A Speech Recognition System for Myanmar Language

Ingyin Khaing[#1], K Zin Lin[*2]

[#]*University of Technology (Yatanarpon Cyber City), ICT Department,*
Pyin Oo Lwin, Myanmar
[1]ingyinkhaing@gmail.com

[*] *University of Computer Studies, Hardware Department,*
Yangon (Bahan Campus), Myanmar
[2]kzinlin78@gmail.com

Abstract— **This paper presents speech processing and recognition system for Myanmar continuous language. The speech segmentation methods are based on two simple speech features, namely time domain features and frequency features. To detect the word boundaries, dynamic thresholding method is applied. Then, important features of speech features are extracted by LPC (linear predictive coding) and GTCC (gamma tone cepstral coefficient) approach. K-means is used in feature clustering and HMM is used in recognition process. All the algorithms used in this work are implemented in Matlab. The system obtained the average word error rate of 1.18 and 0.5 with LPC and GTCC feature extraction techniques respectively.**

***Keywords*— Automatic Speech Segmentation, Short time Energy, Spectral Centroid, LPC, GTCC, 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 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 each word is surrounded by some sort of 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 task is greatly in difficult.

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 and GTCC feature extraction techniques 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. 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 and GTCC is better in noisy environment [1]. 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. Background theory is described in Section 2. The Proposed System Design is detailed in Section 3. Experimental study is presented in Section 4 and the Conclusion is summarized in Section 5.

II. BACKGROUND THEORY

In general, the following steps are executed:

A. Word Level Segmentation

B. Feature Extraction

C. HMM evaluation

A. Word Level Segmentation

1) Speech Acquisition

Speech acquisition is acquiring of continuous Myanmar speech sentences through the microphone. Speech capturing or speech recording is the first step of implementation. Recording has been done by native female speaker of Myanmar. The sampling frequency is 44100 Hz; window length is 512, and both mono and stereo channels are used.

2) Signal Prepocessing

This step includes elimination of background noise, framing and windowing. Background noise is removed from the data so that only speech samples are the input to the further processing. Continuous speech signal has been separated into a number of segments called frames, also known as framing. After the pre-emphasis, filtered samples have been converted into frames, having frame size of 50 msec. Each frame overlaps by half. To reduce the edge effect of each frame segment windowing is done.

3) Speech Feature Extraction

After windowing, compute the short-time energy features and spectral centroid features of each frame of the speech signal.

1) Short-Time Signal Energy:

Short-time energy is the principle and most natural feature

that have been used. Physically, energy is a measure of how much signal there is at any one time. Energy is used to discover voiced sound, which have higher energy than silence/unvoiced, in a continuous speech.

The energy of a signal is typically calculated on a short-time basic, by windowing the signal at a particular time, squaring the samples and taking the average [2]. The square root of this result is the engineering quantity, known as the root-mean square (RMS) value, also used. The short-time energy function of a speech frame with length *N* is defined as

$$E_n = \frac{1}{N}\sum_{m=1}^{N}[x(m)w(n-m)]^2 \quad (1)$$

Where *x(m)* is the discrete-time audio signal and *w(m)* is rectangle window function.

$$w(m) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases}$$

2) Spectral centroid:

The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of gravity" of the spectrum is. This feature is a measure of the spectral position, with high values corresponding to "brighter" sound [3]. The spectral centroid, SC_i , of the i-th frame is defined as the center of "gravity" of its spectrum and it is given by the following equation:

$$SC_i = \frac{\sum_{m=0}^{N-1} f(m)X_i(m)}{\sum_{m=0}^{N-1} X_i} \quad (2)$$

Here, *f(m)* represents the center frequency of *i-th* bin with length *N* and $X_i(m)$ is the amplitude corresponding to that bin in DFT spectrum[4].

4) Speech Segment Detection

After computing speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments.

1. Compute the Mean or average values of smoothed feature sequences.
2. Find the local maxima of histogram.
3. If at least two maxima M_1 and M_2 have been found, then:

Threshold, $T = \frac{W * M_1 + M_2}{W + 1}$ (3)

Otherwise,

Threshold, $T = \frac{Mean}{2}$ (4)

Where W is a user-defined weight parameter [5]. Here, W=10.

The above process is applied for both feature sequences and finding two thresholds: T_1 based on the energy sequences and T_2 based on the spectral centroid sequences. After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed thresholds (for both feature sequences).

5) Post Processing

As a post-processing step, the detected speech segments are lengthened by 5 short term window. Finally, successive segments are merged.

B. Feature 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.

1) Linear Predictive Coding (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:

1) Preemphasis

The digitized speech signal, *s(n)*, is put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The output of the preemphasizer network is related to the input to the network, *s(n)* , by difference equation:

$$\tilde{s}(n) = s(n) - a\tilde{s}(n-1) \quad (5)$$

2) Frame Blocking

The output of pre-emphasis step, $\tilde{s}(n)$, is blocked into frames of *N* samples, with adjacent frames being separated by M samples. If $x_i(n)$ is the l^{th} frame of speech, and there are L frames within entire speech signal, then

$$x_i(n) = \tilde{s}(Ml + n) \quad (6)$$

Where, n = 0,1,…,N – 1 and l = 0,1,…,L – 1

3) Windowing

After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. If we define the window as *w(n)*, 0 ≤ n ≤ N – 1, then the result of windowing is the signal:

$$\tilde{x}_1(n) = x_1(n)w(n) \quad (7)$$

Where, 0 ≤ n ≤ N – 1

Typical window is the Hamming window.

4) Autocorrelation Analysis

The next step is to auto correlate each frame of windowed signal in order to give:

$$r_i(m) = \sum_{n=0}^{N-1-m} \tilde{x}_i(n)\tilde{x}_l(n+m) \quad m = 0,1,..., p \tag{8}$$

Where the highest autocorrelation value, *p*, is the order of the LPC analysis.

5) LPC Analysis

The next processing step is the LPC analysis, which converts each frame of *p* + 1 autocorrelations into LPC parameter set by using Durbin's method. This can formally be given as the following algorithm:

$$\begin{aligned} &E^{(0)} = r(0) \\ &k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|i-j|)}{E_{i-1}} \qquad 1 \le i \le p \\ &\alpha_i^{(i)} = k_i \\ &\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{i-1} \qquad 1 \le j \le i-1 \\ &E^{(i)} = (1-k_i^2)E_{i-1} \end{aligned} \tag{9}$$

For *i*=1,2,…,*p*, the LPC coefficient, a_m, is given as

$$a_m = \alpha_m^{(p)}$$

6) LPC Parameter Conversion to Cepstral Coefficients

LPC cepstral coefficients, is a very important LPC parameter set, which can be derived directly from the LPC coefficient set. The recursion used is

$$\begin{aligned} c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \qquad 1 \le m \le p \\ c_m &= \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \qquad m > p \end{aligned} \tag{10}$$

The LPC cepstral coefficients are the features that are extracted from voice signal and these coefficients are used as the input data of HMM Model. In the HMM model, LPC features are used in building phoneme model.

2) Gamma Tone Cepstral Coefficient (GTCC)

The following steps explain the feature extraction process.

1) Hamming windowing and FFT

The first step of the algorithm is to subdivide a speech sequence into frames. The windowing function is the Hamming window, which aims to reduce the spectral distortion introduced by windowing.

After windowing, Fast Fourier Transform (FFT) is applied to the windowed speech frame. The N-point FFT spectrum, S[k] for 0 ⩽ k ⩽ N-1, of the speech frame is obtained as a result.

2) Gammatone Filterbank

The Gammatone filterbank consists of a series of bandpass filters, which models the frequency selectivity property of the human cochlea. The impulse response of each filter was

$$g(t) = at^{n-1} e^{-2\pi bt} \cos(2\pi f_c t + \phi) \tag{11}$$

Where, *a* is a constant, usually equals to 1, *n* is the order of the filter, *Φ* is the phase shift, f_c and *b* are the center frequency and the bandwidth of the filter in Hz. The next step is to find out the filter output, X_m,

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[K]|^2 |H_m[K]| \tag{12}$$

3) Equal-loudness

The equal-loudness weight of the m^{th} filter, E_m, can be found by evaluating

$$E_m = \frac{(w^2 + 56.8*10^6)w^4}{(w^2 + 6.3*10^6)^2 (w^2 + 0.38*10^9)(w^6 + 9.58*10^{26})} \tag{13}$$

The filter output after equal loudness, $X_{m(e)}$, is simply the product of the filter output and the equal loudness weight

$$X_{m(e)} = E_m X_m \tag{14}$$

4) Logarithmic Compression

The next step of the algorithm is to apply logarithm to each filter output. The aim of this procedure is to simulate the human perceived loudness given certain signal intensity.

Let $X_m(ln+e)$ be the logarithmically-compressed filter output of the m_{th} Gammatone filter.

5) DCT

To de-correlate the filter outputs, Discrete Cosine Transform (DCT) is applied to the filter outputs. Suppose p is the order of GTCC. The feature vector of one speech frame, which has p number of GTCC coefficients, contains the first *p* DCT coefficients of the filter outputs. In mathematical term, the k_{th} GTCC coefficient of the feature vector is defined as follows.

$$GTCC_k = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X_{m(\ln+e)} \cos\left(\frac{\pi k(m-0.5)}{M}\right) 1 \le k \le p \tag{15}$$

C. Hidden Markov Model(HMM)

HMM can be used to model a unit of speech whether it is a phoneme, or a word, or a sentence. 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).The HMMs are suitable for the classification from one or two dimensional signals and can be used when the information is incomplete or uncertain [6]. The following figure shows the HMM structure of the Myanmar phoneme model.

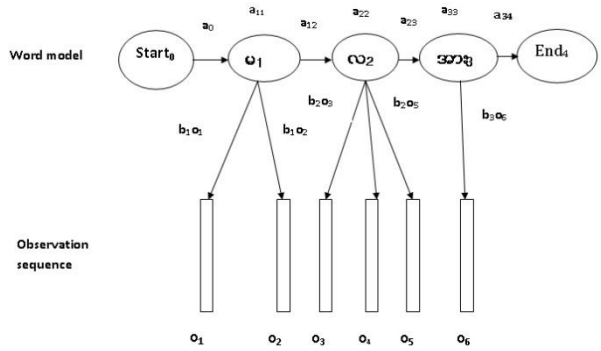

Fig.1 HMM structure of the phoneme "ဖတ်"

III. PROPOSED SYSTEM

In this system, human speech is taken as input to the system. First human speech is decoded into signals for digital processing. For segmenting continuous Myanmar speech sentences into word/sub-words, time-domain and frequency-domain features extraction is used. To detect the word boundaries, dynamic thresholding criterion is applied. Then, important features of speech signals are extracted by Linear Predicted Coding (LPC) and Gammatone Cepstral Coefficient (GTCC) approach. Feature vectors are clustered using K-means. The detailed steps of word segmentation and features extraction techniques are described in Section 2.

A. *Training*

The extracted feature vectors of LPC and GTCC are trained into HMM. The training is done in two steps as

- HMM code book
- HMM training by forward backward re-estimation algorithm

First the code book contains the cluster number specifies to each observation vector, which is obtained by applying the K-means algorithm.

After clustering the training to HMM, parameters begin by applying Forward-backward algorithm. It uses the principle of Maximum likelihood estimation. It returns the state transition matrix A, observation probability matrix B, and the initial state probability vector π. In this phase, the observation vectors being trained in the form of HMM parameters and resulted as the log likelihood of entire speech. This log likelihood is used to store in training database for recognition in real time.

B. *Language Model*

For continuous speech recognition, a pronunciation dictionary was created 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, the system recognizes a sequence of words and that's why it is necessary to incorporate a language model. The Regular grammar modelling technique is used as language model which has the properties like finite state model, small vocabulary and restricted grammar.

C. Recognition

HMM recognition recognizes the words on the basic of log-likelihood. It recalculates the log likelihood of speech vector and compares it to the pre stored value of log likelihood. If it matches the entire log value from the database of specified model, then it can recognize the words. Then, words are composed to get the sentences in text. Finally, input speech is transformed into text structure and displayed as output. The following figure shows the proposed system design.

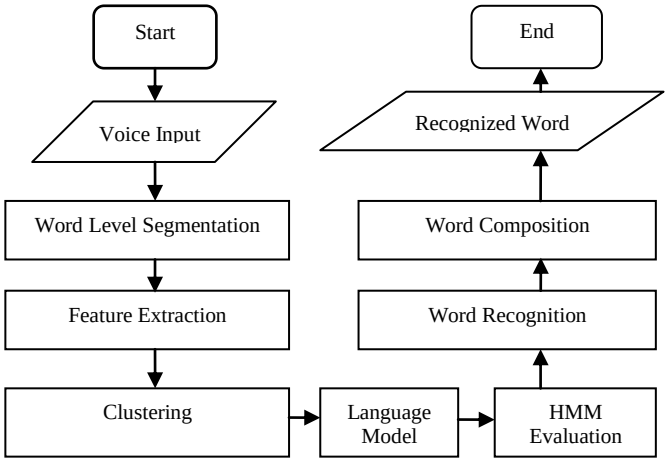

Fig. 2 Proposed System Design

IV. EXPERIMENTATIONS AND RESULTS

For a continuous speech signal to be recognized, the preprocessing and the feature extraction is done first. Then the signal is recognized using the HMM. This work is based on 558 files gathered from Myanmar native female speaker. The words are taken from the Grade 1 Myanmar Text Book. The speech signals were digitized at a sample rate of 44100 Hz using 16 bits and saved in WAV format.

In this experiment, example of 10 spoken sentences was recognized by the system. The system output was recognized Myanmar words. The performance of speech recognition systems is usually specified in terms of accuracy. Accuracy will be measured in terms of performance accuracy which is usually rated with word error rate (WER).Word error rate can then be computed as:

$$WER = \frac{No.\ of\ miss\ recognized\ words}{Total\ No.\ of\ segmented\ words} \qquad (16)$$

An example of Myanmar continuous sentences are given in Table 1. Table 2 shows the output of the system and the detailed recognition results were shown in Table 3.

TABLE I

EXAMPLE OF MYANMAR CONTINUOUS SENTENCES

Sentences ID	Myanmar Continuous Sentences
S1	ခခရေကုံး မလေးပြိုး
S2	ငါးမဖမ်းရ ပန်းမလူးရ ရေကန် အနီးမဆော့ရ စည်းကမ်းလေးစားပါ
S3	မိုးလရာသီ အဘိုးအိုတို့ လယ်တဲ တဲကလေး မိုးယိုနေသလား ကူညီ၍ မိုးပေးပါ
S4	အရှေ့ရွာမှာဘာရှိသလဲလူစည်ကားလှပါသည် ဘုရားပွဲတော်ရှိပါသလား လှည်းစီး၍ လာကြသည် လေ့စီး၍ လည်းလာကြသည်
S5	ရွာလူကြီးများကြွလာပြီကြွေပန်းကန်ယူခဲ့ပါ ကျွဲကောသီးထည့်ထားပါ ကျေးမွေးပြုစုပါရစေ
S6	ကလေးငယ်ငယ် ဘာစားခဲ့သလဲ အမဲသားငါးစားရဲ့လား ကုလားပဲစားပါ
S7	ချွေးတရွဲရွဲ အားကစားပွဲ လွှဲဖယ်၍ မနေပါ ပြေးလွှားကစားကြသည် အားရပါးရ နွှဲပါဦး
S8	စောစော အိပ်ထ ပြုကျင့်က တွင်းပ ကိုယ်ခန္ဓာ ရောဂါခပ်သိမ်း ရှောင်ဆွာတိမ်း ကင်းငြိမ်းလွန်ချမ်းသာ ဥစ္စာ ဓန ပေါကြွယ်ဝ ထွန်းပဥ္ဉာဏ်ပညာ
S9	လယ်ထဲမှာရေလျှံနေပြီလျှော်စည်းများ မျောပါနေကြသည် လျှောကြမည် သတိထား
S10	နေစဉ်မှန်စွာရေချိုးပါ ခေါင်းကိုမှန်စွာ ဖြီးကြပါ ကျန်းမာသန်ရှင်းရောဂါကင်း

TABLE II

OUTPUT OF THE PROPOSED SYSTEM

Sentences ID with different features	Myanmar Continuous Sentences
S1(LPC)	ခ,ခရေ,ကုံးမလေး,ပြိုး;
S1(GTCC)	,ခ,ခရေ,ကုံးမလေး,ပြိုး
S2(LPC)	အ,ဖမ်းရ,ပန်းမ,လူးရ,ရေကန်အနီးမ,ဆော့ရ,စည်း,ကမ်းလေး,လာ,ပါ;
S2(GTCC)	,ဘူး,ဖမ်းရ,ပန်းမ,လူးရ,ရေကန်အနီးမ,ဆော့ရ,စည်း,ကမ်းလေး,လာ,ပါ
S3(LPC)	,အ,အ,ဘိုး,ဘိုး,တိုလဲ,တဲ,တဲကလေး,မိုး,သလား,ကူညီ၍,မိုး,ပေး,ပါ;
S3(GTCC)	,အ,အ,ဘိုး,အို,တိုလဲ,တဲ,တဲကလေး,မိုးယိုနေ,မိုး,ကူညီ၍,မိုး,ပေး,ပါ
S4(LPC)	အရှေ့ရွာမှာ,ဘာရှိ,သလဲ,လူ,စည်ကား,လှ,ပါ,သည်,ပါ,ပွဲတော်,ရှိ,ပါ,သလား,လှည်း,စီး၍,လာ,လေ့,သည်,လေ့,စီး၍,လည်း,လာ,ကြ,ကြ;
S4(GTCC)	အရှေ့ရွာမှာ,ဘာရှိ,သလဲ,လူ,စည်ကား,လှ,ပါ,သည်,ဘုရား,ပွဲတော်,ရှိ,ပါ,သလား,လှည်း,စီး၍,လာ,ကြ,သည်,လေ့,စီး၍လည်း,လာ,ကြ,ကြ
S5(LPC)	ရွာလူကြီးများ,ကြွလာ,ပြီ,ကြွေပန်း,ကန်,ယူ,ကော,ပါ,ကျွဲ,ကော,ကျွဲ,ထည့်, ထည့်,ပါ,ကျေးမွေး,ပြုစု,ကော,ထည့်;
S5(GTCC)	ရွာလူကြီးများ,ကြွလာ,ပြီ,ကြွေပန်း,ကန်,ယူ,ခဲ့,ပါ,ကျွဲ,ကော,သီး,ထည့်,ထား,ပါ,ကျေးမွေး,ပြုစု,ပါရ,စေ
S6(LPC)	ကလေးငယ်ငယ်,ဘာ,စား,ခဲ့,သ,လဲ,အမဲ,သားငါး,စားရဲ့,လား,ကုလား,ပဲ,လား,ပါ;
S6(GTCC)	ကလေးငယ်ငယ်,ဘာ,စား,ခဲ့,လဲ,လဲ,အမဲ,သားငါး,စားရဲ့,လား,ကုလား,လား,လား,ပါ
S7(LPC)	ချွေးတရွဲရွဲ,အားကစား,ပွဲ,လွှဲ,ဖယ်၍,မနေ,ပါ,ပြေးလွှား,ကစား,ကြ,သည်,အားရ,ပါးရ,နွှဲပါ,ဦး;
S7(GTCC)	ချွေးတရွဲရွဲ,အားကစား,ပွဲ,လွှဲ,ဖယ်၍,မနေ,ပါ,ပြေးလွှား,ကစား,ကြ,သည်,အားရ,ပါးရ,နွှဲပါ,ဦး
S8(LPC)	,စောစော,အိပ်,ထ,ပြု,ကျင့်,က,တွင်း,ပ,ကိုယ်,ခန္ဓာ,ရော,ဂါ,ခက်,သိမ်း,ရှောင်,ဆွာ,တိမ်း,ကင်းငြိမ်းလွန်ချမ်းသာ,ဥ,စ္စာ,ဓန,ပေါ်,ကြွယ်ဝ,ထွန်း,ပဥ္ဉာဏ်,ပညာ;
S8(GTCC)	စောစော,အိပ်,ထ,ပြု,ကျင့်,က,တွင်း,ပ,ကိုယ်,ခန္ဓာ,ရော,ဂါ,ခက်,သိမ်း,ရှောင်,ဆွာ,တိမ်း,ကင်းငြိမ်းလွန်ချမ်းသာ,ဥ,စ္စာ,ဓန,ပေါ်,ကြွယ်ဝ,ထွန်း,ပဥ္ဉာဏ်,ပညာ
S9(LPC)	,လယ်ထဲမှာ,ရေလျှံနေပြီ,မျော,မျော,ပါနေကြသည်,လေ့မျော၍,ဆယ်ယူ,ပါ,လျှောကြမည်,သတိ,ထား;
S9(GTCC)	လယ်ထဲမှာ,ရေလျှံနေပြီ,လျှော်စည်းများ,မျော,ပါနေကြသည်,လေ့မျော၍,ဆယ်ယူ,ပါ,လျှောကြမည်,သတိ,ထား
S10(LPC)	နေစဉ်,မှန်စွာ,ရေချိုး,ပါ,ခေါင်းကို,မှန်စွာ,ဖြီးကြ,ပါ,ကျန်းမာ,သန်ရှင်း,ရော,ဂါ,ကင်း;
S10(GTCC)	နေစဉ်,မှန်စွာ,ရေချိုး,ပါ,ခေါင်းကို,မှန်စွာ,ဖြီးကြ,ပါ,ကျန်းမာ,သန်ရှင်း,ရော,ဂါ,ကင်း

TABLE III

RECOGNITION RESULTS

Sentence ID	No. Of Segment	WER by LPC	WER by GTCC
S1	4	0	0
S2	10	0.2	0.2
S3	13	0.23077	0.15385
S4	23	0.21739	0.13043
S5	19	0.26316	0
S6	14	0.071429	0.21429
S7	15	0.066667	0.066667
S8	26	0.038462	0.038462
S9	11	0.090909	0
S10	13	0	0
Total	148	1.18	0.5

The system obtained the average word error rate (WER) of 1.18 and 0.5 based on LPC and GTCC features respectively.

V. Conclusions

This paper presents an automatic speech recognition system for Myanmar language using the appropriate features and recognizer. This paper clearly describes the theory and implementation details of the entire development task using the HMM. Speech segmentation was done using time-domain feature and frequency-domain feature and the recognition accuracies obtained using standard vector comparison method like Euclidean distances. The results also demonstrate the superiority of the features using LPC and GTCC features.

Acknowledgment

I am very grateful to Dr. K Zin Lin for fruitful discussion during the preparation of this paper and also specially thank to Rector, Professors and colleagues from Technology University (Yatanarpon Cyber City), Myanmar.

References

[1] O. Cheng, W. Abdulla, Z. Salcic, "Performance Evaluation of Front-end Processing for Speech Recognition ", School of Engineering Report No. 621

[2] C. T. Zhang and C. J. Kuo, "Hierarchical classification of audio data for acrhiving and retrieving", In International Conference on Acoustics, Speech and Signal Processing, volume VI, pages 3001–3004. IEEE, 1999.

[3] T. Giannakopoulos, "Study and application of acoustic information for the detection of harmful content and fusion with visual information" Ph.D. dissertation, Dept.of Informatics and Telecommunications, University of Athens, Greece, 2009.

[4] C.W. James and T.W. John. "An algorithm for the machine calculation of complex Fourier series", Mathematics of Computation: Journal Review, 19: 297–301, 1965.

[5] T. Giannakopoulos, A. Pikrakis and S. Theodoridis "A Novel Efficient Approach for Audio Segmentation", Proceedings of the 19th International Conference on Pattern Recognition (ICPR2008), December 8-11 2008, Tampa, Florida, USA.

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