ISBN 978-93-84468-20-0
Proceedings of 2015 International Conference on Future Computational Technologies (ICFCT'2015)
Singapore, March 29-30, 2015, pp. 116-122

Myanmar Language Speech Recognition with Hybrid Artificial Neural Network and Hidden Markov Model

Thin Thin Nwe[1], and Theingi Myint[2]
Department of Information Technology Engineering, Yangon Technological University (YTU), Myanmar
thinthinnwe1985@gmail.com,drtgim@gmail.com

Abstract**:** *There are many artificial intelligence approaches used in the development of Automatic Speech Recognition (ASR), hybrid approach is one of them. The common hybrid method in speech recognition is the combination of Artificial Neural Network (ANN) and Hidden Markov Model (HMM). The hybrid ANN/HMM is able to classify the phoneme model and to combine the strength of HMM in sequential modeling structure. Thus, this paper proposed a speaker independent and continuous Myanmar Language speech recognition by using the hybrid ANN/HMM method.*

Keywords: *Artificial Neural Network, Continuous Speech, Hybrid ANN/HMM, Myanmar Language, Speaker Independent, Speech Recognition*

1. Introduction

Automatic speech recognition (ASR) technology allows a computer to identify the words spoken by a person through a microphone or other voice input device. It has long been viewed as a promising alternative for human –computer interaction (HCI) over the traditional keyboard and mouse [1].The Artificial Neural Networks (ANN) models have been used for connectionist speech recognition but with limited success. This is because, although ANN has a good discriminative power and flexible, it is not tailored for sequential data such as speech [2]. In the early of 1970's, the Hidden Markov Model (HMM) was implemented to the speech recognition field by Baker for the Dragon system.. Since then, the HMMs have become the dominant technology in ASR. The main advantages of HMMs-based systems are the statistical representations of the acoustic speech signal and the stochastic processes that capable of modeling sequential data. However, standard HMMs have some drawbacks in building a large vocabulary speaker independent continuous ASR system. It has poor discrimination power due to unsupervised learning [3] where the model parameters are estimated by maximum likelihood (ML estimation). Thus, hybrid ANN/HMM system is proposed to augment ASR performance. The experimental results indicate that the accuracy for hybrid ANN/HMM model outperform the HMM model.

This paper is organized as follows. In Section 2 describes related works. Overview design of the hybrid ANN/HMM speech recognition for continuous Myanmar Language described in section 3. In Section 4, we described the implementation of the proposed system and experimental result. Conclusion is described in section 5.

2. Related Work

Lawrence showed [1] that algorithms for connected word recognition based on whole word reference patterns have become increasingly sophisticated and capable of achieving high recognition performance for small or syntax-constrained, moderate size vocabularies in a speaker trained mode. In particular, it has been demonstrated that for a vocabulary of digits, in a speaker trained mode, very high string accuracy is achievable using either Hidden Markov Models (HMM) or templates as the digit reference patterns.

http://dx.doi.org/10.17758/UR.U0315236

S K Hasnain [4] presented a speech processing and recognition system for individually spoken Urdu language words. The speech feature extraction was based on a dataset of 150 different samples collected from 15 different speakers. The speech recognition feed-forward neural models were developed in MATLAB. In this paper,the author attempted at using an NN(neural network) to recognize spoken Urdu language words. The DFT (Discrete Fourier Transform) of the acquired data was used for training and testing the speech recognition NN. The network made predictions with high accuracy.

K. Roy performed the recognition by Artificial Neural Network (ANN) using back propagation neural Network. They used DSP (Digital Signal Processing) techniques to extract the features of speech signal. M. R. Hassan presented a phoneme recognition approach using ANN as a classifier. A. H. M. Rezaul Karim presented a technique to recognized bangla phonemes using the Euclidian distance measure. Reflection coefficient and autocorrelations have been used as features. K. J. Rahman presented continuous Bangla speech recognition system using ANN. They employed a word separation algorithm to separate the words. They applied fourier transform based spectral analysis to generate the feature vectors from each isolated words. M. R. Islam presented a Bangla ASR system that employed a three layer back propagation Neural Network as the classifier. S. A. Hossain [5] presented a brief overview of Bangla speech synthesis and recognition. A comparative study on the feature extraction methods are presented by M. F. Khan.

Nitin N Lokhande [6] concerned in isolated word recognition systems, accurate detection of the endpoints of a spoken word is important for two reasons, namely: reliable word recognition is critically dependent on accurate endpoint detection and the computation for processing the speech is less, when the endpoints are accurately located. The database used for experimentation was ZERO to NINE digits in English language.

3. Overview Design of the Hybrid Speech Recognition for Myanmar Language

Artificial Neural Network (ANN) consist of a number of interconnected processing units called neurons, which is capable of taking in numbers of input and producing an output.ANN method can be used for estimating posterior probabilities and training the network, while HMM methods can be used for decoding and language modeling. Myanmar speech recognition are still limited especially those using hybrid ANN/HMM for speaker independent and continuous speech recognition system. Therefore this paper aim to apply hybrid ANN/HMM approach for developing a speaker independent continuous speech recognizer with a medium size vocabulary.

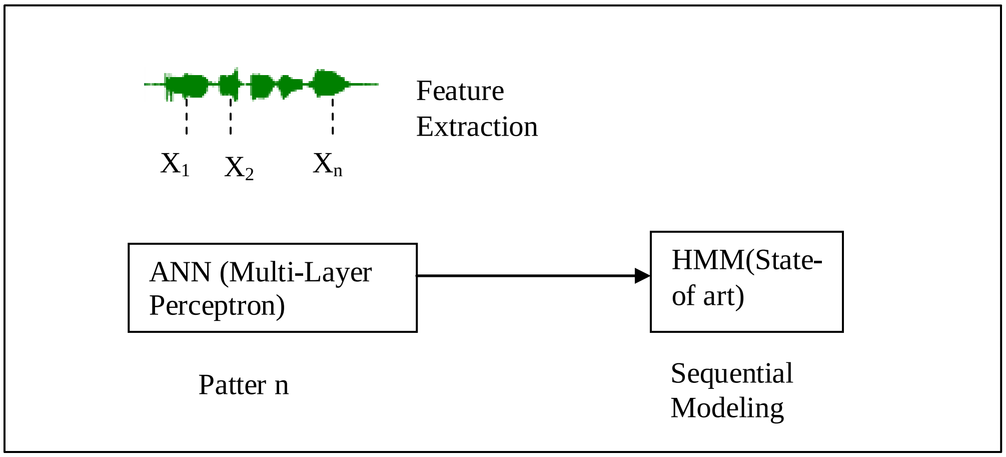


Fig. 1: A typical hybrid ANN/HMM speech recognition system

HMM can be used to model, a unit of speech whether it is a phoneme, a word, or a sentence. HMM is a variant of a finite state machine having a set of hidden states Q, an output alphabet (observations) O, transition probabilities A, output (emission) probabilities B, and initial state probabilities π. The current state is not observable. Instead, each state produces an output with a certain probability (B). Usually the states Q, and outputs O, are understood, so an HMM is said to be a triple (A, B, π). HMM is most easily understood as a generator of vector sequences. In this paper the ANN used to classify the phoneme and HMM method used to recognized words. Fig.1. shows the typical hybrid ANN/HMM speech recognition system.

The main advantage of HMM is rich of mathematical structure, thus it is able to characterize speech signal in a mathematically tractable way. Both Hidden Markov Model (HMM) and Multilayer Perceptron (MLP) based approaches have been developed in the context of a long history of pattern recognition technology. Though specific methods are changing, the pattern recognition perspective continues to be useful for the description of many problems and their proposed solutions.

4. Implementation of the proposed system and Experimental result

Speech recognition is the task of recognizing the spoken word from speech signal. The use of syllables as the basic unit in a speech recognition is very useful and improve the performance of the speech recognition process. Fig. 2 shows the implementation of the proposed system.

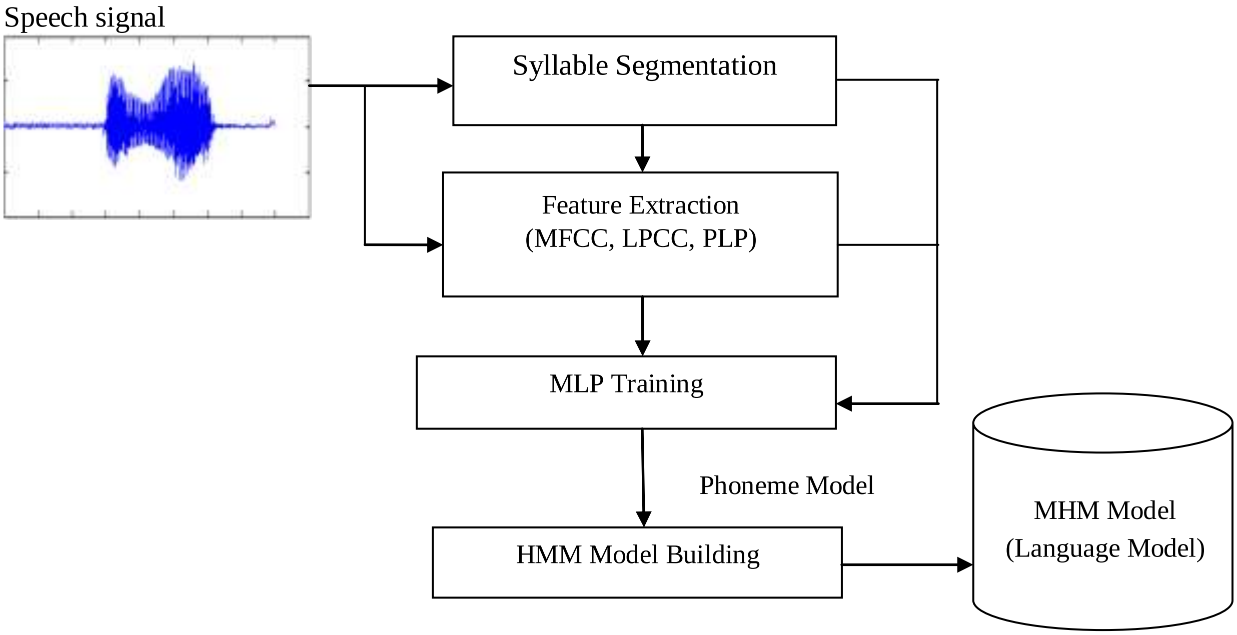


Fig. 2: Implementation of the proposed system

4.1. Speech Signal

Audio files are saved in the encoded format. Speech signal is recorded by a microphone and converted into an electrical signal, where the amplitude of the signal corresponds to the original pressure variation.

4.2. Syllable Segmentation

Syllable segmentation is a process for a sequence of speech sounds. The intra-segment distance d_{i+1} is required because there may be frequently spurious speech segments that satisfy the first criterion. Therefore, speech segmentation is breaking streams of sound into some units like words, phonemes, or syllables that can be recognized. The general idea of segmentation can be used to distinguish different types of audio signals from large amounts of audio data. Segmentation is a process of decomposing the speech signal into smaller units. Segmentation is the very basic step in any voiced activated systems like speech recognition system It will be necessary to merge two such speech segments into one larger segment. This happens frequently with words. If $\lambda_i < \kappa$ and $d_{i+1} > \delta$, then the i^{th} segment is discarded. If (λ_i or λ_{i+1}) $> \kappa$, $d_{i+1} > \delta$ and $\lambda_i + \lambda_{i+1} < \theta$, then the two segments are merged, and anything between the two segments that was previously left, is made part of the speech. Fig. 3 shows the syllable segmentation method.

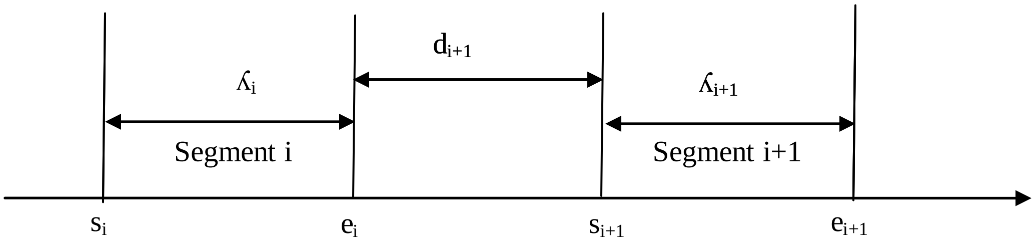


Fig. 3: Syllable segmentation

where *i* is the length of *i* segment, d_{i+1} is the distance between two segments and s_i is start point of *i* segment, e_i is the end point of *i* segment.

4.3. Feature Extraction

The purpose of feature extraction is to convert the speech waveform to some type of parametric representation for further analysis and processing. Feature extraction is process of obtaining different features such as power, pitch and vocal tract configuration from the speech signal. Therefore, feature extraction involves analysis of speech signal. It is also the most important part of speech recognition since it plays an important role to separate one speech from other. This is often referred to as the signal processing front end. The features used in this system are Mel-frequency Cepstral Coefficient (MFCC).

1) Mel Frequency Cepstral Coefficients (MFCC)

The feature in this system used (MFCC) is one of the most commonly used feature extraction front-ends in speech recognition systems. The technique is so-called FFT-based, which means that feature vectors are extracted from the frequency spectra of the windowed speech frames. MFCC extraction procedure as follows:

- Pre-emphasis, Hamming windowing and FFT
- Mel scale Filter Bank
- Logarithmic compression
- Discrete Cosine Transform (DCT)

2) Linear Prediction Cepstral Coefficients (LPCC)

Linear Prediction Cepstral Coefficients (LPCC) has been commonly used in many speech recognition applications for many years. LPCC has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech.The basic steps of LPCC processor include the following:

- Pre-emphasis, Hamming windowing
- Linear Predictive Analysis
- Cepstral Analysis

3) Perceptual Linear Prediction Coefficients (PLP)

Perceptual Linear Prediction (PLP) coefficient is another feature extraction technique, which tries to emulate the human auditory system. The basic step includes the following procedure:

- Hamming windowing and FFT
- Bark scale filter bank
- Equal Loudness curve
- Intensity Loudness compression
- IDFT(Inverse Discrete Fourier Transform) and Linear Predictive analysis
- Cepstral Analysis

4.4. Artificial Neural Network in speech recognition

The inputs of the network are the features extracted from the selected frames. In this paper, a feed-forward multi-layer perceptron with a single hidden layer and trained by gradient descent with momentum and an adaptive learning rate back-propagation method. The neural network is trained by minimizing the mean square error between the outputs HMMs. Fig. 4 shows the simplified Neural Network for continuous speech recognition. Table I show all the row are result of output layer and above the column show number of training target. Table II show the output layer for training target. The database is split into two groups, one for training the neural network, the other for testing the performance of the trained neural network. The first group, training database, comprises 26*10=260 female speaker's utterances.

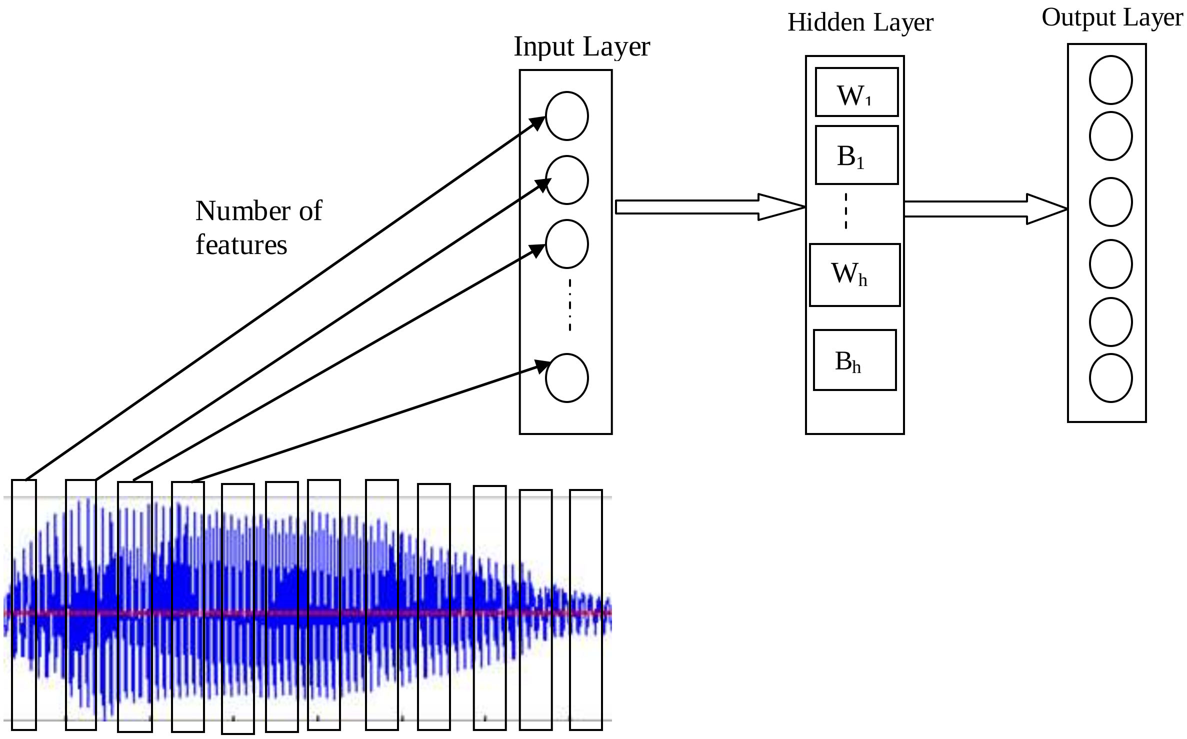


Fig. 4: Simplified Neural Network Architecture for continuous speech recognition

TABLE I: Output Layer of Training Target

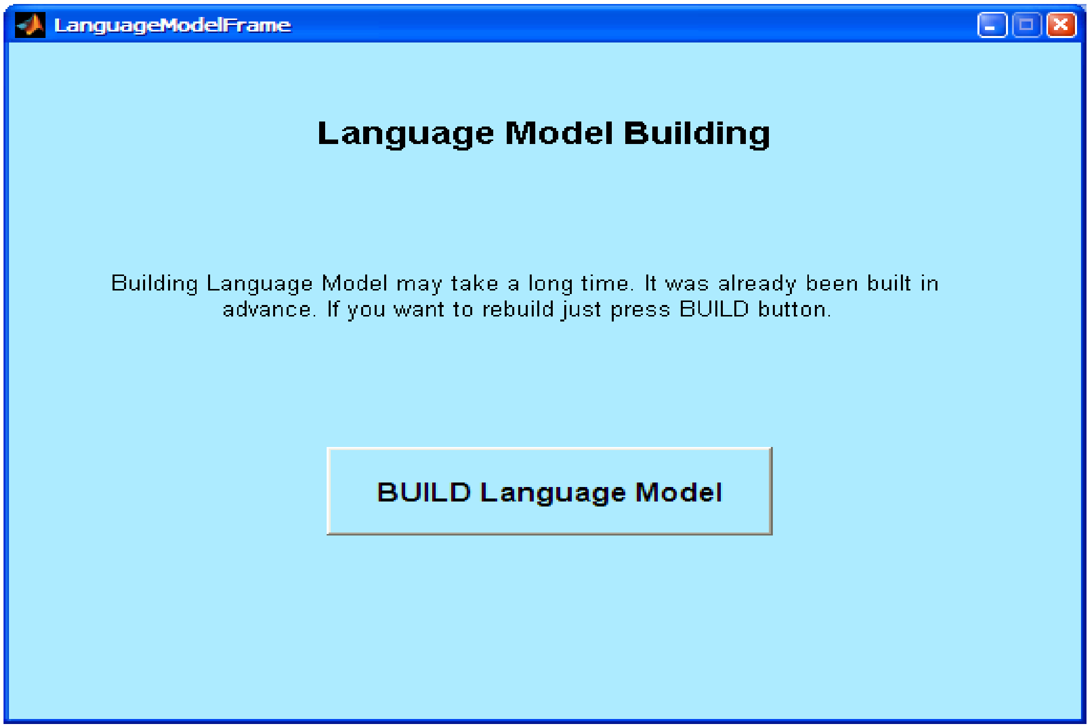
	No. of training target					
Result of output layer	1	2	3	4	5	6
	1	0	0	0	0	0
	0	1	0	0	0	0
	0	0	1	0	0	0
	0	0	0	1	0	0
	0	0	0	0	1	0
	0	0	0	0	0	1

TABLE II: Define the output for phoneme

Output	Definition
1	Start consonant
2	Start vowel
3	Middle consonant
4	Middle vowel
5	End consonant
6	End vowel

4.5. Language Model Building

The purpose of the Language modeling is to provide a mechanism for estimating the probability of some word w_k in an utterance given the preceding word W_1^{k-1} $=w_1\ldots w_{k-1}$.Pronunciation dictionary was created that contains the input-output-pronouncing for each word entry where the pronunciation describes the sequence of HMMs that constitute of each word. Fig. 5 shows the pronunciation dictionary for language model.



Language Model Building

Building Language Model may take a long time. It was already been built in advance. If you want to rebuild just press BUILD button.

Fig. 5: Pronunciation dictionary for Language Model

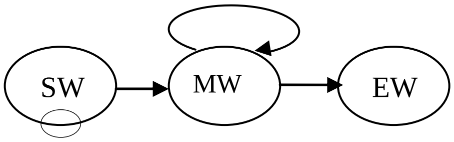


Fig. 6: Show the sub-word for language model

TABLE III: Show the types of words for language model

Type	Words
Start Words(SW)	□□
Middle Words (MW)	□□□□□□
End Words (EW)	□□□

TABLE IV: Show the language model

Greeting words	Definition
□□□□□□□□□□□□□□□□□□	Have u finished meal? (hta min sar pyee pyi lar)
□□□□□□□□□□□	How are you?

4.6. HMM model building

A Hidden Markov model is a type of stochastic model appropriate for non stationary stochastic sequences, with statistical properties that undergo distinct random transitions among a set of different stationary process. Figure 6 show the HMM structure of Myanmar phoneme model for recognition process. Fig.7 show the HMM structure of the phoneme.

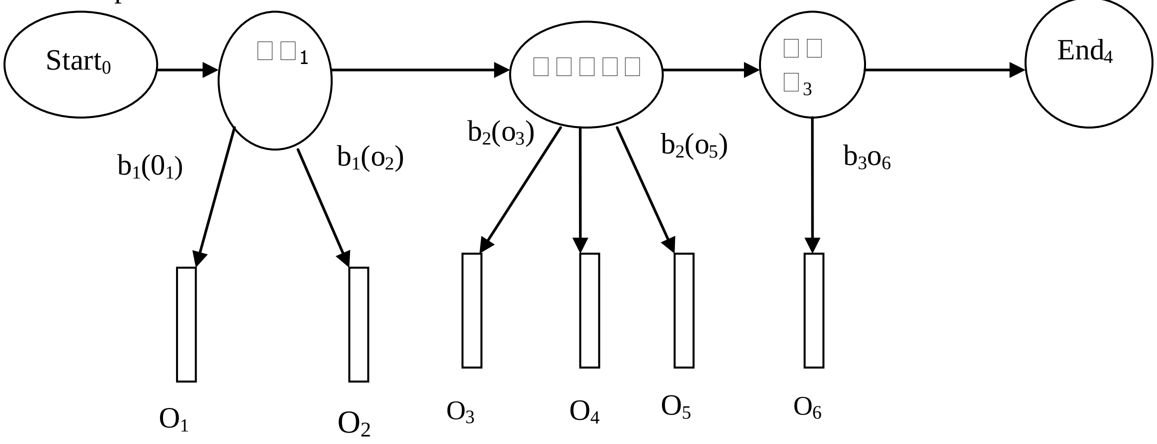


Fig. 7: HMM structure of the phoneme

5. Conclusion

In conclusion, this paper has proposed the use of hybrid ANN/HMM method for developing a speaker independent and continuous Myanmar Language speech recognition. The recognizer is implemented using the HTK toolkit with speech data collected from multiple speakers. Besides, an automatic speech recognition system has been designed using MATLAB programming.

6. Acknowledgements

The author would like to acknowledge her supervisor, brothers, sisters, all of my friends and teachers of life, who provided her with useful comments and helped to improve the quality of this paper and also specially thank to Rector, Professors and colleagues from Yangon Technological University, Myanmar.

7. References

[1] Lawrence R, Rabiner, Fellow, IEEE Transactions on acoustic speech and signal processing, “High performance connected digit recognition using Hidden Markov Model”,vol.37.No.8, August 1989.

[2] Bahl LR, Balakrishnan-Aiyer S, “Performance of the IBM Large Vocabulary Continuous Speech Recognition System” in Pro ICASSP vol1,pp 41-44,Detroit,1995.

[3] Bahl L,Gopa lakrishnan PS, “A Fast Admissible Method for Identifying a short list of candidate words”,Computer speech and language,Vol 6,No.3,pp 215-224,1992.
http://dx.doi.org/10.1016/0885-2308(92)90018-Y

[4] SK Hasnain, A Zam Beg “A Speech Recognition System for Urdu Language” in International Multi-Topic Conference,Pakistan,2008,pp.74-78.

[5] Md.Abul Hasnat, Jabir Mowla, Mumit khan, “Isolated and continuous Bangla Speech Recognition”.

[6] Nitin N Lokhande, Dr.Navnath S Nehe, Pratap S Vikhe, “Voice activity detection Algorithm for Speech Recognition Applications”.

[7] Alleva FA, “Search Organisation for Large Vocabulary speech recognition”, Pro NATO workshop,1990.