SYLLABLE-BASED SPEECH RECOGNITION SYSTEM FOR MYANMAR

Wunna Soe[1] and Dr. Yadana Thein[2]

[1]University of Computer Studies, Yangon (UCSY), Yangon Myanmar
[2]Department of Computer Hardware, University of Computer Studies, Yangon (UCSY), Yangon, Myanmar

ABSTRACT

This proposed system is syllable-based Myanmar speech recognition system. There are three stages: Feature Extraction, Phone Recognition and Decoding. In feature extraction, the system transforms the input speech waveform into a sequence of acoustic feature vectors, each vector representing the information in a small time window of the signal. And then the likelihood of the observation of feature vectors given linguistic units (words, phones, subparts of phones) is computed in the phone recognition stage. Finally, the decoding stage takes the Acoustic Model (AM), which consists of this sequence of acoustic likelihoods, plus an phonetic dictionary of word pronunciations, combined with the Language Model (LM). The system will produce the most likely sequence of words as the output. The system creates the language model for Myanmar by using syllable segmentation and syllable based n-gram method.

KEYWORDS

Speech Recognition, Language Model, Myanmar, Syllable

1. INTRODUCTION

Speech recognition is one of the major tasks in natural language processing (NLP). Speech recognition is the process by which a computer maps an acoustic speech signal to text. In general, there are three speech recognition system; speaker dependent system, speaker independent system, and speaker adaptive system. The speaker dependent systems are trained and learnt based on a single speaker and can recognize the speech of that trained one speaker. The speaker independent systems can recognize any speaker and these systems are the most difficult to develop and most expansive and accuracy is lower than speaker dependent systems, but more flexible. A speaker adaptive system is built to adapt its processes to the characteristics of new speakers.

In other way, there are two types of speech recognition system: continuous speech recognition system, and isolated-word speech recognition system. An isolated-word recognition system performs single words at a time – requiring a pause between saying each word. A continuous speech system recognizes on speech in which words are connected together, i.e. not separated by pause.

Generally, most speech recognition systems are implemented mainly based on one of the Hidden Markov Model (HMM), deep belief neural network, dynamic time wrapping.

Myanmar language is a tonal, syllable-timed language and largely monosyllabic and analytic language, with a subject-object-verb word order. Myanmar language has 9 parts of speech and is

DOI : 10.5121/ijcseit.2015.5201

spoken by 32 million as a first language and as a second language by 10 million. Any Myanmar speech recognition engine has not been before.

In this paper, we mainly focus on the Myanmar phonetic structure for speech recognition system. This paper is organized as follow. In section2, we discuss the related works of the areas of speech recognition system based on syllables models. In section 3, we describe characteristics of Myanmar phones and syllables. In section4, we present the architecture of speech recognition system. In section 5 and 6, we discuss how to build the acoustic model and language model for speech recognition system. In section 7, we mention the phonetic dictionary of speech recognition system. Finally, we conclude the results of proposed system and difficulties and limitations of this system.

2. RELATED WORK

Many researchers have been work for speech recognition based on syllable in other languages. But in our language, Myanmar, there is no one for implementing speech recognition system based on syllable. In the following paragraphs, we present some of the related work in the area of syllable-based speech recognition systems for other languages and speech recognition for Myanmar language.

Piotr Majewski expressed a syllable-based language model for highly inflectional language like Polish. The author demonstrated that syllables are useful sub-word units in language modeling of Polish. Syllable-based model is a very promising choice for modeling language in many cases such as small available corpora or highly inflectional language.[7]

R. Thangarajan, A.M. Natarajan, and M. Selvam expressed the Syllable modeling in continuous speech recognition for Tamil language. In this paper, two methodologies are proposed which demonstrate the syllable's significance in speech recognition. In the first methodology, modeling syllable as an acoustic unit is suggested and context independent (CI) syllable models are trained and tested. The second methodology proposes integration of syllable information in the conventional triphone or context dependent (CD) phone modeling.[8]

Xunying Liu James L. Hieronymus Mark J. F. Gales and Philip C. Woodland presented Syllable language models for Mandarin speech recognition. In this paper character level language models were used as an approximation of allowed syllable sequences that follow Mandarin Chinese syllabiotactic rules. A range of combination schemes were investigated to integrate character sequence level constraints into a standard word based speech recognition system.[9]

Ingyin Khaing presented Myanmar Continuous Speech Recognition System Based on DTW and HMM. In this paper, we found that combinations of LPC, MFCC and GTCC techniques are applied in feature extraction part of that system. 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. [10]

3. MYANMAR SYLLABLE

Myanmar language is a member of the Sino-Tibetan family of languages of which the Tibetan-Myanmar subfamily forms a part. Myanmar script derives from Brahmi script. There are basic 12 vowels and 33 consonants and 4 medial in Myanmar language. In Myanmar language, words are formed by combining basic characters with extended characters. Myanmar syllables can stand one

or more extended characters by combining consonants to form compound words. Myanmar characters 33 consonants are described as the following table.

Table 1. Myanmar Consonants

က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ရ	ယ	လ	ဝ	သ
	ဟ	ဠ	အ	

Basically, there are 12 vowels in Myanmar writing. These vowels are အ(a.), အာ(a), ကြ(i.), ကြီ(i), ဥ(u.), ဦ(u), ဧ(ei), အဲ(e:), ြသ(o:), ြသော်(o), အံ(an), အို(ou). The variation အ(a.), အာ(a), အိ(i.), အီ(i), အု(u.), အူ(u), အေ(ei), အဲ(e:), အော(o:), အော်(o), အံ(an), အို(ou) can also be written. These 12 basic vowels can be extended with the employment of tone markers (့) and (း) and also devowelizing consonants. [2]

The sequential extension of the 12 basic vowels results in 22 vowels listed in the original thinbongyi. These 22 extension vowels are described as the following table.

Table 2. Basic and Extension Vowels

အ(a.)	အာ(a)	အား(a:)
အိ(i.)	အီ(i)	အီး(i:)
အု(u.)	အူ(u)	အူး(u:)
အေ(ei)	အေ့(ei.)	အေး(ei:)
အဲ(e:)	အဲ့(e.)	
အော(o:)	အော့(o.)	အော်(o)
အံ(an)	အံ့(an.)	
အို(ou)	အို့(ou.)	အိုး(ou:)

In the above table, the shaded vowels are 12 basic vowels and other are extended vowels. At first, there remain the original 11 vowels: အ(a.), အာ(a), အိ(i.), အီ(i), အု(u.), အူ(u), ေအ(ei), အဲ(e:), ေအာ(o:), ေအာ္(o), အံ(an). When အို(ou) is added the result is the basic 12 vowels in the Myanmar language.[2]

Myanmar syllables are basically constructed by the combination of consonant and vowel. The combination of အို vowel and က consonant makes one syllable ကို as အို+က=ကို.

There are 10 consonants used for devowerlizing က်, င်, စ်, ဉ်, ည်, တ်, န်, ပ်, မ်, ယ် and there are four basic consonant combination symbols in Myanmar writing ျ, ြ, ွ, ှ. These Myanmar characters may be join together appropriate consonants out of the 33 such as ကျ, ကြ, မွ, မှ. These symbols may also be combined with each other in two or three as in မျှ, လ္ဘ. [2]

4. ARCHITECTURAL OVERVIEW OF SPEECH RECOGNITION SYSTEM

Generally, a speech recognition system takes the speech as input and voice data as knowledge base and then the output result is the text as the below the figure 1. The knowledge base is the data that derives the decoder of the speech recognition system. The knowledge base is created by three sets of data:

- Dictionary
- Acoustic Model
- Language Model

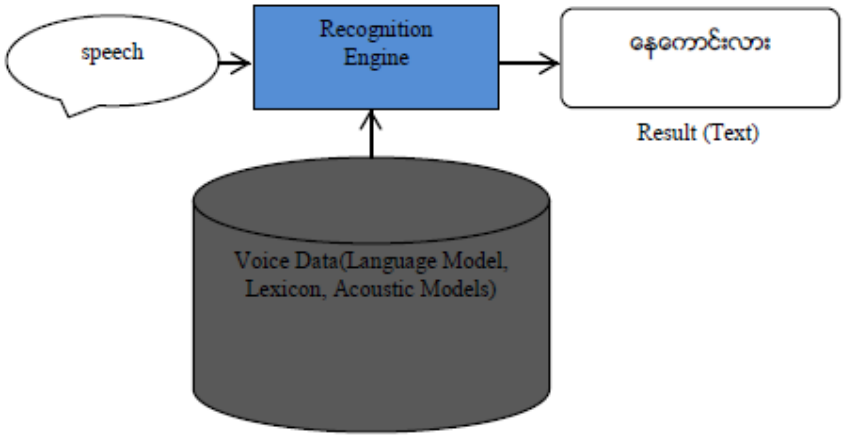

Figure 1. Speech Recognition System

The dictionary contains a mapping from word to phones. An acoustic model contains acoustic properties for each senone, state of the phone. A language model is used restrict word search. It defines which word could follow previously recognized words (remember that matching is a sequential process) and helps to significantly restrict the matching process by stripping words that are not probable.

The proposed speech recognition system has three main components: Feature Extraction, Phone Recognition, and Decoding. The architecture for a simplified speech recognition system is as follow.

In the following architecture, we can compute the most probable sequence W given some observation sequence O. We can choose the sentence which the product of two probabilities for each sentence is greatest as the following equation.[1]

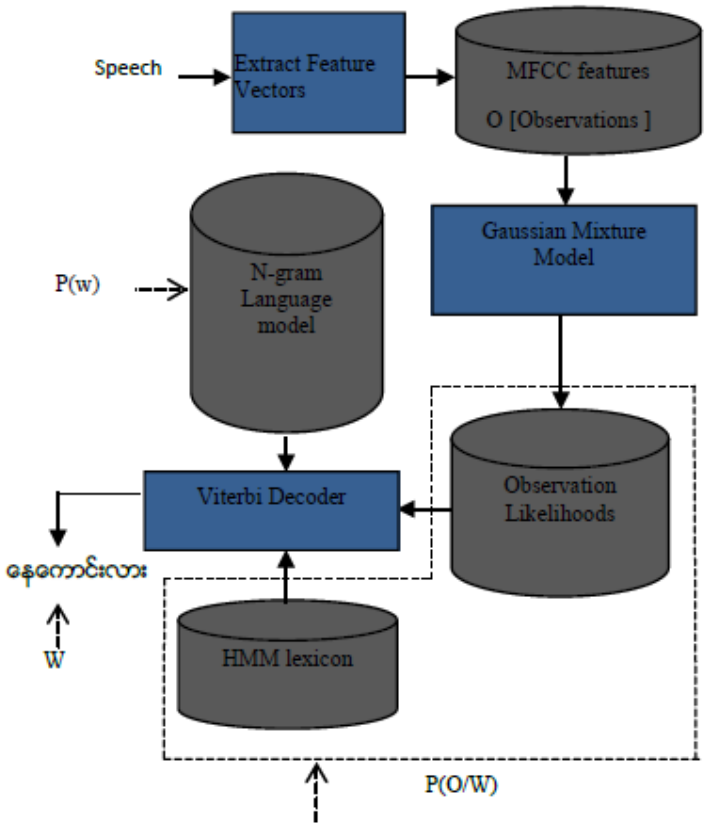

Figure 2. A Simple Discriminative Speech Recognition System Overview

$$\hat{W} = \arg\max_{W \in L} P(O \mid W) P(W) \quad (1)$$

In the equation (1), the acoustic model can computed the observation likelihood, P(O/W). The language model can get for computing the prior probability, P(W).[1]

4.1. Feature Extraction

The feature extraction is the transformation stage of speech waveform into a sequence of acoustic feature vectors. The feature vectors represent the information in a small time window of the

signal. The acoustic waveform is sampled into frames (usually 10, 15, or 20 milliseconds of frame size) that are transformed into spectral features as the following figure. Each time frame (window) is thus represented by a vector of around 39 features representing this spectral information. [1]

There are seven steps in feature extraction process:

1. Pre-emphasis
2. Windowing
3. Discrete Fourier Transform
4. Mel Filter Bank
5. Log
6. Inverse Discrete Fourier Transform
7. Deltas and Energy.

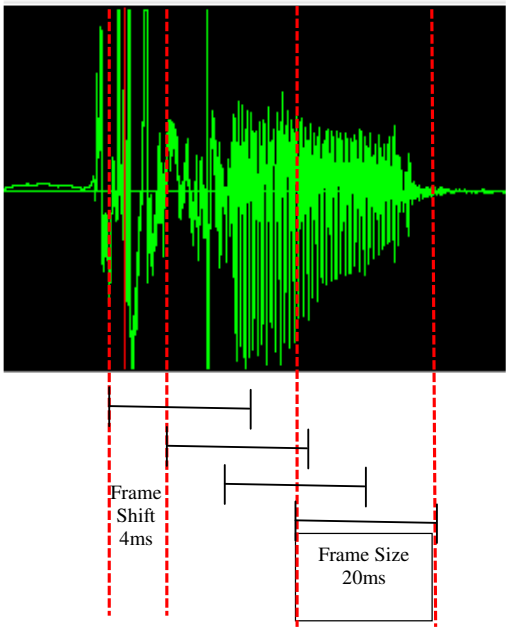

Figure 3. Windowing Process of Feature Extraction

The pre-emphasis stage is to boost the amount of energy in the high frequencies. Information from these higher formats more available to the acoustic model can be made by boosting the high frequency energy and this process can improve phone detection accuracy. The waveform is extracted the roughly stationary portion of speech by using a window which is non-zero inside some region and zero elsewhere, running this window across the speech signal and extracting the waveform inside this window. The method for extracting spectral information for discrete frequency bands for a discrete-time signal is the discrete Fourier transform (DFT).

The form of the model used in Mel Frequency Cepstral Coefficient (MFCC) is to wrap the frequencies output by the DFT onto the Mel. A Mel is a unit of pitch. In general, the human response to signal level is logarithmic; humans are less sensitive to slight differences in amplitude at high amplitudes than at low amplitudes. In addition, the feature estimates less sensitive to variations in input can be made by using a log, such as power variations due to the distance between the speaker and the microphone. The next step in MFCC feature extraction is the

computation of the cepstrum, also called as the spectrum of the log of the spectrum. The cepstrum can be seen as the inverse DFT of the log magnitude of the DFT of a signal.

The extraction of the cepstrum with the inverse DFT from the previous steps results in 12 cepstral coefficients for each frame. The energy in a frame, the 13th feature, is the sum over time of the power of the samples in the frame. And the delta value estimates the slope using a wider context of frames.

4.2. Phone Recognition

The phone recognition stage computes the phone likelihood of the observed spectral feature vectors given Myanmar phone units or subparts of phones. In this proposed system, we use the Gaussian Mixture Model (GMM) classifiers to compute for each HMM state q, corresponding to a phone or sub phone, the likelihood of a given feature vector given this phone p(o/q). We can compute the Gaussian Mixture Model (GMM) as the following equation. [1]

$$f(x|\mu_{jk},\Sigma_{jk})=\sum_{k=1}^{M} c_{jk}\frac{1}{(2\pi)^{D/2}|\Sigma_{jk}|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_{jk})^{T}\Sigma^{-1}(x-\mu_{jk})\right) \quad (2)$$

In the equation (2), M is the number of Gaussian Models, called mixture weights. D is the dimensionality and in this system, it has 39 dimensions.

Most speech recognition algorithms are based on computing observation probabilities directly on the real-valued, continuous input feature vector. The acoustic models are based on the computation of a probability density function (pdf) over a continuous space. By far the most common method for computing acoustic likelihoods is the Gaussian mixture model (GMM) pdfs, although neural networks, support vector machines (SVM), and conditional random fields (CRFs), are also used.

4.3. Decoding

In decoding stage, the proposed system used the Viterbi algorithm as the decoder. The decoder is the heart of the speech recognition process. The task of the decoder is to find the best hidden sequence of states by using the sequence of observations as inputs. First, the decoder selects the next set of likely states and then scores the incoming features against these states. The decoder prunes low scoring states and finally generates the result.

5. HOW TO BUILD ACOUSTIC MODEL

The acoustic model is trained by analyzing large corpora of Myanmar language speech. Hidden Markov Models (HMM) represent each unit of speech in the acoustic model. HMMs are used by a scorer to calculate the acoustic probability for a particular unit of speech. Each state of an HMM is represented by a set of Gaussian mixture density functions. A Gaussian mixture model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. There are many acoustic model training tools. Among them we choose the sphinxtrain tool to build acoustic model for a new language, Myanmar.

To build speech recognition system for a single speaker, we collect recording files for an hour. Each file has 7 seconds average length. The parameters of the acoustic model of the sound units

using feature vectors, are learnt by the trainer. This is called a training database. The file structure of the database is

- /etc
  - /db_name.dic
  - /db_name.phone
  - /db_name.lm.DMP
  - /db_name.filler
  - /db_name_train.fileids
  - /db_name_train.transcription
  - /db_name_test.fileids
  - /db_name_test.transcription
- /wav
  - /speaker_1
    - /file1.wav
    - /file2.wav
    - …

In the above file structure, etc, wav, and speaker_1 are folder names. The db_name.dic file is phonetic dictionary that maps words and phones. The db_name.phone is the phone set file that has one phone per line. The db_name.lm.DMP is language model file. It may be in ARPA format or in DMP format. The db_name.filler file is a filler dictionary that contains filler phones (not-covered by language model non-linguistic sounds like breathe, hmm or laugh). The db_name_train.fileids is a text file listing the names of the recordings one by line for training. The db_name_test.fileids is also a text file listing the names of the recordings one by line for testing. The db_name_train.transcription is a text file that contains the list of the transcription for each audio file for training. The db_name_test.transcription is also a text file listing the transcription for each audio file for testing. The wav files (filename.wav) that we used are recording files that have specific sample rate - 16 kHz, 16 bit, mono. [4]

After training, the acoustic model is located in db_name.cd_cont_<number_of senones> folder. <number_of senones> is the number of senones produced by the training tool. This folder is under the model_parameters folder auto generated by sphinxtrain. In the db_name.cd_cont_<number_of senones> folder, the model should have the following files:

- /db_name.cd_cont_<number_of senones>
  - /mdef
  - /feat.params
  - /mixture_weights
  - /means
  - /noisedict
  - /transition_matrices
  - /variances.

The *feat.params* file contains feature extraction parameters, a list of options used to configure feature extraction. The *mdef* file is the definition file that maps the triphone contexts and GMM ids (senones). The *means* file is Gaussian codebook variances. The *variances* file consists of the Gaussian codebook variances. The *mixture_weights* describes the mixtures for Gaussians. The *transition_matrices* file contains HMM transition matrices. The *noisedict* file is the dictionary for filler words.

6. HOW TO BUILD LANGUAGE MODEL

The language model describes what is likely to be spoken in a particular context. There are two types of models that are used in speech recognition systems - grammars and statistical language models. The grammar-type of language model describe very simple types of languages for command and control, and they are usually written by hand or generated automatically with plain code. The statistical language model uses stochastic approach called n-gram language model. An N-gram is an N-token sequence of words: a b2-gram (bigram) is a two-word sequence; a 3-gram (trigram) is a three-word sequence. N-gram conditional probabilities can be computed from plain text based on the relative frequency of word sequences. By another way, there are two types of statistical language model. The first is the close-vocabulary language model. The close-vocabulary language model assumes that the test set can only contain words from the given lexicon. There are no unknown words in the close-vocabulary model. An open-vocabulary language model is one in which we model the possible unknown words in the test set by adding a pseudo-word called <UNK>. An open-vocabulary model contains the training process for the probabilities of the unknown word model.

There are many approach and tools to create the statistical language models. We use CMU language modeling toolkit to create n-gram language model. The language model toolkit expects its input to be in the form of normalized text files, with utterances delimited by <s> and </s> tags.[4] In this pre-processing step, we make syllable based way for creating normalized text files. The syllable-based normalization as is the following:

<s> မောင်မောင် ကျောင်းသွားသည် </s>
(original sentence)

<s> မောင် မောင် ကျောင်း သွား သည် </s>.
(normalized sentence)

Before normalization we split one syllable by syllable from the sentences. In syllable segmentation process, we use rule-based segmentation to spilt syllables from sentences. The output is a 3-gram language model based on vocabularies given from normalized text file. But in our output language model file is based on Myanmar syllables. In this system, we chose to use the close-vocabulary model.

The output language model file is as the ARPA format or binary format. The ARPA format language model file is shown as the following figure 4.

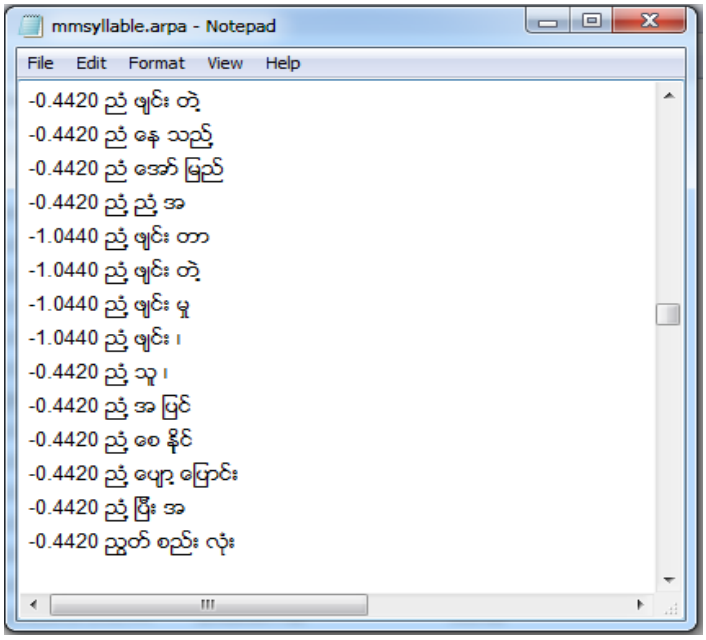

Figure 4. 3-grams (3 syllables sequences) in Language Model File

7. PHONETIC DICTIONARY (LEXICON)

A phonetic dictionary is a text file that contains a mapping from words to phones. It is also a lexicon that is a list of words, with a pronunciation for each word expressed as a phone sequence. The phone sequence can be specified by a lexicon. Each phone HMM sequence is composed of some sub phones, each with a Gaussian emission likelihood model. Example of the phonetic dictionary is as follow in the table 3.

Table 3. Part of Phonetic Dictionary

Syllable	Phonetic
က	ka.
ကာ	ka
ကား	ka:
ကိ	ki.
ကီ	ki
...	...

7. RESULT OF SPEECH RECOGNITION

We simulate how to decode the likelihood sequence using Viterbi algorithm (HMM). In the following figure 5, the two hidden states (ဒန္ and ဂေါင်း) correspond to Myanmar syllables ဒန္ and ဂေါင်း. The observations (O = {1, 2, 3}) correspond to the values of each one dawn from a vocabulary of phonetic dictionary (lexicon). B_1 and B_2 are the sequences of observation that represent the probability of an observation o_t being generated from state i.

In figure 6, hidden states are in circles and observations are in squares. Dotted line(unfilled) circles indicate illegal transitions. For a given state q_j at time t, the vale α_t (j) is computed as follow:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t) \tag{3}$$

In equation (3), v_t (j) is the Viterbi probability at time t and a_{ij} is the transition probability from previous state q_i to current state q_j . $b_j(o_t)$ is the state observation likelihood of the observation o_t given the current state j.[1]

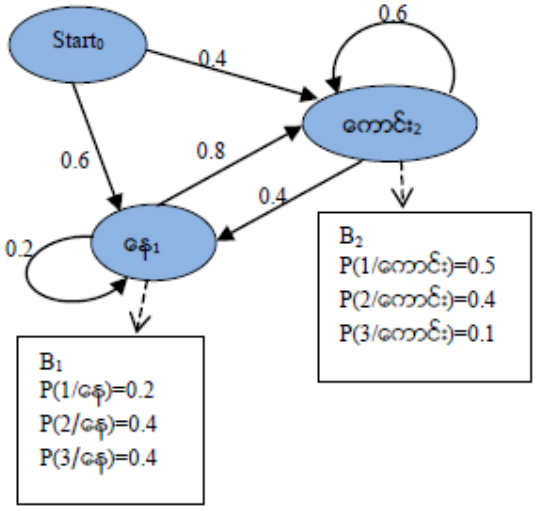

Figure 5. A Hidden Markov Model for relating feature values and Myanmar syllables

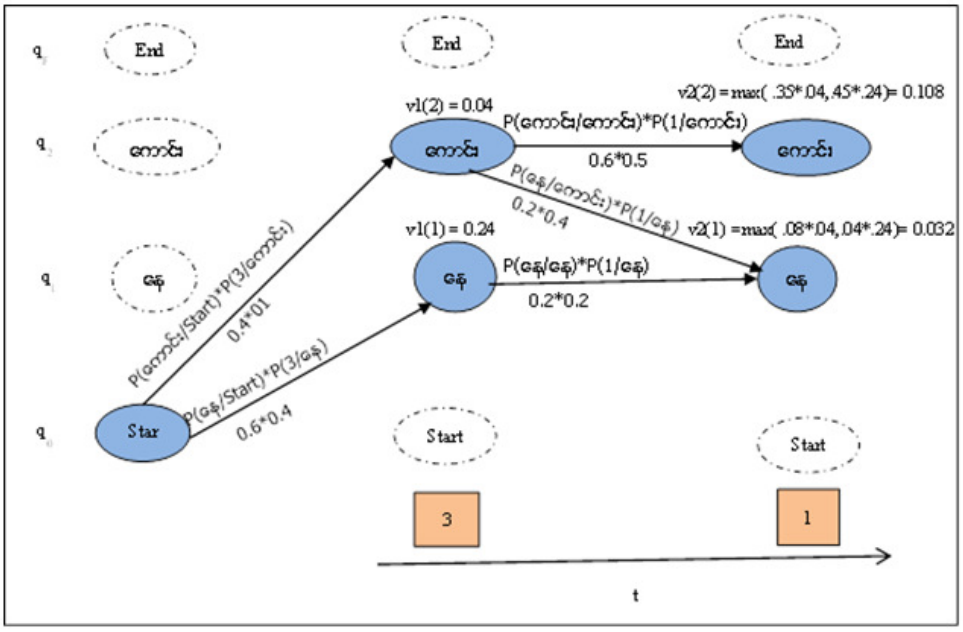

Figure 6. The Viterbi trellis for computing the best sequence of Myanmar Syllable

8. CONCLUSION AND FUTURE WORK

The standard evaluation metric for speech recognition systems is word error rate (WER). The word error rate is based on how much the word string returned by recognizer differs from a correct or reference transcription. But the proposed system use syllable error rate (SER) as evaluation metric instead of word error rate. Therefore, the result is based on how much syllable string returned by recognition engine differs from a correct or reference transcription. Our proposed system is just speaker dependent system at present and language model is also closed-vocabulary type. In the future, this proposed system will be developed as a speaker independent speech recognition system and language model will also hope to be an open-vocabulary type.

REFERENCES

[1] Daniel Jurafsky, and James H. Martin Smith (2009), Speech and Language Processing, Pearson Education Ltd., Upper Saddle River, New Jersey 07458

[2] Myanmar Language Commission (2011), Myanmar-English Dictionary, Department of Myanmar Language Commission, Ministry of Education, Union of Myanmar

[3] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Jole Woelfel (2004), "Sphinx 4: A Flexible Open Source Framework for Speech Recognition", SMLI TR2004-0811, Sun Microsystems Inc.

[4] Hassan Satori, Hussein Hiyassat, Mostafa Harti, and Noureddine Chenfour (2009), "Investigation Arabic Speech Recognition Using CMU Sphinx System", The International Arab Journal of Information Technology, Vol. 6, April

[5] http://en.wikipedia.org/wiki/Speech_recognition

[6] http://cmusphinx.sourceforge.net/

[7] Piotr Majewski (2008), "Syllable Based Language Model for Large Vocabulary Continuous Speech Recognition of Polish", University of Łód´z, Faculty of Mathematics and Computer Science ul. Banacha 22, 90-238 Łód´z, Poland, P. Sojka et al. (Eds.): TSD 2008, LNAI 5246, pp. 397–401

[8] R. Thangarajan, A.M. Natarajan, M. Selvam(2009), "Syllable modeling in continuous speech recognition for Tamil language", Department of Information Technology, Kongu Engineering College, Perundurai 638 052, Erode, India, Int J Speech Technol (2009) 12: 47–57

[9] Xunying Liu, James L. Hieronymus, Mark J. F. Gales and Philip C. Woodland (2013), "Syllable language models for Mandarin speech recognition: Exploiting character language models", Cambridge University Engineering Department, Cambridge, United Kingdom, J. Acoust. Soc. Am. 133 (1), January 2013

[10] Ingyin Khaing (2013), "Myanmar Continuous Speech Recognition System Based on DTW and HMM", Department of Information and Technology, University of Technology (Yatanarpon Cyber City),near Pyin Oo Lwin, Myanmar, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 1 February 2013

[11] Ciro Martins, António Teixeira, João Neto (2004), "Language Models in Automatic Speech Recognition", VOL. 4, Nº 2, JANEIRO 2004, L2F – Spoken Language Systems Lab; INESC-ID/IST, Lisbon

[12] Edward W. D. Whittake, Statistical Language Modeling for Automatic Speech Recognition of Russian and English, Trinity College, University of Cambridge

[13] Mohammad Bahrani, Hossein Sameti, Nazila Hafezi, and Saeedeh Momtazi (2008), "A New Word Clustering Method for Building N-Gram Language Models in Continuous Speech Recognition Systems", Speech Processing Lab, Computer Engineering Department, Sharif University of Technology, Tehran, Iran, N.T. Nguyen et al. (Eds.): IEA/AIE 2008, LNAI 5027, pp. 286–293

Authors

Dr. Yadana Thein is working as an associate professor at department of computer hardware technology in University of Computer Studies, Yangon. She received master degree from University of Computer Studies, Yangon. She received doctoral degree at the same university. She interest in Natural Language Processing.

Wunna Soe is at present Ph.D candidate student from University of Computer Studies, Yangon. He received Master of Computer Science (M.C.Sc.) from University of Computer Studies, Mandalay (UCSM). His current research is Automatic Speech Recognition and Natural Language Processing.