for the most probable word transcription based on the emission probabilities and the search space.

V. SYSTEM ARCHITECTURE

System architecture divided into three phases Noise removal, Training Speech recognition and Testing Speech recognition. After removal of noise the speech signal undergo the following steps segmentation, windowing and FFT, Overlapping and IFFT. From the above steps we applied feature extraction technique for training phase. By using this features generate the GMM model using GMM classifier. This model is stored in database. In Testing phase do the same steps in training upto create GMM model. Then compare the Model with database model and apply the distance calculation method. The recognition is done by means of find probability of the given words with models. From this extract the words consist in each frame of speech.

Data Collection

Data collection involves collection of speech samples in different environments in the presence of different types of noise. Generally, here the samples collected from own speech words by using the wav format.

Preprocessing

There is a need for spectrally flatten the signal. The preemphasizer, often represented by a first order high pass FIR filter is used to emphasize the higher frequency components. In this preprocessing involves three steps: Noise removal, segmentation, Normalization.

Noise Removal

In Noise removal, Spectral Subtraction is one of the popular noise reduction techniques. The basic idea of the spectral subtraction method is to subtract the magnitude spectrum of noise from that of the noisy speech. The noise is assumed to be uncorrelated and additive to the speech signal. An estimate of the noise signal is measured during silence or non-speech activity in the signal.

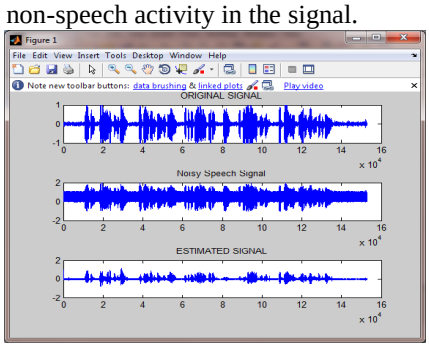

Fig.3. Output of Noise Removal

Segmentation

In segmentation, read the input signal as wav format and find the short term energy of the signal. Invert the value of x of spectrum (x=-x) about y axis and invert the y of spectrum (y=-y) about x axis then identified the x value which is greater than zero and lesser than local maximum is the boundary of each segment. Calculate the Fast Fourier Transform (FFT) for the segmented signal. This is done to convert the signal from time domain to frequency domain. Calculate the initial noise spectrum of the segmented signal by averaging the initial silence segments.

Normalization

In normalization process to regenerate the original signal from the noise removed segments.

AUDIO FEATURE

MEL-Frequency Cepstrum Coefficients (MFCC)

A block diagram of the structure of an MFCC processor is given below. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans.

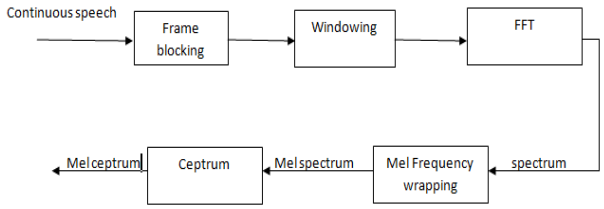

Fig 4: MFCC

FRAME BLOCKING & WINDOWING

Frame blocking

The objective with frame blocking is to divide the signal into a matrix form with an appropriate time length for each frame. Due to the assumption that a signal within a frame of 20 ms is stationary and a sampling rate at 16000Hz will give the result of a frame of 320 samples. In the frame blocking event the use of an overlap of 62.5% will give a factor of separation of 120 samples.

Windowing using Hamming window

After the frame blocking is done a Hamming window is applied to each frame. This window is to reduce the signal discontinuity at the ends of each block. The equation which defines a Hamming window is the following:

$$w(k) = 0.54 - 0.46\cos\left(\frac{2\Pi k}{k-1}\right) \quad Eq\ 4.2$$

FFT AND MFCC

FFT

Use 512 point FFT on each windowed frame in the matrix. To adjust the length of the 20ms frame length, zero padding is used.

Mel spectrum coefficients with filter bank

The fact that the human perception of the frequency content in a speech signal is not linear there is a need for a mapping scale. There are different scales for this purpose. The scale used in this thesis is the Mel scale. This scale is warping a measured frequency of a pitch to a corresponding pitch measured on the Mel scale. The definition of the warping from frequency in Hz to frequency in Mel scale is described in Eq.4.3 and vice versa in Eq.4.4.

$$F_{mel} = 2595.\ log_{10}\left(1+\frac{F_{Hz}}{700}\right) \quad Eq\ 4.3$$

$$F_{Hz} = 700.\ (10^{\,Fmel\,/\,2595} - 1) \quad Eq\ 4.4$$

MEL FILTER BANK

Theoretically it is done according to the following description. The summation is done to calculate the contribution of each filter tap. This will end up in a new matrix with the same number of columns as number of filter taps. The first x_fft frame is multiplied with each of the filter taps and in our case its 20 filter taps. This will result in a 20 sample long vector. Then iterate the same procedure with every other frame and filter taps. The element in *x_mel(1,1)* are obtained by summing the contribution from the first filter tap denoted 1 (MatLab notation. melbank (1:256,:)), then element *x_mel(2,1)* is obtained by summing the contribution from the second filter tap in melbank and so on.

The sample MFCC plot of input

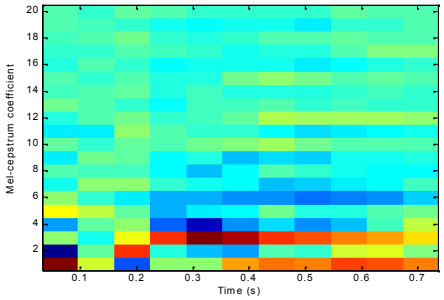

Fig.5. MFCC Plot

VI. CONCLUSION

Thus a speech recognition system with an integral noise removal system has been developed. Different noise removal techniques have been studied, analyzed, implemented and a comparative study is presented which forms the initial part of speech recognition system. The developed speech recognition system is based on syllable segmentation of human speech. The use of segmentation technique improves like phoneme level the accuracy of recognition process greatly and thus enables making of a robust speech recognition system. And to implement the speech(voice) to text conversion using MATLAB.

FUTURE WORK

Historically, there has been an ongoing search for features that are resistant to speaker, noise, and channel variations. In spite of the relative success of MFCCs as basic features for recognition, there is a general belief that there must be more that can be done. One challenge is to develop ways in which our knowledge of the speech signal, and of speech production and perception, can be incorporated more effectively into recognition methods. For example, the fact that speakers have different vocal tract lengths could be used to develop more compact models for improved emotion recognition. Artificial neural networks, which are capable of computing arbitrary nonlinear functions, have been explored extensively for purposes of speech recognition, usually as an adjunct or substitute for Hidden Markov models(HMM). However, it is possible that neural networks may be best utilized for the computation of new feature vectors that would rival today's best features. Our future work includes some further refinements of the speech recognition algorithm, exploration of design space to look for improvements of the hardware–software co-processing system, and encapsulation of the speech recognizer into an intellectual property block that can be easily used in other applications, in one of the application hands free data entry.

REFERENCES

[1] "Hardware–Software Codesign of Automatic Speech Recognition System for EmbeddedmReal-Time Applications", Octavian Cheng, *Member, IEEE*, Waleed Abdulla, *Member, IEEE*, and Zoran Salcic, *Senior Member, IEEE*.

[2] A. Green and K. Eklundh, "Designing for learnability in human robot communication," *IEEE Trans. Ind. Electron.*, vol. 50, no. 4, pp. 644–650, Aug. 2003.

[3] N. Hataoka, Y. Obuchi, T. Mitamura, and E. Nyberg, "Robust speech dialog interface for car telematics service," in *Proc. IEEE Consumer Commun. Netw. Conf.*, 2004, pp. 331–335.

[4] D. Huggins-Daines, M. Kumar, A. Chan, A. Black, M. Ravishankar, and A. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. ICASSP*, 2006, pp. 185–188.

[5] Ibrahim Patel, Dr. Y. Srinivas Rao, " Speech Recognition Using HMM with MFCC – An Analysis using Frequency Spectral Decomposition Technique", *An International Journal(SIPIJ)* Vol.1, No.2, December 2010.

[6] Bjorn Schuller1, Bogdan Vlasenko2, Dejan Arsic1, Gerhard Rigoll, Andreas Wendemuth, "Combininng speech recognition annd acoustic word emotionn models for robust text-independent emotion recognition", 978-1-4244-2571-6/08/$25.00 ©2008 IEEE

[7] Mikael Nilsson, Marcus Ejnarsson, Speech Recognition using Hidden Markov Model, *MEE-01-27*

[8] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signals, *IEEE Press*, New York, 2000.

[9] R. S. Kurcan, "Isolated Word Recognition from In-ear Microphone Data Using Hidden Markov Models," *MSEE Thesis*, Naval Postgraduate School, Monterey, California, March 2006.

[10] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Second Edition, *Academic Press*, San Diego, California, 2003.

AUTHOR'S PROFILE

S. Preethi

was born in Perambalur, Tamil Nadu during 19.05.1990. She had obtained her Bachelor degree in Engineering during 2007 from Roever Engineering College, Perambalur under Anna University. Now pursuing her Masters degree in Embedded Systems and Technologies during 2011 from Easwari Engineering College, Chennai under Anna University, Chennai. Her area of interest is Speech Processing and Embedded System.

She had attended some Workshops AU & AICTE sponsored training Programs in latest trend in engineering and research. She has attended one International Conference.

B. Arivu Selvam

was born in Nagapattinam, Tamil Nadu during 17.05.1981. He had obtained his Bachelor degree in Engineering during 2003 from Barathidasan University and Masters degree in VLSI Design during 2007 from VIT, Vellur. His area of interest is VLSI Design and Embedded System.

He has seven years of teaching experiences in various Engineering Colleges. Currently he is working as Assistant Professor (Senior Grade) in Electronics and Communication Engineering Department at Easwari Engineering college, Chennai, Tamilnadu. He had attended many number of Seminars, Workshops, Faculty Development Programs and AU & AICTE sponsored training Programs in latest trend in engineering and research. He has 6 numbers of intrastate National level Conferences, two International Conference and one International Journal Publication. Motivating Students with Innovative Ideas in Technical field of Engineering is his axiom.