Isolated and Continuous Bangla Speech Recognition: Implementation, Performance and application perspective

Md. Abul Hasnat[1] Jabir Mowla[2] Mumit Khan[3]
[1,2,3] Department of Computer Science and Engineering, BRAC University,
66 Mohakhali, Dhaka, Bangladesh
e-mail: mhasnat@gmail.com, jabir_jr@yahoo.com, mumit@bracu.ac.bd

Abstract

Research on automatic speech recognition has been approach progressively since 1930 and the major advances are since 1980 with the introduction of the statistical modeling of speech with the key technology Hidden Markov Model (HMM) and the stochastic language model (B. H. Juang, 2005). However, the existing reported research works on Bangla speech recognition didn't yet incorporate the HMM technique and language model. This paper presents two different type of Bangla speech recognition from the implementation, performance and application perspective. We used HMM technique for pattern classification and also incorporate stochastic language model with the system. At the signal preprocessing level we perform adaptive noise elimination and end point detection. Spectral feature vectors such as Mel Frequency Cepstral Coefficients (MFCC) with the addition of first and second order coefficients are extracted from each speech wave signal. HMM is used for pattern classification. The system is implemented using the Cambridge Hidden Markov Modeling Toolkit (HTK) (S. Young, 2001-2005).

1. Introduction

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. Research in this area has attracted a great deal of attention over the past five decades where several technologies are applied and the efforts were made to increase the performance up to marketplace standard so that the users will have the benefit in a variety of ways. During this long research period several key technologies were applied where the combination of hidden Markov Model (HMM) and the stochastic language model produces high performance (B. H. Juang, 2005). Most of the research effort on recognizing bangla speech is performed using the ANN based classifier. No research work has been reported yet that uses the DTW technique and HMM based classifier and no language model is included with the existing research works.

An isolated-word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not. The continuous speech consists of continuous utterance which is representative of real speech. On the other hand a sentence constructed from connected words does not represent real speech as it is actually concatenation of isolated words. For Isolated word the assumption is that the speech to be recognized comprises a single word or phrase and to be recognized as complete entity with no explicit knowledge or regard for the phonetic content of the word or phrase. Hence, for a vocabulary of V words (or phrases), the recognition algorithm consisted of matching the measured sequence of spectral vectors of the unknown spoken input against each of the set of spectral patterns for V words and selecting the pattern whose accumulated time aligned spectral distance was smallest as the recognized word (L. Rabiner, 1993). The notion of isolated speech recognition can be extended for connected speech recognition if we consider a small vocabulary and solve the co-articulation problem that arises between words (S. Furui, 2001). In continuous speech recognition, continuously uttered sentences are recognized. The standard approach of continuous speech recognition is to assume a simple probabilistic model of speech production whereby a specified word sequence, W, produce an acoustic observation sequence, so that the decoded string has the maximum a posteriori probability (L. Rabiner, 1993). In continuous speech recognition it is very important to use sophisticated linguistic knowledge. The most appropriate units for enabling recognition success depend on the type of recognition and on the size of the vocabulary. Various units of reference templates/models from phonemes to words have been studied. When words are used as units, word recognition can be expected to be highly accurate; however it requires larger memory and more computation. Using phonemes as units does not greatly increase memory size requirements or computation (S. Furui, 2001). In our experiment we used word as a unit for isolated speech recognition and phoneme as a unit for continuous speech recognition.

Hidden Markov Model is powerful modeling technique for discrete state processes. The basic idea behind the HMM is that the observation sequence generated by the system exists in a finite number of states in the model, and at each time step the model makes a state transition and gives a probability as the output. More precisely, Hidden Markov Model is defined as the triple $\lambda := (\pi, A, B)$:. For real world implementation of Hidden Markov Model, three problems must be solved: the evaluation problem, the decoding problem and the learning problem (S. Furui, 2001). The *forward* algorithm is used to solve the evaluation problem, the *Viterbi* algorithm for the decoding problem, and all parameters are adjusted for solving the learning problem.

The outline of the system is as follows. We begin in section 2 with the related works that describes the past efforts on Bangla speech recognition. Section 3 discusses about the details of the overall system for both isolated and continuous speech recognition. Section 4 describes the implementation details. Section 5 presents the result analysis. Section 6 discusses the applications and at last we end up the discussion with the conclusion at section 7.

2. Related Works

Research on speech recognition has been started since 1930; however research work to recognize bangla speech has been started since around 2000. Here we will mention the research works found so far for Bangla speech recognition. K. Roy (2002) performed the recognition by Artificial Neural Network (ANN) using back propagation neural Network. They used DSP techniques to extract the features of speech signal. M. R. Hassan (2003) presents a phoneme recognition approach using ANN as a classifier. They calculated the RMS energy level as feature from the filtered digitized signal. A. H. M. Rezaul Karim (2002) presents a technique to recognized bangla phonemes using the Euclidian distance measure. Reflection coefficient and autocorrelations have been used as features. K. J. Rahman, (2003) presents 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 (2005) presents a Bangla ASR system that employed a three layer back-propagation Neural Network as the classifier. S. A. Hossain (2004) presents a brief overview of Bangla speech synthesis and recognition. A comparative study on the feature extraction methods are presented by M. F. Khan (2002).

3. Methodology / Overview of the Systems

The block diagram of canonic speech recognition system is shown in figure 1. We can subdivide the entire model into three major parts: speech data extraction or preprocessing, feature extraction, pattern recognition. Although the basic theory of both type of speech recognition system in pattern recognition approach is quite similar, however they applied different strategy for incorporating language model and dictionary into their entire system model. They also used different style of data representation for both of their training and recognition system.

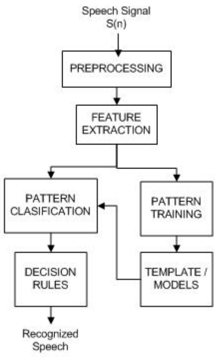

Figure 1: Block diagram for Speech Recognition

3.1 Speech Data Extraction or Preprocessing

In this stage, the first step is to record the speech data by a microphone in a specified format (wav file, 16000Hz and 16 bits). This wav data will be converting into a form that is suitable for further computer processing and analysis through a series of process that involves noise elimination and the speech end point detection process.

3.1.1 Noise Elimination

We used an adaptive filter to eliminate the noise from the recorded speech signal. A sample of the surrounding environment is used as input for the adaptive filter (M. R. Hassan, 2003). The extracted voice data will be simply the subtraction of the predicted noise from the speech signal.

3.1.2 End Point Detection

We used the generalized end point detection algorithm presented at (K. Roy, 2002; K. J. Rahman, 2003; M. R. Islam, 2005) in order to identify the starting and end point of the speech signal. The goal of this process is to detect the presence of voice and remove pauses and silences in the background noise. For continuous speech we use this algorithm only for the start and end point detection. However for isolated speech the intermediate noisy signal and unwanted signal within the speech is also eliminated. A background unvoiced sample of the current recording environment was taken. The sample was split into frames of known size and the energy was calculated for each frame. The frame with the maximum energy was taken as the threshold for the end point detection. Then the recorded voice sample was taken and split into frames in the same manner, each frame of the voice sample was compared with the maximum noise frame and at the point where the voice frame's energy had a lower value than the maximum noise frame, which would be the end point.

3.2 Feature Extraction

In this stage we extract meaningful unique features from the preprocessed speech data. From the comparison among features described at (M. F. Khan, 2002) we decided to use MFCC features. With the addition of the MFCC features we calculate energy, delta coefficient and acceleration coefficient. Finally the total number of features is 39 among those 12 MFCC, 1 energy, 13 first order derivative and 13 second order derivative. Feature vectors are extracted with Hamming window function of window length 25ms. For the elimination of unwanted frequency level a 26-filter bank channel is used with a pre-emphasis coefficient value 0.97.

3.3 Pattern Recognition

The tasks in this step are divided into two phase, Training and Recognition. We used Hidden Markov Model based classifier. An elaborate discussion on HMM model for speech recognition is briefly described at (L. Rabiner, 1989). In our training methodology we created word based HMM model for isolated speech and phoneme based HMM model for continuous speech recognition. We used the following training algorithm for creating the HMM models.

Step 1: Initialize $\lambda = (\pi, A, B)$
Step 2: Compute probabilities using λ.
Step 3: Adjust $\lambda = \lambda'$
Step 4: Repeat 2-3 until converge

3.4 Dictionary and Language model

Our approach towards the isolated speech recognition is quite simple, we used just a simple dictionary contains only the input-output-HMMmodelName and no language model is necessary. However, for continuous speech recognition we created a pronunciation dictionary that contains the input-output-pronouncing for each word entry where the pronunciation describes the sequence of HMMs that constitute each word. For each word the output is provided as Unicode sequence and the pronunciation is given with the consideration of phoneme as a unit. As in continuous speech recognition we recognize a sequence of words and that's why it is necessary to incorporate a language model. We used the Regular grammar modeling technique as our language model which has the properties like finite state model, small vocabulary and restricted grammar. We create a word level network that will typically represent the grammar which defines all the legal word sequence explicitly. The regular grammar model outputs only two values of probability; P(W)=1 where W valid in word network and P(W)=0 otherwise. The Task Grammar which defines all of the legal word sequences explicitly or a Word Loop which simply puts all words of the vocabulary in a loop and therefore allows any word to follow any other word. Word-loop networks are often augmented by a stochastic language model (S. Young, 2001-2005).

4. Implementation

We used HTK as the core engine for our speech recognizer due to its availability, portability and sophisticated facilities for speech analysis, HMM training, testing and results analysis (S. Young, 2001-2005). We used our own implemented algorithms discussed earlier for the preprocessing task and used HTK for the rest of the part with some specified parameters. The preprocessing task is similar for both type of recognition which will output the valid speech data from the recorded speech signal. Here we will briefly discuss the implementation difference of the isolated and continuous speech recognition from pattern recognition point of view. We followed (A K M Mahmudul Hoque 2006) for the implementation of isolated speech recognition.

4.1 Isolated speech recognition

4.1.1 Prepare Training Data

The first step is to label the speech data for each word of the dictionary. The label is same as the text that represents the spoken word. We take 5 to 10 samples for each word and save the labeled file as the word followed by the sample number (ex: "kolom_2.lab" is the labeled file of the word "kolom" for sample number 2). We use HSLab (S. Young, 2001-2005) tool for this task. The second step is to extract the feature vectors from the sample files. All the specifications for feature extraction (described in the previous section) are written into a configuration file. The HCopy (S. Young, 2001-2005) tool is used, which automatically extract the features according to the configuration and save into a file.

The next step is model training. For each word V in the vocabulary, we must build an HMM model i.e. we must estimate the model parameters (A, B, π) that optimize the likelihood of the training set observation vector of the V_{th} word (L. Rabiner, 1989). To create a model, first we have to choose a priori a topology for each HMM model. We choose a HMM prototype with 4 active states and 2 non-emitting states. The prototype is depicted in figure-2. Before beginning the training we initialized the HMM model for each word with the HInit (S. Young, 2001-2005) tool. After initializing, the models are trained with the feature data set. The training is an iterative process until all the models are reached to a convergence. We used HRest tool to re-estimate the model parameters iteratively. After the completion of all these tasks, we will have separate model for each word in the dictionary.

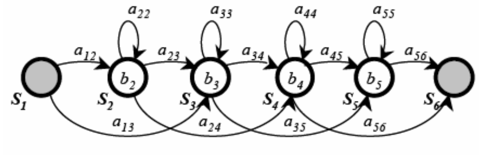

Figure 2: HMM prototype for each model

4.1.2 Recognition

The recognizer can recognize only the words defined in the dictionary. For each unknown word which is to be recognized, the preprocessing and feature analysis must be carried out, that means measurement of the observation sequence via a feature analysis of the speech corresponding to the word; followed by calculation of the model likelihoods for all possible HMM models; followed by selection whose model likelihood is highest. The probability computation step is generally performed using the Viterbi algorithm. We used HVite (S. Young, 2001-2005) tool to perform the recognition task and provide recognized output as text. In addition to the dictionary, extracted features and HMM List, HVite requires a word network as HTK convention. So, to build a word network we include a very simple grammar (see figure-5) with a single state as just the words in the dictionary. We used the HParse (S. Young, 2001-2005) tool to create the word network.

4.2 Continuous speech recognition

4.2.1 Prepare Training Data

Continuous Speech Recognition (CSR) involves connecting HMMs together in sequence. Each model in the sequence corresponds directly to the assumed underlying symbol. For CSR we moved towards the phoneme based HMM model. For this, first we designed the phoneme set for bangla language where we considered mono-phone as the phoneme unit. We have selected 47 mono-phones from our IPA chart (M. A. Hai, 2004). Then we created a set of phonetically balance (PB) sentences consist of total 52 sentences, transcribe those sentences with appropriate phone label and saved that as a HTK specified Master Label File (mlf) format (S. Young, 2001-2005). The recorded speech of the PB labeled sentences is considered as the training data for CSR. In our PB sentences we have listed total 1814 phonemes. Table-1 illustrates the phoneme distribution, where 1st column present the monophone in Bangla, 2nd column present the corresponding IPA symbol, 3rd column present the frequency of the monophone in the PB and the rest of the columns follow the first three columns. Next, acoustic analysis is performed on these training data to extract features. For this, the number of features, configuration parameters and the tool is exactly same as isolated speech recognition.

-	sil	101	খ	k^h	da	দ	d̪	33
অ	ɔ	161	গ	g	dah	ধ	d̪^{h}	12
আ	a	217	ঘ	g^h	p	প	p	39
ই	i	117	ঙ	ŋ	ph	ফ	p^h	12
ঈ	iː	3	চ	c	b	ব	b	67
উ	u	42	ছ	c^h	bh	ভ	b^h	14
ঊ	uː	1	জ	ɟ	m	ম	m	36
এ্যা	æ	1	ঝ	ɟ^h	z	য	ʤ	13
এ	e	165	ঞ	No	r	র	r	135
ও	o	61	ত	t̪	l	ল	l	52
আঁ	ã	5	থ	t̪^{h}	s	শ	ʃ	30
ইঁ	ĩ	2	দ	d̪	sh	স	s	23
উঁ	ũ	2	ধ	d̪^{h}	h	হ	h	28
ওঁ	õ	2	ন	n	ra	ঢ়	ɾ	4
এঁ	ẽ	4	ট	t	y	য়	j	22
ক	k	102	ঠ	t^h	2			

Table 1: Phoneme Distribution

The next step is model training. To create a model, here we choose a HMM prototype with 3 active states and 2 non-emitting states. We initialize the HMM models using HInit. Then we create a HTK specified Master Macro File (mmf) for all monophone using the prototype HMM file. Next we re-estimate the parameters using HRest tool. After the completion of all these tasks, we will have separate model for each phoneme.

4.2.2 Recognition

For a continuous speech signal to be recognized, the preprocessing and the feature extraction (using HCopy tool) is done first. Then the signal is recognized using the HVite tool with the assist of regular grammar based language modeling technique. We create a regular grammar and convert it to an intermediate form of decoding network using the HParse tool. Networks are specified using the HTK Standard Lattice Format (SLF). In the grammar we define the legal word sequences explicitly for constructing valid sentences of the language and also word loop which simply puts all words of the vocabulary.

5. Result Analysis

In this research work, we give emphasis to the inclusion of the HMM technique for recognizing Bangla speech as no such work have been seen and also to evaluate the performance from several aspect. We have taken a vocabulary of 100 words and test samples from 5 different speakers to observe the performance. For Isolated Speech Recognition we recorded the words for training in normal office environment where several samples (5-10) with little variations were taken for each word. The recognizer is capable of recognizing each spoken word existing in the dictionary only when the words are spoken by the same speaker and the mood of the speaker is same. However for different speaker the performance decreases to almost 20%. For continuous speech recognition we used the same 100 words for building the regular grammar. Table-2 shows the performance of both the recognition systems.

SR Type	Speaker Dependent	Speaker Independent
Isolated	90%	70%
Continuous	80%	60%

Table 2: Performance analysis

Recognizing continuous speech with ANN classifier has average accuracy rate of 73.36% (K. J. Rahman, 2003), for three layer Back-Propagation Neural Network the maximum accuracy rate is 86.67% (M. R. Islam, 2005), and spoken letter recognition by measuring Euclidian distance, which can recognize only the vowels, has an 80% accuracy rate (A H M. Rezaul Karim, 2002). In comparison, the recognizer presented in this paper has an average accuracy rate of 85%. The performance analysis reveals the importance about the improvement of the recognition with different speaker. Several studies on SR system emphasizes on the training data with varieties of speakers to increase the performance. So, next we should put our effort on collecting the training data from different speaker and observe the performance.

6. Applications

The entire domain where speech recognition technology can be applied are automatic translation, automotive speech recognition, dictation, hands-free computing: voice command recognition computer user interface, home automation, interactive voice response, medical transcription, mobile telephony, pronunciation evaluation in computer-aided language learning applications and robotics. In our research work we are considering the isolated speech recognition for commands & control, data entry, mobile telephony and home automation task. On the other hand continuous speech recognition can be used for speech to text conversion.

7. Conclusion

In this paper we concentrated on the research and development of a Bangla Speech Recognizer using the appropriate technique and tools. We have studied the past works and to the best of our knowledge this work is the first reported attempt to recognized Bangla speech using HMM Technique with the assist of stochastic language model. We observed that the language specification is not significant for ISR; however it has great importance for CSR specially the language specific issues are constructing the phoneme set, phonetically balance sentences and regular grammar for Bangla Language. This paper clearly describes the theory and implementation details of our entire development task using the HTK tool. This work can be extended to the further research on connected word recognition as an extension of isolated speech recognition and the performance measurement of the diphone or triphone based HMM model as an extension of continuous speech recognition.

8. Reference

A H M. Rezaul Karim, Md. S. Rahman, Md. Zafar Iqbal, “Recognition of Spoken Letters in Bangla”, Proc. of 6th ICCIT, Dhaka, 2002.

A. K. M. Mahmudul Hoque, "Bengali Segmented Speech Recognition System", Undergraduate Thesis Report, Computer Science and Engineering, BRAC University, May, 2006.

B. H. Juang and L. R. Rabiner, “Automatic Speech Recognition-A Brief History of the Technology”, Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

K. J. Rahman, M. A. Hossain, D. Das, A. Z. M. Touhidul Islam and Dr. M G. Ali, “Continuous Bangla Speech Recognition System”, Proc. 6th Int. Conf. on Computer and Information Technology (ICCIT), Dhaka, 2003.

K. Roy, D. Das and M G. Ali, "Development of the Speech Recognition System Using Artificial Neural Network”, Proc. of 5th ICCIT, 2002.

L. Rabiner. A Tutorial on Hidden Markov Model and Selected Applications of Speech Recognition, In Proceedings of IEEE, Vol-77, No-2, February 1989.

L. Rabiner and B. H. Juang, “Fundamentals of Speech Recognition”, 1st edition, Prentice Hall, New Jersey, 1993.

M. A. Hai, “Dhonibiggan O Bangla Dhonitotto”, 8th edition, Mollik Brothers, Dhaka, 2004.

M. F. Khan and R. C. Debnath, "Comparative Study of Feature Extraction Methods for Bangla Phoneme Recognition", Proc. 6th ICCIT, Dhaka, 2002.

M. R. Hassan, B. Nath and M. Ala Uddin Bhuiyan, “Bengali Phoneme Recognition: A New Approach”, Proc. 6th ICCIT, Dhaka, 2003.

M. R. Islam, A. Sayeed M. Sohail, M. W. H. Sadid, M. A. Mottalib, "Bangla Speech Recognition using three layer Back-Propagation Neural Network", Proc. of NCCPB, Dhaka, 2005.

S. A. Hossain, M. L. Rahman, M. F. Ahmed and M. Dewan, "Bangla Speech Analysis, Synthesis and Recognition: An Overview", Proc. of NCCPB, Dhaka, 2004.

S. Furui, "Digital Speech Processing, Synthesis and Recognition", 2nd Edition, Marcel Dekker Inc., New York, 2001.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, “The HTK Book”, 2001-2005 Cambridge University Engineering Departments, Website: http://htk.eng.cam.ac.uk/docs/docs.shtml.