Normal and Whispered Speech Recognition Systems for Myanmar Digits

Nyein Nyein Oo
Computer Engineering and Information Technology Department
Yangon Technological University
Yangon, Myanmar

Masaru Yamashita
Department of Computer and Information Sciences
Nagasaki University
Japan

Shoichi Matsunaga
Department of Computer and Information Sciences
Nagasaki University
Japan

Abstract: Nowadays, Automatic speech recognition (ASR) technology comes as the popular innovation in human machine interaction. This technology allows a computer to recognize the spoken words and convert them to text data. In designing the computer systems that recognize spoken words, one of the challenging tasks is to be recognized spoken Myanmar digits. In this paper we focus on recognizing Myanmar digits spoken by normal voice and whispered voice. Myanmar digits recognition system for both types has been developed by using Hidden Markov Model in HTK tools and Mel Frequency Cepstral Coefficients (MFCC) technique has been used to convert the speech waveform into a set of feature vectors for recognizing the vocalization of a word. In our experiments, HMM-based acoustic and language models are used to evaluate the performance of speech recognizer for both speaker dependent and speaker independent. According to the experimental results, the performance of speaker dependent speech recognition system for normal voice and whispered voice are 90% and 88.7% respectively. The performance of speaker independent speech recognition system for normal voice and whispered voice are 67.3% and 65.7% respectively. We found that the performance of both type of speaker dependent is higher than those of speaker independent.

Keywords: Automatic Speech Recognition, Hidden Markov Model, HTK tools, MFCC, Dependent and Independent Speakers

1. INTRODUCTION

Spoken language is used to communicate information from a speaker to a listener. Thus, speech is a communication method among people. In our daily life, we communicate each other for doing our work via a telephone or using various applications on internet. Moreover, in machine interaction system human can communicate machine and vice vasa. To communicate between machine and human, we need to train the machine to know what we say. Nowadays, ASR technology is emergence in a lot of applications and services in our daily life such as banking system, telephone system, and web-based application system. ASR is a technology that recognizes the spoken words. ASR technology is not yet the ultimate goal that is to allow a computer to recognize in real-time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent. In developed countries, ASR for different languages have been developed by many researchers using different approaches of methods [3]. There is a few researcher to develop the ASR system for Myanmar language and there are some papers which showed auspicious accuracy results. Myanmar ASR system is still in research and the existing work are not far-reaching adequate. The main aim of this paper is to train and test a Myanmar speech recognition system which can recognize the speech signals of isolated digits of Myanmar language, in Linux environment using HTK and MFCC as the feature extraction method and self-recorded database for speaker dependent and independent approach. To build the self-recorded database, we collect the voice data from six persons, five female and one male, for 10 digits spoken by two types of voice: normal voice and whispered voice. 10 sample files for each digit are recorded from them. Thus we collect 1200 files for both voice type. In speech recognition system, there are three speech recognition systems, 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. Thus these systems are more flexible than the speaker dependent systems, but difficult to develop and get the better accuracy than speaker dependent systems. According to the characteristics of new speakers, the processes of speech recognition system are adapted to become a speaker adaptive system. In this paper, we study and evaluate the performance of speaker dependent and independent systems. 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. In this study, we develop the isolated-word recognition system especially for Myanmar digit.

The remaining sections of this paper are organized as follows. Section 2 discusses the related works of the speed recognition systems that have been developed by many researchers using different approaches of methods. Hidden Markov Model (HMM) and MFCC used in this research are described in Section 3. Section 4 presents the overview of speech recognition system and how to build the acoustic model for Myanmar digit speech recognition system. Section 5 describes the experimental results of our system. In section 6, we describe the conclusion of our experimental results.

We ask that authors follow some simple guidelines. This document is a template. An electronic copy can be downloaded from the journal website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website

2. LITERATURE REVIEW

In many countries, speech recognizers for their languages have been developed by many researchers using different approaches and methods. In Myanmar, some researcher has studied and investigated to build a speech recognizer for Myanmar language. But this system does not completely finish to cover 100% accuracy. In this section, we describe some researcher's works that are relevant with our study.

Shaikh Naziya et al. presented the speech recognition system for Urdu digits and they used Linear Predictive Coding (LPC) and Hidden Markov Model (HMM) techniques to analyze the performance of Speech Recognition System for Urdu Digits [4]. They developed the digit corpus for Urdu language and collected Urdu digits spoken by 50 native speakers that contain 50% female and 50% male speakers whose age is above 18 years to develop this corpus. Three utterances are taken from each person and two utterances were used for training phase and one utterance was used for testing. Their performance accuracy is 74% for zero to nine (0-9) digits and when they tested six and nine, they got 100% performance accuracy for them.

P. Prithvi et al. analyzed the recognition rate of English digits ("one" to "ten") in noisy environment and speaker independent recognition system [2]. In this paper, they used Hidden Markov Models is one of the best pattern recognition approaches. In their previous paper, they presented the recognition rate 67% for speaker independent in noisy environment. In this paper they used hybridized model of Vector quantization and Hidden Markov Model to improve the recognition rate. Their experimental results also show the improvement in recognition rate is up to 81.8 % in noisy environment for speaker independent.

MarutiLimkar et al. proposed an approach to recognize spoken English digits zero to nine in an isolated way by different male and female speakers. To accomplish the recognition system, they used the endpoint detection, framing, normalization, Mel Frequency Cepstral Coefficient (MFCC) and DTW algorithm are used to process speech samples and implemented the algorithm to test on speech samples [1]. Their system can recognize the recorded isolated word English digits, that is 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight' and 'nine'. Their algorithm's recognition rates are 80.0% for zero, 95% for one, 80% for two, 100% for three, 90% for four, 100% for five, 80% for six, 100% for seven, 100% for eight and 80% for nine. According to their experimental results the average accuracy test is about 90.5%.

Shabnam Ghaffarzadegan et al. proposed the model and feature based strategies for automatic whispered speech recognition. Their goal is to compensate for the mismatch between neutral-trained recognizer models and parameters of whispered speech and to avoid the degrading performance of speech recognizer [5]. They used Vector Taylor Series (VTS) algorithm to produce a pseudo-whisper generation from neutral speech samples for efficient acoustic model adaptation. In this study they analyzed the efficiency of frequency-based spectral transformations vocal tract length normalization (VTLN) and Shift for whisper speech recognition and proposed a novel approach to pseudo-whisper generation for acoustic model adaptation, requiring only a small amount of whisper samples. It was found that both the spectral transformations and the VTS approach can considerably improve the recognition performance and also perform well when combined together. They described VTS approach can give the better performance only we have small amount of whisper samples.

In this paper, we emphasize to build the Myanmar digit speech recognition system for two types of speech: whispered speech and normal speech and compare their recognition accuracy rates. In future, when we build ASR for Myanmar language, the results acquired from this experiment can be applied in adaptation of ASR to get the better accuracy of recognition system for Myanmar language. The following section shows the methods used in Myanmar digit speech recognition system.

3. METHODS USED IN SPEECH RECOGNITION SYSTEM

In this section we describe the methods used in speech recognition system and these methods can be used to extract the feature from voice signal and build the acoustic model. There are many different techniques to extract the voice feature such as Linear Predictive Coding (LPC), Perceptual Linear Predictive Analysis (PLP) and Mel-Frequency Cepstral Coefficients (MFCC). In this research we use MFCC feature extraction method. To do the speech recognition process, we need to create acoustic model, language model and a pronunciation dictionary that is used to match the speech signal. To create these models we use Hidden Markov Model (HMM) toolkit (HTK) and HMM can hold the statistical representations for phoneme. The following subsections describe the MFCC feature extraction method, HMM and HTK.

3.1 Mel-Frequency Cepstral Coefficients (MFCC)

In 1980's, Davis and Mermelstein introduced Mel Frequency Cepstral Coefficents (MFCCs) which are the main feature type for automatic speech recognition (ASR). Features extraction is to identity the shape of vocal tract in the audio signal. The shape of the vocal tract establishes itself in the envelope of the short time power spectrum, and the work of MFCCs is to exactly characterize this envelope. The characteristics of speech between linearly at low frequencies and logarithmically at high frequencies are captured as the important feature. Each tone in speech has been defined with frequency f (Hz) and a subjective pitch (Mel scale) for describing as a feature. Power spectrum based on the center frequency and bandwidth is used to calculate the cepstral coefficients. MFCCs are commonly used as features in speech recognition systems. The following equation is used for converting the Mel-frequency (m).

$$m=2595\log10\ (1+\frac{f}{700}) \qquad \text{Equation (1)}$$

3.2 Hidden Markov Model (HMM)

Many scientific and engineering applications can be represented as mathematics to implement these applications in computer. In speech recognition systems, we use HMM that applies the probability theory in mathematics to recognize the speech signal. In this system, the speech signal message is coded as a sequence of one or more symbols as shown in Figure 1 [7]. To recognize the symbol sequence specified a spoken utterance, firstly we need to convert the continuous speech waveform to a sequence of equally spaced (10ms) discrete parameter vectors. The parameter vectors represent the speech waveform. Thus, the work of speech recognizer is the mapping between sequences of speech vectors and the sequences of symbols.

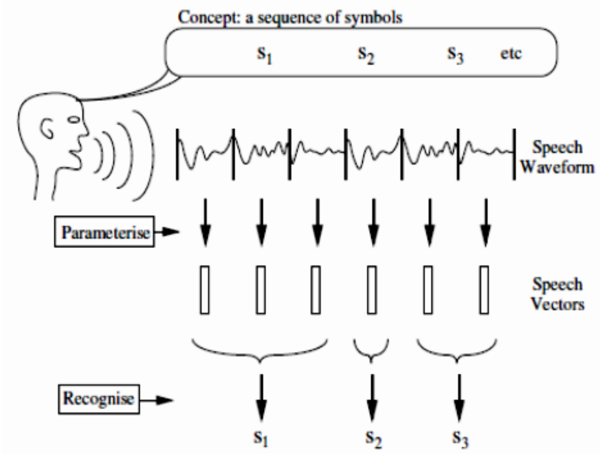

Figure 1. Message Encoding/Decoding

Each spoken word can be represented by a sequence of speech vectors or observations Ot (observed at time t) and the vocabulary word is represented by wi (i’th vocabulary word). In HMM based speech recognition, the sequence of observed speech vectors corresponding to each word is generated by a Markov model as shown in Figure 2 [7]. A Markov model is a finite state machine which changes state once every time unit and each time t that a state j is entered, a speech vector Ot is generated from the probability density bj(Ot). Furthermore, the transition from state i to state j is also probabilistic and is governed by the discrete probability aij.

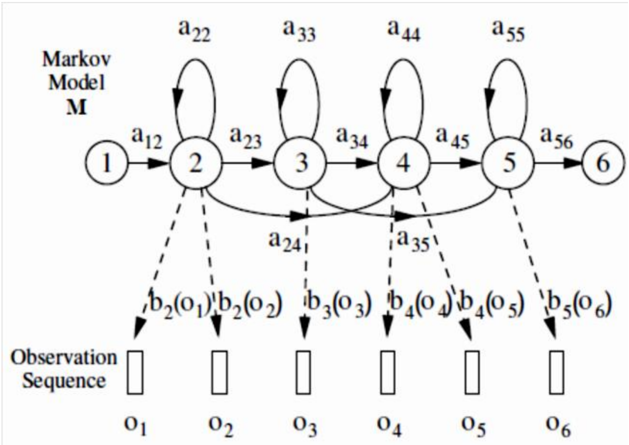

Figure 2. The Markov Generation Model

The following equations (Equation 2 and 3) are used to calculate the values of probability required in Hidden Markov Model. In this HMM, the state sequence X is hidden and observation O is known value so it is called Hidden Markov Model.

$$P(w_i|O)=\frac{P(O|w_i)P(w_i)}{P(O)} \qquad \text{Equation (2)}$$

$$P(O,X|M) = a_{12}b_2(O1)a_{22}b_2(O_2)a_{23}b_3(O_3)... \qquad \text{Equation (3)}$$

3.3 Hidden Markov Model (HMM) Toolkit

Cambridge University Engineering Department (CUED) developed the Hidden Markov Model (HMM) Toolkit that is a HMM-based speech recognition tool for modelling time series. This toolkit can support to build and manipulate Hidden Markov Models (HMMs) and contains many library modules and tools that can be used not only in speech recognition research but also in a lot of potential applications such as speech synthesis, character recognition and DNA sequencing. HTK training tools can be used in estimating the parameters of a set of HMMs based on training utterances and their associated transcriptions. HTK recognition tools can be used to recognize the unknown utterances. Figure 3 shows the software architecture of Hidden Markov Model Toolkit [7].

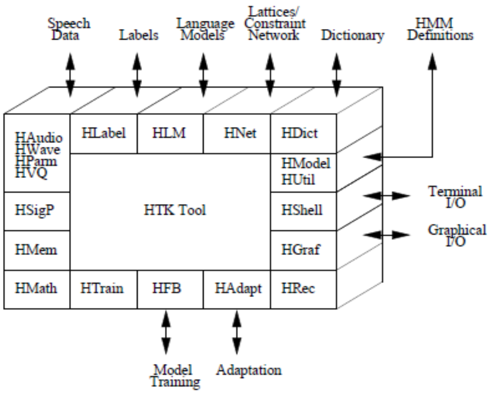

Figure 3. Software Architecture of HMM

In this architecture, there are many library modules such as HShell, HMem, HMath, HSigP, HLabel, HLM, HNet, HDict, HVQ and HModel. HShell controls the interaction between the user input/output and operating system. HMen manages the usage of memory. HMath supports the operations based on mathematics and HSigP supports the required operations in signal processing for analyzing the speech. HLabel provides the interface for label files, HLM for language model files, HNet for networks and lattices, HDict for dictionaries, HVQ for VQ codebooks and HModel for HMM definitions.

4. MYANMAR DIGITS SPEECH RECOGNITION SYSTEM

Myanmar digits speech recognition system is built to recognize the speech signals of isolated Myanmar digits. This system is run and tested in Linux environment using HTK toolkit. To implement this system, we recorded the two types of voice such as normal voice and whispered voice from six persons, five female and one male and 10 sample files for each digit with each type are recorded from them to build the self-recorded database that contains 1200 files for training and testing the Myanmar digits speech recognition system.

Figure 4 shows the overall diagram of Myanmar Digit Speech Recognition System. In Myanmar Digit Speech Recognition System, it contains four main phases: data preparation phase, training phase, testing phase and analysis phase. In data preparation phase, the set of speech data files and their associated transcriptions are required to build a set of HMMs. Thus, we use 1200 recorded speech data files and convert the recorded voice data files (.wav) to (.ad) file with the help of SOX (Sound eXchange).

In this study we use MFCC to extract the features from speech data files. To accomplish the feature extraction process, we use the set of speech data files that are stored in speech folder and configuration file containing configuration parameters that are stored in script folder.

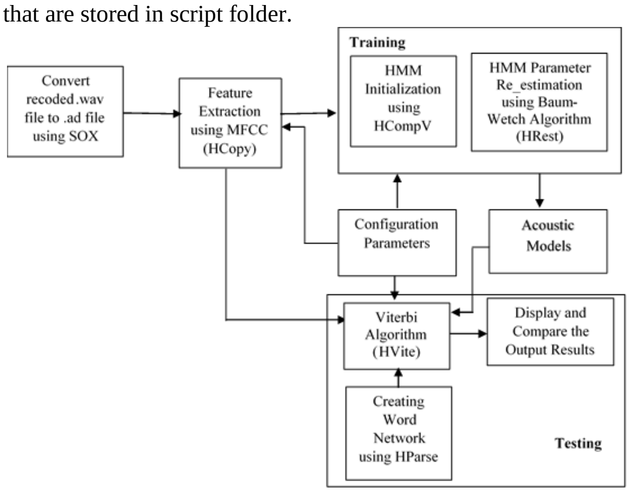
Training
Convert recoded.wav file to .ad file using SOX
Feature Extraction using MFCC (HCopy)
HMM Initialization using HCompV
HMM Parameter Re_estimation using Baum-Wetch Algorithm (HRest)
Configuration Parameters
Acoustic Models
Viterbi Algorithm (HVite)
Display and Compare the Output Results
Creating Word Network using HParse
Testing

Figure 4. Myanmar Digit Speech Recognition System

We also need to write the script file that contains speech data files path (source folder) and destination folder path to store the feature files. This script file is shown in Figure 5. In data preparation phase, the tool HCopy is used to extract the features from speech data files by setting the appropriate configuration parameters. When HCopy is run all speech data files are converted to mfc files under the specified folder. In data preparation phase, the language model, dictionary and word network are needed to build. To accomplish them, the tool HParse can be used.

```
/root/speech/S0001.wav   /root/script/ train/S0001.mfc
/root/speech/S0002.wav   /root/ script//train/S0002.mfc
/root/ speech/S0003.wav  /root/script//train/S0003.mfc
/root/ speech/S0004.wav  /root/script/train/S0004.mfc
```

Figure 5. Script File Containing Speech Data Files Path

In training phase, two processes such as parameter initialization and parameter re-estimating processes are needed to perform. In first process, we need to define the HMM initialization parameters to create acoustic model for each digit. Thus, we use proto definitions to initialize HMM model for each digit. Proto definitions are shown in Figure 6. To train the speech data files, proto_8states, 25 vector size, 8x8 transitional matrix under <TRANSP> and some optional parameters are well-defined in HMM definition. Each state contains mean and variance initialize values.

In initialization process, the tool HCompV can be used to compute the global speech variance and mean values. This command is run with the script file containing the path of mfc files, the configure file containing appropriate configuration parameters and the proto definition file to produce the HMM files for each digit.

In second process, we need to re-estimate the parameters of speech by using Baum-Welch Algorithm. The tool HRest can be used to refine the parameters of existing HMMs using Baum-Welch Algorithm. HRest, a training tool, is used to produce the acoustic model for each digit.

```
<STREAMINFO> 1 25
<VECSIZE> 25 <NULLD><MFCC_E_D_N_Z> <DIAGC>
~h "proto_8states"
<BEGINHMM>
            <NUMSTATES> 8
            <STATE> 2
            <MEAN> 25
            <VARIANCE> 25
            <GCONST>
            <STATE> 3
            <MEAN> 25
            <VARIANCE> 25
            .
            .
            .
<TRANSP> 8
<ENDHMM>
```

Figure 6. Proto Definition for HMM 8_States

In testing phase, HVite, a recognition tool, is used to get the recognition results for each digit. When HVite is run, the configuration file, the language model, dictionary and word network are used to match with the trained data files produced in training phase. In analysis phase, we analyse the output results of testing phase. The following section discuss these output results.

5. EXPERIMENTAL RESULTS

In this experiment, two types of speech (Normal and whispered speech) data files are recorded from six persons for each digit. Thus, we get 1200 speech data files for both types. In this experiment, we test the four ways: normal speech for speaker dependent (NSSD), normal speech for speaker independent (NSNID), whispered speech for speaker dependent (WSSD), and whispered speech for speaker independent (WSSID). In testing and evaluation processes, 600 speech data files are used for each system. The following Table 1 shows the tested results for each digit and their correct percentages for each type.

Table 1. Experimental Results

Type	No. of File	No. of Correct for each digit										Tot	Correct (%)
		0	1	2	3	4	5	6	7	8	9		
NSSD	600	52	53	57	51	54	52	55	58	56	52	540	90
WSSD	600	54	51	53	46	54	51	57	53	60	53	532	88.7
NSSID	600	49	52	52	44	46	38	32	35	31	25	404	67.3
WSSID	600	50	46	38	32	30	31	43	40	52	32	394	65.7

According to the experimental results, the performance of speaker dependent speech recognition system for normal voice and whispered voice are 90% and 88.7% respectively. The performance of speaker independent speech recognition system for normal voice and whispered voice are 67.3% and 65.7% respectively. We found that the performance of both type of speaker dependent is higher than those of speaker independent.

6. CONCLUSIONS

In this research, we use HTK toolkit that supports to build the speech recognizer for Myanmar digits. HTK supports both isolated whole word recognition and sub-word or phone based recognition. Many researcher also use it for the speech recognition research of various languages. According to the studied papers, MFCC and HMM techniques can give higher recognition rate than other methods [6]. In this experiment, we also use MFCC and HMM to recognize Myanmar digits. According to the experimental results, the recognition rates for normal and whispered voice speaker dependent are 90% and 88.7% respectively. The recognition rates of speaker independent system for normal voice and whispered voice are 67.3% and 65.7% respectively. As a conclusion, the performance of both type of speaker dependent is higher than those of speaker independent and the tested results are depend on the way of speaker’s utterance. This experiment can be extended for the recognition of alphabets and words.

7. ACKNOWLEDGMENT

Authors would like to express my sincere gratitude to Japan International Cooperation Agency (JICA) for supporting to enhance Engineering Higher Education in Myanmar and Computer and Information Systems Lab at Nagasaki University for allowing doing this research during staying in Japan.

8. REFERENCES

[1] MarutiLimkara, RamaRaob and VidyaSagvekarc, *Isolated Digit Recognition Using MFCC AND DTW*, International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN (Print): 2278-8948, Volume-1, Issue-1, 2012.

[2] P. Prithvi, Anil Kumar and Dr. T. Kishore Kumar, *Speaker Independent Speech Recognition of English Digits using Hybridized VQ-HMM Model in Noisy Environment*, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181 Vol. 3 Issue 4, April – 2014.

[3] Pooja Prajapati and Miral Patel, *A Survey on Isolated Word and Digit Recognition using Different Techniques*, International Journal of Computer Applications (0975 – 8887) Volume 161 – No 3, March 2017.

[4] Shaikh Naziya S. and R. R. Deshmukh, *LPC and HMM Performance Analysis for Speech Recognition Systemfor Urdu Digits*, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. IV. (Jul.-Aug. 2017), PP 14-18.http://www.iosrjournals.org.

[5] Shabnam Ghaffarzadegan, Hynek Boril and John H. L. Hansen, “Model and Feature Based Compensation for Whispered Speech Recognition”, Fifteenth annual Conference of the International Speech Communication Association, Singapore, September 2014.

[6] Vimala.C and Dr.V.Radha, *A Review on Speech Recognition Challenges and Approaches*, World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 1, 1-7, 2012.

[7] Steve Young, Gunnar Evermann, The HTK Book (for HTK Version 3.4) Available: http://htk.eng.cam.ac.uk/docs/docs.shtml.