Myanmar Continuous Speech Recognition System Based on DTW and HMM

Ingyin Khaing
Department of Information and Technology
University of Technology (Yatanarpon Cyber City),near Pyin Oo Lwin, Myanmar

Abstract- This paper presents automatic speech recognition for continuous speech in Myanmar Language. Actually, a computer or a machine is not expected to understand what is uttered. But it is expected to be controlled via speech or to transcript the acoustic signal to symbols. This system will also address the issue of automatic word/sentence boundary detection in both quiet and noisy environments. Combinations of LPC, MFCC and GTCC techniques are used in feature extraction. MFCC features give the good discrimination of speech signal. LPC provides an accurate estimate of the speech parameters and it is also an efficient computational model of speech. DTW is used in the feature clustering and HMM is used in the recognition process. The HMM method is extended by combining it with the DTW algorithm in order to combine the advantages of these two powerful pattern recognition technique.

Keywords – Speech recognition, feature extraction, DTW, HMM

I. INTRODUCTION

Speech recognition is the process of automatic extracting and determining linguistic information conveyed by a speech wave using computers. Speech recognition has tremendous growth over the last five decades due to the advances in signal processing, algorithms, new architectures and hardware. Speech processing and recognition are intensive areas of research due the wide variety of applications. Speech recognition is involved in our daily life activities like mobile applications, weather forecasting, agriculture, healthcare, video games etc.

Speech recognition systems have been developed for isolated words in Myanmar language. Isolated word recognition, in which each word is surrounded by some sort continuous of pause, is much easier than recognizing continuous speech, in which words run speech into each other and have to be segmented. Continuous speech tasks themselves vary greatly in difficulty.

In speech recognition systems, the pre-processing of the speech signal is a key function for extracting and coding efficiently the meaningful information in the signal. In this system, LPC, MFCC and GTCC feature extraction technique are used. The basic idea of LPC is to predict the current value of the signal using a linear combination of previous samples, each weighted by a coefficient. Mel Frequency Cepstral Coefficients (MFCC) is based on the short-term analysis, and thus from each frame a MFCC vector is computed. Gammatone Cepstral Coefficient (GTCC) technique is based on the Gammatone filter bank, which attempts to model the human auditory system as a series of overlapping bandpass filters.

LPC is used for both noisy and clean environment, MFCC and GTCC is similar recognition in clean environment and GTCC is better in noisy environment [1]. DTW quite efficient for isolated word recognition and can be adapted to connected word recognition. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

This paper is organized as the follows. Section 1 is the introduction, Section 2 explains about the related work. In addition, Section 3 investigates about the overview of the standard Myanmar language. Proposed system design is described in Section 4. The conclusion is summarized in Section 5.

II. RELATED WORK

In [2], the author has used features based on wavelet transform to recognize five Malayalam words recorded by different persons. The feature vector consists of statistical metrics derived from the wavelet coefficients: mean, standard deviation, energy, kurtosis and skewness. The classifier used in this paper is layered feed-forward artificial neural network along with back propagation algorithm.

In [3], Charles and Devaraj presented enhanced speech recognition in Tamil with speaker independent, device independent system with continuous recognition. In their system, the spoken sentence is represented as sequence of independent acoustic phonetic units. The system recognized the spoken queries in the context of many applications

as Web Browsing etc. Initially the speech signal is converted into a sequence of feature vectors based on spectral and temporal measurements. Then the Acoustic models represent the sub-word unit's phonemes as a finite state machine in which, states model spectral structure and transition model temporal structure. HMM in Acoustic Modeling is used in pattern matching process.

In [4] the authors study the problem of speech recognition in the presence of non stationary sudden noise, which is very likely to happen in home environments. As a model compensation method for this problem, they investigated the use of factorial hidden Markov model (FHMM) architecture developed from a clean-speech hidden Markov model (HMM) and a sudden-noise HMM.

In [5], Petrea presents an innovative training strategy for HMM based ASR systems and extensive experimental results obtained for isolated words recognition in Romanian Language. The paper also shows how the HMM system can be tuned up to issue better recognition rates on a 50000 words (among which 10000 are different) database.

III. OVERVIEW OF STANDARD MYANMAR LANGUAGE

3.1 *Tones*

The most important feature of the Myanmar language is its use of tone to convey the lexical meaning of the syllables. Myanmar tones can be divided into two groups: static and dynamic. The static group consists of three tones (mid, low, and high) whereas the dynamic group consists of two tones (falling and rising).

3.2 Stress

The syllable in a word produced with a higher degree of respiratory effort is referred to as "stress." The stressed syllables are usually louder, longer, and higher in pitch than unstressed syllables. The placement of stress on a word in Myanmar is linguistically significant and governed by rules including the monosyllabic word rule and the polysyllabic word rule. For monosyllabic words, all content words are stressed, whereas all grammatical words are unstressed. However, monosyllabic unstressed words when spoken in isolation or emphasized can be stressed as well. For polysyllabic words, stress placements are determined by the number of syllables as well as the structure of the component syllables in the word. The primary stress falls on the final syllable of a word. The secondary stress is determined by the position of the remaining syllables and whether or not they are linker or non-linker syllables.

3.3Vowels and Consonants

The Myanmar alphabet consists of 33 letters and 12 vowels, and is written from left to right. It requires no spaces between words, although modern writing usually contains spaces after each clause to enhance readability. The latest spelling authority, named the *Myanma Salonpaung Thatpon Kyan* (□□□□□□□□□□□□□□□□□□□□□□□□□□□□□), was compiled in 1978 at the request of the Burmese government. Some of the spoken words and their International Phonetic Alphabet (IPA) format are shown in table (1).

Table 1. IPA code for Myanmar words

IPA	Words in Myanmar	IPA format
ʔ	အညာသား	[ʔəɲàðá]
h	ဟုတ်	[hou̯ʔ]
k	ကုန်	[kòʊɴ]
kʰ	ခုန်	[kʰòʊɴ]
l	လုပ်	[lou̯ʔ]
l̥	လျှပ်	[l̥ou̯ʔ]
ə	ခလုတ်	[kʰəlou̯ʔ]
ˋ	ငါ	[ŋà]
˷	ငံ့	[ŋa̰]

IV. PROPOSED SYSTEM

In this system, human speech is taken as input to the system. First human speech is decoded into signals for digital processing. Human speech may contain unvoiced area and each word is separated with pauses (unvoiced area) and endpoint detection is applied to remove unvoiced area between segment words. Then, important features of speech signals are extracted by linear predicted coding (LPC), Mel frequency Cepstral Coefficients (MFCC) and Gammatone Cepstral Coefficient (GTCC) approach. Feature vectors are clustered using DTW to solve the lack of discrimination in the Markov models. Then phoneme model is built with features extracted from speech signals in training database. For the input speech, HMM evaluation is performed to get recognized word. Then words are composed to get the sentence in text. Finally, input speech is transformed into text structure and displayed as output. The proposed system design is shown in Figure 1.

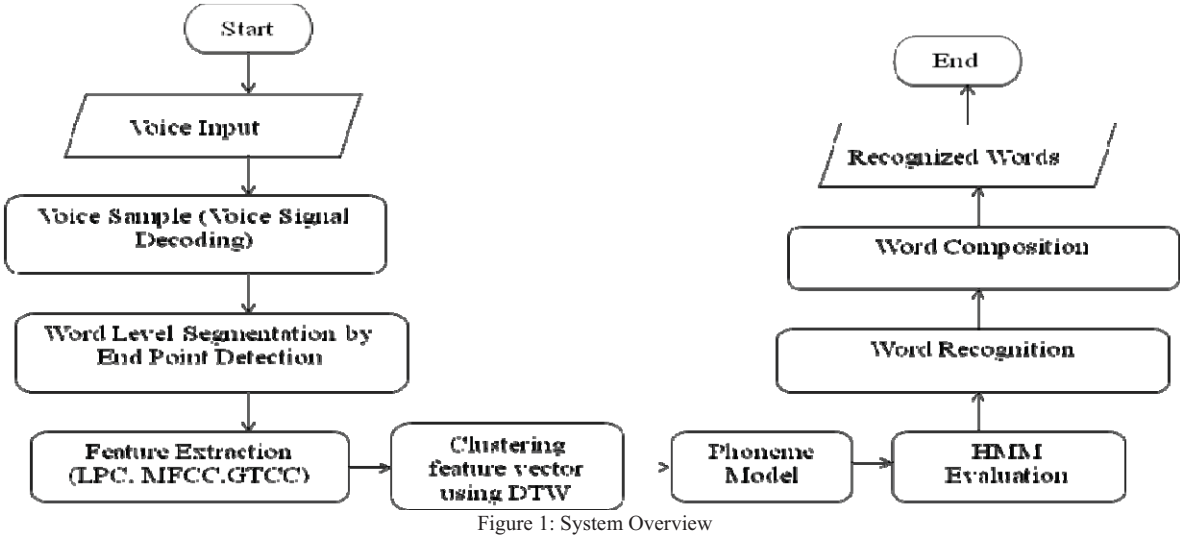

Figure 1: System Overview

4.1Signal Decoding

Audio files are saved in the encoded format. Voice signals are extracted after decoding. Sampling rate may be varied based on the voice level and gender of the speaker. In this system, input samples are changed to the sample rate 8000.

4.2Word Segmentation

Humans generally do not produce very short duration sounds. Therefore each speech segment should have a certain minimum length λ_i. For any i^{th} speech segments $\lambda_i = e_i - s_i$ where
s_i - corresponds to the starting point of the i^{th} frame in the thresholded entropy profile ξ'
e_i - corresponds to the ending point of the i^{th} frame in the thresholded entropy profile ξ'
The lambda corresponds to the shortest phoneme or phone in the vocabulary of the recognition engine and is also a function of the sampling frequency. The intra-segment distance d_{ij} is required because frequently there may be spurious speech segments that satisfy the first criterion. It will be necessary to merge two such speech segments into one larger segment. This happens frequently with words. if $\lambda_i < \kappa$ and $d_{ij} > \delta$, then the i^{th} segment is discarded. if (λ_i or λ_j) $> \kappa$, $d_{ij} > \delta$ and $\lambda_i + \lambda_j < \theta$, then the two segments are merged, and anything between the two segments that was previously left, is made part of the speech [6]. Figure 2 shows the relationship between adjacent speech segments.

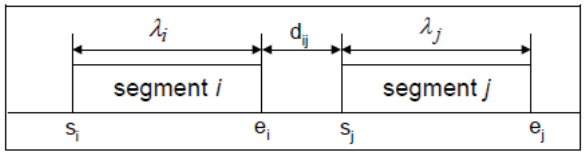

Figure 2. Word Segmentation

4.3Feature Extraction

The purpose of feature extraction is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing. This is often referred as the signal-processing front end. The speech signal is a slowly time varying signal (it is called quasi-stationary). When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. The sample speech signal is shown in Figure 3.

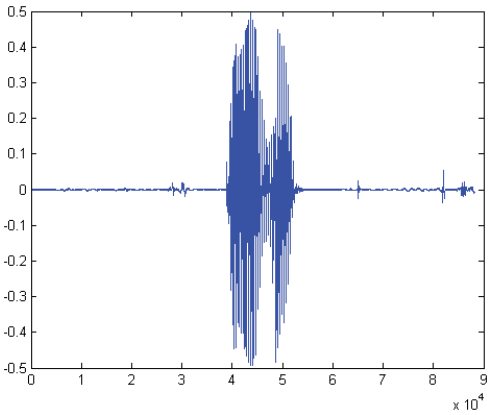

Figure 3. An example of speech signal

1) Mel-Frequency Cepstral Coefficients

The features in this system used are the Mel-frequency Cepstral Coefficients (MFCC) which has been the dominant features for recognition from a long time. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. MFCC includes following features.

- Pre-emphasis, Hamming windowing and FFT
- Mel scale Filter Bank
- Logarithmic compression

Discrete Cosine Transform (DCT)

1) LPC

LPC of speech 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 idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. The basic steps of LPC processor include the following:

- Pre emphasis
- Frame Blocking
- Windowing
- Autocorrelation Analysis
- LPC Analysis
- LPC Parameter Conversion to Cepstral Coefficients

The cepstral coefficients are the features that are extracted from voice signal.

2) GTCC

The basic steps of GTCC processor include the following:

- Hamming windowing and FFT
- Gammatone Filterbank
- Equal-loudness curve
- Logarithmic compression

Discrete Cosine Transform (DCT)

A. Recognition Phase

In the recognition phase, this system uses combination of DTW and HMM in order to increase model discrimination.

1) Dynamic Time Wrapping

Dynamic Time Warping is one of the pioneer approaches to speech recognition. Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed [7].

DTW operates by storing a prototypical version of each word in the vocabulary into the database, then compares incoming speech signals with each word and then takes the closest match. But this poses a problem because it is unlikely that the incoming signals will fall into the constant window spacing defined by the host. For example, the password to a verification system is "HELLLO". When a user utter “HEELLOO”, the simple linear squeezing of this longer password will not match the one in the database. This is due to the mismatch spacing window of the speech “HELLLO”.

Let $X(x_1, x_2, \ldots, x_n)$ and $Y(y_1, y_2, \ldots, y_m)$ be two series with the length of n and m, respectively, and an $n \times m$ matrix M can be defined to represent the point-to-point correspondence relationship between X and Y, where the element M_{ij} indicates the distance $d(x_i, y_j)$ between x_i and y_j. Then the point-to-point alignment and matching relationship between X and Y can be represented by a time warping path $W = <w_1, w_2, \ldots, w_k>$, $\max_{(m, n)} \leq K < m+ n - 1$, where the element w_k (i, j) indicates the alignment and matching relationship between x_i and y_j. If a path is the lowest cost path between two series, the corresponding dynamic time warping distance is required to meet

$$DTW(X,Y) = \min_{W}\{\sum_{K=1}^{K} d_k, W =< w1, w2, ..., w_k >\}$$

Where $d_k = d(x_i, y_j)$ indicates the distance represented as $w_k = (i, j)$ on the path W. Then the formal definition of dynamic time warping distance between two series is described as

$$DTW(<>,<>) = 0$$

$$DTW(X,<>) = DTW(<>,Y) = \infty$$

$$DTW(X<Y) = d(x_i, y_i) + \min\{DTW(X, Y[2:-]), DTW(X[2:-], Y), DTW(X[2:-], Y[2:-])\}$$

where < > indicates empty series,$[2 : -]$ indicates a sub array whose elements include the second element to the final element in an one-dimension array, $d(x_i, y_j)$ indicates the distance between points x_i and y_j which can be represented by the different distance measurements, for example, Euclidean Distance. The DTW distance of two-time series can be calculated by the dynamic programming method based on accumulated distance matrix, whose algorithm mainly is to construct an accumulated distance matrix

$$r(i, j) = d(x_i, y_i) + \min\{r(i-1, j), r(i, j-1), r(i-1, j-1)$$

Any element r (i, j) in the accumulated matrix indicates the dynamic time warping distance between series $X_{1:i}$ and $Y_{1:j}$. Series with high similar complexity can be effectively identified because the best alignment and matching relationship between two series is defined by the dynamic time distance.

2) Hidden Markov Model

A Hidden Markov Model (HMM) 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 processes. In

other words, the HMM models a sequence of observations as a piecewise stationary process. Over the past years, Hidden Markov Models have been widely applied in several models like pattern, or speech recognition. The HMMs are suitable for the classification from one or two dimensional signals and can be used when the information is incomplete or uncertain [7]. The following figure shows HMM structure of the Myanmar phoneme model.

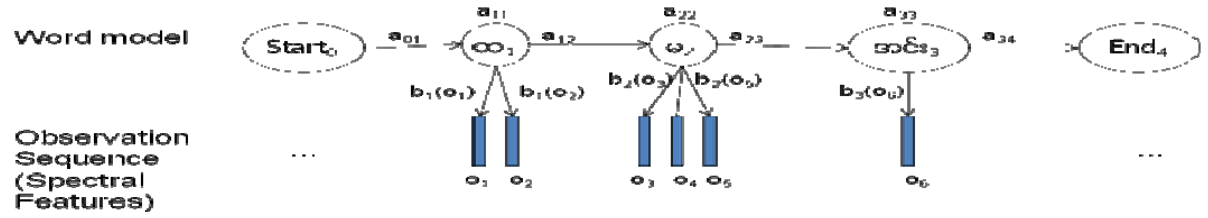

Figure 4. HMM structure of the phoneme

V. CONCLUSION

The interaction between a human and a computer, which is similar to the interaction between humans, is one of the most important and difficult problems of the artificial intelligence. This system presents speech recognition system in Myanmar language. In Myanmar language, there is only isolated word recognition system. In this system continuous word recognition is presented. The use of multiple speech feature extraction algorithms can improve the process for noisy environment. With the combination of HMM and DTW can improve the accuracy of recognition phase since DTW fixes the generalization nature of HMM.

REFERENCES

[1] Abdulla, W., Auditory based feature vectors for speech recognition systems, Advances in Communications and Software Technologies,N. E. Mastorakis & V. V. Kluev, Editor. WSEAS Press. pp 231-236, 2002.

[2] Ambalathody, P. (2010): Main Project-Speech Recognition Using Wavelet Transform.Internet:www.scribd.com/doc/36950981/Main-Project-Speech-Recognition-using-Wavelet-Transform, unpublished.

[3] A.P.Henry Charles & G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore.

[4] Betkowska, A.; Shinoda, K.; Furui, S.(2007):Robust Speech Recognition Using Factorial HMMs for Home Environments, Hindawi.

[5] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.

[6] Octavian Cheng, Waleed Abdulla, Zoran Salcic, "Performance Evaluation of Front-end Processing for Speech Recognition ", School of Engineering Report No. 621

[7] Rabiner, L. and Juang, B., Fundamentals of speech recognition. Prentice Hall, Inc., Upper Saddle River, New Jersey, 1993.