AUTOMATIC SPEECH RECOGNITION FOR RAKHINE LANGUAGE USING HIDDEN MARKOV MODEL AND GAUSSIAN MIXTURE MODEL

HNIN THI DAR KYAW

UNIVERSITY OF COMPUTER STUDIES, YANGON

M.C.Sc **JUNE, 2022**

Automatic Speech Recognition for Rakhine Language using Hidden Markov Model and Gaussian Mixture Model

By

Hnin Thi Dar Kyaw

B.C.Sc

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Computer Science

(M.C.Sc)

University of Computer Studies, Yangon

June 2022

ACKNOWLEDGEMENTS

To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my appreciation and sincere thanks to **Dr. Mie Mie Khin**, the Rector of the University of Computer Studies, Yangon, for her kind permission to submit this dissertation.

I sincerely wish to express my greatest pleasure and special thanks to **Dr. Khin Mar Soe**, Professor, Natural Language Processing Lab, the University of Computer Studies, Yangon, for her kindness, invaluable suggestions and assistances to my effective research completion throughout my study period. I sincerely wish to express deeply and special thanks to **Dr.Win Pa Pa**, Professor Natural Language Processing Lab, the University of Computer Studies, Yangon for her kindness, critical thinking, invaluable suggestions and comments and providing the necessary support for research during the preparation of thesis research.

I would like to mention deeply and special thanks to my supervisor, **Dr. Aye Nyein Mon**, Lecturer, Natural Language Processing Lab, the University of Computer Studies, Yangon, for her keen interest, valuable guidance, supervision, encouragement and constructive comments during the period of study towards completion of this piece of research work.

I would like to thank course coordinator, **Dr. Thi Thi Soe Nyut**, Professor, **Dr. Si Si Mar Win**, Professor, and **Dr. Tin Zar Thaw**, Professor, Faculty of Computer Science, the University of Computer Studies, Yangon for their superior suggestions and administrative supports during my academic study.

I wish to thank **Daw Win Lai Lai Bo**, Assistant Lecturer, Department of English, the University of Computer Studies, Yangon for editing my thesis writing from language point of view.

Finally, I would like to express my special thanks to my teachers of NLP Lab for their assistance and motivation during this period and my beloved parents, my brother and my younger sister. It is their love, understanding and financial support that enable me to accomplish this thesis.

ABSTRACT

The automatic recognition of speech means enabling a natural and easy mode of interaction between human and machine. The process of speech recognition is to translate speech signal into text sequence. Automatic Speech Recognition (ASR) has been carried out by many researchers for their particular languages to provide their nations in language technologies. Therefore, this thesis aims to develop automatic speech recognition for Rakhine language, one of the main ethnic groups in Myanmar. Rakhine language is a low-resourced language and speech data are no freely available. Thus, in this work, speech corpus is built on two domains: broadcast news and daily conversations data. Broadcast news is collected from the web and the conversational data is recorded by uttering with own voice. This corpus is applied to develop the Rakhine ASR. Feature extraction is one of the components of ASR and its function is to extract feature from incoming speech signals. In this work, Mel Frequency Cepstral Coefficient (MFCC) feature extraction technique is used. Because of the phonetic dictionary is essential part for implementing Rakhine speech recognition system, pronunciation lexicon is built for Rakhine language in this work. And, Rakhine language model is also created by utilizing n-gram. In developing ASR, acoustic models is the crucial component and is established the connection between acoustic feature and phonetic. For this Rakhine ASR research, the Gaussian Mixture based Hidden Markov Model (HMM-GMM) is utilized. By using HMM-GMM, Rakhine ASR performance gets promising results.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

Natural Language Processing (NLP) is defined as artificial intelligence methods of interacting with a computer in a natural language. It enables the computer to understand human Language directly. The research of NLP and speech processing have many applications area such as Machine Translation, Speech Recognition, Speech synthesis, Question-Answering Systems, Dictations and Spelling checker. Among them, speech recognition is one of the important types of research carried out in the word of artificial intelligence.

Since the early 1960s, researchers have been working on developing systems that can record, interpret, and understand human utterances. Communicating with computers using voice can help developing countries by implementing language technology in electronic governance systems.

Nowadays, speech recognition is one of the modern technologies for human computer interaction and speech signal is a prominent feature to communicate with natural language. Speech is the most natural form of communication among human beings and there are various languages in the world that are spoken by humans for interaction. As communication among human beings is mainly vocal, it is natural way for people to expect speech interfaces with the computer. The computer system which can understand the spoken language can be very useful in various areas.

Basically, speech recognition is the process which converts speech signal into text output without typing by hand. Automatic Speech Recognition (ASR) is a challenging task because of human speech signals' variability. Many parameters have an impact on the accuracy of speech recognition system such as speaker dependency, vocabulary size, types of speech (isolated, connected, continuous), and recognition time and recognition environment conditions.

1.1 Motivation and Objectives of the Thesis

In recent years, many nations tried to create speech recognition systems for their own languages. They had done many researches in speech recognition with their languages. The accuracy of automatic speech recognition (ASR) systems has been

improved by exploring new architectures or applying particular properties of the target language. Many researches are currently being worked around the world to develop robust automatic speech recognition systems for both well-resourced and low-resourced languages.

For low-resourced languages, they described the development of ASR development in their languages by collecting the data from the beginning. Myanmar language is also one of the low-resourced languages and automatic speech recognition for Myanmar language is being conducted by utilizing different technique However, there is no research in speech recognition for Rakhine language, one of the main ethnic groups in Myanmar. From the above motivation, it is necessary to develop in speech recognition for Rakhine Language.

The main purpose of this thesis is to build Rakhine automatic speech recognition on continuous speech. Moreover, this work aims to provide Rakhine speech processing systems such as dictation, telephone communication, speech to speech translation, automatic question and answering and voice commanding.

Speech processing systems can provide individuals in the disability community. Speech to text processing can help hearing-impaired people in reading online news and it is also convenient and useful for news reporters in transcribing the news. Therefore, this is one of the objectives in the development of Rakhine ASR. Finally, this work is conducted for the development of ASR technology in Rakhine language.

1.2 Contribution of the Thesis

There are three main parts of contribution in this thesis.

The first contribution is the establishment of a speech corpus for Rakhine language. In order to develop Rakhine ASR, speech corpus is needed for training purpose. Although speech data are widely available for well-resourced languages such as English, there is no freely available for low-resourced languages. Thus, speech corpus building is an important step to develop speech recognition system especially for low-resourced languages. Rakhine language can be regarded as a low-resourced language and there is no pre-created speech corpus for speech processing research. Therefore, speech corpus is constructed for Rakhine language by using two ways. The

first way is that daily conversations texts are collected and then these are recorded by reading the transcription. The second one is that the speech data is collected from Broadcasts news and manually transcribe them into texts. Hence, two types of domains (daily conversations and Broadcasts news) are included in this work.

The second contribution is to create Pronunciation lexicon and language model for Rakhine language. Pronunciation lexicon is one of the main components of speech recognition system and language model becomes a vital role in statistical approaches. Therefore, Rakhine pronunciation lexicon and language model is created to develop Rakhine ASR in this thesis.

The final contribution is creating an acoustic model for Rakhine. The acoustic model is a crucial part of the ASR and it is significant to improve the ASR performance. There are many techniques for training and building acoustic model such as Hidden Markov Model-Gaussian Mixture Model (HMM-GMM), Convolution Neural Network (CNN), Deep Neural Network (DNN) and Time Delay Neural Network (TDNN). In this task, HMM-GMM based acoustic model is built for developing Rakhine ASR.

1.3 Organization of the Thesis

This thesis is organized into six chapters, including the basic theory of speech recognition, constructing Rakhine speech corpus, the nature of Rakhine language, the development of an acoustic model based on HMM-GMM, the error analysis of the evaluation results, and conclusion and future work Rakhine ASR.

Chapter 1 presents the introduction to speech recognition, motivations, objectives and the contributions of thesis work. Literature reviews on automatic speech recognition and related works on ASR researches in Myanmar are presented in Chapter 2. In Chapter 3, speech corpus building for Rakhine ASR is described and the speech corpus statistics is also presented in this chapter.

Chapter 4 introduces the nature of Rakhine language and the basic phonemes of Rakhine language. In addition, the structure of Rakhine language is explained and the tone of Rakhine language is also discussed. Chapter 5 describes acoustic models based on Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), Mel Frequency Cepstral Coefficient (MFCC) feature extraction technique, language model

construction for Rakhine language and decoder are explained. The experimental setup and the evaluation results are also shown in this chapter.

Chapter 6 summarizes the Rakhine ASR research work. In this chapter, the advantages and limitations of the work are described. Furthermore, it also indicates future research on Rakhine ASR.

CHAPTER 2

LITERATURE REVIEWS AND RELATED WORK

This chapter describes the literature reviews on automatic speech recognition (ASR) and Myanmar ASR researches found in publications.

2.1 Introduction to Speech Recognition

Speech Recognition is an interdisciplinary subfield of computational linguistics and it incorporates knowledge and research in the linguistics. In the world of science, the computer has always understood human mimics. The creation of recognition system is that humans can interact more easily interact with computers, robots or any machine through speech or vocalization rather than difficult instructions. The basic idea of speech recognition is the processing task of a computer to map an acoustic speech signal to text sequence [19].

2.2 Classification of Speech Recognition System

Speech recognition system can be characterized by a number of parameters. These parameters specify the capability of speech recognition systems and include speaking mode (isolated word, connected words, continuous speech and spontaneous speech), speaking styles (read speech formal style and conversational speech with causal style), speaking situation (human-to-machine-speech and human-to-human-speech), speaker dependency (speaker dependent, speaker adaptive and speaker independent), vocabulary size (small vocabulary, medium vocabulary, large vocabulary).

2.2.1 Classification based on the Speaker Utterances

2.2.1.1 Isolated Words

Isolated word recognizer operates on single word at a time requiring a pause between saying each word. Often, these types of speech have listening time and no time to listen in that states which they require the speaker to pause between utterances.

2.2.1.2 Connected Words

This system requires minimum pause between utterances to flow speech signal smoothly. Connected word recognizer is closely similar to isolated words.

2.2.1.3 Continuous Speech

Continuous speech is the natural way to interact between human and computer and it is normal human speech with no silent pauses between the words. A continuous speech system operates in which words are concatenated together. It is operated without a pause. This kind of speech is difficult to understand.

2.2.1.4 Spontaneous Speech

Spontaneous speech is natural sounding. Spontaneous speech ASR system ability should be able to handle a variety of natural speech characteristic such as words being run at the same time.

2.2.2 Classification based on Vocabulary Size

2.2.2.1 Small Vocabulary

Only a limited number of vocabularies can be recognized in speech recognition system. This is identified as small vocabulary speech recognition system.

2.2.2.2 Medium Vocabulary

Considerable number of vocabularies can be recognized in speech recognition system. It is called medium vocabulary speech recognition system.

2.2.2.3 Large Vocabulary

Large number of vocabularies can be recognized in speech recognition system. These systems are defined as large vocabulary speech recognition system.

2.2.3 Classification based on Speaker Mode

2.2.3.1 Speaker Dependent

Speaker dependent system is developed to operate for a single speaker. These systems are easier in developing, cheaper to buy and more accurate. But, this is not as flexible as speaker independent systems or speaker adaptive.

2.2.3.2 Speaker Independent

Speaker independent system is created in recognizing for any speaker. These systems are more difficult than speaker independent system to develop and more expensive. And, speaker independent accuracy is lower than of speaker dependent system. However, it is more flexible.

2.2.2.3 Speaker Adaptive

Speaker adaptive is created to adapt its operation to the characteristics of news speaker. Its difficulty lies somewhere between speaker independent and speaker dependent systems.

2.3 Performance of Speech Recognition Systems

The performance of speech recognition accuracy and speed are the two most common criteria used to determine speech recognition system performance. The measurement of accuracy in speech recognition is word error rate, and speed is commonly evaluated utilizing Real Time Factor (RTF). Word error rate (WER) can be computed by using in the following equation:

$$\text{WER} = \frac{S+D+I}{N} \tag{2.1}$$

Where, S is the number of substitutions, D is the number of deletions and I is the number of insertions. N is the number of words in the reference.

If inputs of duration I require time P to process, RTF can be calculated by using in the following equation:

$$\text{RTF} = \frac{p}{I} \tag{2.2}$$

Other performance measurements are Concept Error Rate (CER), Single Word Error Rate (SWER) and Command Success Rate (CSR) [17].

2.4 Overview of Automatic Speech Recognition

The automatic speech recognition system mainly involves five stages: feature extraction, acoustic model, language model, pronunciation lexicon and decoding [11].

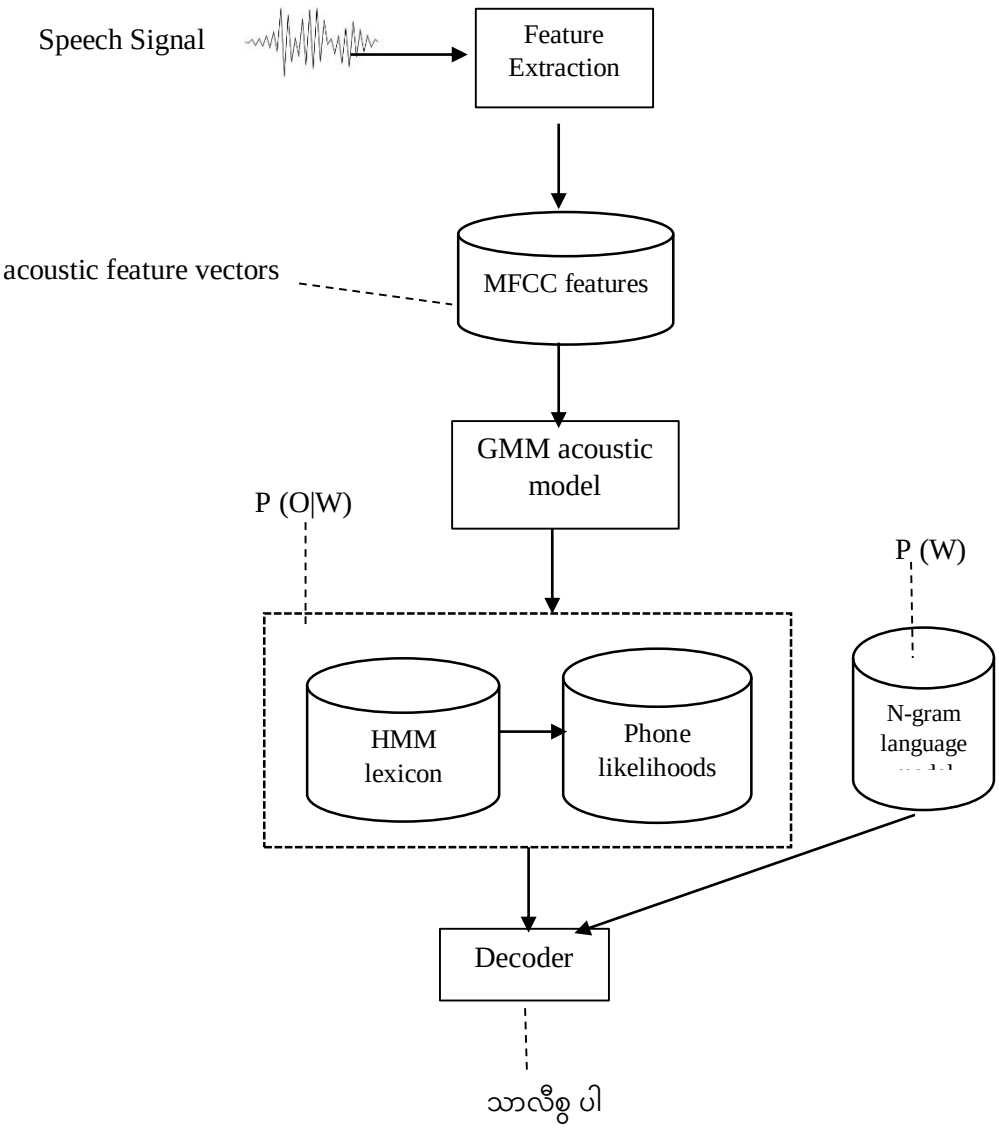


Figure 2.1: Overview of ASR architecture

Figure (2.1) shows the overview of ASR architecture. In the feature extraction stage, speech signal is converted into a sequence of acoustic feature vectors. In the phone likelihood computation stage, Gaussian Mixture Model (GMM) classifier is used to compute for each HMM state, corresponding to a phoneme and likelihood of a given

acoustic features. In Decoding stage, the system takes the acoustic model (AM), which consists of sequence of acoustic likelihood, plus an HMM lexicon combined with the language model (LM) and the system outputs the most likely sequence of words.

The most likely sentence based on some observation sequence O can be computed by applying the product of two probabilities for each sentence and choosing the sentence for which this product is greatest. This is depicted by using the following equation:

$$W = \text{argmax } P(O|W)\, P(W) \tag{2.3}$$

In equation (2.3), the acoustic model can be calculated by the observation likelihood, P (O/W) and language model can be gained for computing the prior probability P (W).

2.5 Myanmar Automatic Speech Recognition Researches

Many researchers have been done for ASR both well-resourced and low-resourced languages. ASR research has been conducted on the Myanmar Language by using different technique to improve Myanmar ASR performance such as Hidden Markov Model and Gaussian Mixture Model (HMM-GMM), Convolutional Neural Network (CNN), Artificial Neural Network (ANN), Time delay neural network (TDNN) etc. However, ASR related to Rakhine Language has not been done yet. In this work, Rakhine ASR is implemented for the first time. This system using HMM-GMM is a speaker independent and continuous speech recognition in Rakhine language.

The related Myanmar ASR researches found in publications is stated in the following:

Thin Thin Nwe et.al, [23] presented by using hybrid artificial neural network and hidden markov model for Myanmar language speech recognition. Syllable-based segmentation was used in this work. In feature extraction stage, Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coding (LPCC) and Perceptual Linear Prediction (PLP) were used. Training database involves 260 female speaker's utterances. In developing acoustic model training, hybrid ANN/HMM method was utilized for developing automatic speech recognition with a medium size vocabulary.

Wunna Soe et.al, [26] proposed syllable-based continuous automatic speech recognition for Myanmar. In this study, Myanmar ASR acoustic model was created using HMM-GMM and syllable based language model is created for Myanmar to predict syllable sequence in speech recognition system. The syllable-based language model was open-vocabulary language model. Moreover, syllable-based phonetic dictionary was also used in this research and the grammatical based word segmentation algorithm was also contributed to create the word-based language model. In addition, the syllable-based language is compared model with word-based language model using different language modeling toolkit (CMU and SRILM). It showed that syllable-based language model achieved better accuracy than word-based language model. Both speaker dependent and speaker independent were analyzed by utilizing three experiments: amount of training data, number of tied states (senones), and number of densities.

Aye Nyein Mon et. al., [1] proposed the effect of tones on both syllable and word-based ASR performance. In this task, the experiments were conducted based on the modeling of tones by incorporating them into the set of phonemes and integrating them into the Convolutional Neural Network (CNN). Additionally, for more effective tone modeling, tonal questions were utilizing to create the phonetic decision tree. The experiments with tone information describe that compared to the Deep Neural Network (DNN) as a baseline and the performance of the CNN model showed an improvement of almost 2% for word-based ASR and more than 2% for syllable-based ASR enhancement using DNN model. As a result, tone information using CNN achieved 2.43% in word error rate (WER) and 2.26% syllable error rate (SER) reductions.

Hay Mar Soe Naing et.al, [10] stated automated speech recognition on spontaneous interview Speech for Myanmar language by utilizing the classical Gaussian Mixture Model based on Hidden Markov Model approach. The speech corpus size was 5 hours long (3.5 hrs for training set, 39 mins for development set, and 47 mins for test set) for Myanmar interview speech data which is collected from web. The duration of each utterance was between 3 sec and 60 sec. It involved average 40 words in one sentence. It explored the effect of variation on the nature of different acoustic features. Moreover, the number of senones and Gaussians were adjusted and the best word error rate of 20.47% was achieved on speaker dependent triphone.

Myat Aye Aye Aung et.al, [13] proposed Time Delay Neural network (TDNN) for Myanmar Speech recognition. In this paper, TDNN was used for acoustic Modeling. Speech corpus contained three domains: names, web news and daily conversations. The speech corpus size is 77 hours 2 mins and 11 sec. It included 233 female speakers and 97 male speakers. By comparing a Gaussian Mixture Model (GMM) as a baseline with Deep Neural Network (DNN) and Convolutional Neural Network (CNN), the performance of TDNN for ASR was shown. The experimental results show that TDNN outperformed HMM-GMM, DNN and CNN.

Khin Me Me Chit, et.al, [12] explored an end-to-end automatic speech recognition (ASR) model for the Myanmar language based on the Connectionist temporal classification (CTC). Experiment was presented on the topology of the model in which the convolutional layers were added and dropped. Different depths of bidirectional long short-term memory (BLSTM) layers were applied and different label encoding methods are also investigated. Speech Corpus size was nearly 26 hours and it achieved 4.72% of character error rate (CER) and 12.38% of syllable error rate (SER) on the test set.

CHAPTER 3

RAKHINE SPEECH CORPUS AND TEXT CORPUS PREPARATION

Nowadays, the development of speech recognition technology is rapidly growing and it becomes popular research field in speech processing technology. Many researchers have been proposed using different techniques in developing speech recognition system. If the system is trained with an unreliable corpus, even the best algorithms with a carefully designed system cannot achieve good speech recognition performance. Thus, the speech corpus becomes a crucial area of research.

The crucial part of automatic speech recognition development is speech corpus that is used for acoustic training. And, Speech corpus preparation is essential step to create automatic speech recognition with own language because it affects the performance of a speech recognizer. Current ASR system uses statistical models constructed on speech data. Therefore, the statistical-based speech recognition system mainly depends on speech corpus. Hence, speech corpus is needed for ASR training.

Rakhine language can be considered as a low-resourced language and pre-collected speech data is not available. And then, there are no many resources to collect speech corpus for Rakhine language. Lack of data becomes the main problem for developing speech recognition research in the Rakhine language. Thus, speech corpus building is needed in developing Rakhine ASR. In this work, a Rakhine speech corpus is constructed by utilizing two domains: daily conversations and Broadcasts News.

3.1 Collection of Speech Data for Low-Resourced Languages

Speech corpus is a collection of speech recordings of spoken language and it also involves transcriptions of the speech. The creation of speech corpus is an important task for the training of any automatic speech recognition system. Especially for low-resourced languages, it is also the first task in developing an ASR system. Because, it has no pre-collected speech corpus in low-resourced languages.

There are some efforts of speech corpus building for low-resourced languages. The speech corpus from Web News for Myanmar Language was developed by Aye Nyein Mon, et.al, [2]. A Myanmar speech corpus, UCSY-SC1, is created for ASR

training purposes. It is built by using two methods. One method is collecting the speech which is already recorded and they are manually transcribed into text. The next method is designing the text corpus first and it is recorded by uttering the collected text. In web news, it involves 25 hrs and 20 mins. It is spoken by 261 speakers which involve 84 males and 177 females. The total number utterances is 9,066. The speech files are segmented using Praat too. In Conversational data, the sentences are recorded by 4 male speakers and 42 female speakers from the faculties and students of the University of Computer Studies, Yangon, Myanmar. The speakers are between 19 and 40 years old. Recording time is 17 hrs and 19 mins and the total utterances is 22,048. For speech recording, Tascam DR-100MKII is used. For speech segmentation, audacity tool is applied and the speech files have formatted mono channel with 16 bits encoding and sampling frequency 16 KHz.

A Spontaneous Interview Speech for Myanmar Language was created by Hay Mar Soe Naing, et.al [10]. This speech corpus building was intended to develop Automatic Speech recognition on Spontaneous Interview Speech. It was collected from the interview of "Let's talk" and "Mizzima Media" that published on Web. The duration of each utterance was between 3 sec and 60 sec which involves average 40 words in one sentence. The total hour is 5 hrs.

The speech corpus for Thai language was created by Virach Sornlertamvanich, et.al, [24]. Thai Language speech corpus had been constructed for the purpose of ASR technology development. There are two types of speech corpus: Thai speech corpus NECTEC-ATR and Thai speech corpus for large vocabulary continuous speech recognition (ORCHID-SPEECH Corpus). The total number of speakers were 20 males and 20 females. The speakers were between 18 to 40 years old. The recordings have been done in a quasi-quiet room. The quality of them were around 20 dB and dynamic microphone (unidirectional microphone: SONY F720) is applied for recording. All utterances are recorded in reading style [25].

Spontaneous Speech database, a phonetically-based multi-purpose database for Agglutinative Hungarian Language was constructed by Tilda Neuberger, et.al, [21]. It contained several types of spontaneous and read speech from 333 monolingual speakers which is about 50 minutes of speech sample per speaker. It involved 184 female and 149 male speakers.

Bengali language speech corpus was created by Sandipan Mandal, et.al, [20]. In this work, a continuous read speech corpus was built as training corpus which involves young and old people. It was built for standard colloquial Bengali language. It is mostly spoken in West Bengal, India. For aligning the speech data, Hidden Markov Model Toolkit (HTK) was utilized. It can be applied for age detection network. Moreover, continuous word recognition and the performance of phoneme recognition were observed to check the speech corpus quality.

3.2 Rakhine Speech Corpus Building

Rakhine language is a low-resourced language and previous work has not been done about speech in language technology. Therefore, pre-collected speech data are not available for Rakhine language. In this work, the purpose of Rakhine speech corpus creation is to develop Rakhine speech recognition system (RASR). Generally, speech data can be created in two methods. One method involves two stages; collecting the text corpus first and recording the speech by reading the collected texts. The second method is collecting the speech data which is already been recorded and then they are manually transcribed them into text. In this work, a Rakhine speech corpus is constructed by utilizing the two ways: daily conversations and broadcast news on two types of domains.

3.2.1 Daily Conversations Recording

The first method (designing of the text corpus and recording of speech by uttering the collected texts) is applied for collecting daily conversations data. It took 4 months for the data recording which is conducted with own voice and 3 native speakers include.

3.2.1.1 Text Corpus Preparation

The first step of corpus design is the text selection. In the designing step, this process is involved the details of text selection or creation are depicted in this section. The domain of text that is used in this corpus which is daily conversations. Firstly, Rakhine daily conversations sentences are constructed with the reference of Rakhine Language guidance book and daily conversations. And then, these sentences are manually segmented into words level and then checked the spelling of the words. It contains Rakhine digits and daily conversations (telephone, hotel, market, street, social, greeting) etc. The sentences involved in this corpus are shorter than that of Broadcast

news data. The shortest sentence has only 1 word and 1 syllable. And, the longest sentence contains 19 words and 28 syllables. It has 5,931 sentences, 3,175 unique words. The format of each sentence is the utterance-id followed by the transcription of each sentence. The sample sentences of daily conversations data are shown in Figure (3.1).

ucsy-record-hninthidarkyaw_00001 ကကောင်း ကောင်း ပါရဲ

ucsy-record-hninthidarkyaw_00002 ကကောင်း ကြာ စာယာ

ucsy-record-hninthidarkyaw_00003 ကကောင်း ကြာ ယာ

ucsy-record-hninthidarkyaw_00004 ကကောင်း ကြိုက် တေ

Figure 3.1: Example sentences of daily conversations data

3.2.1.2 Recording Platform

The recording was made in a laboratory of university of computer studies, Yangon, Myanmar. The utterances are recorded in three environments: Lab1, Digital library, NLP Lab. It is a very quiet environment with no effects from the room such as echo and background noise. It is also a healthy environment for recording because people can breathe freely and feel relaxed. All utterances are reading styles and speech is recorded with the help of Tascam DR-100MKIII. It has a user friendly interface with robust reliability. The speech file can be recorded into .wav, bwf, and mp3 and can be selected mono or stereo channels. In this work, the audio files are recorded with mono type and converted into .wav file format.

3.2.1.3 Speech Segmentation and Recording

The speech file segmentation is done manually by using audacity tool. And, the speech file format is converted into WAV file format with single channel mono type and sampling rate is 16 KHz. Silence portion of each utterance and background noise are removed in segmenting the speech files. Audio file and sentences should be aligned, therefore each recorded sentence is manually listened and checked with their particular text transcription. Moreover, corrections are made to fix the corresponding audio file. If the speaker does not have a clear voice and speech involves noise, the recording is conducted repeatedly until to be correct and smooth.

3.2.2 Data Collection from Online Resources

Nowadays, speech data can be searched on the internet which can be gathered from the web in developing speech corpus. Thus, the second way is used to build Rakhine speech corpus for Broadcasts news. The internet is a source to collect data that has various resource types and video files can be downloaded easily. Furthermore, they are freely available on the internet. However, there are no many online resources for Rakhine Language. In this work, Broadcast news data is collected from the web for Rakhine continuous speech recognition. The duration of collecting Broadcast news data takes 4 months.

3.2.2.1 Speech Corpus Preparation

Rakhine speech corpus preparation is conducted on Broadcast news. Rakhine news is available on web sites but news with Rakhine dialect is very rare. Rakhine language has lack of online resources to construct Rakhine speech corpus. Arakan Princess Media (APM) is Rakhine Broadcast news. This channel broadcasts international news, Myanmar news and Rakhine news. It is spoken in Rakhine Language. Therefore, the speech data is gathered from the site of APM Broadcast news [3]. It includes both local news (political, health, social, business) and foreign news (political, health, sport and weather, social, business). The speech file format is converted into WAV file format with single channel (mono) type and sampling rate is 16 KHz. The range of the speech file is between 2 sec to 18 sec.

3.2.2.2 Speaker Information

The news presenters are professional in speaking, well-experienced and they also have clear voice. Furthermore, recording quality is good in news broadcasting and less noisy. Mostly, female news presenters are found in Broadcast news. Hence, female speakers involve more than male speakers in Broadcasts news data. Rakhine is mainly divided into two dialects. Therefore, Broadcast news speaker involves two dialects (North and South). According to the nature of Rakhine Language, this corpus involves from north and south region. The female (from South) speakers are more than male (from South) speakers. The male (from North) speakers are less than that of male (from

South) speakers. The speaker information of Broadcast news data is shown in Figure 3.2 with respect to gender.

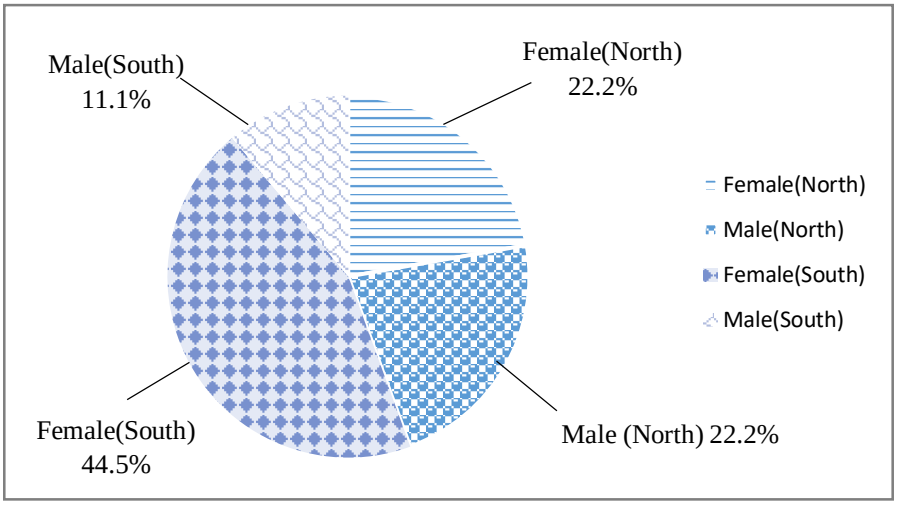


Figure 3.2: Broadcast news data with speaker distribution related to gender

3.2.2.3 Speech Utterance

For Rakhine language, the news from the internet does not already have transcription. The transcription is not available for Rakhine language. Thus, speech utterances are manually transcribed into text. In Rakhine language, segmentation is also needed as there is no space between words when writing. Therefore, the texts are manually segmented into words. Besides, the word segmentation is manually checked again to correct and Rakhine guidance book is used to check the spelling of the words. The longest transcription has 54 words and 75 syllables in one sentence. The shortest transcription has only 2 words and 4 syllables. In broadcast news data, it has 1,439 unique sentences, 2,950 unique words. Pyidaungsu Unicode font is used for the text corpus. The format of each sentence is similar to daily conversations domain (each utterance is followed by utterance id). The example sentences of the Corpus on Broadcasts News is shown in Figure (3.3).

ucsy-apm-dainnjaoo_20001 အေအေ နန့် ဆက်စပ် ဖမ်းထား ရေ လူ သုံး ဆယ့် ခြောက် ယောက် ကို ထပ် လွှတ်ပီး ဖို့ ဆိုရေ သတင်း ကို အယင်ဆုံး ပြောပြ ချင် ပါရေ

ucsy-apm-dainnjaoo_20002 ရခိုင် ပြည် မှာ ရခိုင့် တပ်တော် အေအေ နန့် ဆက်စပ် ဖမ်းထား ရေ လူ သုံး ဆယ့် ခြောက် ယောက် ကို ထပ် လွှတ်ပီး ဖို့ လို့ ရခိုင် ပြည်နယ်ကောင်စီ က ပြောဆို ပါရေ

ucsy-apm-dainnjaoo_20003 ဇာ ရက် မှာ ဇာပိုင် ပုံစံ နန့် လွှတ်ပီး ဖို့ ဆိုစော် ကို အတိအကျ မ ပြော နိုင် သိမ့် လို့ ပြည်နယ်ကောင်စီ က ဆို ပါရေ

ucsy-apm-dainnjaoo_20003 အေအေ နန့် ဆက်စပ် ဖမ်းဆီး ထား ရေ လူ တိ ကို အသုတ်လိုက် ခွဲ ပနာ လွှတ်ပီး လား ဖွယ် ဟိမ့် နိန်ရေ လို့ ပြည်နယ်ကောင်စီ က မှတ်ချက် ပီး ပါရေ

Figure 3.3: Example sentences of broadcast news data

3.2.3 Transcription Normalization

Some of the transcriptions of broadcast news and daily conversations involve non-standard words. They are time, data, numbers, acronyms, abbreviations, symbol and foreign words such as name of person, organization, months, things, social media, and abbreviation. Thus, they are needed to be normalized into Rakhine language. In this task, these words are manually transcribed into Rakhine words by listening the particular speech audio files. Table 3.1 describes the sample words which needed to be normalized.

Table 3.1: Sample of Rakhine Text normalization

Description	Example	Normalization
Time	၁၁ နာရီ ၁၅ မိနစ်	ဆယ့် တစ် နာရီ ဆယ့် ငါး မိနစ်
Date	၂၀၁၆-၂၀၁၇	နှစ် ထောင့် ဆယ့် ခြောက် နှစ် ထောင့် ဆယ့် ခနိုက်
Number	၁၈၅ ယောက်	တစ် ရာ့ ရှိုက် ဆယ့် ငါး ယောက်
Digits	၇.၆ ဧက	ခနိုက် ဒသမ ခြောက် ဧက
Symbols	/ %	မျဉ်းစောင်း ရာခိုင်နှုန်း
Months	September November	စမ်တင်ဘာ နိုဗမ်ဘာ

Person Name	An htonihpalikhan	အန်ထိုနိန်ဖလီခန်
Things	Aircon	အဲယားကွင်း
Social media	Twitter	တွယ်တာ
State	NewYork	နယူးယောက်
Organization	ASEAN	အာဆီယံ
Abbreviation	UN	ယူအမ်

3.3 Statistics of Speech Corpus

The speech corpus involves of two kinds of domains: broadcast news (APM Rakhine Broadcasts news) and daily conversations. Both are read speech data types. The detail information of the speech corpus is depicted in Table (3.2).

In Broadcast news, the speech corpus size is 3 hours 4 min 3 sec spoken by 10 speakers (3 females and 2 males in North and 4 females and 1 male in South) with 1295 utterances. For daily conversations, the duration of the recorded speech size is 3 hours 16 min. The total utterances is 6848 which are recorded by with own voice (North) and include 3 speaker (2 females in north and 1 male in south).

Table 3.2: Statistics of Rakhine speech corpus

Data	**Size**	**Speakers**					**Utterances**
		Female		Male		Total	
		North	South	North	South		
APM Broadcast	3 hours 4 min 3 sec	3	4	2	1	10	1439
Daily Conversations	3 hours 16 min	3	-	-	1	4	6848
Total	6 hours 20 min 3 sec	6	4	2	2	14	8287

3.4 Rakhine Pronunciation Lexicon

Pronunciation lexicon is also one of the main components of ASR system. It contains a list of words which is expressed phoneme for each word. And, it is used to map between words and a phone set. This phones set represents the speech sound for each word. In this task, there are 68 phones unit is used for Rakhine Language. The content of phone set file for Rakhine Language is stated as follows,

Table 3.3: Content of Phoneme set file

Rakhine Phone Symbols											
SIL	a	a-	a:	a.	a’	ai’	an	an.	an:	au’	b
Ch	d	dh	e	e.	e’	ei	ei.	ei:	en	en:	en.
e:	g	gy	h	hp	hr	ht	i	i:	i.	in	in:
in.	j	k	ky	kh	l	m	n	ng	nj	o	o:
o.	ou	ou:	ou’	ou.	p	sh	r	s	t	th	u
U	u.	u:	un	un:	un.	w	z	-	-	-	-

3.4.1 The Nature of Rakhine Phonetic

Some Rakhine syllables do not conform to these standard rules of pronunciation as Myanmar Language. The pronunciation of the syllables can depend on the context of syllables. The differences between standard pronunciation and correct pronunciations of some Rakhine words are described in Table (3.4) as an example,

Table 3.4: Samples of contextually dependent pronunciation

Rakhine words	**Standard**	**Correct**
စိုင်းတင်	sain: ten	sain: den
ကကောင်း	ka. kaun:	ka- gaun:
ပညာ	pa. nja	pain nja

Rakhine language has varied acoustic signal and it also has phonetic variety. Some words have difference between in writing script and their pronunciations in Rakhine Language. The example words of pronunciation difference with writing script are also stated as shown in Table 3.5.

Table 3.5: Samples pronunciation of some words

No	Writing Script (Rakhine)	Standard Pronunciation	Correct Pronunciation
1	နီရောင်	ni raun	nein raun
2	ဂေါင်းမွှိ	gaun: mwi.	gaun: mwein.
3	အမှု	a- mhu.	b- mhoun.
4	ငြိမိ	ngri. mi.	ngrein. mein.

The first word and the fourth word in Table 3.5 is pronounced /as i/ as /ein/ in this case, Rakhine use /i/ pronunciation however some consonants combine with /i/ vowels which is not pronounced as /i/. Instead of /i/, /ein/ is changed. Mostly, when consonants (/န/ /ည/ /မ) combine with /i/ vowels, /ein/ is pronounced in Rakhine. As example, နီကြာ /nein kra/, ညီညာ, /njein nja/, မီးမီး /mein: mein:/, it is not pronounced နိကြာ /ni kra/, ညီညာ, /nji nja/, မိးမိး /mi: mi:/. In the second word, /i./ is pronounced as /ein./. As an another example, နိရက် /ni. re'/ is pronounced as /nein. re'/. In this condition, when /i./ is pronounced by combining with /န/ consonant, sound is changed /ein./ vowels. The third word "/u./ is pronounced as (oun.)". Consonants /မ/ combines with ှု the pronunciation is changed as /oun./ in Rakhine.

3.4.2 Building the Rakhine Phonetic Dictionary

Rakhine Pronunciation lexicon is created in developing Rakhine ASR. In this work, vocabulary are brought from ရခိုင်ဘာသာစကားလမ်းညွှန်, ရခိုင်ဂန္ထဝင်ဝေါဟာရ and Arakan Princess Media (Broadcast news) to develop Rakhine pronunciation lexicon. Although some are based on word in ရခိုင်ဘာသာစကားလမ်းညွှန် and ရခိုင်ဂန္ထဝင်ဝေါဟာရ, some are not based

on word or syllables. When Rakhine phonetic dictionary is created, words are used in mapping of phone list. Generally, the phonetic dictionary is created using the two methods: syllable-based and word-based. Therefore, firstly vocabulary from ရခိုင်ဘာသာစကားလမ်းညွှန်, ရခိုင်ဂန္ထဝင်ဝေါဟာရ were manually broken into words to create word-based Rakhine pronunciation lexicon. The second one was collected from Broadcasts news which contains not only native words but also foreign words. And then, phonemes were tagged by listening to their particular recording files. Words were aligned to their phoneme using human notation.

သာလီစွ ပါ ရခိုင် ပြည်နယ် က ကြိုဆို ပါရေ
tha li zwa. ba ra- khain pre ne ga. krou hsou ba rei

အေ ငသတိုက် တစ် ကောင် ကို ဇာလောက် လဲ
ei nga- dha- dai' ta- gaun g ou za lau' le:

မင်း ကျောင်း က ပြန် လာ စာ ကကောင်းပင် ယင် ရေ
men: kyaun: ga. pran la sa ga- gaun: ben jen rei

Fig 3.4: Phonemes tagged sentences

In developing Rakhine phonetic dictionary, Myanmar phonetic dictionary is referenced [27]. It collected 2134 words from ရခိုင်ဘာသာစကားလမ်းညွှန် [6], 2100 words from ရခိုင်ဂန္ထဝင်ဝေါဟာရ [28] and 2950 words from Arakan princess media (broadcast news) for Rakhine Pronunciation lexicon. The pronunciation for each word is manually annotated. The total vocabulary of Rakhine lexicon contains 7184 words and 68 phonetic units to represent the pronunciation of words.

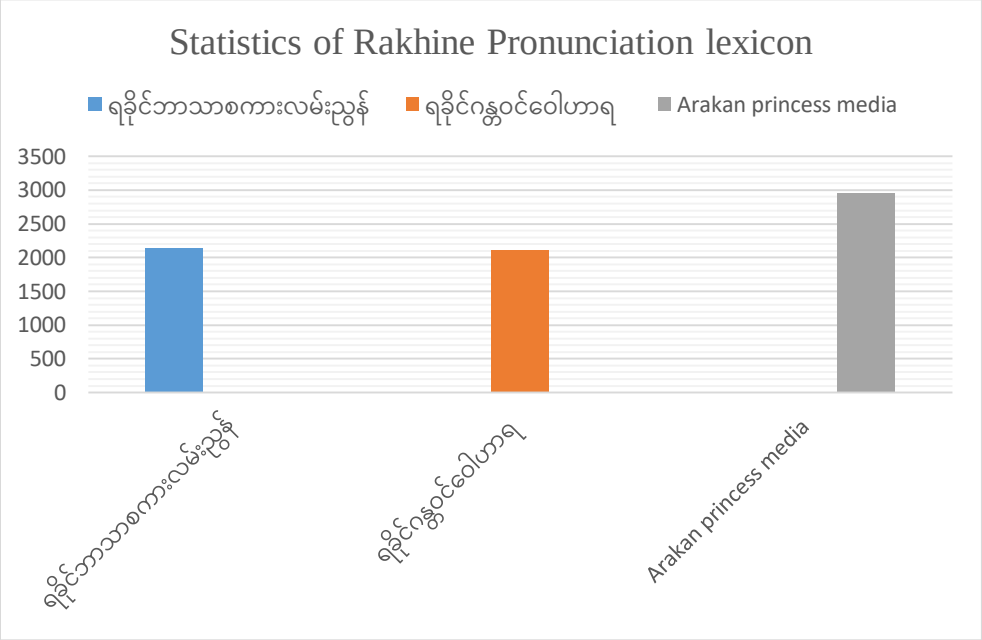


Fig 3.5: Statistics of Rakhine pronunciation lexicon

The sample of phonetic dictionary for Rakhine is described as shown in Table 3.6.

Table 3.6: Example of Rakhine phonetic dictionary

Rakhine Word	Phonetic
ကြိုတင်	k r ou t en
ညမချေ	nj a- m a. ch ei
ကိုဗိုက်	k ou b ai’
ကိုဗိုက်	k ou b i.
ကိုဗိုက်	kh ou b i.
ရှိဆိုင်	hr i. hs a in
ရိတံခွန်	r i t a- kh w an
ချောင်းဆိုး	kh r a un: hs ou:
အီးယောင်	i: j a un
အမင်	a- m en
အဲယားကွင်း	e: j a: k w en:

CHAPTER 4

RAKHINE LANGUAGE

This chapter presents the nature of Rakhine Language and the difference between Rakhine and Myanmar Language. Rakhine sentences structures and consonants and vowels of Rakhine language are also presented. Moreover, basic phoneme of Rakhine Language and Rakhine Speech tone are also discussed in this chapter.

4.1 The Nature of Rakhine Language

Rakhine (Arakanese) is the native of language of Rakhine (Arakanese) and it is a tonal language. Rakhine state is situated in the south of west of Myanmar and it has population of 3,118,963. There are 7 ethnics groups: Rakhine, Mjou, Marama, Dainne, Thet, Khame, Kaman. They also have their spoken languages which are second languages. Rakhine Language is an official language in Rakhine. It is mainly spoken by Arakanese people. Rakhine is mainly divided into two dialects North (Sittwe) and South (Thandwe). They also have different vocabularies and pronunciations. The detailed explanation is described in section (4.6).

Rakhine Language is closely related to Myanmar language. Most Arakanese speak an unusual variety of the Burmese Language which includes significant differences from Burmese pronunciation and vocabulary. Rakhine writing scripts are the same with Myanmar Language. It is written from left to right without any spaces between words or syllables. In Rakhine Language, words are formed by combining basic characters with extended characters. Rakhine syllable can stand one or more extended characters by combining consonants to form compound words.

Rakhine language is similar to Myanmar language and it has basic 33 consonants and 44 vowels and 4 medial in Rakhine language. However, some vowels phonemes have different pronunciations with Myanmar language and some vowels are not used in Rakhine. The following table shows the basic consonants phoneme of Rakhine Language. There are 23 phonemes for 33 consonants scripts. Some scripts share the same pronunciation. As an example, the pronunciations of “ဒ”, “ဓ”, “ဍ” and “ဎ” are the same and they are defined as the same phoneme /d/. And, the pronunciations

of "ꩡ" and "ꩢ" are defined as the same phoneme /ht/ and ဗ and ဘ are defined as the same phoneme /b/.

Table 4.1 Groups of Rakhine consonants phonemes

Grouped of consonants				
Unaspirated	Aspirated	Voiced		Nasal
က/ k	ခ/kh	ဂ/g	ဃ/g	င/ng
စ/s	ဆ/hs	ဇ/z	ဈ/z	ဉ/ည/nj
ꩫ/t	ꩡ/ht	ꩬ/d	ဎ/d	ဏ/n
တ/t	ထ/ht	ဒ/d	ဓ/d	န/n
ပ/p	ဖ/hp	ဗ/b	ဘ/b	မ/m
ယ/j/r	ရ/j/r	လ/l	ဝ/w	သ/th
	ဟ/h	ꩭ/l	အ/a	

In the above table, the shaded consonant pronounces both /j/ sound and /r/ sound. However, Myanmar pronounces only one /j/ sound. In Rakhine Language, /r/ sound is prominently used and /j/ sound is also used. But /r/ sound is not used in Myanmar Language. For instance, ကြိုက် is pronounced as (k r ai') in Rakhine although ကြိုက် is pronounced as (k j ai') in Myanmar. Some pronunciations are different for /j/ sound and /r/ sound in Rakhine which has north and south. In Rakhine Language, for example, ကယား is pronounced as (k a- j a:) in north and (k a- r a:) in south. In this condition, /j/ sound is used in north however /r/ sound is used in south. Although both /j/ sound and /r/ sound are used in Rakhine, there has different pronunciations in north and south. The other example is that ကြိုတင် is pronounced as (k r ou t en) in north though (k j ou t en) in south. In this example, /r/ sound is used in north but /j/ sound is used in south. Therefore, both /j/ sound and /r/ sound are used in Rakhine however using is different.

4.2 Rakhine Vowels Phoneme

Rakhine use 44 vowels phoneme. Although vowels writing script are all the same as Myanmar Language which have different pronunciations in Rakhine. Most of the vowel phonemes are similar with Myanmar. Example, အိ (i.), အီ (i), အီး (i:), ေအ (ei), ေအ့ (ei.), ေအး (ei:), အယ်(e), အယ့်(e.), အဲ (e:), အာ့ (a.), အာ (a), အား (a:), ေအာ် (o), ေအာ့(o.), ေအာ (o:), အူ (u), အု (u.), အူး (u:). However, some vowels phonemes are not used in Rakhine. For instance, the pronunciations for အစ် (i') and အင် (in), အွန် (un), အွတ် (u') are not used in Rakhine. For example, စစ်တွေ is pronounced (s i' t we) in Myanmar however စစ်တွေ is pronounced (s ai' t we) in Rakhine. In this example, writing script of အစ် is used and it does not pronounced အစ် (i'). Instead of အစ် (i'), အိုက် (ai') is pronounced in Rakhine. The next one is that ခင် is not pronounced as (k in) in Rakhine. But, it pronounces as ခင် (k en). In this example, although writing script အင် is used, pronunciation for အင် is (en) in Rakhine. Other examples are that ပွတ် is pronounced as (p u') in Myanmar however (p wa') in Rakhine and မွန် (m un) is also pronounced as (m wan) in Rakhine which does not pronounce as (m un) in Myanmar.

Table 4.2 and Table 4.3 shows basic vowels of Rakhine in writing script. The shaded vowels phonemes are different with Myanmar which is explained in above sentences.

Table 4.2: Vowels phoneme of Rakhine Language

အီ (i)	အိ (i.)	အီး (i:)	အစ် (ai')
ေအ (ei)	ေအ့ (ei.)	ေအး (ei:)	အိတ် (ei')
အယ်(e)	အယ့်(e.)	အဲ (e:)	အိုက် (ai')
အာ (a)	အာ့(a.)	အား (a:)	အတ်(a')
ေအာ်(o)	ေအာ့(o.)	ေအာ (o:)	ေအာက် (au')

အူ(u)	အု (u.)	အူး (u:)	အွတ် (u’)
အိုု(ou)	အိုု့ (ou.)	အိုုး (ou:)	အုပ် (ou’)

Table 4.3: Vowels phoneme of Rakhine language

အင် (en)	အင့် (en.)	အင်း (en:)
အိန် (ein)	အိန့် (ein.)	အိန်း (ein:)
အိုုင် (ain)	အိုုင့် (ain.)	အိုုင်း (ain:)
အန် (an)	အန့် (an.)	အန်း (an:)
အောင် (aun)	အောင့် (aun.)	အောင့် (aun:)
အုုန် (oun)	အုုန့် (oun.)	အုုန်း (oun:)
အွန် (un)	အွန့် (un.)	အွန်း (un:)

4.3 Rakhine Phonology

Speech can be produced if there are just place of articulation, articulator, manner of articulation. The basic consonants of Rakhine are depicted in Table (4.4) according to the three parameters of place of articulation, manner of articulation, and articulator.

Table 4.4: Rakhine phonology

Manner of Articulation	Place of Articulation						
	Bilabial	Dental	Alveolar	Palato-alveolar	Palatal	Velar	Glottal
Nasal (stop)	မ		န	ည		င	
	မ့		န့	ည့		င့	
Stop Voiced	ဘ (ဗ)		ဒ			ဂ	
Voiceless	ပ ဖ		တ ထ			က ခ	
Fricative Voiced		သ	ဇ				
Voiceless		သ	စ ဆ	ရှ			
Affricate Voiced				ဂျ			
Voiceless				ကျ ချ			
Central Approximant Voiced	ဝ		(ရ)		ယ		
Voiceless	ဝ့		သျ/ရှ				ဟ
Lateral Approximant Voiced			လ				
Voiceless			လ့				

Note that some consonants share the same phonemes in Rakhine, ဒ, ဓ, ဥ and ဗ with phoneme (d), န and ဏ with phoneme (n), ဇ and စျ with phoneme (z), ရှ and သျ with phoneme (sh) and ည and ဉ with phoneme (nj).

Since Rakhine is tonal language, there are four tone level in Rakhine phonology. These tones are long vowel, long vowel with falling tone, short vowel and glottal stop. The basic phoneme of Rakhine vowels is described by using these four tones level in the following table.

Table 4.5: Rakhine vowels with tone level

Non-nasalized Vowels				Nasalized Vowels		
Tone I	Tone II	Tone III	Tone IV	Tone I	Tone II	Tone III
1.အီ	2.အီး	3.အိ	4.အစ်	5.အိန်	6.အိန်း	7.အိန့်
8.အေ	9.အေး	10.အေ့	11.အိတ်	12.အိုင်	13.အိုင်း	14.အိုင့်
15.အယ်	16.အဲ	17.အယ့်	18.အက်	19.အန်	20.အန်း	21.အန့်
22.အင်	23.အင်း	24.အင့်	25.အိုက်	26.အောင်	27.အောင်း	28.အောင့်
29.အာ	30.အား	31.အ	32.အတ်	33.အုန်	34.အုန်း	35.အုန့်
36.အော်	37.အော	38.အော့	39.အောက်	-	-	-
40.အို	41.အိုး	42.အို့	43.အုပ်	-	-	-
44.အူ	45.အူး	46.အု	47.အုတ်	-	-	-
47.အွန်	48.အွန်း	49.အွန့်	-	-	-	-

In the table above, Tone I is long vowel, Tone II is long vowel with falling tone. In the third tone level, Tone III is short tone and tone IV is glottal stop. The shaded vowels tone /အင်, အင့်, အင်း/ and /အွန်, အွန့်, အွန်း/ are Non-nasalized in Rakhine however they are nasalized vowels in Myanmar. In tone IV, the pronunciations of အစ် and အိုက် are the same and they are defined as the same phoneme /ai'/. In the above table, there is not just postscript double dot -: in the tone level II. The vowel number 16 /အဲ/ and 37 /အော/ are also contained in this tone II since the nature of their phoneme are other vowels with postscript double dot in this column.

The phonology is the system that comprises the vowel and the consonant. Rakhine phonology can be composed by just one vowel, or one vowel and consonant,

and consonant combination symbols. In Rakhine language, the vowels have their own sounds. Therefore, just only one vowel can produce clear အ, အာ, အား sound such as Rakhine consonants have no clear own sound. If it combines with a vowel, it can produce the clear sound.

Table 4.6: Pairs of two consonants combinations symbols in Rakhine

First Combination symbols	Second Combination symbols	Results of two combination symbols
◌ျ	◌ော	◌ျော
◌ြ	◌ော	◌ြော
◌ျ	◌ာ	◌ျာ
◌ြ	◌ာ	◌ြာ
◌ော	◌ာ	◌ောာ

There are the pronunciations of four basic consonant combination symbols in Rakhine (-◌ျ / -◌ြ / -◌ော / -◌ွ). These Rakhine character can be joined with appropriate out of the 33 consonants. These symbols can be comprised with each other in two characters or three characters. The phonetic of four basic consonants combination symbols are depicted as in Table 4.7. There are six phonology method in Rakhine Language.

1. Vowel phonology
2. Combining consonants and vowels phonology
3. Combining consonants combination symbol /-◌ျ/ and vowels
4. Combining consonants combination symbol /-◌ြ/ and vowels
5. Combining consonants combination symbol /-◌ော/ and vowels
6. Combining consonants combination symbol /-◌ွ/ and vowels

Table 4.7: Phonetic consonants combination symbols of Rakhine

Combination symbols	Phonetic	Example
-ျ	-y-/-r-	k y a./g y a./kh r a.
-ြ	-j-/-r-	p j e/ p r e
-ွ	-w-	z w a.
-ှ	-h-/sh-/hr-	l h a./sh a./ hr a.

1. Vowel phonology

In writing script, there are 49 vowels but some vowels are the same phoneme. Therefore, the basic phonemes 44 are contained in vowel phonology.

Examples:

အီ, အို, အီး

အိုင်, အိုင့်, အိုင်း

အို့, အို့, အို့း ..... etc

2. Combining consonants and vowels phonology

The original vowels can be combined with consonants. By combining original vowel with consonants, it can produce one syllable or meaningful one word. Examples are described in Fig (4.1).

Example:

Consonants		Vowel	Result
က	+	အား	ကား
ဇ	+	အီ	ဇီ
ပင်	+	အင်	ပင်

Figure 4.1: Combining consonants and vowels phonology

3. Combining consonants combination symbol /-ျ/ and vowels

ျ (ယပင်း) combines with 10 consonants. They are ကျ, ချ, ဂျ, ပျ, ဖျ, ဗျ, မျ, လျ, သျ. Some words are formed by combing two symbols က (consonant) + -ျ (first symbol) + ျ (second symbol) = ကျား. The following figure shows the example combination of consonants combination symbols ျ (ယပင်း) and vowels of Rakhine with phonetic.

Example:

Consonants combination symbols		Vowels	Results	Phonetic
ကျ	+	အား	ကျား	ky a:
ချ	+	အီ	ချီ	ch i
ချ	+	အောင်း	ချောင်း	kh r aun:
ဂျ	+	အင်	ဂျင်	gy en
ဗျ	+	အင်း	ဗျင်း	b j en:

Figure: 4.2: Rakhine phonology combination with consonants, consonants combination symbols ျ (ယပင်း) and vowels with phonetic

4. Combining consonants combination symbol /-ြ/ and vowels

ြ (ရရစ်) combines 11 consonants. These are ကြ, ခြ, ဂြ, ငြ, ဒြ, ပြ, ဖြ, ဗြ, မြ. Some words are formed by combing two symbols က (consonant) + -ြ (first symbol) + -ွ (second symbol) = ကြွ. The following figure shows the combination of consonants combination symbols -ြ (ရရစ်) and vowels of Rakhine with phonetic.

Example:

Consonants combination symbols		Vowels	Results	Phonetic
ကြ	+	အား	ကြား	k r a:
ဂြ	+	အင်း	ဂြင်း	g r en:
ဖြ	+	အား	ဖြား	hp r a:
ဗြ	+	အာ	ဗြာ	b r a
မြ	+	အီ	မြီ	m r ein

Figure 4.3: Rakhine phonology combination with consonants, consonants combination symbols ြ (ရရစ်) and vowels with phonetic

5. Combining consonants combination symbol /ွ / and vowels

ွ (ဝဆွဲ) combines consonants. Those consonants are က, ခ, ဂ, ဍ, စ, ဆ, ဇ, ည, တ, ထ, ဒ, န, ပ, ဖ, ဘ, ဗ, မ, ယ, ရ, လ, သ, ဟ, အ.

Consonants combination symbols		Vowels	Results	Phonetic
ကွ	+	အာ	ကွာ	k w a
ဇွ	+	အဲ	ဇွဲ	z w e:
ထွ	+	အီး	ထွီး	ht w i:
ယွ	+	အံ့	ယွဲ့	j w e.
ဟွ	+	အေး	ဟွေး	h w ei:

Figure 4.4: Rakhine phonology combination with consonants, consonants combination symbols ွ (ဝဆွဲ) and vowels with phonetic

6. Combining consonants combination symbol /-ှ/ and vowels

ှ (ဟထိုး) combines 6 consonants. Those consonants are င, န, မ, ရ, လ, ဝ.

Example:

Consonants combination symbols		Vowels	Results	Phonetic
ငှ	+	အား	ငှား	ng h a:
လှ	+	အူ	လှူ	l h u
ရှ	+	အိ	ရှိ	hr i.
ရှ	+	အိ	ရှိ	sh i.
ဝှ	+	အန်	ဝှန်	h w an.

Figure 4.5: Rakhine phonology combination with consonants, consonants combination symbols ့ (ဟထိုး) and vowels with phonetic

4.4 Rakhine Speech Tone

Both Rakhine and Myanmar are tonal languages however the tone of voice differs between Rakhine language and Myanmar language. For instance, some words share same writing script however Rakhine speech tone is different with Myanmar pronunciation. There are difference speech tones between Rakhine and Myanmar Language.

Myanmar	**Rakhine**
အူ/အာ	အ
ဧအ	အီ
အစ်	အစ်
အည်	အိုင်
(ဧအ/ဧအ့/ဧအး)	(အီ/အိ/အီး)
(အယ်/အယ့်/အဲ့)	(ဧအ/ဧအ့/ဧအး)

အူ is pronounced as အ in Rakhine, for example လူပျို /l u b jou/ in Myanmar လပျို /l a- b jou/ in Rakhine.

ေအ is pronounced as အီ in Rakhine, for example လေ /l ei/ in Myanmar လီ /l i/ in Rakhine and ဆေး /hs ei:/ in Myanmar ဆီး /hs i:/ in Rakhine.

အစ် is pronounced as အိုက် in Rakhine. As an instance တစ် /t i'/ in Myanmar is pronounced as တိုက် /t ai'/ in Rakhine and ပစ် /p i'/ in Myanmar ပိုက် /p ai' in Rakhine. အည် is pronounced as အိုင်, for example လည်ပင်း /l ei b in:/ in Myanmar လိုင်ပင်း /l ain b en:/ in Rakhine.

(ေအ, ေအ့, ေအး) is as pronounced as (အီ, အိ, အီး) in Rakhine. For example, (လေ, လေ့, လေး) is pronounced as (လီ, လိ, လီး).

(အယ်, အယ့်, အဲ့) is pronounced as (ေအ, ေအ့, ေအး). For example, (တယ်, တယ့်, တဲ့) is pronounced as (ေတ, ေတ့, ေတး)

4.5 Rakhine Grammar

The grammar is part of the language and indicates the structure of that language. As grammar is an important part of linguistics, it is the rules behind languages. Exactly, the grammatical aspect which does not directly concern the meaning is called syntax while the aspects which concern the meaning include semantics and pragmatics. Rakhine grammar and sentences are closely related with Myanmar language. In Rakhine grammar, there are nine parts of speech. They are noun, pronoun, verb, adverb, adjective, postpositional marker, conjunction, particle, and exclamation. The part of speech tag-set for Rakhine is described in Table 4.8.

Table 4.8: Part of speech tag- set of Rakhine language

POS Tag	Brief Definition	Examples
Noun	names, activities, events, persons, objects, abstract, ideas	နိ(နေ), ဟင်းဇား(ဇလုံ), စျိး(ဈေး), အဘောင်သျှင်(အဘွား)
Pronoun	takes the place of a noun	အကျွန်(ကျွန်မ), ကျနော်(ကျွန်တော်), ငါ, သူ
Adjectives	can be before or after a noun and usually starts with တဝ, တဗျင်း, ကောင်း and does not end with သော, သည့်, မည့် as Myanmar	တဝ(အရမ်း) ရှည်, တဗျင်း(အရမ်း) ချို, ကောင်း (အရမ်း) လှ

Adverb	always before a verb and does not end with စွာ as Myanmar	အယင်(မြန်မြန်), အမှန်(စင်စစ်), ခခင်မမင်(ခင်ခင်မင်မင်), ကကောင်း (အလွန်)
Verb	always suffixed with one or more particles to show the tenses	ရှိုး(ရေး), ဟိမ့်ရေ (ရှိတယ်), လား(သွား)ဖို့
Conjunction	joins words, phrases or sentences	ယေကေသင့်(ဒါပေမဲ့), ယားတွက်နန့်(ဒါကြောင့်)
Particle	suffixed or prefixed to nouns, verbs, adjectives, and adverbs	တိ(တွေ), ရို့(တို့), ခ (ခဲ့), ဖို့ (မည်)
Postpositional Marker	after pronoun or noun, similar to preposition in English but no exact words to express English	က, မှာ, ကို, နန့် (ဖြင့်), စာ (သည်)
Exclamations	Expresses of emotion	အဘာလေး (အမလေး)

4.5.1 Sentences Structure of Rakhine Language

There are two kinds of Rakhine sentences defining their structure; simple sentence and compound sentences. On the other hand, there are five kinds of Rakhine semantics that define sentences. These are statement sentence, question sentence, negative sentence, urge sentence, desire sentence.

There are two types of words in the Rakhine languages; isolated word and supplied word. Isolated Rakhine word is the meaningful word that does not depend on other words. Isolated words may be noun, pronoun, adjective, verb, adverb and exclamation from Rakhine part of speech. Examples, အကျွန် (a- ky w an) is pronoun, ကဇပ် (k a- z a') is verb, ကဒိုးကဒိုင် (k a- z ou: k a- z ain) is adverb. However, exclamation can be isolated words and sometimes it may be a word that combines a supplied word. The supplied words provide and combine to be a meaningful word in building phrases

or sentences. The supplied words are postpositional marker, particle and conjunction as an example က, ကို, နှင့်, ယောက်.

In Rakhine language, most sentences are comprised with phrases. Phrases are created with words and words are comprised with syllables. In a Rakhine sentence, a syllable is the smallest semantic unit. Figure 4.6 shows how to comprise a simple Rakhine sentence from the syllable level to the sentence level.

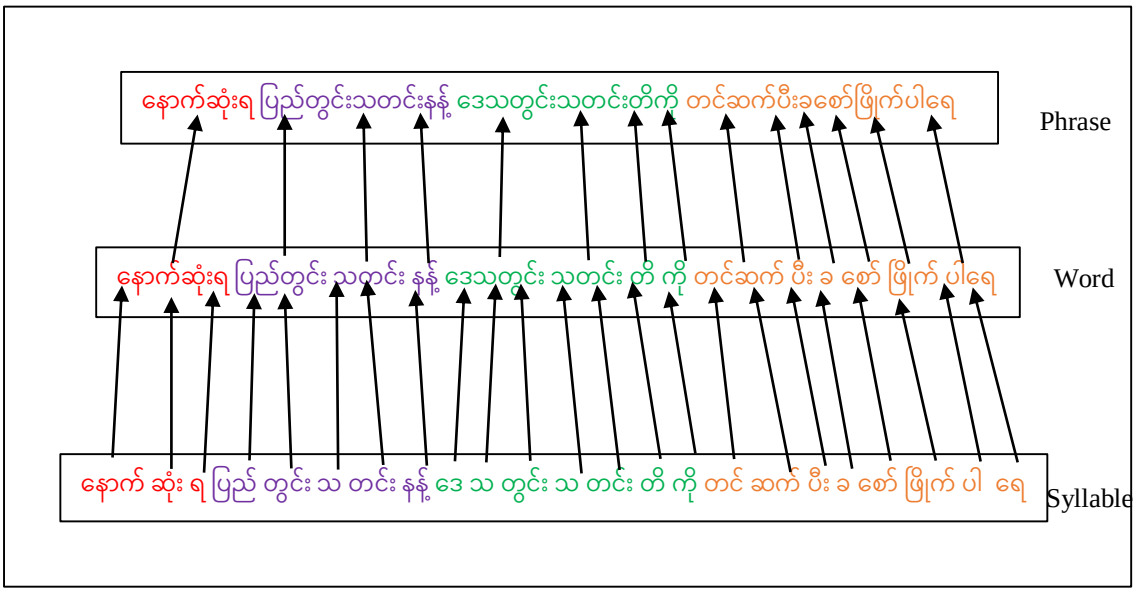


Figure 4.6: Example of grammatical hierarchy of Rakhine sentence

Basically, Rakhine syllables are combined by consonant and vowel. Rakhine syllable involves either a vowel by itself or a consonant combined with a vowel. Some syllables compose one or more characters combinations. The combination of အီ vowel and လ consonant produce one syllable လီ as လ + အီ = လီ. There are 13 consonants used for devowelizing က်, င်, စ်, ည်, ဉ်, တ်, န်, ပ်, မ်, ယ်, လ်, ဏ်, ဒ် in Rakhine Language. Moreover, the combination of consonant လ and က် creates one syllable.

Rakhine syllable structure is as shown in Table 4.9 and the example of Syllable-based sentences is described in Table 4.10.

Table 4.9: Rakhine syllable structure

Initial Consonants	Glide Consonants	Vowel	Final Consonants	Tone
C	(G)	V	(N/)	T

Table 4.10: Example of syllable sentence

တ	နား	စိုင်း	စား	ကြည့်	လိုက်	ပါ	မေ

In general, there are four patterns of syllable structure in every language. Syllable structure patterns are described by symbols as follows:

(1) -V-

(2) CV-

(3) -VC

(4) CVC

Among them, CV- pattern can find in all language. There are many different languages, and some languages have both /CV-/ and /-V-/. Some languages have /CV-/, /-V-/, and /-VC/. Moreover, some have /CV-/, /-V-/, /-VC/ and /CVC/. In these above patterns, C means consonant and V means vowel.

4.6 Rakhine dialects difference in North and South

According to the nature of Rakhine Language, it can be divided North and South in Rakhine. They also have changing the sounds and vocabulary in Rakhine. These vocabulary difference and pronunciations changing is explained in the following Figure (4.7) and Figure (4.8).

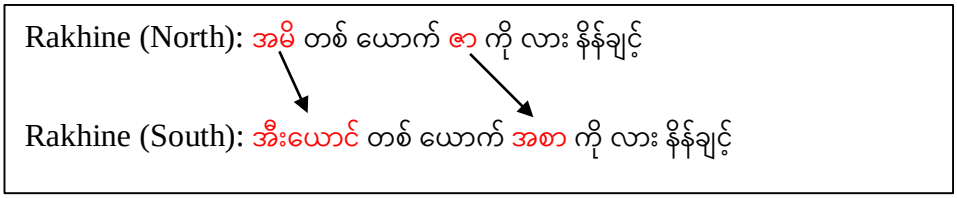


Figure 4.7: Example of vocabulary difference in north and south

In the above figure, the example sentences mean အမေ တစ် ယောက် ဘယ် ကို သွား နေ တာလဲ in Myanmar. In Rakhine North, အမေ /a m ei/ in Myanmar means အမိ /a- m ein./. For South, အမေ /a m ei/ means အီးယောင် /i: j aun/. The next one is that North uses ဇာ /z a/ however south use အစာ /a- s a/instead of ဇာ that means ဘယ် /b e/ in Myanmar. In this example, the first sentence is Rakhine (North) conversation and the second one is Rakhine (South) conversation.

Rakhine (North): ဆက်လက် ပြီးကေ ပြည်တွင်း သတင်း တိ ကို တင်ဆက် လား ဖို့ ဖြုက် ပါရေ

(North Phonetic): hs e' l e' b ri: g ei p re d wen: dh a- d en: t i. g ou t en hs e' l a: hp ou. hp r ai' p a r ei

Rakhine (South): ဆက်လက် ပြီးကေ ပြည်တွင်း သတင်း တိ ကို တင်ဆက် လား ဖို့ ဖြုက် ပါရေ

(South phonetic): hs a' l a' b ri: g ei p j e d wan: dh a- d an: t i. g ou t an hs a' l a: hp ou. hp j ai' p a j ei

Figure 4.8: Example of pronunciations difference in north and south

In above figure, all vocabularies are the same but they have different pronunciations between North and South. In this example, although ဆက်လက် is pronounced /hs e' l e'/ in North, South pronounces /hs a' l a'/. For instance, အ က် (e') pronunciation is changed as အက် (a') in South. ပြည်တွင်း is pronounced /p r e d w en:/ in North however /p j e d w an:/ in South. The phonetic changing between North and South Rakhine is described in Figure 2.

4.7 Difference between Rakhine and Myanmar Language

Rakhine language is the closest language to Myanmar though there are many words which differ between the two languages. There are vocabulary differences between Rakhine and Myanmar Language. The vocabulary difference is shown in Table 4.11.

Table 4.11: Vocabulary difference between Rakhine and Myanmar language

Rakhine (Arakanese)	Myanmar (Burmese)
ယေဒု/ယားတွက်နန့်	ထို့ကြောင့်
ဇာဂရာ	ဂါဝန်
အမိုင်း/အီးယောင်	အမေ
လောင်ဗွမ်း	ပန်းကန်
စူးဇွန်	ခက်ရင်း
ဘုသျှေ	ကလေး
လလှပပါ့	လှလှပပ
အထက်ပေါ်ရီ	အပေါ်ယံ

Rakhine vocabulary is significantly changed with Myanmar language. Vocabularies described in table above, (ယေဒု/ယားတွက်နန့်, ဇာဂရာ, အမိုင်း/အီးယောင်, အမိုင်း/အီးယောင်, လောင်ဗွမ်း, စူးဇွန်, ဘုသျှေ, လလှပပါ့, အထက်ပေါ်ရီ) are not found in Myanmar vocabulary.

Moreover, Rakhine language has not only significant vocabulary differences but also differs pronunciation with Myanmar language. For example, အာကာ (a g a) in Rakhine, ကောင်းကင် (g aun; g in) in Myanmar. Some words are same words but same voice have difference meaning eg. ပုဆိုး (p a- hs ou:) means စောင် in Rakhine which is not လုံချည် (l oun gy i) called ဒယော (d a- j o:) in Rakhine. Some are same words but difference pronunciation eg, ကြိုတင် pronounce (k r ou t en) in Rakhine but ကြိုတင် pronounce (k j ou t in) in Myanmar. Others are the same both meaning and pronunciation eg, ပန်းသီး (pan: thi:). These words are native words အဘုသျှေ (a- b u. ch

ei), ပဒါကာသီး (b a- d a g a th i:), ချောဒေါင့်သီး (ch o: d aun. th i:), ငသတိုက် (ng a- dh a- d ai'), အမင် (a- m en), ဘားဘာ (b a: b a) which are not found in Myanmar. Others are foreign words which are ကိုဗိုက် (k ou b ai'), စမ်တမ်ဘာ (s a- t en b a). Most of the foreign words are pronounced one or more sound in Rakhine. Example, ကိုဗိုက် (k ou b ai'), ကိုဗိုက် (k ou b ain), ကိုဗိုက် (kh ou b i.), စမ်တမ်ဘာ (s a- t en b a). စမ်တမ်ဘာ (s an t en b a), စမ်တမ်ဘာ (s en t en b a).

Table 4.12: Difference between Rakhine and Myanmar

Rakhine (Arakanese)	Myanmar (Burmese)
အာကာ /a g a/	ကောင်းကင် /g aun: g in/
ပုဆိုး / p a- hs ou:/	စောင် /s aun/
ဒယော / d a- j o:/	လုံချည် /l oun gy i/
ကြိုတင် / k r ou t en /	ကြိုတင် /k j ou t in/
ပန်းသီး /p an: th i:/	ပန်းသီး (p an: th i:)
အဘုသျှေ /a- b u. ch ei/	ကလေး /kh a- l ei:/
ပဒါကာသီး /b a- d a- g a th i:/	သဘော်သီး /th en: hp o th i:/
ကိုဗိုက် /k ou b ai'/, /kh ou b i. /	ကိုဗစ် /k ou b i'/

CHAPTER 5

BUILDING RAKHINE ACOUSTIC MODEL, LANGUAGE MODEL AND EVALUATIONS

In this chapter, Hidden Markov Model and Gaussian Mixture Model HMM-GMM based acoustic models are presented to develop Rakhine automatic speech recognition system (RASR). MFCC feature extraction technique and decoder applied in the experiments are also described. Moreover, word segmentation and building language model for Rakhine language is also presented.

The creation of Rakhine phonetic dictionary is also stated in this chapter. Finally, the training and test data used in this experiment is described. The evaluation of Rakhine Automatic Speech Recognition (RASR) performance is conducted on four kinds of acoustic models by using HMM-GMM.

5.1 Hidden Markov Model (HMM) and Gaussian mixture model (GMM) based Acoustic Models

Acoustic modeling is the main component part of ASR system. The connection between the acoustic information and phonetics is established in this model. It plays an essential role in accuracy of the system and responsible for computational load. There are many acoustic modeling techniques such as Hidden Markov Mode (HMM), Artificial Neural Network (ANN), Deep Neural Network (DNN) and Time Delay Neural Network and so on. Among them, HMM-GMM based acoustic models is widely used in traditional speech recognitions and known as it is an effective algorithm for training and recognition [16]. HMM is the most powerful parametric model at acoustic level and GMM is used to model the distribution of the acoustic characteristic of speech. They can be modeled for a wide range of time series data and this technique is widely used because of its high recognition accuracy.

5.1.1 Hidden Markov Model

The hidden Markov model is a popular machine learning model in speech recognition. Hidden Markov models are generative model based on stochastic finite state networks. Markov models are stochastic state machines that have a finite set of N states. Given a pointer to the active state at time t, the selection of the next state has a

constant probability distribution. Therefore, the sequence of states is a stationary stochastic process. An n-order Markov assumption is that the probability of entering a given state depends on the occupation in the previous n states.

HMM is a simple network that can generate speech (sequence of cepstral vectors) using the number of states for each model. HMM model $\lambda = (A, B)$ and an observation sequence $O = O_1, O_2 \ldots\ldots O_T$ which determine the likelihood $P(O \mid \lambda)$. And, an observation sequence O and in HMM, learn the HMM parameters A and B. In HMM, an observation sequence O and HMM $\lambda = (A, B)$, discover the best state sequence $Q = q_1, q_2, \ldots, q_n$.

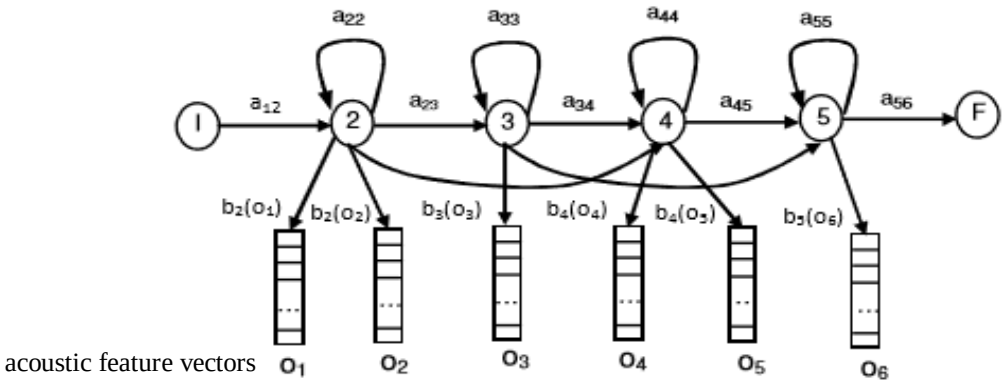


Figure 5.1: Typical Hidden Markov Model Architecture

In the above figure, a set of N states and $a_{11}, a_{12}, \ldots a_{nn}$ represents a transition probability matrix, each state a_{ij} representing the probability from state i to state j and $O_1, O_2, \ldots, O_T$ (T = 1,2, …., n) is a sequence of T observations (acoustic features vectors). $b_j (O_t)$ is a sequence of observation likelihoods, also called emission probability of an observation O_t that generated from state j.

To recognize a small number of words in a speech processing task, applying an HMM state to define a phone is sufficient. However, generally continuous speech recognition tasks, a more fine-grained representation is required. To account for information regarding the non-homogeneous nature of phones over time in CSR, a phone is usually modeled with more than one HMM state [8]. Mostly, common configuration uses three HMM states, begin, middle, and end states to represent a phone. Therefore, each phone has 3 transmitting HMM states instead of one (plus two non-emitting states at either end).

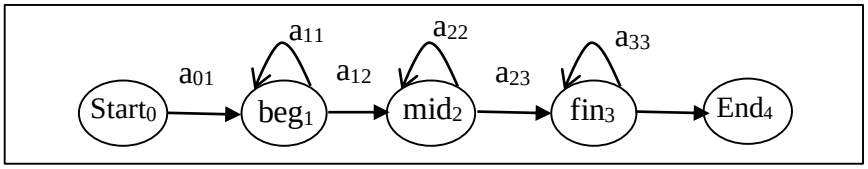


Figure 5.2: A standard 5-state HMM model for a phone

To create an HMM for an entire word using these more complex phone models, each phone of the word model is simply substituted with a tri-state phone HMM. The non-emissive start and end states for each phone model are substituted by transitions directly to the previous and next phone emissive state, leaving only two non-emissive states for the entire word, as shown in Figure (5.3).

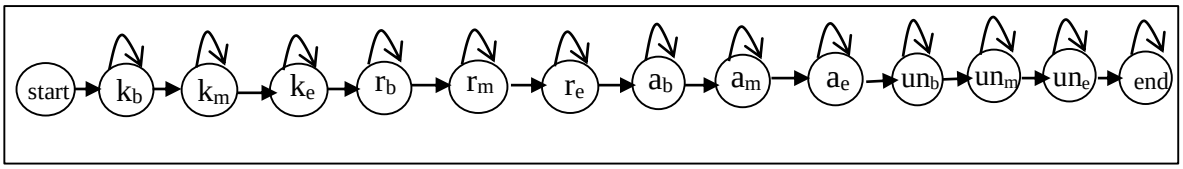


Figure 5.3: A composite word model for "ဧကုန်" [k r a un]

5.1.2 Gaussian Mixture Model

Modern acoustic models are based on the calculation of observation probabilities directly on the real valued, continuous input acoustic feature vector. These acoustic models are based on computing of a probability density function (pdf) on a continuous space. For calculating acoustic probabilities, the most common method is Gaussian mixture model (GMM) pdfs [15].

A Gaussian distribution is a function parameterized by a mean, or average value, and a variance. μ indicates the mean, σ^2 indicates the variance which characterizes the average spread or dispersal from the mean. The following formula is for a Gaussian function:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5.1}$$

$$\mu(mean) = {}^{1}/_{N} \sum_{i=1}^{N} x_i \tag{5.2}$$

$$\sigma^2(\text{variance}) = {}^{1}/_{N} \sum_{i=1}^{N} (x_i - u)^2 \tag{5.3}$$

For a given HMM state with mean vector and covariance matrix and a given observation vector, O_t the multivariate Gaussian probability is estimated using the following equation (5.4). In this equation, D is defined as the number of dimensions and the covariance matrix ∑j means the variance between each pair of feature dimensions.

$$bj\ (Ot) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}\ (\mathrm{Ot} - \mu\mathrm{j})^T \sum_j^{-1} (\mathrm{Ot} - \mu\mathrm{j})) \quad (5.4)$$

A multivariate Gaussian model assigns a likelihood score to an acoustic feature vector observation. The system does not always model the observation likelihood for a non-normal distribution, with a single multivariate Gaussian; instead, it is modeled with a weighted mixture of multivariate Gaussians. This model is called a Gaussian mixture model (GMM). The equation for the GMM is described in the following:

$$f(x|\mu, \Sigma) = \sum_{k=1}^{M} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp[(x - u_k)^T \Sigma^{-1} (x - \mu_k)] \quad (5.5)$$

In this equation, x is an observations O for phone likelihood computation and μ is mean and Σ is covariance matrix. The output likelihood function $b_j(O_t)$ as the GMM is shown in the following:

$$\mathrm{b_j}(\mathrm{O_t}) = \sum_{k=1}^{M} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp[(x - u_k)^T \Sigma^{-1} (x - \mu_k)] \quad (5.6)$$

In training the GMM likelihood function, Baum-Welch is used to express the probability of a certain mixture accounting for the observation. And then, it iteratively updates this probability. It is a computationally efficient algorithms also known as (forward backward algorithms). In speech recognition system, Baum-Welch algorithm is used to train its parameters and to search likelihood of speech sample. In this process, to find P (O/λ), given the observation sequence O = O_1, O_2, ……., O_T. It is applied to calculate the maximum likelihoods and estimation of posterior mode for the parameters for HMM in training process.

5.2 Feature Extraction

Feature extraction is one of the integral parts of speech recognition [18]. It is a basic and fundamental pre-processing step to pattern recognition and machine learning problem in which relevant data are extracted from the speech. It transforms speech

waveform into a sequence of acoustic feature vectors, each vector representing the information in a small-time window of the signal. It's a special form of dimensionality reduction technique which is used to reduce the data which is very large to be processed by an algorithm.

5.2.1 Mel Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficient (MFCC) is a popular feature extraction technique. Basically, it includes Pre-emphasis, Framing and Windowing, Discrete Fourier Transform, Mel Filter Bank, Inverse Discrete Fourier Transform. The steps of Mel Frequency Cepstral coefficient feature extraction are shown as in Fig (5.4).

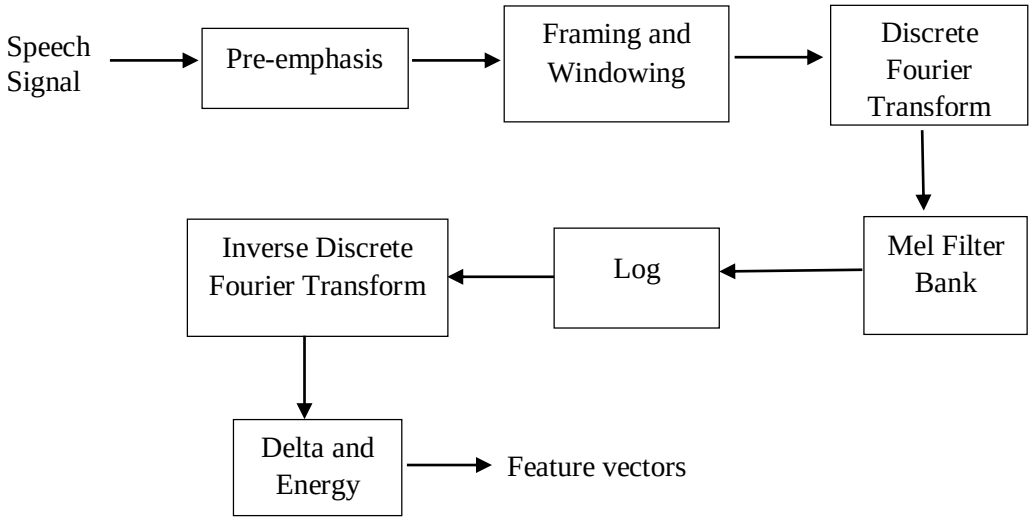


Figure 5.4: MFCC feature extraction steps

5.3 Language Model

Language modeling is used in many application fields of natural language processing. It also plays an essential part in creating speech processing tasks such as ASR and in natural language processing applications (part-of-speech tagging, word segmentation, etc.). There are two ways of language models such as grammars and statistical language models which are utilized in speech recognition tasks. The grammar type language model states very simple types of languages for command and control. They are typically written manually or generated automatically with plain code.

The statistical language model applies stochastic approach called n-gram language model [7]. An n-gram is an n-token words sequences. There are 2-gram

(bigram) that is a pair of two-word sequence and a 3-gram (trigram) is a pair of a three-word sequence. Basically, n-gram language model is applied to find for correct word sequence by predicating the likelihood of the n^{th} word, using the n-1 preceding words. The probability of occurrence of a word sequence W is computed as;

$$P(W) = \prod_{i=1}^{n} P(W_i \mid W_{i-1}) \tag{5.7}$$

Estimate the probability (Maximum Likelihood Estimation),

$$P(W_i \mid W_{i-1}) = \frac{C(Wi-1\ Wi)}{C(Wi-1)} \tag{5.8}$$

Where, Wi is current state i of W, $Wi-1$ is the previous words from current i of w and w is the number of words in the vocabulary.

For statistical-based language modeling, many software packages are available. In these software packages, the CMU-Cambridge Statistical Language Modeling toolkit has been widely applied and popular language modeling toolkits in research work also it has been provided the creating and testing of statistical language models. Other widely applied statistical language modeling toolkit is SRI Language Modeling (SRILM) [4]. In this work, SRI Language Modeling (SRILM) is applied for developing Rakhine Language model.

5.3.1 Creating Rakhine Language Model using SRILM

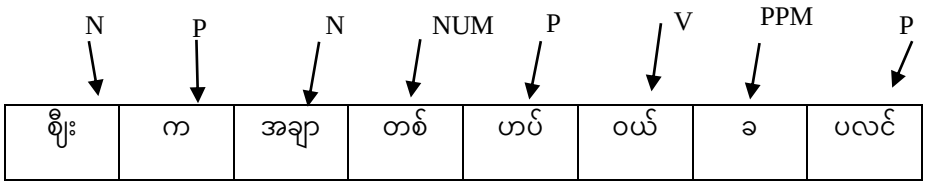
SRILM is used to create statistical-based language model (LM). It is mostly applied in speech recognition, machine translation and statistical tagging. SRILM is an open source toolkit and it is available for non-commercial use. For estimating n-gram models in SRILM command is ngram-count. The ngram-count command manipulates the n-gram counts and estimates the back off models. The -text option defines the input text file to generate the n-gram count. The input text file is normalized text file that has finished word segmentation. The -lm option is to estimate a backoff n-gram model from the total counts and writes it to ARPA file. The ARPA file is the output language model file that predicts the backoff n-gram model from the total counts. And, these language model file contains unigram, bigram and trigrams of words.

5.3.1.1 Rakhine Language Model

In this work, the word-based language model is created for Rakhine language. In developing language model, word segmentation is done on the training and testing data. The text file that contains manually segmented word sequence is trained by using SRILM toolkit. The segmentation of Rakhine sentences is described in the next section.

As the nature of writing system of Rakhine Language, word segmentation is the main point for creating n-gram language model. Generally, it contains Subject, Verb, and Object in writing system. There are nine parts of speech in Rakhine grammar: noun, pronoun, verb, adverb, adjective, conjunction, Post-positional Marker, Particle and exclamation. However, there is no-standard rule for segmenting the Rakhine sentences and word segmentation tool is not available for Rakhine language. Although there is no standard rule for word segmentation in Rakhine, almost of the sentences use spaces between phrases or words. Therefore, Rakhine sentences are manually segmented based on part of speech tagging. The POS of Rakhine Language is also presented in section (3.4). The following sentence is an example of word segmented Rakhine sentence.

Word-based Rakhine sentences



To create language model, the training texts are brought 1439 sentences from the Arakan Princess Media (Rakhine Broadcasts News) by transcription text and 6848 sentences from daily conversation text data. Moreover, 1854 sentences are taken from training corpus of Neural Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese) Myanmar [22]. Therefore, the total language model training corpus contains 10,141 sentences.

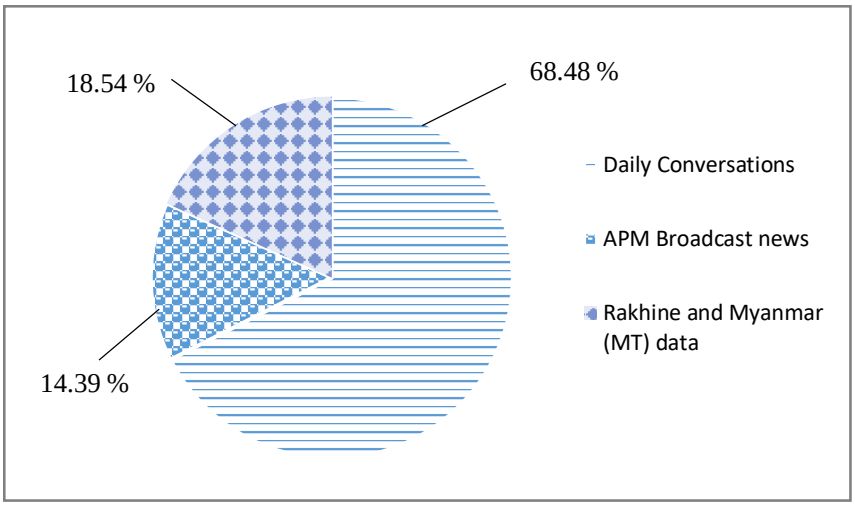


Figure 5.5: Statistics of text data used in the language model

In the above figure, the statistics of text corpus used in the language is described. The sample of input word sequence text file is also described in figure (5.6).

ရန်ကုန် ကွန်ပျူတာ တက္ကသိုလ် က ကြိုဆို ပါရေ

ရခိုင် အသံ က နိန် စာသား ပြောင်း ပီး ရေ စနစ် ဖြိုက် ပါရေ

ကွန်ပျူတာ တက္ကသိုလ် ဇာ နား မှာ လေ

ကောင်း ရီသောက် ချင် နိန်ယာ

နောက်ဆုံးရ ပြည်တွင်း သတင်း နန့် ဒေသတွင်း သတင်း တိ ကို တင်ဆက် ပီး ခ စော် ဖြိုက် ပါရေ

သာလီစွ ပါ ပြည်တွင်း သတင်း တိ ကို တင်ဆက် ပီး ခ စော် ဖြိုက် ပါရေ

နောက်ဆုံး ရရှိ ထား ရေ ထိပ်တန်းရောက် နိုင်ငံတကာ သတင်း တိ နန့် အားကဗေပ် သတင်း တိ ကို တင်ဆက် ပီး ခ စော် ဖြိုက် ပါရေ

Figure 5.6:Example training sentences in language model

An example of n-gram language model file for Rakhine language describes in Figure 5.7 with ARPA format. There are 5,147 1-gram, 24,588 2-gram, and 9,013 3-gram. In the header of ARPA format for n-gram backoff models, it shows how many unique n-gram types were observed of each order n up to the maximum order of the model. After that, n-grams are listed one per line and they are grouped by n-gram order.

```
\data\
ngram 1=5147
ngram 2=24588
ngram 3=9013

\1-grams:
-0.9792872    </s>
-99    <s>    -0.6931928
-1.778639    က    -0.244402
-2.556077    ကကောင်း    -0.4518867
-3.687481    ကကောင်းကြီး    -0.2880004
….
\2-grams:
-1.780482    <s> ကကောင်း    -0.00785114
-2.770342    <s> ကကောင်းကြီး    0.008232918
-2.522558    <s> ကကောင်းမမွန်    0.04962574
….
\3-grams:
-0.551042    ကာလ က ထုတ်ပြန်
-0.3749507    ကုန်ကျစရိတ် က မ
-0.3749507    ကော်မတီ က ဆို
\end\
```

Figure 5.7: Sample of n-gram language model ARPA language model file

The best language model is one that best predicts an unseen test set and the lower perplexity gives better model. Therefore, in this work, language model perplexity is tested with testset1 (Daily Conversation), testset2 (Broadcast News), testset3 (Broadcast News and Daily Conversations) which is described in Table 5.1.

Table 5.1: Perplexity for Rakhine Test Sentences

TestData	TestSet1(Daily Conversations)	TestSet2(Broadcast News)	TestSet3(Daily conversation and Broadcast news)
Perplexity	21.88567	19.40986	38.76969

5.4 Decoding for ASR

Decoding is the process of combination all probability estimators to generate the most probable sequence of words. The best sequence of words is the one that maximizes the product of two factors, acoustic model likelihood, P (O|W) and a language model prior P (W).

Most of the speech recognition system is based on models such as HMM, tree lexicons, or n-gram language models which are finite-state. The system can be characterized by weighted finite state transducers [14].

Weighted finite-state transducers (WFSTs) is a finite-state automata with state transitions labeled with input and output symbols and each transition having an associated weighting. A transition can be defined by an arc from the source state to the target state. It composes the input label, the output label and the weight. The output label of a path is the concatenation of the output labels of its transitions. It integrates different models on composition operations (lexicon, grammar, phonetics) into a single model.

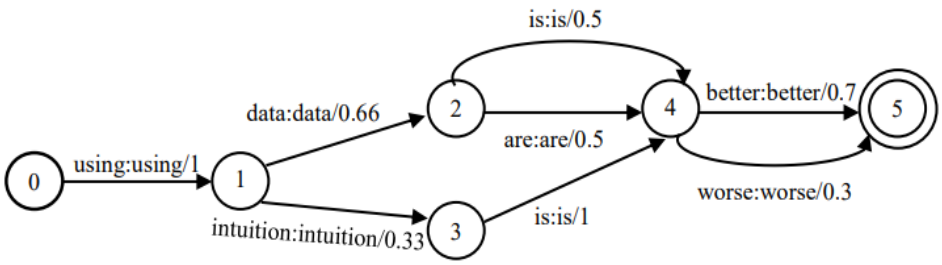


Figure 5.8: Finite-State Language Model

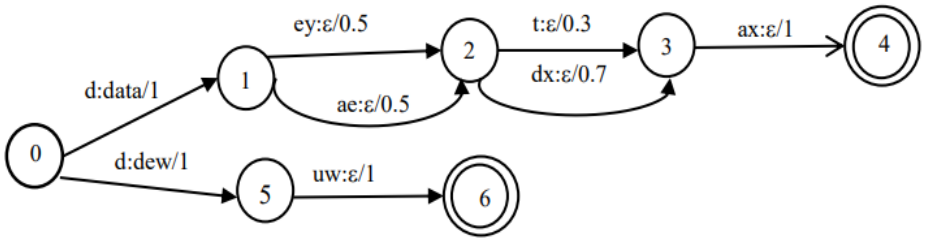


Figure 5.9: Finite-State Pronunciation Lexicon

5.5. Experimental Setup and Evaluation Results

The experiments are done using Kaldi [9] toolkit. The detailed training data and test data used in the experiment is shown in Table (5.2).

5.5.1 Experiment Setup

The speech corpus consists of two types of domains: web news (APM Rakhine Broadcasts news) and conversational data. Both are read speech data types. Rakhine language can be divided into two dialects: Sittwe (North) and Thandwe (South). Therefore, this corpus consists of speakers from North (N) and South (S) regions.

For the baseline GMM-based acoustic model training, the standard Mel-Frequency Cepstral Coefficients (MFCC) features with its first and second derivatives are applied. Then, cepstral mean and variance normalization (CMVN) is applied on MFCC features. After that, splicing 9 frames of MFCCs together and linear discriminant analysis (LDA) is used to project down to 40 dimensions. A maximum likelihood linear transform (MLLT) is applied for estimation on the LDA features. After that, speaker adaptive training is conducted with feature-space Maximum Likelihood Linear Regression (fMLLR) on the top of LDA and MLLT model. There are an average of 44 Gaussian components per state with 2050 context dependent (CD) triphones in GMM-HMM model.

Table 5.2: Training data and test data used in the experiments

Data	Size	Speakers					Utterances
		Female		Male		Total	
		North	South	North	South		
Train Set	5 hrs 46 mins 4 sec	6	4	2	2	14	7723
Test Set 1	15 mins	3	-	-	1	-	418
Test Set 2	15 mins 11 secs	2	1	1	1	5	146
Test Set 3	3 mins	1	1	-	-	-	50

In the above table, TestSet1 is Daily Conversations, TestSet2 is Broadcast News and TestSet3 is daily conversations and broadcast news. In Broadcasts news, the speech corpus size is 3 hours 4 mins 3 secs spoken by 10 speakers (3 females and 2 males in North and 4 females and 1 male in South) with 1439 utterances. For daily conversational data, the duration of the recorded speech is 2 hours 16 mins. Hence, the total size of training size is about 6 hours. TestSet 1 includes 1 speaker with recorded conversational data. TestSet2 involves with native 5 speakers. TestSet1 and TestSet2 is close dataset. TestSet3 is open data set which involves Daily Conversations and Broadcasts news data. It includes 2 speakers (1 female and 1 male).

5.5.2 Evaluation Result

The evaluation of speech recognition system is measured by word error rate (WER). The following formula is used to compute WER,

$$\mathbf{WER} = \frac{Insertions(I)+Substitutions(S)+Deletions(D)}{Total\ Words} * \mathbf{100} \qquad (5.9)$$

Word Error Rate (WER) is calculated by dividing the total number of insertions, substitutions, deletions in the hypothesis text by the total number of word in the

reference text. And, the result is multiplied by one hundred. The example of hypothesis text, reference text and word error rate is depicted as follow;

Reference Text

တစ်ဆက်တည်း မှာ ပင် အာဖကန် အရီး ကူညီ ဆောင်ရွက် ဖို့ မူဆလင် နိုင်ငံ တိ တွိဆုံ ဆွေးနွေး ဖို့ ဆိုရေ သတင်း ကို လေ့ ပြောပြ ချင် ပါရေ

Hypothesis Text

S S I

တစ်ဆက်တည်း မှာ ပင်း အာဖကန် အရီး ကူညီ ဆောင်ရွက် ဖို့ မှာ မူဆလင် နိုင်ငံ တိ တွိဆုံ ဆွေးနွေး ဖို့ ဆိုရေ သတင်း ကို လေ့ ပြောပြ ချင် ပါရေ

Insertions (I) = 1, Substitutions (S) = 2, Deletions (D) = 0

WER = (1+2+0)/21*100 = 14.29%

Rakhine ASR performance is evaluated by using HMM-GMM acoustic modeling technique. The result of different experiment as shown in Table 5.3.

In this implementation, four kind of acoustic model training process is used: Mono – a context independent monophone model, Tri1 – a context dependent triphone model with MFCC feature (Δ+ΔΔ), Tri2 – the triphone model with speaker independent transformation used MFCC + Linear Discriminant Analysis (LDA) + Maximum Linear Likelihood Transform (MLLT), Tri3 - the triphone model with speaker dependent transformation added MFCC + (LDA) + (MLLT) + feature space Maximum Likelihood Linear Regression (fMLLR).

Table 5.3: Evaluation of Rakhine ASR performance in terms of WER

HMM-GMM Model	WER%		
	TestSet1	TestSet2	TestSet3
Mono (Δ+ΔΔ)	27.35	28.81	35.75
Tri (Δ+ΔΔ)	24.11	19.89	33.72
Tri (LDA+MLLT)	22.24	18.68	32.99
Tri (LDA+MLLT+SAT)	20.56	17.77	30.77

For context independent (CI) monophone training, word error rates (WERs) of 27.35 on TestSet1 (Conversational data), 28.81% on TestSet2 (Broadcast news data) and 35.75% on TeSet3 are achieved. When context dependent triphone model with MFCC+Δ+ΔΔ features are applied, it can be reduced 3.24% on TestSet1, 8.92 % on TestSet2 and 2.03% on TestSet3 than the baseline monophone model. When the triphone model with speaker independent transformation (MFCC+ Linear Discriminant Analysis (LDA) + Maximum Linear Likelihood Transform (MLLT)) are used, WERs of 22.24% on TestSet1, 18.68% on TestSet2 and 32.99 % on TestSet3 are obtained. With speaker adaptive training (MFCC + LDA + MLLT + SAT), it can be decreased 6.79% WER on TestSet1 and 11.04% on TestSet2 and 4.98% on Testset3 in comparison with the baseline monophone model.

As a result, the lowest WERs of 20.56% on Test set1 (Conversational data) and 17.77% on Testset2 (APM Broadcast news) and 30.77 % on Testset3 (broadcast news and daily conversation) are attained with the speaker adaptive training. When comparing the evaluation result of Testset1 (Conversational data) and Testset2 (Broadcasts news data), Testset2 has lower error rate than TestSet1 because broadcast news has clear voice and less noisy than the recording data.

5.6.3 Error Analysis

Error Analysis is done based on the recognition hypothesis text. The following errors are found in Rakhine ASR hypothesis text.

5.6.3.1 Similar Pronunciation Error

Some words are falsely recognized that have similar pronunciations. For instance, Rakhine word အဂင့် (a- g en.) is incorrectly recognized အခင်း (a- kh en:). Another example is the word လယ်သမား (l e th a- m a:) is wrongly output as လျှသမား (l h e th a- m a:).

5.6.3.2 Vowel Error

This system misrecognizes some vowels in the Rakhine ASR output text. As an example, the Rakhine word ဧလ (l ei) is incorrectly recognized as လယ် (l e). In this case

the vowel ‘e’ is flasely recognized as (ei). The another example is that the word “ပေါက်တော” (“p au’ t o:”) is produced as “ပေါက်တောင်” (“p au’ t aun”).

5.6.2.3 Tone Error

Some words were also occurred tone error in this experiment. For example, the Rakhine word “မှ” (mha.) gave incorrect output “မှာ” (“mha”). The next sample of tone error is that the word “စာ” (“sa”) is wrongly recognized as “စား” (“sa:”).

5.6.2.4 Ambiguous Error

Some ambiguous cases were not clearly defined in this work. For instance, the Rakhine word “အကယောင်း” (“a- g a- j aun:”) was confused as “အကောင်း” (“a- k aun:”). Both words are different word and different meanings however both words appear as same pronunciation.

CHAPTER 6

CONCLUSION

This chapter summarizes the research work of Rakhine ASR and presents the advantages and limitations of the system. Furthermore, this chapter also describes future work on Rakhine ASR.

6.1 Thesis Summary

In this research, a speaker independent continuous speech recognition for Rakhine language is developed by utilizing the classical acoustic model, HMM-GMM. Rakhine language can be considered as a low-resourced language and speech corpus are not freely available for Rakhine language. Thus, in this task, speech corpus is constructed by utilizing two types of domains. They are broadcast news and daily conversations. In constructing speech corpus, the news is collected from the site of Arakan Princess Media and daily conversations data is recorded by ourselves (which involves me and other 3 native speaker). The total collection speech data size is nearly 6 hours 20 mins which includes 14 speakers. By using these collected speech corpus as training corpus, the baseline GMM-HMM acoustic model is created. For evaluation of the system, two open test sets are conducted which are testset1 and testset2. Testset1 is recording data which involves 1 speaker and Testset2 is the broadcast news data which involves 5 speakers. The experiments are conducted according to Mono (Δ+ΔΔ), Tri (Δ+ΔΔ), Tri (LDA+MLLT), Tri (LDA+MLLT+SAT).

As a result, word error rate is achieved 20.56% on conversational data (Testset1) and 17.77% on broadcast news data (Testset2). The better accuracy of Rakhine ASR is achieved on speaker adaptive training. In this work, although the training data set is a small data set, the ASR performance for Rakhine language gets the promising result because of the Rakhine lexicon containing words which covers the most frequent words of the web text and daily conversations data. When comparing the evaluation result of Testset1 (Conversational data) and Testset2 (Broadcasts news data), Testset2 has lower error rate than TestSet1 because broadcast news has clear voice and less noisy than the recording data.

In addition, error analysis is done Rakhine ASR recognition outputs, hypothesis texts. It is noted that that five significant types of errors are found which are similar pronunciation error, ambiguous error, tone error, vowel error, and foreign error.

7.2 Advantages of the system

This Rakhine automatic speech recognition gets good quality on read speech and it can recognize both daily conversations and broadcast news. The recognition accuracy of broadcasts news data obtains better than that of daily conversations in this work because broadcast news has clear voice and less noisy than the recording data.

According to nature of Rakhine, the training corpus involves north speaker and south speaker. Therefore, when comparing north and south, recognition accuracy of north speaker is better than that of south speaker. It is because the number of north speakers is more than the number south speakers in the training data.

Moreover, this is a system a speaker independent and the people who does not include in training corpus can recognize by using this system. Furthermore, this system can correctly recognize utterance until average 40 words in one sentence.

By utilizing this system, it can assist individuals in the disability community and hearing-impaired people in reading online news. And it’s also useful for news presenters to transcribe automatically recorded audio.

7.3 Limitations and future extension of the System

The system is based on continuous read speech. So, ASR performance can reduce for spontaneous speech. It can also reduce the Rakhine ASR performance in a noisy environment because speech data are recorded in clean environment. Moreover, the system cannot produce directly for English words. Instead of English words, Rakhine words is produced because training was conducted on Rakhine language.

There are still many interesting avenues of discriminative acoustic modeling approaches such as Maximum Mutual Information (MMI), boosted Maximum Mutual Information (bMMI). Therefore, sequence discriminative training will be performed to improve Rakhine ASR performance. Moreover, the size of lexicon and language model will be extended and will be trained for acoustic model with many speakers for future work.

REFERENCES

[1] A.N.Mon, W.P.Pa, "*Exploring the Effect of Tones for Myanmar Language Speech Recognition using Convolutional Neural Network (CNN)*", in the 15th International Conference of the Pacific Association for Computational linguistics (PACLING), August 16-18, 2017, Yangon, Myanmar.

[2] A.N.Mon, W.P.Pa , Y.K.Thu, and Y.Sagisaka, "Developing A Speech Corpus From Webs News for Myanmar (Burmese) Language", In Proceedings of the 20th Conference of the Oriential Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (OCOCOSDA 2017), Seoul, R.O.Korea, pp. 1-6, November 1-3,2017.

[3] Arakan Princess Media, (Rakhine Broadcast news)

http://www.facebook.com/ArakanPrincessMedia

[4] A.Stolcke, "Srilm-An Extensible Language Modeling Toolkit", pp. 901--904 (2002).

[5] Author (အသျှင်စကိ္ခန္ဒ), ရခိုင်ဘာသာစကားလမ်းညွှန်, published in 1994, October. (Rakhine Guidance Book)

[6] B.H.Juang and L.R.Rabiner, "Automatic Speech Recognition–A Brief History of the Technology Development", Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, January, 2005.

[7] C.Martins, A.Teixeira, J.Neto, "Language Models in Automatic Speech Recognition", REVISTA DO DETUA, Vol. 4, No. 2, 2004

[8] D.Jurafsky and J.H.Martin, "Speech and Language Processing: An Introduction to Natural Language Processing", Computational Linguistics, and Speech Recognition, Prentice Hall PTR, Upper Saddle River, NJ, USA, 1 st edition, 2000

[9] D.Povey, et al., "The Kaldi Speech Recognition Toolkit," Idiap, 2011.

[10] H.M.S.Naing, W.P.Pa, "*Automatic Speech Recognition on Spontaneous Interview Speech*", 16th International Conference on Computer Application 2018, Yangon, Myanmar.

[11] HMM-GMM for Speech Recognition (http://medium.com/@jonathanhui/speech-recognition-gmm-hmm-8bb5eff8b196)

[12] K.Me.Me.chit, Laet Laet Lin, "*Exploring CTC Based End-To-End Techniques for Myanmar Speech Recognition*", International Conference on Intelligent Computing & Optimization ICO 2020, pp 1038-10 46.

[13] M.A.Aye Aung, W.P.Pa,*"Time Delay Neural network for Myanmar Speech Recognition*", In proceeding of the IEEE 18th International Conference on Computer Applications ,27th -28th February,2020.

[14] M.Mohri, F.Pereira and M.Riley, "Weighted Finite-State Transducers in Speech Recognition", International Journal of Computer Speech & Language, Vol. 16, No. 1, pp. 69-88, 2002.

[15] N.Singh, A.Agrawal, R. A. Khan ''Gaussian Mixture Model: A Modeling Technique for Speaker Recognition and its Component'' Advanced Computing and Communication Techniques for High Performance Applications (ICACCTHPA-2014).

[16] P. Bansal, A. Kant, S. Kumar, A. Sharda, S. Gupta, "IMPROVED HYBRID MODEL OF HMM/GMM FOR SPEECH RECOGNITION," International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008.

[17] P.K.Kurzekar, R.R.Deshmukh, V.B.Waghmare and P.P.Shrishrimal, "Continuous Speech Recognition System A Review", Asian Journal of Computer Science and Information Technology, Vol. 4, No. 6, pp. 62-66, 2014.

[18] P.K.Kurzekar, R.R.Deshmukh, V.B.Waghmare, P.P.Shrishrimal, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 12, pp. 18006-18016, December 2014.

[19] S.K.Saksamudre, P.P.Shrishrimal, and R.R.Deshmukh, "A Review on Different Approaches for Speech Recognition System", International Journal of Computer Applications (0975 – 8887), Vol. 115, No. 22, April 2015.

[20] S.Mandal, B.Das, P.Mitra, and A.Basu, "Developing Bengali Speech Corpus for Phone Recognizer Using Optimum Text Selection Technique", in International

Conference on Asian Language Processing, IALP 2011, Penang, Malaysia, pp. 268-271, November 15-17, 201.

[21] T.Neuberger, D.Gyarmathy, T.E.Gráczi, V.Horváth, M.Gósy, and A.Beke, "Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language", in Text, Speech and Dialogue - 17th International Conference, TSD 2014, Brno, Czech Republic, pp. 424-431, September 8-12, 2014.

[22] T.M.Oo, Y.K.Thu and K.M.Soe, "Neural Machine Translation Between Myanmar (Burmese) and Rakhine (Arakanese)", In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 80–88, Ann Arbor, Michigan. Association for Computational Linguistics.

[23] T.T.Nwe and T.Myint, "Myanmar Language Speech Recognition with Hybrid Artificial Neural Network and Hidden Markov Model", In Proceedings of 2015 International Conference on Future Computational Technologies (ICFCT'2015), Singapore, pp. 116–122, March 29-30, 2015.

[24] V.Sornlertlamvanich, N.Thatphithakkul, "Thai Speech Recognition Corpus"

[25] Wutiwatchai, P. Cotsomrong, S. Suebvisai, S. Kanokphara, 2002, Phonetically Distributed Continuous Speech Corpus for Thai Language, Third International Conference on Language Resources and Evaluation (LREC2002), pp. 869-872.

[26] W.Soe and Y.Thein, "Syllable-based Myanmar Language Model for Speech Recognition", in Proceedings of IEEE/ACIS 14th International Conference on Computer and Information Science(ICIS)-2015, pp. 291-296, 2015.

[27] Y.K.Thu, W.P. Pa, A.Finch, J.Ni, E.Sumita and C.Hori, "The Application of Phrase Based Statistical Machine Translation Technique to Myanmar Grapheme to Phoneme Conversion''

http://github.com/ye-kyaw-thu/myG2p

[28](ရခိုင်ဂန္ထဝင်ဝေါဟာရ)

http://drivegoogle.com/file/d/1A23zRnYjzoBBFnratpM06Dj3xkTRHAA/view?usp=drivesdk