A Novel Speech to Text Converter System for Mobile Applications

R.Sandanalakshmi
Asst.Professor, Dept of ECE
Pondicherry Engineering College

V.Martina Monfort
Dept of ECE
Pondicherry Engineering College
Puducherry, India

G.Nandhini
Dept of ECE
Pondicherry Engineering College

ABSTRACT

In this paper an efficient implementation of speech to text converter for mobile application is presented. The prime motive of this work is to formulate a system which would give optimum performance in terms of complexity, accuracy, delay and memory requirements for mobile environment. The speech to text converter consists of two stages namely front-end analysis and pattern recognition. The proposed method uses effective methods for voice activity detection in preprocessing, feature extraction and recognizer. The energy of high frequency part is separately considered as zero crossing rate to differentiate noise from speech. RASTAPLP feature extraction method is used in which RASTA filter suppresses the spectral components that change more slowly or quickly than the typical range of change of speech thus avoiding unnecessary information in the extracted features. In the proposed system Generalized Regression Neural Network is used as recognizer in which syllable level recognition is used that reduces memory requirement and complexity for mobile application. Thus a small database containing all possible syllable pronunciation of the user is sufficient to give recognition accuracy closer to 100%. Reduction in 50% with respect to delay and memory requirement is proved in the proposed system. Thus the proposed technique entertains realization of real time speaker dependant applications like mobile phones, PDAs etc.

Keywords

Preprocessing, Voice activity detection, RASTAPLP, Neural network, syllable based recognition

1. INTRODUCTION

Speech recognition technology is the budding technology where new services can be created to open a door for human-machine interaction. It has been an attractive research field for the last decades which still has a number of unsolved problems. The acquired speech needs to be preprocessed before extracting information from it for recognition. The most important step in preprocessing is speech detection in the acquired signal. The effectiveness of speech recognizer is crucially dependent on its performance. This is an algorithm to detect silence parts of a speech signal and remove it as it does not provide any information. Speech pause detection algorithm [1] is used which detects speech pause minima by adaptively tracking minima in a noisy signal power envelope. An ideal voice activity detector needs to be independent from application area and noise condition. But at the same time, the VAD algorithm [2] should be of low complex to facilitate real time application. Therefore simplicity and robustness against noise are two essential characteristics of a practicable voice activity detector which is incorporated in the proposed system. VAD should be most sensitive as the unwanted part of the speech also has some energy and appears to be speech. This demands to use a VAD that also calculates energy of high frequency part separately as ZCR to differentiate noise from speech even in noisy environments [3][4].

The preprocessed signal needs to be represented in a compact manner for further process. Speech carries much information of which linguistic information is required for speech recognition task. Conventional short term spectrum based feature extraction technique like MFCC, LPC etc blindly represents most information in the signal. RASTAPLP [5] overcomes this limitation. The linear microphone characteristics appear as a convolutional component in the signal and it form as additive component in the logarithm spectrum of speech. However this characteristic of microphone is slow varying in time. So feature extraction technique should be made invariant to slow changes in the logarithmic spectrum of speech. Similarly the rate of change of nonlinguistic components in speech often lies outside the typical rate of change of the vocal tract shape. So RASTA filter used in this technique avoids unnecessary information in the extracted features.

In recent years Neural Network is beginning to gain importance. This paved a path for efficient deployment of recognizer. The usability of neural network for speech recognizer proved to be a suitable technology [6] and it outperforms traditional recognizers like Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ) etc.

The promising results were obtained using Radial Basis Neural Network as recognizer [7]. It also is found that RBF trains and tests faster than Multilayer Perceptron (MLP) Neural Networks. Moreover RBF gives more accurate result than MLP [8].

In the proposed system GRNN [9] is used as recognizer. A Generalized Regression Neural Network is often used for function approximation. It is similar to Radial Basis Neural Network but has a special linear layer in addition to RBF network. It is found that the performance of GRNN is superior to the other classifiers namely Linear and MLP Neural Networks [10].

The article is organized as follows: The initial steps to be performed on the acquired speech are described in speech preprocessing section. The feature extraction part deals about RASTAPLP method which is used to represent the speech. The speech recognizer section elaborates about the

implementation of a well-organized recognizer for speech to text conversion application. The analysis of GRNN recognizer and its architecture is explained under proposed NN recognizer, GRNN analysis and architecture. The suitability of GRNN for speech recognizer is also elucidated. The results obtained for the proposed technique is discussed in evaluation of proposed method segment. The solution for reducing large memory requirement in speech to text converter is provided using syllable based recognition which is explained in syllable level recognition [11].

2.SPEECH PREPROCESSING

The speech signal is recorded as single channel (mono) PCM waveform with a sampling rate of 8 KHz and 16 bit sample resolution using 'Sony Sound Forge 8.0' software. Preprocessing is the primary stage in any speech recognition system. Certain operations like pre emphasis, Framing, windowing, Voice Activity Detection are required to be performed on the acquired speech signal before obtaining the features necessary to represent the speech signal. They are described as follows:

2.1 Pre emphasis:

Due to physiology of human speech production system, speech signal is attenuated approximately 20 db per decade. Thus high frequency signals have less amplitude. This gives rise to a negative spectral slope and is compensated by appropriate pre-emphasis filter.
The z-transform of the filter is given by:

$$H(z) = 1 - az^{-1}, \quad 0.9 \leq a \leq 1.0 \qquad (1)$$

where 'a' is the filter coefficient and is chosen as 0.9375
The output signal after pre-emphasis is given by:

$$y(n) = s(n) - a*s(n-1) \qquad (2)$$

where y(n) - output signal
s(n) - input speech signal
This pre-emphasis filter boosts the signal spectrum approximately 20 dB per decade.

2.2 Framing:

Speech is a highly non-stationary signal. Hence speech analysis must be carried out on short segments across which the speech signal is assumed to be stationary. For this the speech signal is divided into frames of small duration typically 20 to 40ms with overlap of 10 to 15ms for short-time spectral analysis.
Here we use frame size = 200 samples (25ms), step size = 80 samples (10ms)
Hence overlap = 120 samples (15ms)

2.3 Windowing:

To avoid problems due to truncation of signal, a weighting window with appropriate spectral properties must be applied. Windowing minimizes the discontinuities by tapering the signals to quite small values (nearly zeros) at the edges of a frame. Typically Hamming and Hanning windows are used because they have very less spectral leakage..In this system hamming window is used which is given as

$$w(n) = 0.54 - 0.46\cos(2\pi n), \quad 0 \leq n \leq N$$
$$= 0, \quad \text{otherwise} \qquad (3)$$

The output speech signal can be described as

$$x(n) = w(n).s(n), \quad 0 \leq n \leq N \qquad (4)$$

Each frame s(n) is multiplied by hamming window w(n) of 25ms duration.

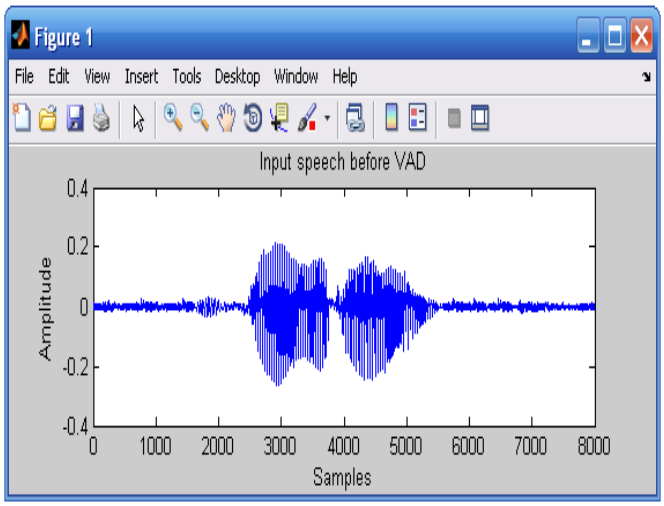
2.4 Voice Activity Detection:

Voice activity detection is a technique to detect the un-silenced part of the incoming speech signal.

The steps involved in voice activity detection algorithm are as follows:[2]
1. Normalize the speech signal sample values.
2. Set frame size =150ms and frame increment size = 40ms.
3. Assign the highest sample value as maximum amplitude threshold (amp_1) and one third of that value as the minimum amplitude threshold (amp_2).
4. Set maximum silence duration as 30ms (assuming silence between two words to be greater than this value). Maintain a silence count and increment it on encountering silence.
5. Compute the zero crossing rate for each frame. This helps to differentiate speech and noise part. The product and difference of nth sample and (n+1)th sample is calculated. If the product is negative and the difference is nearer to zero then that frame consists of unwanted small disturbances. On receiving such a frame, silence count is incremented.
6. For i =1 to number of frames (n)

Compute energy for each frame.Maintain a count value to track the number of un-silenced frames traversed.For the first instance of energy>amp_1, mark the start of the frame as x (x = n−count).Increment count value when one of the three conditions, energy>amp_2 or zcr>5 or energy>amp_1, occurs.Otherwise increment silence count. When silence count > maximum silence duration then end of speech is detected and the iteration is stopped abruptly.
End
7. Extract speech alone by taking apart the frames from x upto (x+count) separately. The snapshot of the VAD output is shown below

Figure 1. Sample speech before VAD



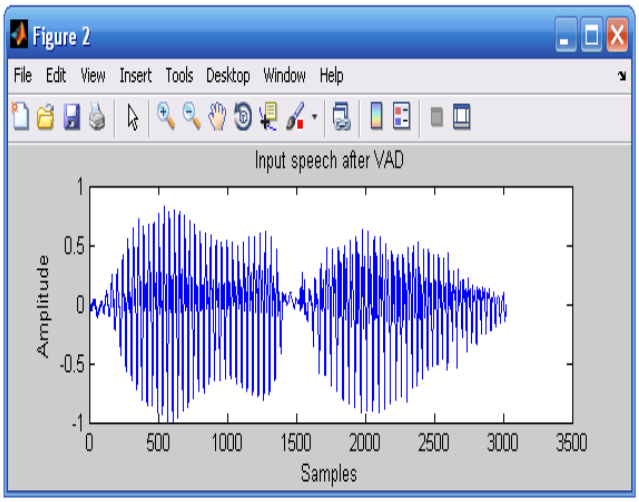


Figure 2. Input speech waveform after VAD algorithm

Number of samples in the acquired speech = 8000
Number of samples after Voice Activity Detection = 3021

3.FEATURE EXTRACTION

Feature extraction is the process of obtaining distinguishing features from the input signal. It is about reducing the dimensionality of the input-vector while still maintaining the uniqueness of the signal.Several techniques are available for feature extraction like MFCC (Mel Frequency Cepstral Coefficients), LPC (Linear Predictive Coding), PLP (Perceptual Linear Prediction) and RASTAPLP (RelAtive SpecTrA PLP)[13].RASTAPLP outperforms other methods because it derives most distinct linguistic information alone from speech by converting the speech spectrum into human auditory-like spectrum considering spectral resolution, sensitiveness and perceived loudness of human hearing. This technique also utilizes LPC analysis at final stage to characterize features in concise form. Log-RASTA filtering[12] performed in RASTAPLP takes care of mismatch between training and testing conditions and removes phonetically unimportant information through bandpass filtering. Thus it provides compact features that are insensitive to channel mismatch and additive background noise. It also gives promising recognition accuracy when compared to other methods.

3.1 RASTAPLP:

RASTAPLP is a feature extraction technique in which features are obtained from standard all pole modeling or linear predictive analysis of short term speech spectrum. The speech spectrums are modified by a set of transformations that are based on models of human auditory system. The step by step procedure followed for obtaining features from incoming speech signal is given below. [5]

1. Power spectrum is calculated for each frame.
2. Power spectrum is then converted into auditory spectrum. The spectral resolution of human hearing is roughly linear up to 800 or 1000 Hz, but it decreases with increasing frequency above this linear range. PLP incorporates critical-band spectral resolution into its spectrum estimate by remapping the frequency axis to the Bark scale and integrating the energy in the critical bands to produce a critical-band spectrum approximation.

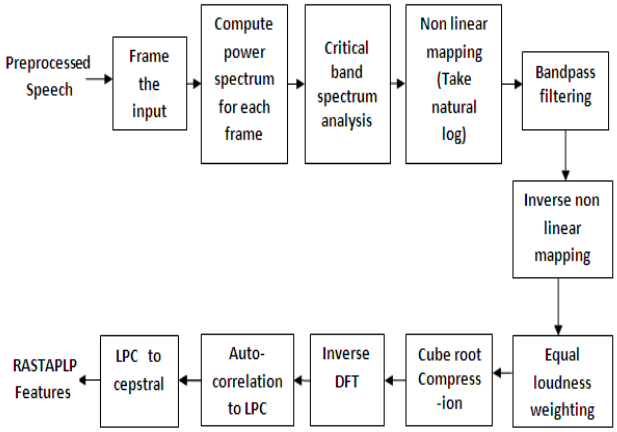


Figure 3. RASTAPLP block diagram

3. Log RASTA filtering is performed on the critical band spectrum
   a. Natural logarithm of each spectral component is performed. This is done to convert multiplicative distortion in frequency domain into an additive distortion, which can be filtered.
   b. Then band pass filter is applied to each spectral component in the critical-band spectrum. The system function of the filter [4],[12] is :

$$H(z) = z^4 * \frac{0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4}}{1 - 0.94z^{-1}} \quad (5)$$

RASTA high pass filtering removes low frequency components introduced by microphone, and RASTA low pass filtering removes speaker generated high frequency components that are not phonetically important. Thus log RASTA addresses input channel mismatch and speaker variation.

4. The critical-band spectrum is again brought back to normal domain by exponential operation.
5. Equal-loudness hearing curve: At conversational speech levels, human hearing is more sensitive to the middle frequency range of the audible spectrum. PLP incorporates the effect of this phenomenon by multiplying the critical-band spectrum by an equal

loudness curve that suppresses both the low and high frequency regions relative to the midrange from 400 to 1200 Hz.

6. Intensity-loudness power law of hearing: There is a nonlinear relationship between the intensity of sound and the perceived loudness. PLP approximates the power law of hearing by using a cube-root amplitude compression of the loudness equalized critical-band spectrum estimate.

The spectrum derived by combining all three of the above described transformations, has less detail and a smaller dynamic range than the original spectrum. So it can be modeled well by a low-order all-pole model, which is effective in suppressing speaker specific details of the spectrum. This lower analysis order results in better estimates of recognition parameters for a given amount of training data.

7. The next step is the LPC analysis. The LPC technique described above is used.

8. The LPC coefficients are then converted into cepstral domain. This gives the RASTAPLP coefficients.

The PLP order is smaller than is typically needed by LPC-based speech recognition systems, and this lower analysis order results in improved estimates of recognition parameters for a given amount of training data. The RASTA-PLP technique is found to be advantageous over LPC.

4. SPEECH RECOGNIZER

The role of the recognizer is to assign the feature vector provided by the feature extractor to a category. It aims at measuring the similarity between an input speech and a reference pattern or model which is obtained during training. Conventionally used speech recognizers are based on Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) approaches. DTW works well for small vocabulary but does not give promising results as vocabulary size increases. While, HMM makes modeling assumptions like Gaussian and Markovian assumptions which limits its performance. Moreover it requires large training set to learn parameters for recognition.

The limitations of the conventional recognizers have been overcome with the advent of neural networks which has opened up a new era in forming speech recognizers. Thus, in this project neural network has been found suitable for developing the recognizer.

4.1 Proposed NN Recognizer:

Speech is a randomly varying nonlinear process whose functional form can be approximated by regression. The goal of speech regression analysis is to model a recognizer by providing a finite set of observations of speech and its associated target values such that it gives most probable output value for an unknown speech input [7].

In order to approximate a function, it is necessary to assume some functional form with unknown parameters. The values of these parameters are chosen to make the best fit of the observed data. But the approach used here is to express the function as probability density function which is empirically determined using non parametric estimators. The resulting regression equation can be implemented in a parallel neural network like structure. Since the parameters of the structure are determined directly from examples the structure learns and begins to generalize immediately.

General Regression Analysis:

Let f(x**,** y) represents the joint continuous probability density function of a vector random variable x and a scalar random variable y. Let X and Y be a particular measured value of the random variable x and y. y is the variable on which regression is performed to find its most probable value for a given X based on observations of X and its associated Y which is supplied during training. In speech regression process the dependent variable Y is the system output and the independent variable X is the speech input.

The conditional mean of y given X (also called the regression of y on X) is given by [9]

$$E\left(\frac{Y}{X}\right) = \frac{\int_{-\infty}^{\infty} y.f(X,y)}{\int_{-\infty}^{\infty} f(X,y)} \qquad (6)$$

When the density f(x, *y)* is not known, it must be estimated from a sample of observations of x and y. For a nonparametric estimate of f(x, y) a class of consistent estimators proposed by Parzen is utilized.

The probability estimator f(X, Y) is based upon sample values X_i and Y_i of the random variables x and y, where n is the number of sample observations and p is the dimension of the vector variable x**:[9]**

$$\hat{f}(X,Y) = \frac{1}{(2\pi)^{(p+1)/2}\,\sigma^{(p+1)}}\,\frac{1}{n}\sum_{i=1}^{n} \exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right] \exp\left[-\frac{(Y-Y_i)^2}{2\sigma^2}\right] \qquad (7)$$

Substituting the joint probability estimate $\hat{f}(X,Y)$ in (4.9) into the conditional mean (4.8) gives the desired conditional mean estimate of Y given X designated as $\hat{Y}$ [9].

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{D_i^2}{2\sigma^2}\right)} \qquad (8)$$

where

$$D_i^2 = (X-X_i)^T(X-X_i) \qquad (9)$$

The estimate $\hat{Y}(X)$ can be visualized as a weighted average of all of the observed values Y_i where each observed value is weighted exponentially according to its Euclidean distance from X**.** When the smoothing parameter σ is made large, the estimated density is forced to be smooth and becomes a

Gaussian. On the other hand, a smaller value of σ allows the estimated density to assume non-Gaussian shapes.

4.2 Neural Network Architecture:

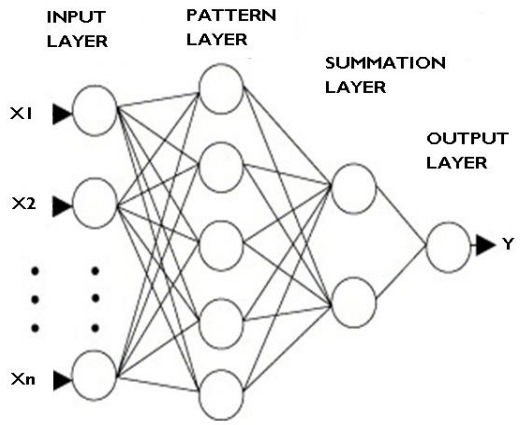


Figure 4. Regression neural network architecture

The regression equation (8) can be expressed as

$$y_j = \frac{\sum_{i=1}^{n} w_{ij} h_i}{\sum_{i=1}^{n} h_i} \quad (10)$$

$$h_i = \exp\left(-\frac{D_i^2}{2\sigma^2}\right) \quad (11)$$

where w_{ij} is the target output corresponding to input training vector x_i and j^{th} output.
The equations (4.12) and (4.13) can be implemented as the above neural network structure.

1. The input layer has neurons which represent the input patterns. It passes the input vector to each of the neuron in the pattern layer.
2. The pattern layer has one neuron for each pattern. During training all variations of input vectors and its associated desired output or target (w_{ij}) will be provided to the network. The neuron in the pattern layer compute C_i for the input provided to it.

During testing the neuron will compute h_i using C_i and X_i as shown below:[8]

$$D_i^2 = \| C_i - X_i \| \quad (12)$$

$$h_i = \exp\left(-\frac{D_i^2}{2\sigma^2}\right) \quad (13)$$

where σ is the spread (width) of the kernel function.

Spread denotes the distance an input vector must be from a neuron's C_i in order to be correctly classified. The choice of spread value for the network has to be done very carefully. The value of spread chosen for the implemented network is 0.5

3. The summation layer has two units N and D where N computes $\sum_{i=1}^{n} w_{ij} h_i$ and D calculates $\sum_{i=1}^{n} h_i$.
4. The output unit divides N by D which gives the prediction result.

The above implementation is illustrated in the figure shown below:

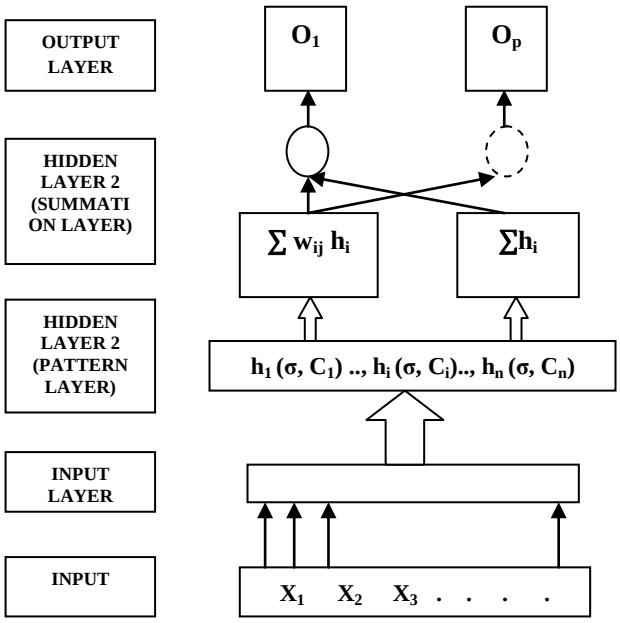


Figure 5. GRNN recognizer

The above neural network is known as General Regression Neural Network (GRNN). It is one of the radial basis networks. The hidden layer neurons of this network are referred as radial basis neurons as they employ radial basis kernel function. The proposed network does superior function approximation and thus is well suited for speech recognition.

Evaluation of Proposed method:

To evaluate the performance and efficiency of the proposed speech to text converter recognition of digits zero to nine are performed .The speech samples of 'zero' to 'nine' are collected from 20 speakers consisting of 5 male and 15 female speakers. Each of the words zero to nine has 194 samples leading to 1940 words. The test set was formed with 200 words with 100 words each from trained and untrained speakers.The following results are obtained:

Table 1. Performance table for digit recognition using LPC coefficients

Speaker Details	Testing Set	Accuracy*
10 trained speakers	10 ×10 =100 words	60%
10 untrained speakers	10 ×10 =100 words	50%

Table 2. Performance table for digit recognition using RASTAPLP coefficients

Speaker Details	Testing Set	Accuracy*
10 trained speakers	10 ×10 =100 words	85%
10 untrained speakers	10 ×10 =100 words	70%

* Accuracy = Number of words correctly recognized / Total number of words

Network training and testing time requirements are as follows

Network training time excluding feature extraction : 2.67 sec
Memory requirement of network : 4.5 MB
Time taken for feature extraction of a single word : 0.87 sec
Time taken to recognize a single word by the network : 0.86 sec
Overall time taken to recognize a single word : 1.73 sec

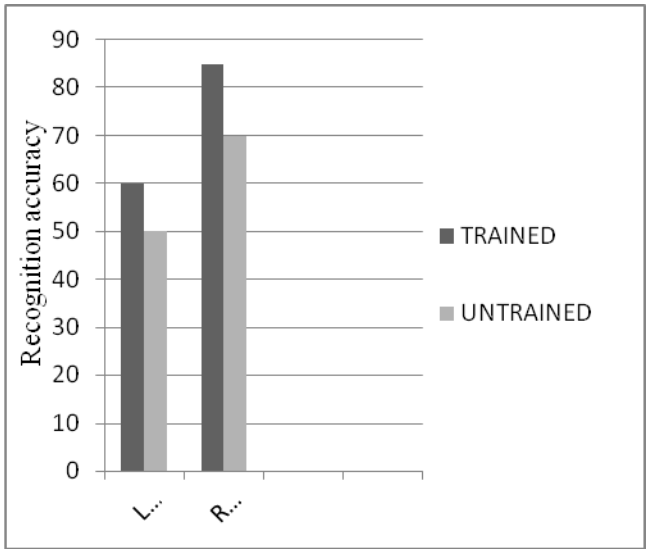


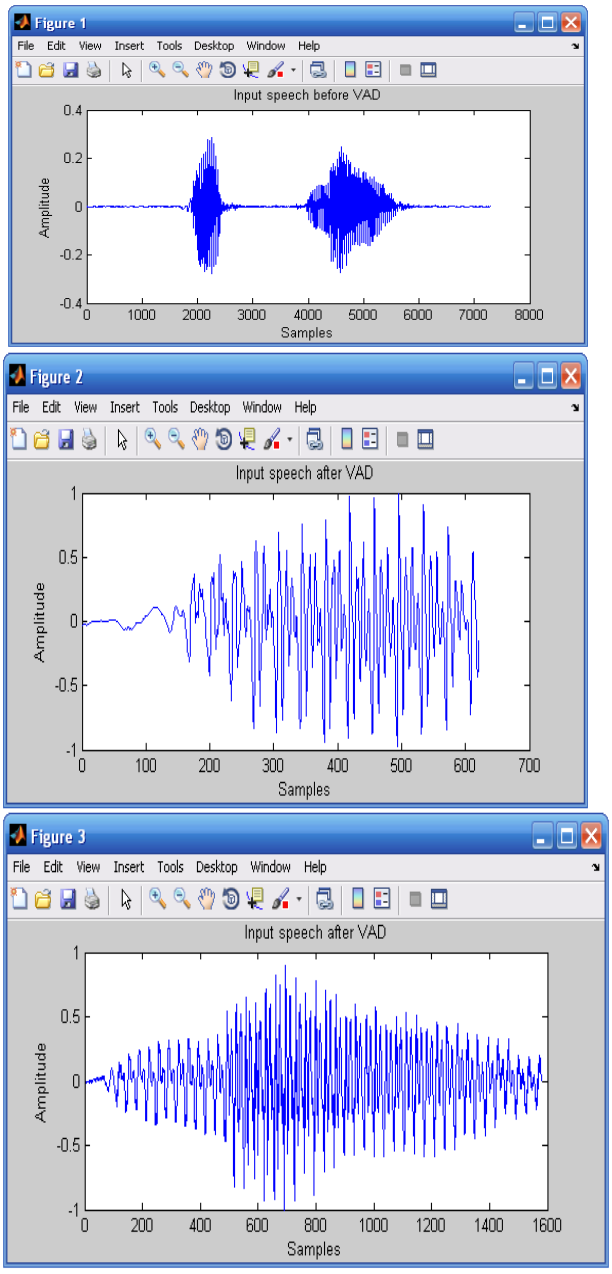
Figure 6. Comparison chart of digit recognizer performance using LPC and RASTAPLP

Syllable level recognition:

Memory requirement plays a vital role for practical implementation of speech to text converter in mobile environment. It is not feasible to train a network for every word of a language as it requires huge memory. This can be overcome by employing syllable based word recognition as syllables are limited in a language than words. The technique is tested using 9 syllables namely 'be', 'com', 'in', 'out', 'ply', 'pose', 'press', 'side' and 'sup'. The list of 20 words that can be formed using 9 syllables are : be, become, beside, come, compose, comply, compress, in, income, inside, out, outcome, outside, ply, pose, press, side, suppose, supply, suppress. Thus instead of separately training 20 words it is sufficient to train 9 words which tremendously reduces memory requirement.

From the incoming speech, syllables are separated using VAD algorithm which is depicted in fig. 1. The acquired speech waveform for the word 'supply' is shown in Fig 1
2. The separated waveforms of 'sup' and 'ply' after passing through VAD algorithm.

Figure.6 The output of VAD for the sample sup & ply



To test accuracy of syllable based implementation of word recognition speech samples of 9 syllables are collected from 2 female speakers .30 samples were taken for each word from each speaker. The test set was formed with 4 speakers with 2 trained and 2 untrained voices.

The network is trained with 540 samples taken from 2 female speakers.

Table 3. Performance table for word recognition

Speaker Details	Testing Set	Accuracy*
2 trained speakers	10 ×10 =100 words	80%
2 untrained speakers	10 ×10 =100 words	70%

* Accuracy = Number of words correctly recognized / Total number of words

Memory requirement of network : 450 KB
Time taken to recognize a single word by the network : 0.876 seconds.

5. CONCLUSION

In the proposed method, speech to text converter for mobile application is developed. The incoming speech signal will consist of not only the spoken utterance but also silence periods and noise. In order to separate the word alone, proper segmentation of speech waveform is required which has been carried out with the help of suitable VAD algorithm. The features for representing speech are obtained using LPC and RASTAPLP technique. In LPC analysis representation of vocal tract configuration is made by assuming an all pole model for speech production system. While in RASTAPLP most distinct linguistic information alone is derived from speech by converting the speech spectrum into human auditory-like spectrum considering spectral resolution, sensitiveness and perceived loudness of human hearing. This technique also utilizes LPC analysis at final stage to characterize features in concise form. Log-RASTA filtering performed in RASTAPLP takes care of mismatch between training and testing conditions and removes phonetically unimportant information through bandpass filtering. Thus it provides compact features that are insensitive to channel mismatch and additive background noise. It also gives promising recognition accuracy when compared to LPC.In this work, GRNN has been proposed for recognition. The use of non-parametric estimator with appropriate smoothing factor (spread) improves the generalization capability of the network. The proposed model has several advantageous characteristics such as fast learning, flexible network size and robustness to speaker variability (ability to recognize the same words pronounced in various manners). GRNN promises to be a successful and powerful alternative to the conventional speech recognizers.
The designed recognition system with all the above salient features has been utilized for developing a digit recognition system and a syllable level word recognition system. From digit recognition, the large memory requirement was found to be a disgusting factor. This has been overcome by envisaging yet another usefulness of VAD algorithm. The recognizer was trained for syllables. These syllables were separately recognized and concatenated to output the complete word as text. The syllable based recognition has given encouraging results. Good recognition accuracy has been achieved in both cases. These implementations illustrate the potential of optimal configurations of key ASR components.

6. REFERENCES:

[1] M.Marzinzik and B.Kollmeir, "Speech Pause Detection For Noise Spectrum Estimation By Tracking Power Envelope Dynamics*", IEEE Transactions On Speech And Audio Processing*, Barcelona, Vol.10, No.2, pp. 109-117, Feb.2002.

[2] M.H.Moattar and M.M.Homayounpour, "A Simple But Efficient Real-Time Voice Activity Detection Algorithm", *17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland, pp. 2549-2553, August 24-28, 2009.

[3] Namgook Cho and Eun-Kyoung Kim, "Enhanced Voice Activity Detection Using Acoustic Event Detection And Classification", *IEEE Transactions On Consumer Electronics*, Vol. 57, No. 1 pp. 196-202, Feb.2011,.

[4] Varela, Indra S.A., Madrid, Spain San-Segundo. R., Hernandez, L.A., "Robust speech detection for noisy environments", *IEEE Aerospace and Electronic Systems Magazine*, vol. 26, Issue. 11, pp. 16 - 23, Nov. 2011.

[5] Hynek Hermansky and Nelson Morgan, "RASTA Processing Of Speech", *IEEE Transactions On Speech and Audio Processing*, Vol.2, No.4, pp. 578-589, October 1994.

[6] Chin Luh Tan and Adznan Jantan, "Digit Recognition Using Neural Networks", *Malaysian Journal Of Computer Science*, Vol.17, No.2, pp. 40-54, Dec. 2004.

[7] R.L.K.Venkateswarlu, R.Vasantha Kumari and G.Vani Jayasri, "Speech Recognition Using Radial Basis Function Neural Network", *3rd International Conference On Electronics Computer Technology (ICECT) 2011*, Vol.3, pp. 441-445, 2011.

[8] Wouter Gevaert, Georgi Tsenov and Valeri Mladenov, "Neural Networks Used For Speech Recognition", *Journal Of Automatic Control*, University of Belgrade,Vol.20, pp.1-7, 2010.

[9] L.K.V.Revada, V.K.Rambatla and K.V.N.Ande, "A Novel Approach To Speech Recognition By Using Generalized Regression Neural Networks", *IJCSI International Journal Of Computer Science Issues*, Vol.8, Issue 2, pp. 484-489, March 2011.

[10] Abderrahmane Amrouche and Jean Michel Rouvaen, "Efficient System For Speech Recognition Using General Regression Neural Network", *World Academy Of Science, Engineering And Technology*, Vol.1, N0.6, pp. 271-277, 2006.

[11] George E. Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition", *IEEE Trans. Audio Speech Language Process*., vol. 20, no. 3, Jan. 2012.

[12] N.Morgan and H.Hermansky, "RASTA extensions, Robustness to additive and convolutional noise," in *Proceeding of workshop speech processing Adverse Environments*, Cannes, France,Nov.1992.

[13] N.Morgan and H.Hermansky,H.Boulard,P.Kohn, and C.Wooters, " Continuous speech recognition using PLP analysis with multilayer perceptrons," *in proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, Toronto ,Canada,1991, pp. 49 -52.