Study on Extremely Low-Resource Automatic Speech Recognition (ASR) with Burmese, Shan, and Pa'O Languages

Khaing Zar Mon[†] Ye Kyaw Thu[†‡] Hay Man Htun[†]
Zun Hlaing Moe[†] Thida San[†] Hnin Aye Thant[†] Reenu[†]
[†]University of Technology (Yatanarpon Cyber City), Myanmar
[‡]National Electronics and Computer Technology Center (NECTEC), Thailand
{khaingzarmon,yekyawthu,haymanhtun}@utycc.edu.mm
{zunhlaingmoe,thidasan,hninayethant,reenu}@utycc.edu.mm

Abstract

The technology of automatic speech recognition (ASR) has progressed greatly over the past few years. Speech and voice recognition is the process of extracting the speech and voice attributes and specifying the same characteristics with the pre-recorded dataset. Research and development of speech technology applications in low-resource languages are challenging due to the difficulty of data collection, lack of available resources of proper speech corpus, and linguistic knowledge in the low-resource languages. In this paper, we propose a low-resource ASR systems for Burmese and ethnic languages of Myanmar. To the best of our knowledge, this is the first speech recognition experiment for ethnic languages of Myanmar. The Burmese speech corpus contains 3 male and 13 female speakers, with a total duration of almost 10 hours of weather forecasting news. Shan speech corpus has a total length of over 4 hours and is made up of 1 male and 2 female speakers. Pa'O's data is over an hour long and includes only one female speaker. We apply word-level and phrase-level segmentation to each language to evaluate the ASR performance. The experimental results show that the Burmese speech recognition based on word-level segmented corpus leading the best result which achieves 12.95% of word error rates.

Keywords: speech recognition, low-resource languages, Burmese and ethnic languages of Myanmar, Shan, Pa'O, speech and text corpus

1 Introduction

Speech recognition, also known as automatic speech recognition or speech-to-text, is a technology which enables a program to process human speech into a written format [1]. A typical ASR system receives acoustic input from a speaker through a microphone, analyzes it using some pattern, model, or algorithm, and produces an output, usually in the form of a text. Traditional automatic speech recognition (ASR) systems are composed of multiple components, including an acoustic model, a language model, a lexicon, and possibly other components, and each of these is trained independently and combined during decoding. Automatic speech recognition technologies require a large amount of annotated data for a system to work reasonably well [2]. For well-resourced languages like English, there are a rich amount of available resources that can be used for speech processing. But for low-resourced languages, the lack of speech corpus is the main difficulty for speech researchers, and they need to build the corpora by themselves to develop ASR systems. Burmese, Shan, and Pa'O languages can be recognized as low-resourced languages because of the lack of available resources for Natural Language processing. The fundamental challenge is a lack of speech corpus, which is essential for developing speech recognition systems. We present Burmese and Shan speech corpora for the weather domain and Pa'O speech corpus for the travel domain in this paper.

Aye Nyein Mon et al. introduced a Myanmar speech corpus for automatic speech recognition. In this paper, a speech corpus named UCSY-SC1 (University of Computer Studies Yangon - Speech Corpus1) is created for Myanmar ASR research. The corpus consists of two types of domain: news and daily conversations. Experiments were conducted on different data sizes and evaluation is done by two test sets: TestSet1, web news, and TestSet2, recorded conversational data. The Myanmar ASR using this corpus is leading to word error rates of 15.61% on TestSet1 and 24.43% on TestSet2 [3]. Hay Mar Soe Naing et al. proposed a Myanmar large vocabulary continuous speech recognition system. In this system, 3 kinds of acoustic models; 1 Gaussian Mixture Model (GMM) and 2 Deep Neural Networks (DNNs) were explored by only utilizing the developed phonemically balanced corpus consisting of 4K sentences and 40 hours of speech. An open evaluation set containing 100 utterances, spoken by 25 speakers, were experimented. Concerning the sequence discriminative training DNN, the results reached up to 15.63% in word error rate (WER) or 10.87% in SER [4].

The primary goal of our paper is to create speech corpora and conduct research on speech recognition performance for Burmese, Shan, and Pa'O languages. At present, the state-of-the-art ASR systems are generally developed with corpus-based approaches in which a large-scale speech data for the acoustic model (AM) and textual data for the language model (LM) are necessary. In most cases, the more data corresponding to the recognition task, the better the ASR system will be. However, creating such data is usually time-consuming and costly [4]. For the ethnic languages of Myanmar, besides insufficient data, there is no prior system to be referred to, therefore, to construct a system in a short period of time with a limited budget would be a challenge.

This paper is organized as follows. In the next section, we present a brief review of Burmese, Shan, and Pa'O languages. Sections 3 and 4 address data preparation and the ASR architecture, respectively. Section 5 presents the findings as well as some discussion, and Section 6 concludes the paper.

2 Myanmar language, Shan language, and Pa'O language

2.1 Tones and Syllable structure of Myanmar Language

Burmese is the official language of Myanmar and it is also the most widely spoken language in Myanmar. About 32 million people speak Burmese as their first language and 10 million people speak it as a second language [5]. Burmese has a simple syllable structure consisting of an initial consonant followed by a vowel with an associated tone. There are no final consonants. Burmese is a tonal language. This means that all syllables have prosodic features that are an integral part of their pronunciation and that affect word meaning. Prosodic contrasts involve not only pitch, but also phonation, intensity (loudness), duration, and vowel quality. According to one analysis, Burmese has 4 tones [6]. In the following table, the four tones are marked on the vowel /a/ as an example.

Table 1: Characteristics of Myanmar Tones

Tone	Notation	Description	Example
Low	à	low pitch	k^{h}à - ခါ
High	á	slightly breathy, high pitch	k^{h}á - ခါး
Creaky	a∼	Tense or creaky, high pitch	k^{h}a - ခ
Checked	aʔ	final glottal stop, high pitch	k^{h}aʔ - ခတ်

Myanmar script is composed of 33 consonants, 11 basic vowels, 11 consonant combination symbols, and extension vowels, vowel symbols, devowelizing consonants, diacritic marks, specified symbols, and punctuation marks [7] [8]. Myanmar script represents sequences of syllables where each syllable is constructed from consonants, consonant combination symbols (i.e. Medials), vowel symbols related to relevant consonants, and diacritic marks indicating tone level [9]. The following is an example Burmese sentence, its pronunciation, and translation of English:

Burmese	: မင်္ဂလာ နံနက်ခင်းပါ ရှင်
Pronunciation	: min ga- la nan ne' khin: pa shin
English	: Good morning

2.2 Tones and Syllable structure of Shan Language

Shan is one of the main eight ethnic groups of Myanmar, with a population of nearly 6 million in Shan state according to the 2017 Myanmar population. Hence, Shan is said to be the second-largest ethnic group in Myanmar. The Shan language (Shan written: လိၵ်ႈတႆး, pronounced [lik táj]; Shan spoken: ၵႂၢမ်းတႆး, pronounced [kwá:m táj]; Burmese: ရှမ်းဘာသာ) is the native language of the Shan people and is mostly spoken in Shan State, Myanmar. There are also Shan speakers in Kachin, Kayah, Mandalay, and in the Sagaing regions. Shan is a tonal language with five tones, plus a sixth which is used for emphasis. (တူၼ်းသဵင် ႖ တူၼ်း ၊ "– ၊ ◌ႇ ၊ ◌ႈ ၊ ◌း ၊ ◌ႉ ၊ ◌ႊ" in Shan) [10]. Shan has phonemic contrasts among the tones of syllables. Table 2 shows an example of Shan phonemic tones [11].

Table 2: Characteristics of Shan Tones

Tone	Shan	IPA	Transliteration	English
rising	ၼႃ	nǎ:	na	thick
low	ၼႃႇ	nà:	na,	very
mid	ၼႃႈ	nā:	na;	face
high	ၼႃး	ná:	na:	paddy field
creaky	ၼႃႉ	na̰	na.	aunt, uncle

The basic components of Shan language are 19 consonants (တူဝ်မႄႈလိၵ်ႈ ႑႙ တူဝ် in Shan), 10 basic vowels (မႄႈၵပ်းငဝ်ႈ ႑႐ တူဝ် in Shan), 15 diphthongs (မႄႈၵပ်းသွၼ်ႉ ႑႕ တူဝ် in Shan), 3 medial diacritics (မႄႈသဵင်သွၼ်ႉ ႓ တူဝ် ၊ “ျ” ၊ “ြ” ၊ “ႂ” in Shan), 60 final vowels (တူဝ်ယႅတ်းသဵင်လင်း ၊ မ-ၼ-င ႖႐ တူဝ် and တူဝ်ယႅတ်းသဵင်ၶၢတ်ႇ ၊ ပ-တ-ၵ ႖႐ တူဝ် in Shan). Writing the Shan numbers is very easy. For example, 1 equals ႑ , 10 equals ႑႐, 100 equals ႑႐႐, etc [10].

The word order of Shan sentence is Subject-Verb-Object (SVO). The grammatical order of Shan language is like English. The following is an example Shan sentence, its pronunciation, and translation of English:

Shan : မႂ်ႇသုင်ၵၢင်ၼႂ်ၶႃႈ
Pronunciation: maɯ2 s^{h}uŋ1 kaaŋ1 naɯ k^{h}aa3
English : Good morning

2.3 Tones and Syllable structure of Pa'O Language

Pa'O (also spell Pa-O, Pa Oh) is a Central Karenic language spoken by half a million Pa'O people in Myanmar. It is also the family of the Tibeto Burman Language. The Pa'O people live mostly in Shan State, Kayin State, Kayah State, Mon State, Bago Division and Mae Hong Son Province, in northern Thailand. Pa'O people are the seventh largest ethnic nationality in Myanmar [12].

The Pa'O languages are written using the Burmese script and the same alphabet with the Burmese. The Pa'O languages mainly use a system of phonetics. In the Pa'O alphabet, “◌ႇ” as “Mine Ngar” and “◌ႈ” as “Mine Paat Ngar” make the original pronunciation a little shorter and longer, giving a special meaning. Moreover, the medial “◌်” as “Athat”, “◌ျ” as “Yapint” and “◌ြ” as “Layit or Rayit” pronunciations have different pronunciations in some places. In the Pa'O script, the medial “◌ြ” in “က” as “Ka”, “ပ” as “Pa” and “ဗ” as “Ba” alphabets has a “Layit” pronunciation and the medial “◌ြ” in “ခ” as “Kha” and “ဖ” as “Pha” alphabets has a “Rayit” pronunciation. Thus, for example, in the Pa'O script, “ကြ” is pronounced as “Kla” and “ခြ” is pronounced as “Khra”. Similarly, ပြ” is pronounced as “Pla” and “ဖြ” is pronounced as “Phra”. “◌ြ” as “Layit or Rayit” sounds like “ကြ” as “Kla”, “ပြ” as “Pla”, “ခြ” as “Khra” and “ဖြ” as “Phra” are the most common sounds and alphabets in the Pa'O language and script. Compared to Burmese language, the speech of the Pa'O language is likely to be closer to the written form. Although the Pa'O alphabets are similar to the Burmese alphabets, some alphabets have different pronunciations. The alphabets “ရ”, “သ” are pronounced as “Ya”, “Tha” in the Burmese script, but they are pronounced as “Ra”, “Sa” in the Pa'O script [12].

There are four contrastive tonemes in Pa'O language spoken at Huay Salop village, three contour tones: high rising, high falling, and low falling, and one level tone: mid-level. The open syllables and closed syllables with final nasals / m, n, ŋ / can bear all four of them while the closed syllables with final stops / p, t, k, ʔ / occur only with the high rising and the low falling tones. Pa'O tones function, together with vowels, as the syllable-nucleus [13]. Table 3 shows the phonemic notation of Pa'O languages. /1/ is a high-rising tone. Its tonal figure starts below high-level and then moves up quickly to high-level. It occurs with the open syllable and the closed syllable. /2/ is a high-falling tone. Its tonal figure starts from high-level and then moves down quickly to low-level. It occurs only with the open syllable and the closed syllable with final nasals / m, n, ŋ /. /3/ is a mid-level tone. Its tonal figure starts from mid-level, and continues and ends at the same range. It occurs only with the pre-syllable, open syllable and the closed syllable with final nasals / m, n, ŋ /. /4/ is a low-falling tone. Its tonal figure starts slightly below mid-level and then moves down softly to low-level. It occurs with the open syllable and the closed syllable [13].

Table 3: Characteristics of Pa'O Tones

Phonemic Notation	Description	IPA	English
/1/	high-rising	do$ʔ^1$	cover
/2/	high-falling	$k^h am^2$	rain
/3/	mid-level	ham^3	ground
/4/	low-falling	cop^4	play

In Summary, there are grammatical differences and the most significant differences between Pa'O and Burmese languages are in their pronunciations and their vocabularies. The basic components of Pa'O language are 33 consonants, 8 independent vowels, 3 medial diacritics, 16 dependent vowels. In 33 consonant scripts of Pa'O language, some scripts share the same pronunciations. For example, "ဌ", "ဎ", "ထ", "ဓ" are pronounced as "Hta" and "ဖ", "ဘ" are pronounced as "Pha". Word order of Pa'O sentence is Subject-Verb-Object (SVO), which has same order as English language [12]. Some example Pa'O words and their pronunciation, and translation of English are as follows:

Pa'O : ညာႉ, ရှဥ်, ဗွာ, ထီ, သီ, ဖြေင်း, မွိုး, ဆား, နွိုႉ, ထော့ႉ
Pronunciation : ŋja, rʌn, bwa, t^hi, si, p^hreŋ, mɣ, c^ha, nɨʔ, t^hɔʔ
English : far, silver, white, water, die, black, mother, star, enter, pig

3 Data Preparation

Building speech corpora is the primary step for developing automatic speech recognition (ASR) systems, especially for low-resourced languages, and it is critical for the statistical speech recognition system. In addition, the performance of a speech recognizer depends on the speech corpora. Speech corpora for well-resourced languages such as English are publicly available for speech processing research. However, being low-resourced languages, Burmese, Shan, and Pa'O languages have few public resources for speech processing. In our experiment, we planned to develop Burmese speech recognition system for the weather domain with the purpose of testing the performance of the speech recognition models as weather domain in low-resourced languages is still challenging due to the small amount of data. There was an open-source crowd-sourced multi-speaker speech corpus for Burmese named "Crowd-sourced high-quality Burmese speech data set" [14]. The corpus is a crowd-sourced corpus so, we needed to build Burmese speech corpus for the weather domain. Development of the speech corpora for Burmese, Pa'O, and Shan languages is crucial for improving and promoting Myanmar speech recognition activities. A speech corpus can be built mainly in two methods. The first method is to collect the speech that has already been recorded and manually transcribe it into text. The second method is to create the text corpus first and record the speech by reading the collected text [3].

3.1 Burmese speech and text corpus preparation

To build a corpus associated with Burmese weather forecasting news, weather forecasting news videos are downloaded from the website and Facebook page of the Department of Meteorology and Hydrology, Myanmar. At first, we convert the news videos to wave file format. All the audio files are formatted with sample frequency 16 kHz and mono channel. The audio files are segmented with "NowSmart Cut" audio cutter software and the silent portion of each audio file is discarded. The total audio files used are 3,575 which make up a total duration of 10 hours, 19 minutes and 17 seconds. The news presenters have a clear voice in news broadcasting because they are professional, well-trained and well-experienced. In our corpus, 3 male speakers and 13 female speakers are involved. The data collection process for Burmese lasts about 6 months.

As we did not have any reference text of any of the audio files, the transcriptions are manually done and Myanmar3 Unicode is used for that purpose. Word segmentation is done by hand as Myanmar language has no word boundary. Our text corpus contains 960 unique words. The example news sentence is shown as below:

file name: Daw_Nant_Hsan_Phaung_1(10-2-2020_7am)
Burmese: မင်္ဂလာ နံနက်ခင်းပါ ရှင်
English : Good morning

file name: Daw_Thet_Mar_Soe_8(7-

2-2020_7pm)
Burmese: လှိုင်းအမြင့်မှာ မြန်မာ့ ကမ်းရိုးတန်းတစ်လျှောက် နဲ့ ကမ်းလွန်ပင်လယ်ပြင် တို့မှာ လေး ပေမှ ရှစ် ပေခန့် ရှိနိုင်ပါတယ်
English : Wave height will be about (6-8) feet off and along Myanmar Coasts

Some of the words such as numbers, dates, time, and range cannot be found in the dictionary. So, it is necessary to do text normalization and transliteration into Myanmar language. Table 4 shows the example words that need to be normalized.

Table 4: Example of text normalization

Description	Example	Normalization
Date	၂၀၂၀ (2020)	နှစ်ထောင့်နှစ်ဆယ်
Time	၇ နာရီ ၃၀ မိနစ် ၁၅ စက္ကန့် (7 hours 30 minutes 15 seconds)	ခုနစ် နာရီ သုံးဆယ် မိနစ် ဆယ့်ငါး စက္ကန့်
Range	၃၀ မီလီမီတာ မှ ၄၀ မီလီမီတာ အတွင်း (from 30 millimeters to 40 millimeters)	သုံးဆယ် မီလီမီတာ မှ လေးဆယ် မီလီမီတာ အတွင်း

In our system, we use the "myG2P (Myanmar Grapheme-to-Phoneme) dictionary for VoiceTra (Multilingual Speech Translation Application) Myanmar language project of NICT, Japan (during 2014-2015)" for pronunciation mapping. In this dictionary, the Myanmar Language Commission (MLC) Pronunciation Dictionary is used as a basis for pronunciation mapping. In order to deal with the problem of out-of-vocabulary (OOV) words, we extend the grapheme to phoneme mapping table. Myanmar consonant scripts are grouped by their pronunciation types; un-aspirated, aspirated, voice, and nasal. There are 23 phonemes for 33 consonant scripts, some scripts share the same pronunciations, for example, "ဒ", "ဓ", "ဍ" and "ဎ" [4]. The pronunciation of the syllables can depend on the context of the syllables.

Some Myanmar syllables do not conform to these standard rules of pronunciation. The pronunciation of the syllables can depend on the context of the syllables, for example, the character "စ" in "ဖိတ်စာ" (invitation letter) is pronounced as /s/ and in "သတင်းစာ" (newspaper) is pronounced as /z/. Differences between standard pronunciations and correct pronunciations of some words are shown in Table 5 as examples.

Table 5: Examples of contextually dependent pronunciations of some Myanmar words

Words	Standard	Correct
သုံးဆယ့်ခုနစ် (37)	thoun: hse khu. nhi'	thoun: ze. khun nhi'
သတင်း (news)	tha. tin:	dha- din:
အမြင့်ဆုံး (highest)	a- mjin. hsoun:	a- mjin. zoun:

3.2 Shan speech and text corpus preparation

For Shan language, there are no available resources of weather forecasting news, and we manually prepare the speech corpus and text corpus associated with weather information in Shan language. Firstly, we create text corpus by translating Burmese text corpus into Shan language. Text corpus contains 46,482 words in total and 471 unique words. To develop the speech corpus, the speaker needs to utter a prescribed piece of text. We use commodity mobile phones running the Android platform to record and store high-quality speech in various recording environments. Then we associate a transcript with each utterance. The duration of the audio utterances are between 4.4 and 25.5 seconds and the total duration of these files is 4 hours, 11 minutes and 43 seconds. In our corpus, 1 male speaker and 2 female speakers are involved and all speakers are volunteer participants and native Shan speakers.

In our system, "SEAlang Library Burmese Dictionary" and "Wiktionary, the free dictionary" are used as the basis for phonetic transcriptions in the International Phonetic Alphabet (IPA). The SEAlanglibrary Burmese dictionary resources are primarily based on the Myanmar-English Dictionary (1993, Myanmar Language Commission; republished 1996, Dunwoody Press ISBN 1-881265-47-1). Wiktionary is a collaborative project to produce a free-content multilingual dictionary.

3.3 Pa'O speech and text corpus preparation

We didn't have any resources or references for the weather domain in Pa'O, therefore we created a corpus for the travel domain. We used 1,000 Burmese sentences (without name entity tags) of the ASEAN-MT Parallel Corpus, which is a parallel corpus in the travel domain [16]. Manual translation into Pa'O language was done by native Pa'O monks from Taunggyi, Shan State, Myanmar, and native Pa'O students from Myanmar universities. We create the text corpus in Pa'O language first and record the speech by reading the collected text. Transcriptions contain 1,258 unique words and 4,707 words in total and the duration of the audio files are between 4.5 and 9.9 seconds. In building the speech corpora, if the speech is recorded by ourselves, the professional recording devices are very expensive and it is very time consuming. In this system, we collected the speech data with an easy-to-use, built-in Android audio-recording app on Redmi Note 7. In our corpus, we used 1,000 sentences in the travel domain and it involves only 1 female speaker. We used "epitran", a library and tool for transliterating orthographic text as IPA.

4 ASR Architecture

There are three main stages of the automatic speech recognition. In the feature extraction stage, the sound waveform is sampled into frames that are transformed into spectral features. This step is required for classification of sounds because the raw speech signal contains information besides the linguistic message and has high dimensionality. These characteristics of the raw speech signal would be unfeasible for the classification and result in high WER [16] [17]. In the phone likelihood stage, the system computes the likelihood of the observed spectral feature vectors given linguistic units (words, phones, subparts of phones). The final stage of ASR, the decoding is the process of searching huge HMM network for determining the most likely path given the acoustic observations [18]. The model of speech λ can be divided into the acoustic model λ_{A} and language model λ_{L} and the probability calculations for each can be performed separately [19].

$$arg_w max P(w|o, \lambda_A, \lambda_L) = \\ arg_w max P(o|w, \lambda_A) P(w|\lambda_L)$$

4.1 Feature Extraction

The first step in any automatic speech recognition system is to extract features. Feature extraction is the process of computing a sequence of features for each short-time frame of the input signal, with an assumption that such a small segment of speech is sufficiently stationary to allow meaningful modeling [20]. Feature extraction involves the analysis of speech signals. Commonly used feature extraction techniques are Mel-Frequency Cepstral Coefficients (MFCC), Discrete Wavelet Transform (DWT), Linear Predictive Coding (LPC), Relative Spectral (RASTA-PLP), Linear Prediction Cepstral Coefficients (LPCC), Principal Component analysis (PCA), Linear Discriminant Analysis (LDA) and Perceptual Linear Prediction (PLP).

4.2 Acoustic Model

The acoustic model typically deals with the raw audio waveforms of human speech, predicting what phoneme each waveform corresponds to, typically at the character or sub-word level [17]. The model is trained from a set of audio recordings and their corresponding transcripts. The acoustic model is usually based on Hidden Markov Models and Artificial Neural Networks, mapping the relationship between the audio signal and the phonetic units in the language [21]. Each state of HMMs is usually represented by a Gaussian Mixture Model (GMM) to model the distribution of feature vectors for the given state [18]. Hidden Markov Models are natural candidates for acoustic models since they are great at modeling sequences.

4.3 Language Model

The language model is used to calculate the likelihood of the word sequence. For example, "I drink coffee" will be more likely than "I coffee you" or "I drink waffles". It predicts the next word given the previous words. There are two types of language models that are used in speech recognition systems - grammars and statistical language models [18].

4.4 Decoding

The final step of speech recognition is decoding, the process to calculate which sequence of words is most likely to match the acoustic signal represented by the feature vectors [18].

5 Experimental Setup

We developed traditional automatic speech recognition models for Burmese, Shan, and Pa'O languages with different types of corpus to test the performance of the models. We use Kaldi, a free, open-source toolkit for speech recognition research [22]. Kaldi provides a speech recognition system based on finite-state transducers (using the freely available OpenFst), together with detailed documentation and scripts for building complete recognition systems. Kaldi is written in C++, and the core library supports modeling of arbitrary phonetic-context sizes, acoustic modeling with subspace Gaussian mixture models (SGMM) as well as standard Gaussian mixture models, together with all commonly used linear and affine transforms [22].

5.1 Data Statistics

In our experiments, we apply manual data segmentation to each corpus to increase the robustness of speech recognition on language modeling level by the detection of word and phrase boundaries and it can lead to significant performance improvements of the models. Table 6 shows the training and testing sets for Burmese, Shan, and Pa'O languages.

For Burmese, the total audio files used are 3,575 which make up a total duration of 620 minutes from 16 speakers (13 female speakers and 3 male speakers). We define 350 sentences as testing data and 3,225 sentences as training data. We apply phrase-level and word-level segmentation to the Burmese corpus to enhance the performance of the Burmese speech recognition model.

For Shan language, we have 1,403 sentences which make up a total duration of 251 minutes and it involves 2 female speakers and 1 male speaker. For Pa'O language, we collected 1,000 sentences in total 78 minutes and these sentences are recorded by 1 female speaker. All of the speakers are students of the University of Technology (Yatanarpon Cyber City), Myanmar. As our corpora are relatively small, we apply word-level, and phrase-level data segmentation to boost the performance of the model. For Shan ASR, we use 306 sentences as testing data and 1,097 sentences as training data. For Pa'O ASR, we use 100 sentences as testing data and 900 sentences as training data.

5.2 MFCC Feature Extraction

In our system, we use Mel Frequency Cepstral Coefficients (MFCCs), one of the popular audio feature extraction methods. The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFCC, the frequency bands are positioned logarithmically (on the Mel scale) which approximates the human auditory system's response more closely than the linearly-spaced frequency bands obtained directly from the first Fourier transform or discrete Fourier transform [23].

The steps involved in MFCC feature extraction are pre-emphasis, windowing, Discrete Fourier Transform, Mel scale filter bank analysis, logarithmic compression, and Discrete Cosine transform. A pre-emphasis stage is used to enhance the intensity of higher frequencies within the signal. Next, windowing involves the slicing of the audio waveform into sliding frames. These are often computed using window functions, sometimes called Hamming and Hanning Windows, which more gently taper the boundaries of each window's signal, allowing for higher downstream processing quality [21].

After this, Discrete Fourier Transform (DFT) is applied on each frame of N samples of speech to convert from time domain to the frequency domain. The outputs of the DFT are squared, giving the power of speech at each frequency. Finally, filter banks are computed by applying triangular filters to these power frequencies using a Mel scale. The Mel scale basically maps the signal in a way that is more consistent with the way humans naturally perceive sound. Our ears are more sensitive to lower than to higher frequencies, so the Mel scale adjusts the signal accordingly. The outputs of this step are called filter banks, and they themselves are often used as features to acoustic models [21]. Finally, we take the log of the Mel filter bank outputs and then apply a Discrete Cosine Transform (DCT) to the result.

5.3 HMM-GMM Acoustic Model

We use the open-source Kaldi toolkit [22]. Adopting the Kaldi's standard scripts, we used MFCC+Δ+$\Delta\Delta$ features with standard cepstral mean and variance normalization (CMVN) to train the acoustic model.

Table 6: Training and Testing sets for Burmese, Shan, and Pa’O languages

Language	Types of Corpus	Utterances	Length	Speakers	Training Data Size	Testing Data Size
Burmese	Phrase-level	3,575	10 hr, 19 min, 17 sec	16	3,225	350
Burmese	Word-level	3,575	10 hr, 19 min, 17 sec	16	3,225	350
Shan	Phrase-level	1,403	4 hr, 11 min, 43 sec	3	1,097	306
Pa’O	Phrase-level	1,000	2 hr, 17 min, 54 sec	1	900	100

5.4 N-gram Language Model

Our language model was trained using the SRI Language Modeling (SRILM) toolkit. SRILM is a toolkit for building and applying statistical language models (LMs), primarily for use in speech recognition, statistical tagging and segmentation, and machine translation [24].

5.5 Result and Discussion

In this section, we will present evaluation results for Burmese, Shan, and Pa’O languages. To evaluate the performance of speech recognition models, we used automatic evaluation of word error rate (WER). Several factors can impact word error rate, such as pronunciation, accent, pitch, volume, and background noise [1].

Table 7 shows the evaluation results of speech recognition models for Burmese, Shan, and Pa’O languages. In our experiment, we compare the word error rates of phrased-based, word-based ASRs and syllable error rates of syllable-level segmented data. For fair comparison, the phrases and words of the hypothesis text of the phrase-based and word-based ASRs are changed to syllable levels and the WERs of that words are calculated again. Syllable is a basic unit of Myanmar, Shan and Pa’O languages. We used the syllable segmentation tool named “sylbreak4all” to change phrases and words to syllable levels [25]. “sylbreak4all” is a syllable breaking tool for nine ethnic languages of Myanmar. “sylbreak4all” can be used for Burmese, Shan, PaO, Sgaw Kayin, Pwo Kayin, Dawei, Beik, Mon, and Rakhine languages.

Myanmar language are made up of a large proportion of monosyllable words where a single word has a meaning in itself. For example, “က” means to dance, and “ကျ” means fall down. Those words are sometimes used as root words to form longer words or phrases, such as the word “မ” which negates the verb that follows it. Some of the words can be grouped to become a phrase or can be used as the single words. For example, the two words “ကစားကွင်း” (“the playground” in English) and “သို့” (“to” in English) can be grouped to become a phrase “ကစားကွင်းသို့” (“to the playground” in English). Here, we will compare the phrase-based and word-based Burmese ASRs in terms of word error rate (WER%) and syllable error rate (SER%). We got 16.07% of WER for phrase-based Burmese speech recognition model and 5.8% of SER (syllable error rate) when we apply syllable-level segmented data. We achieved 12.95% of WER for word-based Burmese speech recognition model (substitutions 6.95%, deletions 3.5%, insertions 2.5%) and the error rate reduces to 7.1% of SER (substitutions 2.8%, deletions 1.6%, insertions 2.7%) when we use syllable-level segmented corpus.

Table 7: Evaluation results of GMM ASR models for Burmese, Shan, and Pa’O languages

Language	Types of Corpus	WER%	SER%
Burmese	Phrase-level	16.07	5.8
Burmese	Word-level	12.95	7.1
Shan	Phrase-level	22.84	15.6
Pa’O	Phrase-level	50.4	46.8

According to the evaluation result, the Burmese ASR system was wrongly recognized the words that have similar pronunciation. For example, Myanmar word "ကိုးဆယ့်သုံး"("kou: ze. thoun:") ("93" in English) was incorrectly recognized as "သုံးဆယ့်သုံး" ("thoun: ze. thoun:") ("33" in English). Tone mistakes were also found in this experiment. For example, the Myanmar word "သာယာပါ့မယ်" ("tha ja ba. me") gave incorrect result "သာယာပါမယ်" ("tha ja ba me"). Some ambiguous errors can also be found in the system. As an example, the Myanmar word "ရှိနေပါတယ်" ("shi. nei ba de") was confused as "ဖြစ်နိုင်ပါတယ်" ("hpji nain ba de").

In our experiment, the syllable-based Shan ASR is developed and the evaluation results of the syllable-based Shan ASR and phrase-based Shan ASR are compared. With phrase-level segmented data, Shan ASR achieves 22.84% of WER. The error rate reduces to 15.6% of SER (substitutions 3%, deletions 5.4%, insertions 7.2%) when we use syllable-level segmented corpus. The results demonstrate that the type of segmentation applied to the corpus affects the word error rates. Pa'O speech recognition system achieves 50.4% of WER and 46.8% of SER because the corpus includes only 1 female speaker and 1,000 utterances. It can be clearly seen that SER% are significantly less than the WER%.

6 Conclusion

In this paper, we introduced Burmese, Shan, and Pa'O speech corpora and we evaluated the performance of the ASR models for very-low resourced languages. As Burmese, Shan, and Pa'O languages are low-resourced languages, developing the speech corpus is essential and we believe that these speech corpora will be useful for future Myanmar speech processing research. To the best of our knowledge, this is the first system for Shan and Pa'O speech recognition. As the size of the Burmese speech corpus is larger than Shan and Pa'O speech corpora, the Burmese speech recognition system outperforms Shan and Pa'O speech recognition systems. In our work, Burmese speech recognition based on word-level segmented data reaches the best performance and achieves 12.95% of WER. In the future, we plan to increase the size of each corpus and the number of speakers to enhance the performance of speech recognition models.

Acknowledgment

We would like to express our special thanks to students of the Myanmar-Shan translation team from the University of Technology, Yatanarpon Cyber City (UTYCC), Myanmar. Namely, Sai Kwang Kham, Nang Aeindray Kyaw, Nang Vo Hein, and Nang Ying Mo Hom for participating in the data collection task.

References

[1] By: IBM Cloud Education. (n.d.). What is speech recognition? IBM. Retrieved from https://www.ibm.com/cloud/learn/speech-recognition.

[2] Scharenborg, Odette, Francesco Ciannella, Shruti Palaskar, Alan W. Black, Florian Metze, Lucas Ondel and Mark Hasegawa-Johnson. "Building an ASR System for a Low-resource Language Through the Adaptation of a High-resource Language ASR System: Preliminary Results." (2017).

[3] Aye Nyein Mon, Win Pa Pa & Ye Kyaw Thu (2019). UCSY-SC1: A Myanmar speech corpus for automatic speech recognition. International Journal of Electrical and Computer Engineering (IJECE). 9. 3194. 10.11591/ijece.v9i4.pp3194-3202.

[4] Hay Mar Soe Naing et al., "A Myanmar large vocabulary continuous speech recognition system", 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015, pp. 320-327, doi: 10.1109/APSIPA.2015.7415529.

[5] Nag, Oishimaya Sen (2017, August 1). What languages are spoken in Myanmar (Burma)? WorldAtlas. Retrieved from https://www.worldatlas.com/articles/what-languages-are-spoken-in-myanmar-burma.html.

[6] Burmese language - structure, writing &amp; alphabet - mustgo. MustGo.com. (n.d.). Retrieved from https://www.mustgo.com/worldlanguages/burmese.

[7] MLC. 2002. Myanmar-English Dictionary. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

[8] Ye Kyaw Thu and Yoshiyori URANO, "Text Entry for Myanmar Language

SMS: Proposal of 3 Possible Input Methods, Simulation and Analysis", Proceedings of the 4th International Conference on Computer Applications (ICCA 2006), February 23 24, 2006, Yangon, Myanmar, pp. 199-207

[9] Hla Hla Htay, Kavi Narayana Murthy, "Myanmar Word Segmentation using Syllable Level Longest Matching," In proc. of the 6th Workshop on Asian Language Resources, pp.41-48, 2008

[10] Nang Aeindray Kyaw, Ye Kyaw Thu, Hlaing Myat Nwe, Phyu Phyu Tar, Nandar Win Min and Thepchai Supnithi, "A Study of Three Statistical Machine Translation Methods for Myanmar (Burmese) and Shan (Tai Long) Language Pair," 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1-6, doi: 10.1109/iSAI-NLP51646.2020.9376832.

[11] Wikimedia Foundation. (2021, August 13). Shan language. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Shan_ language.

[12] Hay Man Htun, Ye Kyaw Thu, Hlaing Myat Nwe, May Thu Win, Naw Naw, "Statistical Machine Translation System Combinations on Phrase-based, Hierarchical Phrase-based and Operation Sequence Model for Burmese and Pa'O Language Pair", Journal of Intelligent Informatics and Smart Technology, Oct 2nd Issue, 2021, pp. 1-9.

[13] Thanamteun, Orranat. 2000. A phonological study of Pa-O (Taungthu) at Ban Huay Salop, Tambon Huay Pha, Muang district, Mae Hong Son province. (MA thesis, Mahidol University; xii+154pp.)

[14] Yin May Oo, Theeraphol Wattanavekin, Chenfang Li, Pasindu De Silva, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson and Alexander Gutkin. "Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech." LREC (2020).

[15] Prachya, Boonkwan and Thepchai, Supnithi, "Technical Report for The Network-based ASEAN Language Translation Public Service Project", Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013

[16] Bourlard, Hervé, Hynek Hermansky and Nelson Morgan. "Towards increasing speech recognition error rates." Speech Commun. 18 (1996): 205-231.

[17] O'Shaughnessy, Douglas. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition. 41. 2965-2979. 10.1016/j.patcog.2008.05.008.

[18] Aye Nyein Mon, Win Pa Pa and Ye Kyaw Thu. "Building HMM-SGMM Continuous Automatic Speech Recognition on Myanmar Web News." (2017).

[19] Siivola, Vesa. "Language models for automatic speech recognition: construction and complexity control." (2007).

[20] A, Sithara & Thomas, Abraham & Mathew, Dominic. (2018). Study of MFCC and IHC Feature Extraction Methods With Probabilistic Acoustic Models for Speaker Biometric Applications. Procedia Computer Science. 143. 267-276. 10.1016/j.procs.2018.10.395.

[21] Taylor, Ryan (2021, June 24). What is an acoustic model in speech recognition? Rev. Retrieved from https://www.rev.com/blog/resources/what-is-an-acoustic-model-in-speech-recognition.

[22] Povey, Daniel & Ghoshal, Arnab & Boulianne, Gilles & Burget, Lukáš & Glembek, Ondrej & Goel, Nagendra & Hannemann, Mirko & Motlíček, Petr & Qian, Yanmin & Schwarz, Petr & Silovský, Jan & Stemmer, Georg & Vesel, Karel. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

[23] Mikhael, Wasfy & Premakanthan, Pravinkumar. (2002). Speaker recognition employing waveform based signal representation in nonorthogonal multiple transform domains. 2. II-608 . 10.1109/ISCAS.2002.1011426.

[24] A. Stolcke. Srilm - an extensible language modeling toolkit. pages 901–904, 2002.

[25] GitHub - ye-kyaw-thu/sylbreak4all: Syllable Breaking Tool for Nine Ethnic Languages of Myanmar, 2021. Retrieved from https://github.com/ye-kyaw-thu