Speech Analysis for Alphabets in Bangla Language: Automatic Speech Recognition

Asm SAYEM
School of Computer Science and Software Engineering, University of Wollongong, Australia
softwarist@yahoo.com

Abstract: This paper presents a technique for recognizing spoken letter in Bengali Language. We first derive feature from spoken letter. Mel-frequency cepstral coefficient (MFCC) has been used to characterize a feature. Dynamic time warping (DTW) employed to calculate the distance of an unknown letter with the stored ones. K-nearest neighbors (KNN) algorithm is used to improve accuracy in noisy environment.

Keywords: Automatic Speech Recognition (ASR), Bengali Alphabets, DTW, KNN, MFCC

1. Introduction:

Automatic recognition of spoken letter is one of the most challenging tasks in the field of computer speech recognition. The difficulty of this task is due to the acoustic similarity of many of the letters. Accurate recognition requires the system to perform fine phonetic distinctions. English and Bangla, two languages belonging to Indo-European Family, have spectacular similarities as well as differences in their phonemic systems. This paper is an overview of the status of Bangla speech processing.

Bangla (can also be termed as Bengali), which is largely spoken by the people all over the world, has been performed a very little research where many literatures in automatic speech recognition (ASR) systems are available for almost all the major spoken languages in the world. Although Bangla speakers' number is about 250 million today, which makes Bangla the seventh language (banglapedia, 2013), a systematic and scientific effort for the computerization of this language has not been started yet. The Bengali alphabet is a syllabic alphabet in which consonants all have an inherent vowel which has two different pronunciations, the choice of which is not always easy to determine and which is sometimes not pronounced at all. Some efforts are made to develop Bangla speech corpus to build a Bangla text to speech system (Hossain et al., 2007). However, this effort is a part of developing speech databases for Indian Languages, where Bangla is one of the parts and is spoken in the eastern area of India (West Bengal). But most of the natives of Bangla (more than two thirds) reside in Bangladesh, where it is the official language. Although the written characters of standard Bangla in both the countries are same, there are some sounds which are produced differently in different pronunciations of standard Bangla. Therefore, there is a need to do research on main stream of Bangla, and we confined this study on recognition of spoken letter only.

Some developments on Bangla speech processing or Bangla ASR can be found in (Hasnat et al., 2007; Stuckless, R 2006). For example, Bangla vowel characterization is done in (Hasnat et al., 2007); isolated and continuous Bangla speech recognition on a small dataset using hidden Markov models (HMMs) is described in (Karim et al., 2002); recognition of Bangla phonemes by Artificial Neural Network (ANN) is reported in (Roy et al.,2002; Rahman et al., 2003). Continuous Bangla speech recognition system is developed in (Hossain et al., 2007), while (Roy et al., 2002) presents a brief overview of Bangla speech synthesis and recognition. Most of the research effort on recognizing Bangla speech is performed using the ANN and LPC based classifier. No research work has been reported yet that uses the DTW and K-NN technique and MFCC based feature extraction.

In this paper, we build an ASR system for Bangla alphabets. We first develop a small size of database for bangle alphabets to achieve the goal. Then mel-frequency cepstral coefficient (MFCC) is used to extract feature from the input speech. After that extracted feature are saving as reference templates. Then real time input coefficient are compared to the reference templates using dynamic time warping (DTW) algorithm. And finally the output of DTW are inserted into the k-nearest neighbors (K-NN) based classifier for obtaining the word recognition performance.

2. Brief Description of Bangla:

Bangla, one of the more prominent Indo-Aryan languages, spoken by a population that now exceeds 250 million. Geographical distribution of Bangla-speaking population percentages are as follows: Bangladesh (over 95%), West Bengal (85%), the Indian States of Andaman & Nicobar Islands (26%), Assam (28%), and Tripura (67%). The global total includes those who are now in diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, United Arab Emirates, United Kingdom, and United States. Bangla has two literary styles: one is called Sadhubhasa (elegant form) and the other Chaltibhasa (commonly used form). The differences between the two styles are not huge and involve mainly forms of pronouns and verb conjugations.

The origin of modern Bangla derives from Vedic Sanskrit (1500 BC – 1000 BC). Bangla writing system evolved from the Brahmi script, which is closely related to the Devanagari alphabet, from which it started to diverge in the 11th century AD. The current printed form of Bangla alphabet first appeared in 1778 when Charles Wilkins developed printing in Bangla. Its script includes two types of symbols, the letters (Varnas) and signs (Cinhas). The letters are vowels, consonants and conjunct consonants. English has 36 phonemes while Bangla has 37. The signs used are vowel signs and prosodic signs. When the vowel is used with the consonant except in the first position, it is written in sign symbols.

3. Analysis of Bangla Alphabets:

Bangla is an intonation language having no tone, accent or stress (Rahman, 1992). Normal Bangla speech is heard more or less on a monotone with slight rise and fall of pitch and loudness within the sentence. Just as the written form of a language is a sequence of elementary alphabet, speech is also a sequence of elementary acoustic sounds or symbols known as phonemes that convey the spoken form of a language (Hai, 1985). Bangla is heard more or less in monotone. But the interesting point in Bangla is nasality. Moreover, there are pure vowels, reduced vowels, diphthongs, semivowel and triphthongs. There are 7 vowels and all the vowels in Bangla have their corresponding nasal counterparts. Nasalization of vowel changes meaning of some words in Bangla. There are 29 simple consonants, 19 diphthongs and two semivowel phonemes. There are monosyllabic to seven syllabic words in Bangla. Vowel is the nucleus of a syllable. Some vowel itself has meaning and is treated as a word.

4. Research propositions and Analytical fields:

These ASR systems could not be able to provide enough performance at any time and everywhere. One of the reasons is that the Acoustic Models (AMs) of a Hidden Markov Model (HMM)-based classifier include many hidden factors such as speaker-specific characteristics that include gender types and speaking styles. It is difficult to recognize speech affected by these factors, especially when an ASR system contains only a single acoustic model. One solution is to employ multiple acoustic models, one model for each type of gender. Though the robustness of each acoustic model prevails to some extent, the whole ASR system can handle gender effects appropriately. Most of these ASR systems use Mel Frequency Cepstral Coefficients (MFCCs) of 39 dimensions (12-MFCC, 12-ΔMFCC, 12-ΔΔMFCC, P, ΔP and ΔΔP, where P stands for raw energy of the input speech signal). Here, Hamming window of 25 ms is used for extracting the feature. The value of pre-emphasis factor is 0.97. Although these standard MFCCs are prevalent to current ASR system, but these features do not provide better performance because frequency domain information are not incorporated within the feature vector during the extraction process.

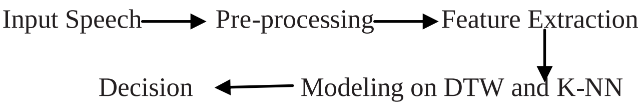
Recently, dynamic parameters such as velocity and acceleration coefficients of speech showed its necessity for embedding them as features to resolve the coarticulation effect due to widening the context window size. Though the coarticulation effects can be solved by incorporating the triphone models (Young et al., 2005), but a large-scale speech corpus is required to negotiate all the triphones. Besides, the training of triphone models incurs many complexities in HMM based classifiers. Contemporary Bangla automatic speech recognition suffers from some difficulties: (1) lack of large scale speech corpus, (2) unavailability of labelled speech data, and (3) insufficient research opportunities though more than 250 million people speak in Bangla as their native language. These problems should be reduced immediately for constructing an ASR system for recognizing the voice. The followings explicate the objectives of the chapter in details:

- To construct a phoneme recognizer based on standard MFCC features
- To incorporate time and frequency domain information, new feature called local feature instead of standard MFCC is extracted from an input speech for an ASR system
- To extract phoneme probabilities based on time delay neural network by using MFCCs as input feature

Conventional Automatic Speech Recognition (ASR) systems use stochastic pattern matching techniques, where a word candidate is matched against word templates represented by Hidden Markov Models (HMMs) (Young et al., 2005). Although these techniques have a fair performance in limited applications, they suffer from huge computational cost at classifier stages, and also they always reject a new vocabulary or so-called Out-Of-Vocabulary (OOV) word. On the other hand, a traditional segmentation-based phone decoding technique can be used to solve these problems, but, until now, its recognition accuracy is far from sufficient performance.

5. Speech Recognition System:

Speech is input via microphone and its analog waveform is digitized. The job of the recognition system is to derive necessary information from the waveform needed to make the correct decision. The process of recognition system operation typically consists of two phases first one is training and finally recognition. In the training phase, data for known classes are fed to the system. In the recognition phase, the system computes the features of pattern for unknown input and identifies the input with the class whose reference pattern matches these features most closely.



There are several alternative candidates those can be used as feature for speech. In the current research, mel-frequency cepstral coefficient (MFCC) is used for extracting features. Patterns are vectors of features. In pattern recognition DTW and K-NN are used for take decision.

5.1 Processing:

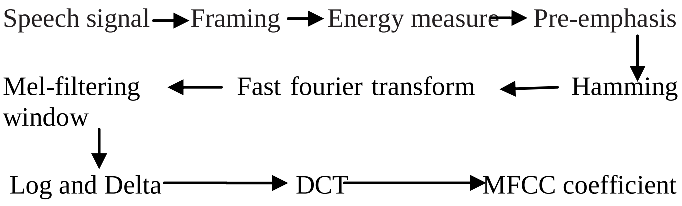
In this stage, the first step is to record the speech data by a microphone in a specified format (wav file, 16000Hz and 16 bits). This wav data will be converting into a form that is suitable for further computer processing and analysis through a series of process that involves noise elimination and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

5.1.1 End Point Detection:

We used the generalized end point detection algorithm which accepts an audio sample as input and returns a trimmed down version with non-speech sections trimmed off. Also known as voice activity detection, it utilizes the algorithm due to (Rabiner & Sambur, 1975).

5.2 Feature Extraction:

In this paper the most important thing is to extract the feature from the speech signal. The speech feature extraction in a categorization problem is about reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. In this project we are using the mel-frequency cepstral coefficient (MFCC) technique to extract features from the speech signal. Figure below shows the complete pipeline of Mel Frequency Cepstral Coefficients (MFCC).



5.2.1 Framing:

The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. In this study we used 25 ms fo frame and overlap of 10 ms. i.e. 400 samples in each frame and 160 samples of overlap. For each of the frame we calculate 39 MFCC coefficient ((Energy+12) MFCC + Delta MFCC + Delta-Delta MFCC)

5.2.2 Energy Measure:

The logarithmic frame energy measure (logE) is computed after the offset compensation filtering and framing for each frame:

$$\mathrm{logE} = \ln \{\Sigma^{N}_{1} S(i)^{2}\}$$

Here N is the frame length and S_{of} is the offset-free input signal. A floor is used in the energy calculation which makes sure that the result is not less than -50. The floor value (lower limit for the argument of ln) is approximately 2e-22.

5.2.3 Pre-Emphasis:

The speech signal s(n) is sent to a high-pass filter: $s_2(n) = s(n)-0.97.s(n-1)$

Where s2(n) is the output signal and the value of a is usually between 0.9 and 1.0. We used 0.97 as the value of a. The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.

5.2.4 Windowing:

When signal or any other function is multiplied by a window function, the product is zero valued outside the interval. The windowing is done to avoid problems due to truncation of the signal. Frame signal is tapered by hamming window to avoid discontinuities at the ends. If the signal in a frame is denoted by s(n), n = 0,…N-1, then the signal after Hamming windowing is s(n)*w(n), where w(n) is the Hamming window defined by:

$$w(n,\alpha) = (1-\alpha) - \alpha\cos(2\pi n/(N-1)),\ 0 \le n \le N-1$$

In practice, the value of α is 0.46.

5.2.5 FFT & Mel-Filtering:

A Fast Fourier Transform (FFT) is an efficient algorithm to compute the Discrete Fourier Transform (DFT) and its inverse. A DFT decomposes a sequence of values into components of different frequencies but the frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. We multiple the magnitude frequency response by a set of 24 triangular bandpass filters to get the log energy of each triangular bandpass filter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency f by the following equation:

$$mel(f)=1125*\ln(1+f/700)$$

Mel-frequency is proportional to the logarithm of the linear frequency, reflecting similar effects in the human's subjective aural perception. The reasons for using triangular bandpass filters are Smooth the magnitude spectrum such that the harmonics are flattened in order to obtain the envelop of the spectrum with harmonics. This indicates that the pitch of a speech signal is generally not presented in MFCC. As a result, a speech recognition system will behave more or less the same when the input utterances are of the same timbre but with different tones/pitch. And it Reduce the size of the features involved.

5.2.6 Log & Delta Cepstrum:

The energy within a frame is also an important feature that can be easily obtained. Hence we usually add the log energy as the 13rd feature to MFCC. Delta Cepstrum is used to catch the changes between the different frames. Delta Cepstrum is used to catch the changes between the different frames. It is also advantageous to have the time derivatives of (energy+MFCC) as new features, which shows the velocity and acceleration of (energy+MFCC). The equations to compute these features are:

$$\Delta Cm(t)= [\Sigma_{\pi=-M}^{M} C_m(t+\pi)\pi] / [\Sigma_{\pi=-M}^{M} \pi^2]$$

The value of M is usually set to 2. If we add the velocity, the feature dimension is 26. If we add both the velocity and the acceleration, the feature dimension is 39.

5.2.7 DCT:

In this step, we apply DCT on the 24 log energy Ek obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients. The formula for DCT is shown next.

$C_m=\Sigma_{k=1}^{N}\cos[m*(k-0.5)*\pi/N]*E_k$, m=1,2, ..., L Where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. Usually we set N=24 and L=12. Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC.

5.3 DTW & K-NN:

Speech is a time-dependent process. Hence the utterances of the same word will have different durations, and utterances of the same word with the same duration will differ in the middle, due to different parts of the words being spoken at different rates. To obtain a global distance between two speech patterns (represented as a sequence of vectors) a time alignment must be performed. In this type of speech recognition technique the test data is converted to templates. The recognition process then consists of matching the incoming speech with stored templates. The template with the lowest distance measure from the input pattern is the recognized word.

If D(i,j) is the global distance up to (i,j) and the local distance at (i,j) is given by d(i,j)

$$D(I,j)=\min[D(i-1,j-1),D(i-1,j),D(i,j-1)]+ d(i,j)$$

The final global distance D(n,N) gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score. In order to improve the recognition accuracy in noisy environment we use K-NN algorithm. K-nearest neighbor algorithm (K-NN) is a method for classifying objects based on closest training examples in the feature space. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). We use k=3 in our study.

6. Experiments and Results:

In this research work, we give emphasis to the inclusion of DTW and K-NN technique for recognizing Bangla speech as no such works have been seen and also to evaluate the performance from several aspects.

The experiment was done using digital computer with 2.8GHz speed and 512MB memory machine. The sound samples were taken in a room environment and different speech processing techniques were applied to make the samples suitable for feature extraction and recognition. We have taken a vocabulary of 10 consonants and test samples from 2 different speakers to observe the performance. The recognizer is capable of recognizing each spoken consonants existing in the database 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 10 to 15%. 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 90%. Spoken letter recognition by using only DTW is 80% but use of K-NN it increases almost 10%. Table-1 shows the performance of using K-NN.

Table. 1 Accuracy rate

Classifier	Speaker dependent	Speaker independent
DTW	78%	60%
DTW & KNN	86%	75%

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.

7. Application:

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.

8. Summary:

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 DTW and K-NN Technique with the assist of template language model. Scientists achieved remarkable success in speech recognition for many languages. In English, continuous speech can be recognized with accuracy rate more than 95% (Hai, 1985). Unfortunately in Bengali accuracy is about 85% (Karim et al., 2002). In this paper, we discussed how a spoken letter can be recognized. And the level of accuracy is almost 90%. By ensuring the perfection (i.e. noise free) of the recorded signal it is possible to increase the accuracy of recognition. Hope our effort will help to take the research on Speech Recognition one step toward continuous speech recognition in Bangla with higher accuracy.

Acknowledgments:

I do acknowledge tremendous support of Professor Jim Tognolini, Senior Vice President, Research & Assessment Office of the Chief Education Advisor, Pearson PLC & Senior Research Fellow, University of Oxford, and Professor Peter Croll, Pro Vice Chancellor (Research), Southern Cross University & Head of School, School of Computer Science and Software Engineering often beyond their obligation. A special thanks to Akramul Kabir, Professor of Linguistics in Bangladesh, for the linguistic discussions related to the Bangla script. Thanks also to Dr. Omar Ogarcia, Sociologist Farjana Rahman and Dr. Michael Cowling for their true inspiration and technical guidance.

References:

i. Hai, A. 1985, Dhvani-vignan o bangla dhvani tatta. Mullick Brothers, Dhaka, Bangladesh.

ii. Hasnat, M. A., Mowla, J. & Khan, M. 2007, "Isolated and continuous bangla speech recognition: Implementation performance and application perspective", In proceedings of the International Symposium on Natural Language Processing (SNLP), Hanoi, Vietnam.

iii. Hassan, M. R., Nath, B. & Bhuiyan, M. A. 2003, "Bengali phoneme recognition: a new approach", In proceedings of the 6th International Conference on Computer and Information Technology, Dhaka, Bangladesh.

iv. Hossain, S. A. 2008, Analysis and synthesis of bangla phonemes for computer speech recognition, PhD dissertation, University of Dhaka, Dhaka, Bangladesh.

v. Hossain, S. A., Rahman, M. L. & Ahmed, F. 2007, "Bangla vowel characterization based on analysis by synthesis", In proceedings of WSEAS 5th International Conference, Corfu, Greece.

vi. Hossain, S. A., Rahman, M. L., Ahmed, F. & Dewan, M. 2004, "Bangla speech synthesis, analysis, and recognition: an overview", In proceedings of the NCCPB, Dhaka, Bangladesh.

vii. Houque, A. K. M. M. 2006, "Bengali segmented speech recognition system", Undergraduate thesis, BRAC University, Dhaka, Bangladesh.

viii. Karim, R., Rahman, M. S. & Iqbal, M. Z. 2002, "Recognition of spoken letters in bangla", In proceedings of the 5th ICCIT conference, Dhaka, Bangladesh.

ix. Khan, Sameer ud Dowla 2010, Bengali (Bangladeshi Standard) Journal of the International Phonetic Association, 40:2, pp 221-225.

x. Mukherjee, S. & Das Mandal, S.K. 2012, "A bengali speech synthesizer on android", In proceedings of the 1st workshop on Speech and Multimodal interaction in Assistive Environments, Association for Computational Linguistics, ACL, Jeju, Republic of Korea.

xi. *Odden, D. 2005, Introducing Phonology, Cambridge University Press, Cambridge.*

xii. *Online, as accessed in October 2013 from http://www.banglapedia.org*

xiii. *Quatieri, T. F. 2002, Discrete-time speech signal processing, Prentice Hall, New York.*

xiv. *Rabiner, L.R. & Sambur, M.R. 1975, 'An Algorithm for Determining the Endpoints of Isolated Utterances', Bell System Technical Journal, 54: 2, pp 297-315.*

xv. *Rahman, K. J., Hossain, M. A., Das, D., Islam, T. & Ali, M. G. 2003, "Continuous bangla speech recognition system", In proceedings of the 6th ICCIT conference, Dhaka, Bangladesh.*

xvi. *Rahman, M. 1992, Power spectrum and formant extraction of bangla speech, Rajshahi University, Rajshahi, Bangladesh.*

xvii. *Roy, K., Das, D. & Ali, M. G. 2002, "Development of the speech recognition system using artificial neural network", In proceedings of the 5th ICCIT conference, Dhaka, Bangladesh.*

xviii. *Seddiqui, M. H., Azim, M.A., Rahman, M.S. & Iqbal, M.J. 2002, Algorithmic approach to synthesize voice from bangla text, In proceedings of 5th ICCIT conference, Dhaka, Bangladesh.*

xix. *Sen, Sukumar 2002, Bhashar Itibritto, Kolkata, Ananda Publishers, India.*

xx. *Stuckless, R. 2006, Special Applications of Automatic Speech Recognition (ASR) with Deaf and Hard-of- Hearing People, American Annals of the Deaf.*

xxi. *Young, S., Evermann G., Gales M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland P. 2005, HTK Book, University Engineering Department, Cambridge, UK.*