International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979, Impact Factor: RJIF 5.72
www.allsubjectjournal.com
Volume 3; Issue 7; July 2016; Page No. 122-127

Bangla speech sentence recognition using hidden Markov models

[1] Md. Ashraful Kadir, [2] Md. Mijanur Rahman

[1] Research Student Dept. of Computer Science and Engineering Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh.

[2] Associate Professor, Dept. of Computer Science and Engineering Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh

Abstract
This paper aims to build a Bangla speech sentence recognition system by Hidden Markov Model (HMM). This system includes two phases; such as, a feature extraction phase to generate speech features from the Bangla speech sentence and a recognition phase to identify the Bangla speech sentence. The Mel Frequency Cepstral Coefficients (MFCCs) have been used to generate the features from the input Bangla speech sentences. These MFCCs features were used in the HMM based speech recognizer to identify the Bangla speech sentences. Train data with the different types of HMM training algorithm. In both training and testing process, one hidden Markov model for each sentence has been implemented. The models were trained with labeled training data, and the classification was performed by passing the features to each model and selected the best match. The development and experiments were done on MATLAB 2010 and the learning behavior of the algorithms was tested on different Bangla speech sentences from different speakers.

Keywords: Feature Extraction, Hidden Markov models, Speech Recognition, MFCC

1. Introduction
Speech recognition is simply transcribing the speech without necessarily knowing the meaning of the utterance. Automatic speech recognition has a long history of being a difficult problem from about 1950 [1]. The Hidden Markov Model (HMM) currently shows the best performance of the available techniques [2]. Speech recognition is widely researched topic and many researchers are busy with doing works on speech recognition in the world. Worldwide speech recognition is mostly done in different languages mostly English. But our mother tongue Bangla is not enriched with a speech recognizer. This research deals with Bangla speech recognition. This paper will discuss the speech recognition problem and how HMM are used to identify speech sentence.

2. Background Study
The speech processing is an important and popular field for research in digital signal processing for last 60 years and the automatic speech recognition by machine was made in 1950s. In 1952 Davis, Biddulph and balashek built a system for isolated digit recognition for single speaker [3]. In 1956, Olson and Belar tried to recognize 10 distinct syllables of a single talker, as embodied in 10 monosyllabic words [4]. Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants in 1959 [5]. In 1961 Suzuki and Nakata of the radio Research Lab in Tokyo described a Hardware Vowel Recognizer [6]. In 1970s speech recognition research achieved a number of significant milestones. Speech research in the 1980s was template based approach of statistical modeling method especially Hidden Markov Model (HMM). Some popular feature extraction technique, such as Linear Prediction coding (LPC), Mel Frequency Cepstral Coefficients (MFCC), Linear prediction Cepstral Coefficients (LPCC) etc. and some pattern recognition approach such as Acoustic phonetic approach, Pattern recognition approach, Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Artificial Intelligence Approach (Knowledge based approach) etc. are very important milestones in digital speech recognition [3]. In 2010s, M M Rahman tried to implement a system for isolated speech recognition system for Bangla words which introduces Bangla speech recognition system that works with speaker independent, isolated and subword-unit-based approaches [7], to develop a continuous speech segmentation, classification, feature extraction [8], to develop a simple and novel feature extraction approaches for segmenting continuous Bangla speech sentences into words/sub-words [9]. In 2007, M A Hasnat implemented a speech recognition system for isolated and continuous bangla speech [10]. This paper shows how to increase the quality of audio signal and demonstrates the end-point detection algorithm. M F Khan has done a comparative study of different feature extraction methods for Bangla Phoneme Recognition in 2000s [11]. The article of English sentence recognition based on Hidden Markov Model and Clustering by Xinguang Li that explores the segment mean algorithm for dimensionality reduction of speech feature parameters in 2013 [12]. From the above study it is concluded that a very few works have been done in Bangla speech sentence recognition. Therefore, we have tried to develop a speech recognition system using hidden Markov model to identify Bangla speech sentences in this research.

3. Hidden Markov Models
Hidden Markov models (HMMs) provide an efficient algorithmic solution to certain problems involving Markov processes. The 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)*.

A schematic view of an HMM is given in Figure 1, where $A = \{a_{ij}\}$ is the (row stochastic) matrix driving the underlying Markov process, the X_i are the states of the Markov process, the 0_i are the observations, T is the length of the observed sequence, $B = \{b_{ij}\}$ is a (row stochastic) matrix that relates the states to the observations, and the dashed line is a "curtain" between the observer and the underlying Markov process.

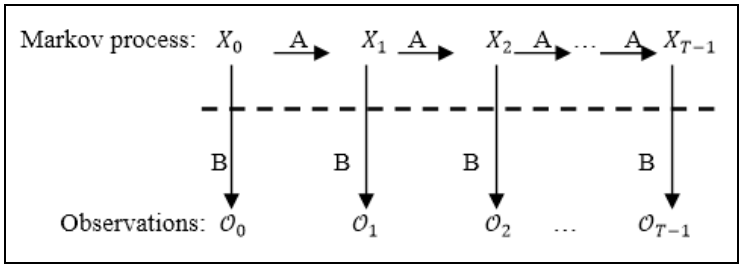

Fig 1: A schematic view of an HMM

There are three basic problems for HMM; such as evaluation problem, hidden state determination and learning. It is vital to solve these problems using different efficient algorithms. The problems and their solutions are described in the following sub-sections.

3.1 Problem 1: Evaluation problem

Given the observation sequence, $O = O_1 O_2 \cdots O_T$, and model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of observation sequence given the model. The evaluation problem for HMM is solved by using forward and backward algorithm.

3.1.1 Forward Algorithm

We can eventually calculate $\alpha_t(i)$, and then summing them over all states, we can obtain the required probability. The formal definition of the algorithm is given below:

Initialization

$\alpha_1(i) = p_i\ b_i(o(1)),\ i=1,\ldots,N$

Recursion

$\alpha_{t+1}(i) = [\sum_{j-1}^{N} \alpha_t(j)\ a_{ji}]\, b_i(o(t+1))$

where $i = 1,\ldots,N, t=1,\ldots,T-1$

Termination

$P(o(1)o(2)\ldots o(T)) = \sum_{j-1}^{N} \alpha_T(j)$

Fig 2: Forward algorithm definition.

3.1.3 Backward Algorithm

In a similar manner, we can introduce a symmetrical backward variable $\beta_t(i)$ as the conditional probability of the partial observation sequence from $o(t+1)$ to the end to be produced by all state sequences that start at i-th state:

$\beta_t(i) = P(o(t+1), o(t+2), \ldots, o(T) \mid q(t) = q_i)$.

The Backward Algorithm calculates recursively backward variables going backward along the observation sequence.

3.2 Problem 2: Hidden State Determination

Given the observation sequence $O = O1\ O2 \cdots O_T$, and model $\lambda = (A, B, \pi)$, how do we choose corresponding state sequence $Q = q_1\ q_2 \cdots q_T$ which is optimal in some meaningful sense. To solve this problem, the Vitrebi algorithm is used.

3.1.2 Viterbi algorithm

The Viterbi algorithm chooses the best state sequence that maximizes the likelihood of the state sequence for the given observation sequence. Let $\delta_t(i)$ be the maximal probability of state sequences of the length t that end in state i and produce the t first observations for the given model.

$\delta_t(i) = \max\{P(q(1), q(2), \ldots, q(t\text{-}1)\ ; o(1), o(2), \ldots, o(t) \mid q(t) = q_i).\}$

Initialization

$\delta_1(i) = p_i b_i(O(i))$

$\psi_1 = 0, i = 1, 2, \ldots, N$

According to the above definition, $\beta_T(i)$ does not exist. This is a formal extension of the recursion given below.

Recursion

$\delta_t(j) = max_i[\delta_{t-1}(i)a_{ij}]b_j(O(t))$

$\psi_t(j) = \arg max_i[\delta_{t-1}(i)a_{ij}]$

Termination

$p^* = max_i[\delta_T(i)]$

$q_T^* = \arg max_i\,[\delta_T(i)]$

Path (state sequence) backtracking

$q^*_t = \psi_{t+1}(q^*_{t+1}),\ t = T\text{-}1,\ T\text{-}2, \ldots, 2, 1$

Fig 3: The Viterbi Algorithm.

3.3 Problem 3: Learning

How do we adjust the model parameter λ = (A, B, π) to maximize P (O|λ)? In which we try to optimize model parameter so as to best describe as to how given observation sequence comes out.

3.1 Baum-Welch algorithm

Now we want to find the parameters λ that maximize the likelihood of the observations. This will be used to train the hidden Markov model with speech signals. The Baum-Welch algorithm is an iterative expectation-maximization (EM) algorithm that converges to a locally optimal solution from the initialization values. Let us define $\xi_t(i, j)$, the joint probability of being in state q_i at time t and state q_j at time $t+1$, given the model and the observed sequence:

$\xi_t(i, j) = P(q(t) = q_i, q(t+1) = q_j \mid O, \Lambda)$

Figure-4 below illustrates the calculation of $\xi_t(i, j)$.

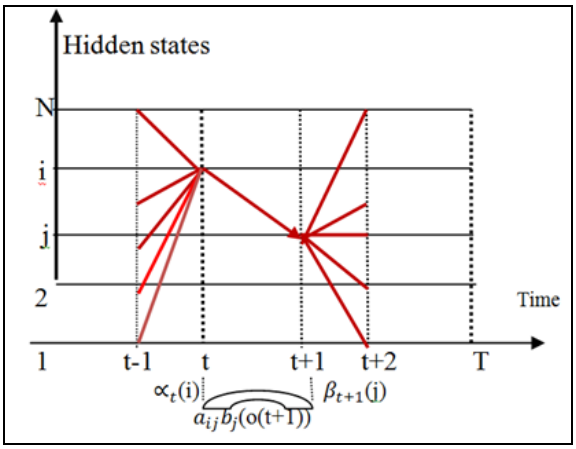

Fig 4: Joint probability paths.

Therefore, we get

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o(t+1)) \beta_{t+1}(j)}{P(O \mid \Lambda)}$$

The probability of output sequence can be expressed as

$$P(O \mid \Lambda) = \sum_{i=1}^{N} \sum_{J=1}^{N} \alpha_t(i) \alpha_{ij} b_j(o(t+1)) \beta_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i)$$

The probability of being in state q_i at time t:

$$\gamma_t(i) = \sum_{J=1}^{N} \xi_t(i, j) = \frac{\alpha_t(i) \beta_t(i) b}{P(O \mid \Lambda)}$$

4. Proposed System Design

Speech recognition consists of two main modules, feature extraction and feature matching. The feature extraction module takes the recorded speech signal as input and produces the speech feature vector as output. In feature matching, the extracted feature vector from unknown voice sample is scored against acoustic model, the model with max score wins, and recognized the speech sentences as output. The process for build up the HMM based speech recognition system includes the following five major parts, as shown in Figure-5 and The HMM based speech sentence recognizer is used in this research, summarized by a flowchart in Figure-6.

Module-1. Speech Acquisition and Preprocessing
Module-2. Speech Features Extraction
Module-3. HMM Training
Module-4. HMM matching to determine the probability
Module-5. Speech Sentence Recognition to identify the recognized command.

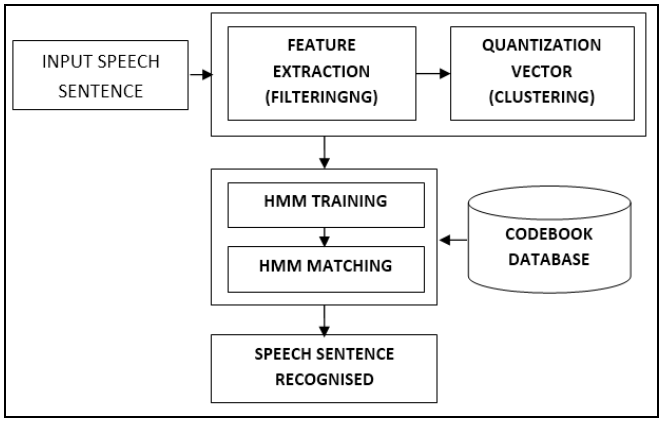

Fig 5: Block Diagram of the Proposed HMM Based Speech Recognizer.

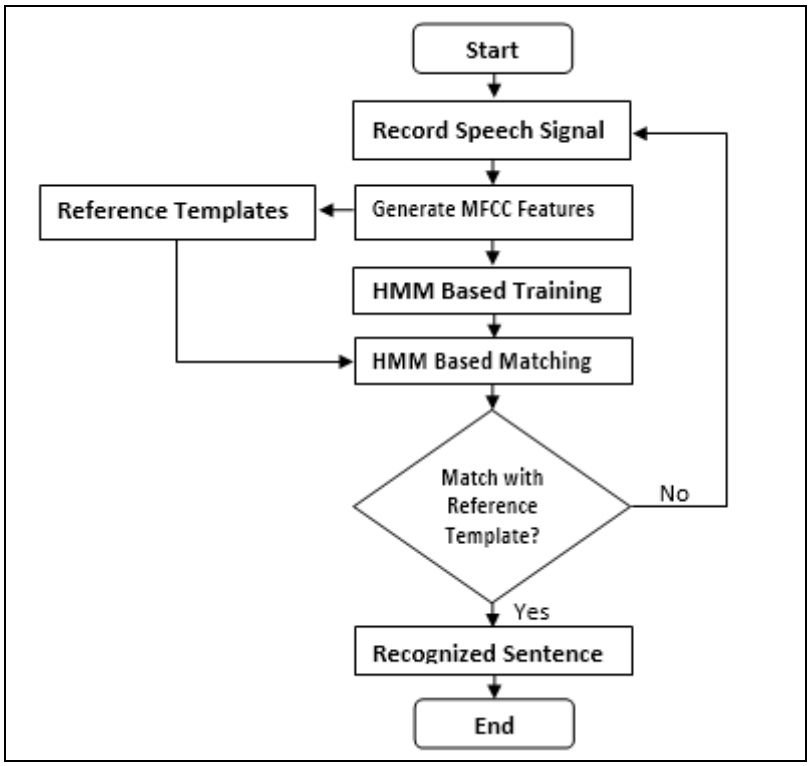

Fig 6: Flowchart of HMM Based Speech Recognition.

4.1 Speech Acquisition and Preprocessing

The first step is to record the speech data by a microphone in a specified format (32 bits wav file format). This wav data is converted into a form that is suitable for further computer processing and analysis through a series of process that involves noise elimination and the speech detection process. The source Bangla speech sentence is sampled at 8000 Hz and quantized with 32 bits.

4.2 Speech Features Extraction

Feature extraction converts the speech waveform to some type of parametric representation (a collection of meaningful features). A good feature may produce a good result for any recognition system. Mel Frequency Cepstrum Coefficient (MFCC) is the most evident and popular feature extraction technique for Bangla speech recognition and is used in this research. It approximates the human system response more closely than any other system because frequency bands are placed logarithmically here.

The computation technique of MFCC is based on the short-term analysis and thus from each frame MFCC vector is computed [3]. The signal is split up in short frames of 80 samples corresponding to 5 second of speech. The frames overlap with 20 samples on each side. Computation of MFCC includes a series of operations e.g. Fast Fourier Transform (FFT), Mel Frequency Warping, Discrete Cosine Transform (DCT) and finally the computation of Mel Frequency Cepstrum Coefficient (MFCC). The sequence is shown in Figure-7 [13].

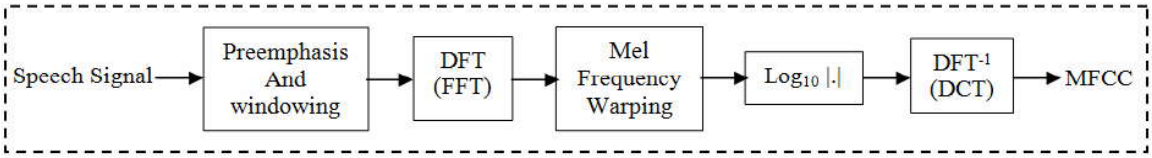

Fig 7: Calculation of MFCC feature

4.3 HMM Training

After processing speech data and setting up all system components, the next step is to train the system with speech data of different speakers. The training is a combination of both supervised and unsupervised techniques. In the proposed system, one hidden Markov model for each Bangla sentence is trained with the previously classified speech signals. The number of different states in each model is very important for the training phase. In a model, each state should represent a phoneme in the sentence. From the given Bangla sentence speech signal the observation sequence of length *T*, at each time step there are *N* different states that can generate a given observation, therefore there are N^T different state sequences that can generate an observation sequences, such a procedure is called the forward procedure. The algorithm is belongs to the class of dynamic programming algorithms, instead of considering each state sequence in turn it calculates the values for all subsequences at each time step in parallel, using the results from the previous time step. This is highly efficient as many paths share the same sub-paths. For this procedure the probability $\alpha_t(i)$ is defined as the probability of being in state *i* at time t and the observation sequence $o_1 o_2 \; o_t$; so we can define the probability $\alpha_t(i)$ as follows:

$\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = i)$

The clustering of the Gaussians is unsupervised and will depend on the initial values used for the Baum-Welch algorithm. For this research, totally random guesses (that obey the statistical properties) for *A* and π were used as initial values. For Σ_s, the diagonal covariance matrix for the training data was used for all states. For each state a random training data point was chosen as μs. The training examples for each word are concatenated together, and Baum-Welch is run for 5 iterations.

4.4 HMM Matching

So far from the parameter model, the transition probabilities and the observation probabilities have been considered as a given. But we usually do not know these values and we therefore would like to have a method that can automatically determine these parameters, $\lambda = (A, B, \pi)$ in such a way that the model best matches the observation data it is supposed to represent.

$$\hat{\lambda} = \arg\max P(0 \mid \lambda)$$

Here $\hat{\lambda}$ is the best match for the observed data O. There is no known way to analytically solve for the model parameter set that maximizes the probability of the observation sequence. This will be used to train the hidden Markov model with speech signals. The Baum-Welch algorithm is an iterative expectation-maximization (EM) algorithm that converges to a locally optimal solution from the initialization values.

4.5 Speech Recognition

There can be several possible sequences of states arising from the same frame sequence. The sequences of states with the highest overall probability is sought. The Viterbi algorithm (already explained in section 3) is used in this research. The state sequence with the highest probability is recognised.

5. Experiments and Result Analysis

For each of the sentences, 5 utterances by this author were recorded. The performance of the system was measured by five-fold cross-validation on the recorded data set of 5 utterances. Experimentation indicated that the two most important parameters were the number of hidden states, N, and the number of frequencies extracted from each frame, D. The cross-validation was therefore run with different values for these parameters, and the results are shown in table 1 and chart 1.

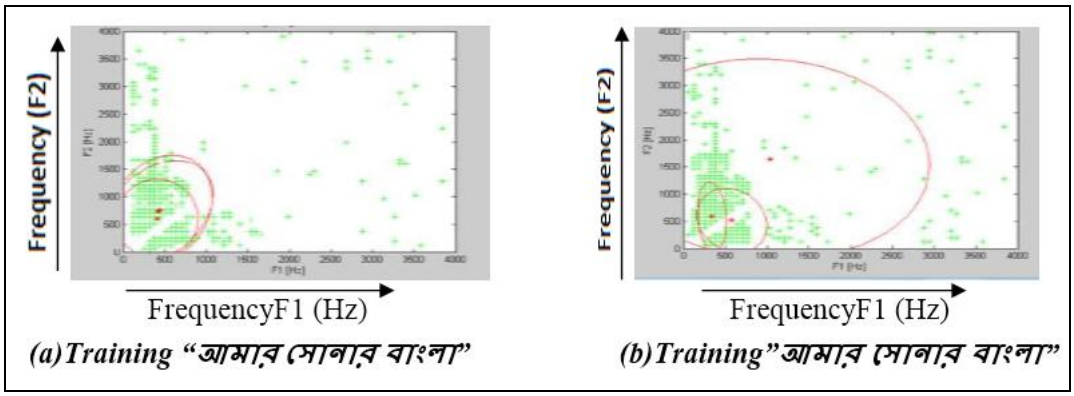

Fig 8: Graphical representation for different bangla sentence to show the confidence level of each utterances.

Figure 8 shows the confidence level of the bangla sentence during training period. After iteration of the Baum-Welch algorithm here we analyze with graphical representation to show the confidence level from different utterance. In the figure 8, each green plus represents a frame from a training speech signal, the star means of each Gaussian and the ellipse indicate their 75% of confidence interval, F1 and F2 are the frequency label in x-axis and y-axis for given sentence which maximum frequency in both axis is 400 Hz Higher frequencies present in the sentence containing unvoiced phonemes in 8(b). So from the bangla sentence "*আমার সোনার বাংলা*" we see that the following graphical representation in figure 8(a) each green plus are nearly connected to each other and maximum are inside the ellipse. Which clearly indicate the training speech signal of the sentences recognize properly.

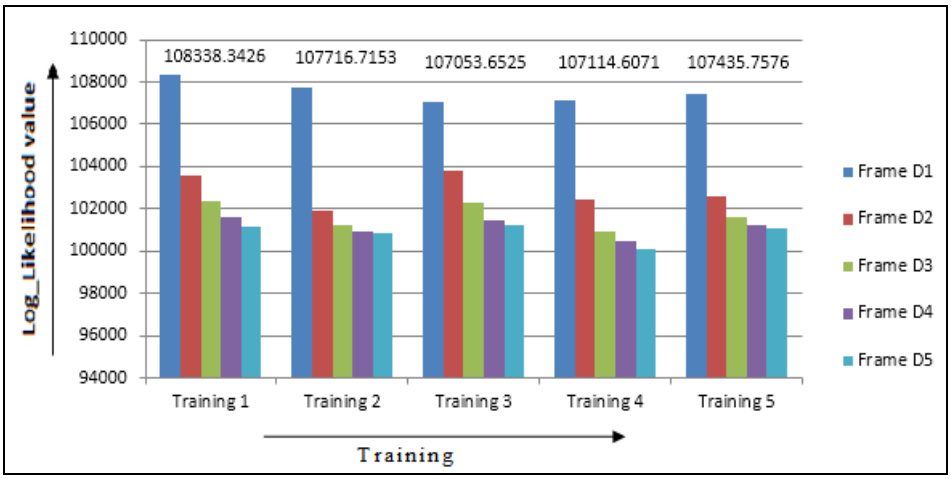

Fig 1: Show log_likelihood value for five-fold cross-validation with different values for the number of hidden states, N, and the number of frequencies extracted from each frame, D.

The results are quite good compared to the simple approach taken, especially in the feature extraction phase. More advanced features like Mel-frequency cepstral coefficients were considered, but we decided on simple frequencies due to the low misclassification rates achieved. It should be noted that this system would not perform well if trained and tested with different speakers. This is because of the different frequency characteristics of different voices, especially for speakers of different gender. We also experimented with increasing the number of training iterations for the Baum-Welch algorithm, including setting a threshold on the likelihood difference between steps. That, however, proved to have little benefit in practice; neither the execution time nor the misclassification rate showed any mentionable improvements over just fixing the number of iterations to 5.That means 5 utterances iterated (5X5) 25times for low misclassification rates and result show in table 1.

Table 1: Log likelihood value for bangla sentence "*আমার সোনার বাংলা*" with different hidden state N and extracted for each frame D.

N (states) / D(frames)	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
State 1	0.00%	4.51%	5.63%	5.43%	6.85%
State 2	0.58%	1.62%	1.17%	0.61%	0.32%
State 3	1.12%	0.24%	0.14%	0.08%	0.05%
State 4	1.14%	1.07%	1.48%	1.09%	1.05%
State 5	0.84%	0.94%	0.81%	0.35%	0.10%

In table 1, this system is implemented by using the MFCC for feature extraction and HMM as the recognizers. In speech database, audio files are recoded and these are analyzed to get feature vectors. These features are initially modeling in the HMM. After that, the test spoken sentence is addressed by forward algorithm of HMM. From the simulation results, it can be clearly seen the misclassification rate, and its better

match in every states. But, if the number of states is too large, there are no enough observations per state to train the model. So, this may degrade the performance of the system. Thus, the choice of the number of states in the HMM also plays an important case in recognition. In this work, the performance of the system is more accurate and reliable by using Baum-Welch algorithm in iteration stage.

From the above analysis of chart 1 we see that miss classification is minimum in different state and using different frame which clearly indicate the results are quite good and recognition accuracy of bangle sentence is 100%, especially in the feature extraction phase.

6. Conclusion

The main objective of this research is to develop a system for Bengla speech sentence recognition by using hidden Markov model with higher recognition accuracy. This research measured the performance of several training algorithm. This research proposed to use Baum-Welch update algorithm for training the network and the Vitrebi algorithm to find the best match output. In this research, MFCC features of speech sentences have been used. The system is able to recognize long free speech utterances. The cross validation results are good for a single speaker compare to the multiple speakers. This system is developed only for isolated bangla speech sentence recognition. In future, this research will be employed in continuous speech system to detect words and sentence boundaries. The proposed system will be integrated to interactive voice response system for developing a real time speech recognition in Bangla language.

7. References

1. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 1989; 77:257-286.
2. Hakon Sands mark. Isolated-word speech recognition using hidden Markov models, 2010.
3. idhi Desai N, Prof. Kinnal Dhameliya, Prof. Vijayendra Desai. Feature Extraction and Classification Techniques for Speech Recognition: A Review, International Journal of Emerging Technology and Advanced Engineering. 2013; 3(12).
4. Olson HF, Belar H. Phonetic Typewriter, J Acoust. Soc. Am, 1956; 28(6):1072-1081,
5. Fry DB. Theoretical Aspects of Mechanical Speech Recognition"; and P. Denes, The Design and Operation of the Mechanical Speech Recognizer at University College London, J. British Inst. Radio Engr. 1959; 19(4):211-229.
6. Suzuki J, Nakata K. Recognition of Japanese Vowels-Preliminary to the Recognition of Speech, J Radio Res. Lab, 1961; 37(8):193-212.
7. Md. Mijanur Rahman, Fatema Khatun, Dr. Md Al-Amin Bhuiyan. Development of Isolated Speech Recognition System for Bangla Words, International Journal of Applied Research on Information Technology and Computing (IJARITAC), Indianjournals.com, 2010; 1(3):272-278.
8. Md. Mijanur Rahman, Md. Farukuzzaman Khan, Md. Al-Amin Bhuiyan. Continuous Bangla Speech Segmentation, Classification and Feature Extraction, International Journal of Computer Science Issues (IJCSI). 2012; 9(2-1):67-75.
9. Md. Mijanur Rahman, Md. Al-Amin Bhuiyan. Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches", International Journal of Advanced Computer Science and Applications (IJACSA), Vol-3, No-11, pp.131-138, November 2012.
10. Md. Abul Hasnat, Jabir Mowla, Mumit Khan, Isolated and Continuous Bangla Speech Recognition: Implementation, Performance and application perspective, 2007.
11. Md. Farukuzzaman Khan, Ramesh Chandra Debnath. Comparative Study of Feature Extraction Methods for Bangla Phoneme Recognition, Rajshahi University, 2000.
12. Xinguang Li, Jiahua Chen, Zhenjiang Li. English Sentence Recognition Based on HMM and Clustering, American Journal of Computational Mathematics. 2013; 3:37-42.
13. Rahman M M, Bhuiyan MA. On segmentation and Extraction of Features from Continuous Bangla Speech Including Windowing, International Journal of Applied Research on Information Technology and Computing (IJARITAC), 2011; 2(2):31-40.