A Speech Recognition System for Myanmar Digits

Zin Zin Tun and Gun Srijuntongsiri

***Abstract*—Automatic speech recognition (ASR) is a technology that allows a computer to recognize the words spoken by a person through a telephone, microphone or other input devices. This paper focuses on recognizing numbers or digits because of their importance in banking system, customer queuing system, phone dialing system, and so on. We describe a speech recognition system for Myanmar digits based on the Hidden Markov Model (HMM) using HTK Tools. Our system recognizes the speech utterance by converting the speech waveform into a set of feature vectors using Mel Frequency Cepstral Coefficients (MFCC) technique. We use HMM-based acoustic and language models. Experiment is carried out to evaluate the performance of our speech recognizer with both context independent and context dependent models and yields significant results.**

***Index Terms*—Hidden Markov model, MFCC, Myanmar digits, speech recognition.**

I. Introduction

Speech is the natural form of communication for humans to exchange the information in their daily lives. In speech processing, automatic speech recognition (ASR) is a technology that converts speech signal to a sequence of words which is spoken by human. Digit recognition system is crucial in speech recognition area because it has wide range of applications, such as voice-operated interface for phone dialing, speech information retrieval (SIR), banking system, numerical data entry, language translation, and so on.

There have been many literatures in automatic speech recognition (ASR) system for the major languages of the world such as English, Japanese, Chinese, etc. However, only a few of works have been done in ASR for Myanmar language. The major difficulty in the research process of Myanmar language ASR is the lack of Myanmar speech corpus. Generally it is not easy to build the speech corpus because it requires a huge amount of speech data, time, and efforts.

In speech recognition system, many parameters may affect the performance of recognition such as type of speech, vocabulary size, speaker dependence mode, speaker independence mode, and environment. A speaker-dependent is the system which trained on specific speaker to recognize with high accuracy. A speaker-independent system is capable of recognizing speech from any speaker. Therefore, speaker-independent system is more difficult and practical than the speaker-dependent system. Nowadays, there are various advancements in this field like multi-lingual/cross-lingual ASR using statistical techniques such as neural network, Hidden Markov Model (HMM), and so on [1], [2].

The aim of this paper is to implement the Myanmar digits speech recognizer that is capable of recognizing and responding to digits speech inputs. The Myanmar digits speech recognizer would be applicable for various digits-based applications, such as phone dialing system, banking system, customer service queuing system, and other systems. We utilize the statistical modeling method based on HMM to recognize the digits of Myanmar language. In this work, we propose the context-independent model and context-dependent model. We compare their effects to the recognition performance and some significant experiment results have been obtained.

The remainder of this paper is organized as follows: Section II gives the overview of Myanmar language and the architecture of the proposed system is explained in section III. Section IV describes the Myanmar digit pronunciation. Section V presents the experimental results. Finally, the conclusion is in Section VI.

II. Nature of Myanmar Language

Myanmar Language (formerly known as Burmese) is the official language of Myanmar. The Myanmar script was adapted from the Mon Script, and descends from Brahmi script of South India [3], [4]. It is syllabic script and one of the languages which have complex structures and unique. Myanmar words are formed by collection of syllables and each syllable consists of up to seven different sub syllabic elements. Myanmar is written from left to right without any spaces between syllables or words. Nowadays, modern writing sometimes contains space after a sentence in order to enhance readability.

TABLE I: Basic Consonants

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

Myanmar script has 34 consonants (known as “Byee”) as described in Table I. They are served as the base characters for Myanmar words [4]. Vowels are known as “Thara”. In Myanmar language, vowels are the basic building blocks of syllable formation. However, some syllable or words can be formed from consonants only. On the other hand, multiple vowel characters can exist in a single syllable like in other

Manuscript received December 2015; revised March 19, 2016.

The authors are with the Department of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand (e-mail: zin2tun@gmail.com, gun@siit.tu.ac.th).

languages. Table II lists Myanmar vowels. In addition, Myanmar has medials which are known as "ByeeTwe". There are four basic medials and six combined medials in the Myanmar script [4]. Lastly, there are ten basic digits for counting the numbers in Myanmar language as shown in Table III.

TABLE II: VOWELS

Vowels				
◌	◌	◌	◌	◌
◌	◌			

TABLE III: NUMERALS

Numerals				
၀	၁	၂	၃	၄
၅	၆	၇	၈	၉

III. System Architecture

The system architecture of our Myanmar digits speech recognition system based on HMM. The main steps of HMM based speech recognition system are shown in Fig. 1.

A. Signal Analysis

The first step in speech recognition system is digitalization of the input analog speech signals that are recorded through microphone or telephone. The input speech is known as the utterance and an utterance may be isolated words, connected words, and continuous sentences [5]. This step is a process of speech preprocessing for further process. This process composes of signal pre-emphasis and windowing.

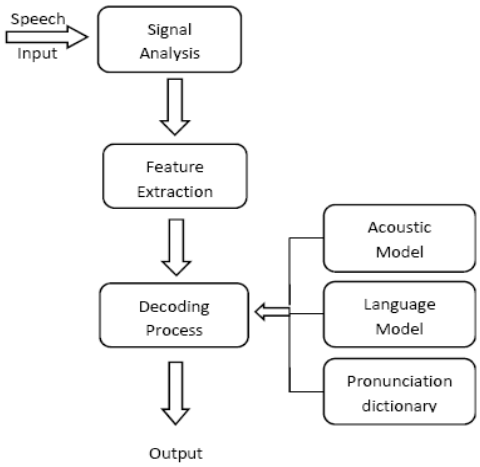

Fig. 1. Architecture of our Myanmar digit speech recognition system.

B. Feature Extraction

The main objective of feature extraction is to extract characteristic from speech signal that are unique, discriminative, robust and computationally efficient to each word which are then used to differentiate between different words [6]. The purpose is to extract relevant information from the speech frames, as feature parameter or vectors. In other words, it is responsible for converting the speech waveform to parametric representation for analysis and processing. The shape of the speech signal depends on the speed of the spoken speech and emotion of the speaker. Moreover, we also need to eliminate the presence of environmental noise in the speech signal. Therefore, preprocessing stage should be done on the original speech signal to extract meaningful features. This stage is referred as signal-processing front end. Many different techniques have been developed for feature extraction such as Linear Predictive Coding (LPC) and Mel-Frequency Cepstral Coefficients (MFCC). However, we use MFCC Feature extraction method in this work.

First, the waveform of each input string is pre-emphasized by using a high-pass filter whose transfer function is $H_{preem}(z) = 1\text{-}0.97z^{-1}$. After that, the speech signal is divided into a sequence of frames. Each frame is used Hamming window whose length is 25 ms long. Spacing between each frame interval is 10ms.

For each frame, Mel-Frequency Cepstral Coefficients (MFCCs) plus an energy feature are computed. MFCC is one of the most effective feature parameters in speech recognition. Moreover, it is based on the human ear's non-linear frequency characteristic and has a high recognition rate in practical application [5], [7]. We use MFCC feature extraction method to compute 39 features consisting of 13 mel-scaled cepstral, 13 delta, and 13 delta-delta features from each frame.

C. Modeling

Using feature vector as described above, modeling phrase generates the modelling technique for speech recognition. Generally, there are two models, namely acoustic model and language model that are used in the speech recognition process. Besides the two models, a pronunciation dictionary is also used to do the matching process of the speech recognition system.

1) Acoustic model

Acoustic Model is used to recognize the speech in speech recognition system. It takes audio recording of the speech and their text transcriptions to create a statistical representation of the sound for each word. It must contain the sounds for each word in the grammar. The words in the grammar provide the sequence of the sounds that make up particular words. These statistical representations are called Hidden Markov Model (HMM). This model contains statistical representation and each of this statistical representation is assigned a label called a phoneme. Each of the phonemes in the word has its own HMM and associated sound with symbols.

2) Language model

Language model helps a speech recognizer figure out how likely a word sequence is independent of the acoustics. The model tries to capture the properties of a language and to predict the next word in a speech sequence. The most common language model is the *n-gram* language model which contains statistics of the word sequence. This *n-gram* language model is used to search for the correct word sequence by predicting the likelihood of the *n*th word on the basis of *n*-1 preceding words [5].

3) Pronunciation dictionary

A pronunciation dictionary contains the phoneme list and is used to map words to their corresponding phonemes. That is, the different pronunciations are noted depending on the

corresponding phoneme in the dictionary.

D. Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a powerful tool in speech recognition. It provides an efficient algorithm for state and parameter estimation [1], [8]. HMM is used not only in acoustic models but also in many other applications. It is a statistical model that is assumed to be a Markov process with unknown parameters. HMM can be considered as the simplest dynamic Bayesian network. On the other hand, in Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. However, in a hidden markov model, the state is not directly visible and hence is called hidden.

IV. Phonetic of Myanmar Digits

Myanmar digits written form and their corresponding IPA used in this research are as following in Table IV.

TABLE IV: Myanmar Digit Pronunciation

English Digit	Myanmar Digit	Pronunciation (Myanmar)	IPA
0	၀	သုည	θṵɲa̰
1	၁	တစ်	tɪʔ
2	၂	နှစ်	n̥ɪʔ
3	၃	သုံး	θóʊɴ
4	၄	လေး	lé
5	၅	ငါး	ŋá
6	၆	ခြောက်	tɕʰaʊʔ
7	၇	ခုနစ်	kʰṵnɪʔ
8	၈	ရှစ်	ʃɪʔ
9	၉	ကိုး	kó

V. Experimental Results

In this section, we carry out experiments to evaluate the performance of the system and discuss the results. All experiments presented in this paper were done using HTK Toolkit [9] for building Hidden Markov Models (HMMs) from training data and evaluating with the testing data. The total number of recorded speakers is ten, four males and six females. Each speaker was prompted to utter 13 randomly-generated digit strings for per record. There were a total record of 320 digit strings for training and testing. Of these, 202 audio utterances from eight speakers: three males and five females were used as the training data and the remaining utterances are used as the testing data. For the testing phrase, we test with two new speakers; one male and one female who are not being included in the training phase for speaker independent system. On the other hand, we also test one female speaker who is already trained in training phase for speaker dependent model. Then, the correctness and accuracy percent are calculated and shown in Table V.

Then, we trained two acoustic models namely context-independent model and context-dependent model. The context-dependent model is also known as the triphone model, i.e., the context consists of one previous phone and one succeeding phone. We used a context-dependent acoustic model in which different HMMs are used for the same phone that occurs in different contexts. We compare the performances of these two models in Table V. During the training phase, we used Baum-Welch algorithm for re-estimating the HMM models and Viterbi search algorithm to compute the state sequence in HMM for a sequence of observed outputs in recognition phase [9]. Then we evaluate the results using HTK analysis tool. The results are shown in Table V.

Correctness is calculated by

$$\text{Correctness} = \frac{N - D - S}{N} \times 100\%,$$

where N is the total number, S is the number of substitutions, and D is the number of deletions. Accuracy is calculated by

$$\text{Accuracy} = \frac{N - D - S - I}{N} \times 100\%$$

where I is the number of insertions.

TABLE V: Results of System Performance

Model	Context-independent Model (Monophone)		Context-dependent Model (Triphone)	
	Correctness (%)	Accuracy (%)	Correctness (%)	Accuracy (%)
Speaker dependent	87.73	54.03	93.91	89.29
Speaker independent	71.64	38.25	62.02	53.49

VI. Conclusions

We present a speech recognition system for Myanmar digit here. In our experiments, the recognition of context-dependent model is better than the context independent model and it supports the existing theory to be more significant. Moreover, when we also compare with speaker dependent system and speaker independent system in the context dependent model, we found that the accuracy of speaker dependent system is 89.29% while in the accuracy of speaker independent model 53.49%. Although there are only ten digits to recognize, major difficulties lie on the pronunciation of a digit and the lack of grammar restriction of digit strings. Thus, language models cannot be used to improve recognition accuracy anymore. Consequently, the system performance lies entirely on speech characteristics such as environmental noise, speed of the speech, and mood of the speaker during recording. Although it is difficult to avoid these factors, effort should be taken to minimize the effect.

Acknowledgment

This research was partially supported by the Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University.

REFERENCES

[1] M. Aymen, A. Abdelaziz, S. Halim, and H. Maaref, "Hidden Markov models for automatic speech recognition," 2011.

[2] A. A. Abushariah, T. S. Gunawan, and O. O. Khalifa, "English digits speech recognition system based on hidden Markov models," in *Proc. International Conference on Computer and Communication Engineering*, 2010.

[3] Myanmar Language Commission, *Myanmar-English Dictionary*, 11th ed. Yangon, Myanmar, 2014.

[4] T. T. Thet, J. Na, and W. K. Ko, "Word segmentation for the Myanmar language," *Journal of Information Science*, 2008.

[5] S. Boruah and S. Basishtha, "A study on HMM based speech recognition system," in *Proc. IEEE International Conference on Computational Intelligence and Computing Research*, 2013.

[6] M. A. M. Abu Shariah, R. N. Ainon, R. Zainuddin, and O. O. Khalifa, "Human computer interaction using isolated-words speech recognition technology," in *Proc. the International Conference on Intelligent and Advanced System*, Kuala Lumpur, Malaysia, 2007.

[7] N. A. Meseguer, "Speech analysis for automatic speech recognition," 2009.

[8] P. Saini and P. Kaur, "Automatic speech recognition: A review," *International Journal of Engineering Trends and Technology*, 2013.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridege University, Engineering Department, 2008.

Zin Zin Tun was born in Yangon Province, Myanmar, in 1990. She received her B.C.Sc (Hons.) degree from the University of Computer Studies in Yangon (UCSY) in 2011. She is also a certified IT professional by IT Professional Examiniation Council (ITPEC). She is currently pursuing her master degree in the School of Information, Computer, and Communication Technonology, Sirindhorn International Institute of Technology, Thammasat University, Thailand. She is currently working in the research of speech recognition for Myanmar Language. Her research interests include optimization, data mining, natural language processing and liguistic research.

Gun Srijuntongsiri is an assistant professor in the School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand. He received his B.Sc. and Ph.D. degrees from Cornell University, USA, in 2002 and 2008, respectively. His research interests are in computer-aided design, computational geometry and optimization.