Speech Recognition Based Web Browsing

Vinitkumar Dongre
EXTC, TCET, Kandivali (E)
Mumbai, Maharashtra

Sanjeev Ghosh
EXTC, TCET, Kandivali (E)
Mumbai, Maharashtra

Ashwini Katkar
EXTC, VCET, Vasai (W)
Thane,Maharashtra

Abstract

The technology of voice browsing is rapidly evolving these days. Listening and speaking are the natural modes of communication and information gathering. As a result all are now heading towards a more voice based approach of browsing rather than operating on textual mode. The results of a case study carried out while developing an automatic speech recognition system for web browsing are presented in this paper. A generalized coding is done to make the system compatible for 'n' number of samples without any change in basic coding. Speech data is collected from independent speakers and pre-processed to extract the features needed in this research. Feature extraction is carried out using Mel Frequency Cestrum Coefficient (MFCC) technique. In [1], it was demonstrated that MFCC outperforms than other feature extraction techniques. After the training session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. The centroid based neural network Adaptive Resonance Theory (CNN-ART) approach is used for mapping vectors from a large vector space to a finite number of regions in that space. For comparison purpose, the distance between each test codeword and each codeword in the master codebook is computed. The difference is used to make recognition decision. The prototype can recognize the word as well as sentences by concatenating the words stored in the database to form a sentence. The recognition accuracy of the system is 85% in speaker dependent environment while 70% in speaker independent environment. Also, system provides 70% accuracy for sentence recognition while for isolated word, recognition accuracy is 80%.

1. Introduction

Speech is a natural mode of communication for people. Human beings learn all the relevant skills during early childhood, without instruction, and continue to rely on speech communication throughout lives. It comes so naturally that they don't realize how complex a phenomenon speech is. Yet people are so comfortable with speech that they would also like to interact with computers via speech for applications such as web browsing, language translation rather than having to resort to primitive interfaces such as keyboards and pointing devices. One of the major challenges with existing browsers is to provide a simple navigation framework that demands user-friendly interaction. For people who usually use computer may face with bad healthy syndrome called Carpal-Tunnel Syndrome. Carpal-Tunnel Syndrome is an uncomfortable feeling on body especially on hand and fingers after doing any particular job or activity repeatedly. So, by implementing speech in using computer became more efficient rather than just using keyboard and mouse alone. By combining these tools, accessing activities became more efficient. Browsers with Speech capability provide simple and effective user interaction, which can be rightly called as hands-free browsers. The ultimate goal is to ease the user interactivity with browser while surfing the net. Here the user instead of traversing the web pages by clicking on hyperlinks, he/she reads out the hyperlink and the corresponding page automatically gets loaded.

2. Speech Recognition Module

The General scheme for Speech Recognition is shown in Figure 1. Test and reference patterns (feature vectors) are extracted from speech utterances statistically or dynamically. At the training stage, reference models are generated (or trained) from the reference patterns by various methods. A reference model (or template) is formed by obtaining the statistical parameters from the reference speech data. A test pattern is compared against the reference templates at the feature matching stage. The comparison may be conducted by probability density estimation or by distance measure. After comparison, the test pattern is labelled to a speech model at the decision stage. Different stages are explained as follows.

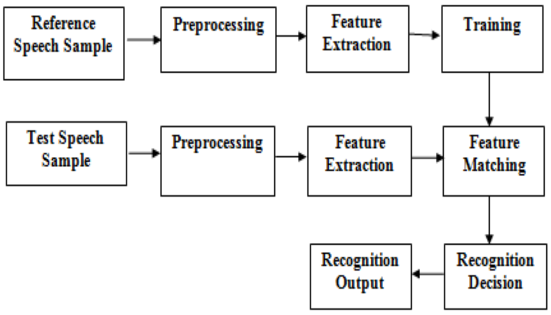

Figure 1: Speech Recognition Module

2.1 Speech Input from the User

The speech signals are recorded in a low noise environment with good quality recording equipment. The signals are sampled at 11kHz.Reasonable results can be achieved in isolated word recognition when the input data is surrounded by silence.

2.2 Pre-processing

Various approaches exist for the creation and administration of large corpora of data to be used for Speech Signal Processing (SSP). Such data include speech files, pitch files, spectrogram files, text files and the like. Once the speech is recorded, then for the further processing it is required to be stored in the computer memory. While the speech is recorded the speech file may be stored in many different formats i.e. .wav, .mp3, .mp4, .mp6, .midi, .au, .voc, .wma etc and many more. Pre-processing is the important step which makes the signal suitable for further processing. Once a speech signal is digitalized in both time and in amplitude, it can be stored and processed by a computer.

2.3Feature Extraction

The goal of feature extraction is to represent speech signal by a finite number of measures of the signal. In non-metric spectral analysis, Mel frequency Cepstral Coefficients (MFCC) is one of the most popular spectral features in ASR [2]. Feature extraction involves the mining of useful amount of information required to describe a large set of data accurately. When captured by a microphone, speech signals are seriously distorted by background noise and reverberation. Fundamentally speech is made up of discrete units. The units can be a word, a syllable or a phoneme. Each stored unit of speech includes details of the characteristics that differentiate it from the others. Apart from the message content, the speech signal also carries variability such as speaker characteristics, emotions and background noise. A method of generating feature signals from speech signals comprises of the following steps:

1. Receive the speech signal which is in analog format and stored as a .wav file.
2. Convert the .wav file to a data file for further processing.
3. Eliminate the background noise using End Point Detection.
4. Subdivide the speech sequence into frames and form the frequency domain representation of the said framed speech sequence.
5. Pass the said frequency domain representation through Mel-filter banks to generate Mel Frequency Cepstral Co-efficient (MFCC).

2.3.1 End point Detection

After accepting the speech input from user, the next step is to process it. An important problem in speech processing is to detect the presence of speech in a background of noise. This problem is often referred to as the end point location problem. Endpoint detection, which aims at distinguishing speech and non-speech segments using signal processing and pattern recognition, is considered as one of the key preprocessing components in automatic speech recognition (ASR) systems. The incorrect determination of endpoints for an utterance results in at least two negative effects [3]:

1. Recognition errors are introduced;
2. Computations increase.

To determine the end point of the speech signal, energy approach is considered, which considers both energy and frequency characteristics of the speech signal. It can be very effective when the speech signal is very weak but has frequency components different than background noise. Therefore, it is a very useful tool to locate the beginning and end point of an utterance.

2.3.2 Generating Mel Frequency Cepstral Coefficients (MFCCs)

Considering the known variation of the ear's critical band-width frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This suggests that a compact representation would be provided by a set of Mel-Frequency Cepstrum Coefficients (MFCC). These coefficients collectively make up an MFC (Mel Frequency Cepstrum). They are derived from a type of Cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are positioned logarithmically (on the Mel scale) which approximates the human auditory system's response more closely than the linearly-spaced frequency bands obtained directly from the FFT or DCT.

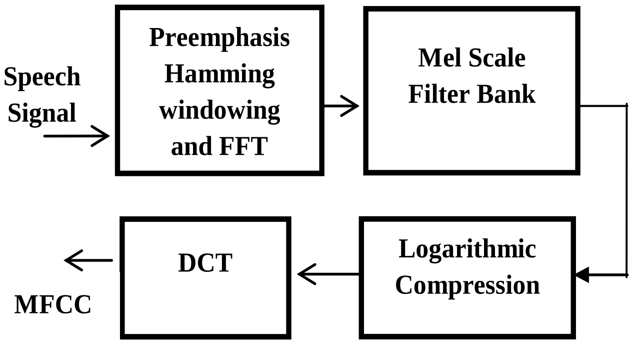

Figure 2 MFCC steps for feature extraction

2.3.2.1 Pre-emphasis, Hamming windowing and FFT

In the first step, speech sequence is passed through a 1st order digital filter for pre-emphasis. After pre-emphasis, the speech sequence is subdivided into frames using Hamming windows. Fast Fourier transform (FFT) is then applied to each speech frame in order to obtain its speech spectrum.

2.3.2.2Mel scale Filter Bank

The Mel frequency filter bank is a series of triangular band pass filters, which mimics the human auditory system. The filter bank is based on a non-linear frequency scale called the Mel scale. The Mel frequency scale is a psychoacoustic measure of pitches judged by human. A 1000Hz tone, with 40dB above the listener's threshold, is defined as having a pitch of 1000 mels. Below 1000Hz, the Mel scale is approximately linear to the linear frequency scale. We know that human ears, for frequencies lower than 1 kHz, hear tones with a linear scale instead of logarithmic scale for the frequencies higher that 1 kHz. The mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The voice signals have most of their energy in the low frequencies. It is also very natural to use a mel-spaced filter bank showing the above characteristics. The following equation describes the mathematical relationship between the mel scale and the linear frequency scale,

$$fmel = 1127.01In\left(1+\frac{f}{700}\right) \qquad (1)$$

Where $fmel$ is the mel frequency in mels and f is the linear frequency in Hz. As previously mentioned, the mel frequency filter bank consists of a series of triangular band pass filters. The filters are overlapped in such a way that the lower boundary of one filter is situated at the centre frequency of the previous filter and the upper boundary is situated at the centre frequency of the next filter. The maximum response of a filter, that is, the top vertex of the triangular filter, is located at the filter's centre frequency and is normalised to unity.

2.3.2.3 Logarithmic Compression

In order to model the perceived loudness of a given signal intensity, the filter outputs are compressed by a logarithmic function

$$X_{m\,(\mathrm{ln})} = \ln(Xm)1\leq m\leq M \qquad (2)$$

In the above equation, Xm (ln) is the logarithmically-compressed output of the mth filter.

2.3.2.4 Discrete Cosine Transform (DCT)

The final step of the algorithm is to decorrelate the filter outputs. Discrete Cosine Transform (DCT) is applied to the filter outputs and the first few coefficients are grouped together as a feature vector of a particular speech frame. Suppose p is the order of the Mel scale cestrum. The feature vector is obtained by considering the first p DCT coefficients. Mathematically, the k^{th} MFCC coefficient can be expressed by the following formula.

$$MFCCk = \sqrt{\frac{2}{M}}\sum_{m=1}^{M} Xm(In)cos\left(\frac{\Pi k(m-0.5)}{M}\right) \qquad (3)$$

$$where\ 1\leq k\leq p$$

The static feature vector is often appended by either a log energy component or a zerothorder coefficient or both. The zeroth order MFCC coefficient is the zeroth order DCT coefficient of the filter outputs. The following equation is the expression of the zeroth order MFCC coefficient

$$MFCCo = \sqrt{\frac{1}{M}}\sum_{m=1}^{M} Xm(In) \qquad (4)$$

2.4 Feature Matching

The state-of-the-art in feature matching techniques used in speech recognition includes Dynamic Time Warping (DTW), Hidden Markov Modelling (HMM), and Vector quantization [10].

2.4.1 Dynamic Time Warping Algorithm

Dynamic Time Warping algorithm (DTW) is an algorithm that calculates an optimal warping path between two time series [5]. The algorithm calculates both warping path values between the two series and the distance between them. Suppose we have two numerical sequences ($a1$, $a2$... an) and ($b1$, $b2$... bm). As we can see, the length of the two sequences can be different. The algorithm starts with local distances calculation between the elements of the two sequences using different types of distances. The most frequently used method for distance calculation is the absolute distance between the values of the two elements (Euclidian distance). That results in a matrix of distances having n lines and m columns of general term:

$$dij = \mid ai - bi \mid, i = 1, \text{n and j = 1, m} \qquad (5)$$

Starting with local distances matrix, then the minimal distance matrix between sequences is determined using a dynamic programming algorithm and the following optimization criterion:

$$dij = \mid ai - bj \mid, i = 1, \text{n and j = 1, m} \qquad (6)$$

Where a_{ij} is the minimal distance between the subsequence (a_1,a_2, ..., a_i) and (b_1, b_2, ..., b_j). A warping path is a path through minimal distance matrix from a_{11} element to a_{ij} element consisting of those a_{ij} elements that have formed the a_{nm} distance.

The global warp cost of the two sequences is defined as shown below:

$$GC = \frac{1}{p} \sum_{i}^{p} wi \qquad (7)$$

Where w_i are those elements that belong to warping path, and p is the number of them.

2.4.2 Hidden Markov Models

Until now this is the most successful and the most used method for pattern recognition. It's a mathematical model derived from a Markov Model. Speech recognition uses a slightly adapted Markov Model. Speech is split into the smallest audible entities (not only vowels and consonants but also conjugated sound like ou, ea, au).All these entities are represented as states in the Markov Model. As a word enters a Hidden Markov Model it is compared to the best suited model (entity). According to transition probabilities there exists a transition from one state to another. E.g. the probability of a word starting with "xq" is almost zero. A state can also have a transition to its own if the sound repeats itself. Markov Model seems to perform quite well in noisy environments because every sound entity is treated separately. If a sound entity is lost in the noise the model might be able to guess that entity based on the probability of going from one sound entity to another.

2.4.3 Vector Quantization

The problem of speech recognition belongs to a much broader topic in scientific and engineering so called pattern recognition. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in this case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual words. Since the classification procedure in this case is applied on extracted features, it can be also referred to as feature matching [11]. The patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are known, then one can evaluate the performance of the algorithm. Vector quantization (VQ) is a method of compressing vector data by partitioning the continuous vector space into non-overlapping subsets and representing each subset with a unique codeword. The set of available codewords is termed the Codebook. It is an efficient and simple approach for data compression [7].The design of a codebook is the typical application in speech recognition based on vector quantization. The well-known k-means clustering algorithm is one of the most popular competitive learning vector quantization schemes. Although the k-means algorithm is simple and appealing, it has some inevitable problems. Intuitively, the mean square error (MSE) seems to monotonically decrease with an increasing k. However, the MSE may sometimes increase even when the value of k increases. Another problem is that the initial codebook strongly affects the performance of the k-means algorithm. Still another problem is that the algorithm may not converge towards an optimal solution. A variation of the k-means algorithm, known as the LBG algorithm, still suffers from these problems. The self-organizing feature map (SOFM) is one of the most popular competitive neural network algorithms. During the training procedure, the SOFM algorithm decides a winner neuron and updates the weights of both the winner and its neighbours. In order to obtain the best results from SOFM, the updated neighbourhood, the learning rate and the total number of iterations for a given set of data should be chosen carefully.

2.4.3.1Centroid Neural Network Artificial Resonance Theory (CNN-ART)

The conventional method of speech recognition insist in representing each word by its feature vector & pattern matching with the statistically available vectors using neural network. Neural networks are composed of simple computational elements operating in parallel [4]. There also have several architectures applied to previous speech recognition by using neural network research. Examples are by using Radial Basis Function (RBF), Learning Vector Quantization (LVQ), Feed forward networks, recurrent connection and time delay and others. Every architecture have its own strength and weakness.RBF network uses hidden unit with localized receptor fields because they only respond to input that are close to the center. It consists of two layers. In first layer, it does not use the weighted sum of inputs and the sigmoid transfer function. Instead, the output of the first layer neuron, each of which represents a basis function, that is determined by the distance between the network input and the center of the basis function. The second layer is linear and produced by the weighted sum of outputs of the first layer.An unsupervised competitive learning algorithm, called the centroid neural network adaptive resonance theory (CNN-ART) algorithm, has been used in this project. Since the neurons of CNN-ART grow dynamically until the size of the codebook, the set of synaptic weights can be treated as a codebook. Due to this significant property of the CNN-ART algorithm, the appropriate initial codebook can be easily obtained in contrast to the conventional algorithms mentioned earlier. One of the best features is that CNN-ART does not require a schedule for the learning rate. After iterations, the weights can converge fast towards the centroid of clusters for the expected codebook size.

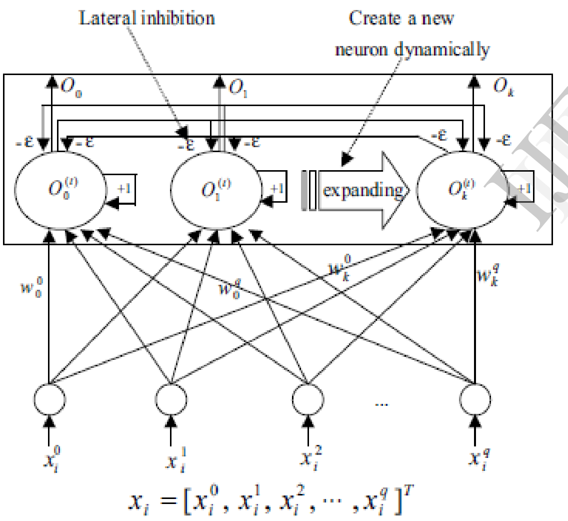

Figure 3 Architecture of CNN-ART

The adaptive resonance theory (ART) algorithm is a special type of neural network that can realize classification without supervision. The CNN-ART network consists of an input layer and an output layer, which is a simple net for determining the nearest cluster exemplar. Each node in the input layer is connected directly to the neurons in the output layer. A synaptic weight w_p; p=0, 1, 2,3....k; that has the same dimension as the input data xi, is associated with each neuron in the output layer. As a result, the set of weights wp can be treated as a codebook. In the subnet, the number of neurons starts with one and grows by one each time until the desired codebook size is reached; that is, the number of neurons is proportional to the size of the codebook. Due to this property, the CNN-ART algorithm does not require the selection of any appropriate initial codeword in advance. In the CNN-ART competitive learning algorithm, at first, an input training vector x0 is selected as the centroid of the first cluster, and then the next input training vector is compared to the first cluster centroid. It is classified as part of the first cluster if its Euclidean distance is smaller than the threshold. Otherwise, it forms the centroid of a new cluster. This process is repeated for all the training input vectors. As the input training vectors are reiterated in the CNN-ART algorithm, input training vector xi changes the cluster from the old cluster to a new one. In this case, the weights of the new neuron are updated and the weights of the old neuron are updated. Input training vector xi is classified to belong to the same cluster as before. In this case, no learning action is performed. The CNN-ART algorithm is reiterated until a stable cluster formation occurs, so it is not necessary to decide the total number of iterations to process in advance. The final set of synaptic weights can be treated as a codebook.

3. System Flow

Initially a user pronounces a word. All the pronounced words are then compared with the pre-stored set of words in the database. If the match is found, those are concatenated to form a sentence. Then the corresponding web-page automatically gets loaded.

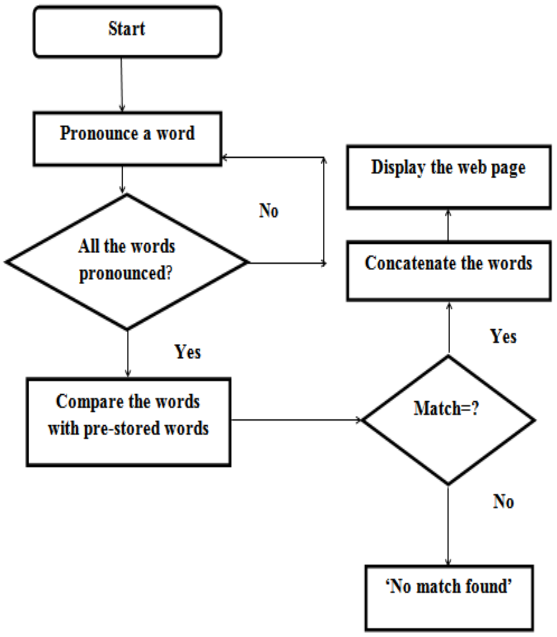

Figure 4 System Flow

4. Comparative Study

A comparative study for determining the accuracy of the system based on speaker dependent vs speaker independent environment and an isolated vs continuous sentence gives the following results. Table I give the recognition accuracy for comparative study for speaker dependent vs speaker independent testing. It is observed that when testing is made in speaker independent environment, accuracy is less i.e. 70% compare to when testing is made in speaker dependent environment where accuracy obtained is 85%. Table II give the recognition accuracy for comparative study for isolated vs continuous sentences. It is observed that isolated word recognition gives best results on an average 80% than the continuous word speech recognition with 70%

(i)Comparative study based on the speaker dependent and speaker independent testing:

Table I

Sr.No.	Testing	Accuracy%
1	**Speaker Dependent**	85
2	**Speaker Independent**	70

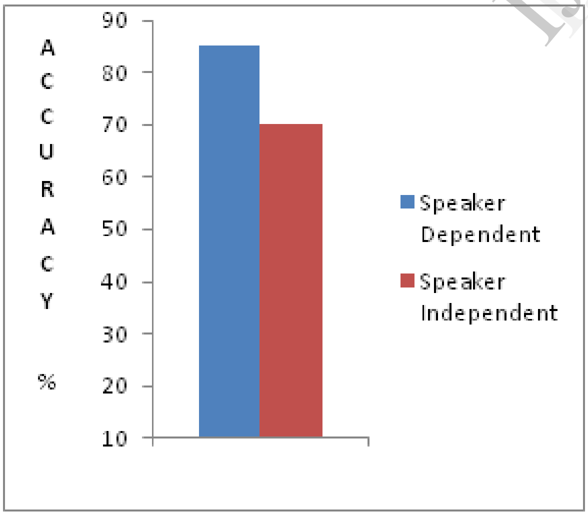

Figure 5 Comparison Graph for speaker dependent and independent environment

(ii)Comparative study based on the type of the Speech:

Table II

Sr.No.	Speech Type	Accuracy%
1	**Isolated Word**	**80**
2	**Continuous Sentences**	**70**

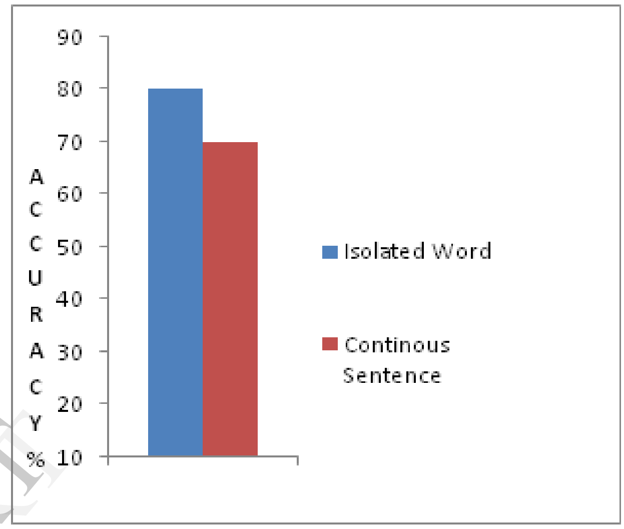

Figure 6 Comparison Graph for Isolated word Vs Continuous sentences

5. Conclusion

In order to make technology more familiar to the user its access should be made easier. As internet access experiences various limitations such as people who are physically handicapped cannot use keypads or touch screens for giving instructions. Above all these limitations today's generation demands to use internet independent of keyboards and also hands free access to it. For this, voice browsing is an intelligent idea. This allows user to access web even in situations like driving etc. where user operate web just speaking rather than typing.

The speech recognition system contains two main modules (i) feature extraction and (ii) feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speech sample. Feature matching involves the actual procedure to identify the known speech by comparing extracted features from voice input with the ones from a set of known speech samples. In this project, a very efficient speech recognition system for web browsing is designed. A special care is taken at pre-processing stage, as most of the success of speech recognition systems lies in pre-processing stage. MFCC technique is used for feature extraction and feature matching is carried out using centroid based neural network approach. Then comparison is made based on isolated words and continuous words, and speaker dependent and speaker independent environment. It is found that isolated word recognition gives best results on an average 80% than the continuous word speech recognition with 70% accuracy. Also it is observed that when testing is made in speaker independent environment, accuracy is less i.e. 70% compare to when testing is made in speaker dependent environment where accuracy obtained is 85%. The system was tested many times with various databases and found to be very reliable. Also, the system is designed to be speaker independent. Hence, it provides a greater flexibility in terms of usage by any person and also eliminates the time required for training. Further improvement can be obtained by increasing the reference database size. Implementing recognition by recognizing speech sounds can increase the accuracy greatly thus enabling the usage of this system in high-end systems. Increasing the scope of this system can reduce the usage of mouse of keyboard to a greater extent.

6. References

[1] Davis, S.B., Mermelstein, P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on Acoustic, Speech and Signal Processing, 28(4):357–366 (1980).

[2] Lakshmi Kanaka Venkateswarlu Revada, Vasantha Kumari Rambatla and Koti Verra Nagayya Ande,'A Novel Approach to Speech Recognition by Using Generalized Regression Neural Networks' IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011

[3] Lingyun Gu and Stephen A. Zahorian, 'A new robust algorithm for isolated word end point detection'*ICASSP, page 4161. IEEE, (2002)*

[4] Dr.R L K Venkateswarlu,Dr.VasanthaKumari,A K V Nagayya, 'Efficient Speech Recognition by Using Modular Neural Network' International Journal on Computer Technology and Applications, Vol 2 (3), pp.463-470

[5]RabinarL,BingHwangJ."Fundamentals On Speech Recognition",Prentice Hall,1993

[6] John Coleman, "Introducing Speech and language processing", Cambridge university press, 2005
.

[7] Tzu-Chao Lin ,Ching –Yun Chang "A Survey of VQ Codebook Generation"Jorrnal of information hiding and Multimedia Signal Processing,Volume 1, No.3,July 2010

[8]Tzu-Chao Lin ; Pao-Ta Yu 'A new unsupervised competitive learning algorithm for vector quantization' Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on (Volume:2) 944 – 948,2002

[9]Abderrahmane Amrouche, and Jean Michel Rouvaen'Efficient System for Speech Recognition using General Regression Neural Network, International Journal of Electrical and Computer Engineering 1:6 2006

[10] Rita Singh, Bhiksha Raj, and Richard M. Stern, Member, IEEE, "Automatic Generation of Sub word Units for Speech Recognition Systems", IEEE Transactions on speech and audio processing, VOL. 10, NO. 2, Feb. 2002.

[11] Dipmoy Gupta, Radha Mounima C. Navya Manjunath, Manoj PB' Isolated Word Speech Recognition Using Vector Quantization (VQ)' International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012