Improving Rakhine Automatic Speech Recognition with Subspace Gaussian Mixture model (SGMM)

Aye Nyein Mon
Natural Language Procesing Lab
University of Computer Studies,Yangon
Yangon, Myanmar
ayenyeinmon@ucsy.edu.mm

Hnin Thida Kyaw
University of Computer Studies,Yangon
Yangon, Myanmar
hninthidar.kyaw@ucsy.edu.mm

Win Pa Pa
Natural Language Processing Lab
University of Computer Studies,Yangon
Yangon, Myanmar
winpapa@ucsy.edu.mm

Abstract—Automatic speech recognition (ASR) researches have been developed by many nations for their ethnic languages. ASR researches for Myanmar ethnic languages are very rare and there is only a few research for Rakhine language, which is one of the ethnic groups in Myanmar. It is a low-resourced language and this work aims to improve ASR performance for Rakhine language by utilizing smoothing techniques of language model and Subspace Gaussian Mixture Model (SGMM). Experiments are carried out on about 6 hours of Rakhine speech corpus which consists of broadcast news and recorded conversational speech. The findings show that the impact of language model is significantly great for ASR especially for low-resourced languages and the SGMM-based model outperformed the baseline Gaussian Mixture Model (GMM)-based approach.

Keywords—Rakhine, Subspace Gaussian Mixture Model (SGMM), Automatic Speech Recognition (ASR)

I. Introduction

Speech recognition is one of the modern technologies for human-computer interaction and speech signal is a prominent feature of communicating with natural language. Many researchers have tried to build speech recognition for their languages. They have been proposed by using different techniques for both well-resourced and low-resourced languages. ASR development for languages with limited resources requires starting from scratch with data collection.

Rakhine (Arakanese) is the native language of Rakhine (Arakanese) and it is a tonal language. It is also one of the main ethnic groups in Myanmar and is closely related to the Myanmar language. And, there is only a few ASR research for the Rakhine language. Most Arakanese speak an unusual variety of the Myanmar language which includes significant differences from Myanmar pronunciation and vocabulary. It is one of the low-resourced languages and pre-collected speech data is not available. Therefore, in previous work, the speech corpus for Rakhine ASR was built and it was developed by using the Gaussian Mixture Model-Hidden Markov Model(GMM-HMM) approach[1].

In this work, a Subspace Gaussian Mixture Model (SGMM)-based technique is applied to enhance ASR performance for the Rakhine language. It is an acoustic model that is especially suitable for applications with limited resources [2][3]. The only parameters that are unique to acoustic states are some relatively low-dimensional (for example, 40-dimensional) vectors that characterize fewer parameters than a typical GMM-based system. The common of the SGMM trainable parameters are, in typical configurations, globally shared and not specific to any individual acoustic state. Therefore, SGMM-based automatic speech recognition is developed for Rakhine ASR in this work. In addition, the ASR performance will be explored on language models with different smoothing techniques.

This paper is structured as follows. The introduction to Rakhine language is described in Section II. A speech corpus and pronunciation lexicon developed for Rakhine language are explained in Section III. The experimental setup is depicted in Section IV. The evaluation result is discussed in Section V. Conclusion and future work are summarized in Section VI.

II. The Nature of Rakhine Languag

This section presents the nature of the Rakhine language. Rakhine (Arakanese) is the native language of Rakhine (Arakanese), and there are 7 ethnic groups: Rakhine, Mjou, Marama, Dainne, Thet, Khame, and Kaman. They also have their own spoken languages, which are second languages. Rakhine is mainly divided into two dialects: North (Sittwe) and South (Thandwe). They also have different vocabularies and pronunciations.

Rakhine writing scripts are the same as in Myanmar. It is written from left to right without any spaces between words or syllables. In the Rakhine language, words are formed by combining basic characters with extended characters. Rakhine syllables can stand for one or more extended characters by combining consonants to form compound words. Moreover, it is similar to the Myanmar language, and it has 33 basic consonants, 44 vowels, and 4 medials in the Rakhine language. However, some vowel phonemes have different pronunciations in Myanmar, and some vowels are not used in Rakhine. There are 23 phonemes for 33 consonant scripts. Some scripts share the same pronunciation.

III. A Speech Corupus and Pronunciation Lexicon for Rakhine ASR

In this work, a Rakhine speech corpus and pronunciation lexicon for ASR development were used [1], and the corpus consists of daily conversations and broadcast news. For daily conversations, the sentences are collected from Rakhine language guidance book. It contains Rakhine digits and daily conversations. All utterances are reading styles, and speech is recorded with the help of the Tascam DR-100MKIII. The sample sentences of daily conversations are shown in Figure 1.

ucsy-record-hninthidarkyaw_04563 ထမင်း ကို အအီးစစစက် မ စား ပါကေ့
ucsy-record-hninthidarkyaw_04564 ထမင်း က ပပူလလောင် စား လို့ မ ရ သိ ပါ

Figure 1: Samples Sentences of Daily Conversations

The broadcast news was collected from Arakan Princess Media (APM), spoken in Rakhine. It includes both local news and foreign news. Rakhine is mainly divided into two dialects, and therefore, broadcast news speakers use two dialects (North and South). The speech utterances are manually transcribed into text. The example sentences of the broadcast news are shown in Figure 2.

ucsy-apm-sheinhtwenjou_21375 ဘင်္ဂလာ ဘက် မှာ မြီး ကတ် တေပိုင် ကန် ထဲ ရိုထုတ် ရိသွင်း စနစ် နှန့် မြီး စော် ပါ
ucsy-apm-sheinhtwenjou_21376 အယင်က ပင်လယ် မြစ်ချောင်း တိ မှာ သဘာဝ သားပေါက် တိ ပေါ လို့ အောင်မြင် ခ ကတ် တေ

Figure 2: Sample Sentences of Broadcast News

The statistics of the speech corpus are shown in Table 1. The speakers consist of the north (N) and south (S) regions of Rakhine State. The total vocabulary of the Rakhine lexicon contains 7,184 words. In the training dataset, there are 68 phonetic units.

Table 1: Statistics of Rakhine Speech Corpus

Data	Size(hh:mm:ss)	Speakers					Utterances
		Female		Male		Total	
		N	S	N	S		
APM Broadcast	03:04:03	3	4	2	1	10	1,439
Daily Conversation	03:16	3	-	-	1	4	6,848
Total	06:20:03	7	4	2	1	14	8,287

IV. Experimental Setup

In this section, the experimental setup for training, test data, language model, and acoustic model training are described. The training data and test data are described in Table 2. The training data consists of 7,723 utterances, and it takes about 6 hours. The TestSet1 (conversational data) contains 418 utterances spoken by one speaker. The TestSet2 (broadcast news) has 146 utterances and consists of 5 speakers.

Table 2: Training Data and Test Data Statistics

Data	Size (hh:mm:ss)	Speakers					Utterances
		Female		Male		Total	
		N	S	N	S		
Train Set	05:46:04	6	4	2	2	14	7,723
Test Set 1	00:15:00	1	-	-	-	-	418
Test Set 2	00:15:11	2	1	1	1	5	146

A. GMM and SGMM Acoustic Model

Kaldi open source toolkit[4] is applied for building Rakhine ASR development. The experimental setup for the GMM training was the same as the previous research [1]. It has 2050 context-dependent (CD) triphone states with the average of 14 Gaussian components per state. SGMM is built on top of Linear Discriminant Analysis (LDA)+ Maximum Likelihood Linear Transform(MLLT)+Speaker Adaptive Training (SAT) features. For the SGMM experiment, the phonetic subspace S=40 dimension and the diagonal Gaussians that were produced from the HMM are clustered to I=400 in order to initialize the Universal Background Model (UBM). The 40-dimensional LDA features were used to train the SGMM.

For language model building, SRILM toolkit [5] is applied, and 3-gram language model is used in this work. Moreover, experiments are performed with five different smoothing techniques (good-turing discounting, absolute discounting, kneser-ney, witten-bell, and natural discounting).

V. Evaluation Result

Rakhine ASR performance is evaluated by using Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) and Subspace Gaussian Mixture Model (SGMM) acoustic models with different discounting techniques. The results of the experiments are shown in Table 3.

Table 3: Evaluation of Rakhine ASR performance in terms of WER

Model	Discounting Techniques	WER%	
		TestSet1	TestSet2
GMM-HMM	**Good Turing**	**20.56**	**17.77**
	Absolute	21.58	20.25
	Kneser-Ney	19.94	15.81
	Witten-Bell	19.24	14.31
	Natural	**19.00**	**14.01**
SGMM	Natural	**16.07**	**12.40**

From the experiment, using GMM-HMM [1], it achieved a word error rate (WER) of 20.56% on TestSet1 and 17.77% on TestSet2 with default good-turing discounting. Moreover, the effect of language models with different smoothing methods is explored. According to the

experimental results, the best WER is obtained with the natural discounting technique.

The SGMM model gains an absolute word error rate reduction of 2.93% on TestSet1 or 1.61% on TestSet2 over GMM by using the best natural discounting technique. Therefore, it is observed that SGMM outperformed GMM on both test sets. As a result, the SGMM model leads to the lowest WERs of 16.07% on TestSet1 and 12.40% on TestSet2. Although the training data set is small, the ASR performance for the Rakhine language gets a promising result using SGMM. When the evaluation result of Testset1 (conversational data) is compared to Testset2 (broadcast news), Testset2 obtains a lower word error rate than TestSet1 as broadcast news has a clear voice and less noise than recorded data.

VI. Conclusion

In this paper, the automatic speech recognition for Rakhine language was improved by using SGMM and language model smoothing techniques. The ASR performance was evaluated on about 6 hours of Rakhine speech corpus, and it achieved the best WER of 16.07% on recorded conversational data and 12.40% on broadcast news. It is observed that smoothing techniques have a significant impact on reducing WERs for Rakhine language. Moreover, it can be concluded that SGMM significantly improved the Rakhine ASR performance compared to GMM in cases where the amount of available training data is limited.

In the future, the quality of Rakhine ASR will be promoted with state-of-the-art end-to-end deep learning approaches.

References

[1] H.D.Kyaw, A.N.Mon, “Automatic Speech Recognition for Rakhine Language”, National Journal of Parallel and Soft Computing, Volumn 03, Issue 02, December 2022

[2] A.N.Mon, W.P.Pa and Y.K.Thu, “Building HMM-SGMM Continuous Automatic Speech Recognition on Myanmar Web News”, In Proceedings of the 15th International Conference on Computer Applications (ICCA 2017),Yangon, Myanmar, pp. 446-453, February 16-17, 2017.

[3] D. Povey, Lukas Burget, et al., "The Subspace Gaussian Mixture Model-A Structured Model for Speech Recognition", Computer Speech and Language, vol. 25, Issue 2, pp:404-439, 2011.

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

[5] A.Stolcke, “Srilm - An Extensible Language Modeling Toolkit,” 2002, pp. 901--904.