2017 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique(O-COCOSDA)
1-3 November 2017, Seoul, Korea

DEVELOPING A SPEECH CORPUS FROM WEB NEWS FOR MYANMAR (BURMESE) LANGUAGE

Aye Nyein Mon[‡], *Win Pa Pa*[‡], *Ye Kyaw Thu*[Λ] , *Yoshinori Sagisaka*[†]

[‡]Natural Language Processing Lab., University of Computer Studies, Yangon (UCSY), Myanmar
[Λ]Artificial Intelligence Lab., Okayama Prefectural University (OpU), Japan
[†]Language and Speech Science Research Lab., Waseda University, Japan
ayenyeinmon, winpapa@ucsy.edu.mm, ye@c.oka-pu.ac.jp, ysagisaka@gmail.com

ABSTRACT

Speech corpus is important for statistical model based automatic speech recognition and it reflects the performance of a speech recognizer. Although most of the speech corpora for resource-riched languages such as English are widely available and it can be used easily, there is no Myanmar speech corpus which is freely available for automatic speech recognition (ASR) research since Myanmar is a low resource language. This paper presents the design and development of Myanmar speech corpus for the news domain to be applied to convolutional neural network (CNN)-based Myanmar continuous speech recognition research. The speech corpus consists of 20 hours read speech data collected from online web news and there are 178 speakers (126 females and 52 males). Our speech corpus is evaluated on two test sets: TestSet1 (web data) and TestSet2 (news recording with 10 natives). Using CNN-based model, word error rate (WER) achieves 24.73% on TestSet1 and 22.95% on TestSet2.

Index Terms--- Speech corpora, automatic speech recognition (ASR), Myanmar, convolutional neural network (CNN)

1. INTRODUCTION

Automatic speech recognition (ASR) system converts the speech signal into words. The recognized words can be the final output or the input to natural language processing (NLP). ASR is a very challenging task because speakers may have different accents, dialects, or pronunciations, and speak in different styles, at different rates, and in different emotional states. Many researchers have carried out research on automatic speech recognition and most of the progress in the field was done for English language. Modern ASR system uses statistical models based on speech data. The statistical-based ASR systems highly rely on corpora of speech data. Therefore, speech corpora are necessary for analysis, modeling, training, and evaluation.

Speech corpus is a large collection of audio recordings of spoken language and it also has text files containing transcriptions of the words spoken. Speech corpora can be divided into two types: Read speech and Spontaneous speech. For example, excerpts from books, news broadcasts, word lists, number sequences, etc., are included in read speech. Spontaneous speech type contains dialogs and meetings, narratives, etc.

Speech corpora creation is essential for developing any automatic recognition. In order to build ASR systems, a large amount of speech data are needed for training. Furthermore, ASR performance is also depended on the speech data. Speech corpora have been created for many resource-riched languages. For example, in English, such resources are popular and widely available—examples are the TIMIT corpus or the Switchboard corpus. It has abundant of collected speech data. For low resource languages including most of Non-Latin languages, it has to build speech data sets from the scratch since most of them do not have pre-created speech corpora [1].

It has been shown that there are attempts in developing the speech corpus for low resource languages. For example, AGH corpus of Polish speech was created by Piotr Zelasko, et.al, [1] in 2016. Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language was built by Tilda Neuberger, et.al, [2]. Speech corpus for Bengali language was developed by Sandipan Mandal, et.al, [3]. Building the speech corpus for Bulgarian language was done by Neli Hateva, et.al, [4]. Spontaneous speech corpus for European Portuguese was created by Tiago Freitas, et.al, [5].

This work aims to present building the speech corpus for Myanmar language and evaluation on the created corpus using state-of-the-art acoustic model, convolutional neural network (CNN). If the speech data is recorded by ourselves, the professional recording devices are very costly and it is time-consuming. Today, a lot of speech data are highly available on the Internet and therefore,

978-1-5386-3333-5/17/$31.00

the speech data is developed from the Web in building the speech corpus. The data is collected from websites of broadcast news. The speech database is evaluated by applying in CNN-based Myanmar continuous speech recognition research.

The paper is organized as follows. In Section 2, about Myanmar language is described. Building speech and text corpus for Myanmar ASR is presented in Section 3. In Section 4, statistics of the speech corpus is shown. Experimental Setup for data sets and acoustic models are explained in Section 5. Evaluation on the speech corpus is done in Section 6. In Section 7, conclusion and future work are summarized.

2. MYANMAR LANGUAGE

Myanmar language, which is formerly known as Burmese, is a tonal, syllable-timed language and largely monosyllabic, and analytic language. It is the official language of Myanmar. It has subject-object-verb word order with 9 parts of speech. There are four nominal tones transcribed in written Myanmar: low, high, creaky and checked. The different tones carry in different meanings. The Myanmar tones with their different meanings are described in Table 1.

Table 1. Different types of Myanmar tones

Tone	Myanmar Word	Description
Low	ကာ	k a [shield]
High	ကား	k a: [car]
Creaky	က	k a. [dance]
Checked	ကပ်	k a'[disaster]

There are basics 12 vowels, 33 consonants, and 4 medials in Myanmar language. A word is formed by one or more syllables which are composed of an initial component followed by zero or more medials, zero or more vowels with an associated tone. Myanmar words are divided into simple words, and compound words or loan words. A simple word is regarded as a syllable, a compound word as a combination of several simple words, and loan words as transliterations mainly used for foreign words. A syllable is a basic unit of Myanmar language. Myanmar syllables are basically constructed by the combination of consonant and vowel. For example, the combination of ◌ော vowel and က consonant makes one syllable as က (k a.) + ◌ော (ou) = ကော (k ou) [6].

3. BUILDING SPEECH AND TEXT CORPUS FOR MYANMAR ASR

This section explains how to build the speech and text corpora for news domain.

3.1. Collecting Data from the Online Resources

A speech corpus can be created mainly in two ways. One way is to collect existing speech data (speech that is already been recorded) and manually transcribe them into text. The second way is to design the text corpus first and record the speech by reading the collected text. The first approach is used to build our speech corpus for news domain.

Nowadays, the Internet is a source that seems to be almost unlimited in size, and it also offers a wide variety of resource types: social media such as Facebook or Twitter offer videos files and short, colloquial texts, while blogs and news portals might offer more formal and longer texts news, and audio files. Moreover, they are freely available on the Internet and can be downloaded easily. So, our speech corpus is developed by collecting the data from the Internet [1].

There are many websites that Myanmar news is available and for our corpus building, the speech data is collected from the site of Myanmar Radio and Television (MRTV)[1]. In addition, the speech data also gather from social media, Facebook, of Eleven broadcasting[2], 7days TV[3], and ForInfo news[4].

Our speech corpus includes both local and foreign news. They are about politics, health, speech, crime, sports, weather, education, and business news.

3.2. Speech Converting and Segmentation

The format of the speech files from online is .FLV and .MP4. Hence, firstly, the audio files are converted to .WAV file format. Then, the speech wave files are set to single channel (mono) type and 16 kHz is used for sampling rate.

The long wave files are needed to segment into short length audio files. Therefore, the speech segmentation is done using Praat [7] tool. When segmenting the files, any portion that includes silence and background noise is discarded. The duration of the segmented speech wave files is between 2 sec and 30 sec.

3.3. Speaker Distribution

Most of the Myanmar broadcast news presenters are females and only a few males are available. Therefore, the distribution of speakers with regard to gender is that 126 females and only 52 males are found in this corpus. Moreover, the age of the speakers is under 35.

[1] http://www.mrtv.gov.mm/
[2] https://www.facebook.com/elevenbroadcasting/
[3] https://www.facebook.com/7DayOnlineTV/
[4] https://www.facebook.com/forinfo/

လွတ်တော် ရုံး များ ၏ သစ်ပင် စိုက်ပျိုး ပွဲ အခမ်းအနား ကို မနေ့ က လွတ်တော် ရုံးဝင်း အတွင်း ကျင်းပ ခဲ့ ပါတယ်
[Hluttaw offices held a ceremony to plant trees in the compound of the Hluttaw yesterday.]

ဧရာဝတီ စစ်တောင်း ဝဏ် တိုး မြစ် များ ၏ ရေမျက်နှာပြင် များ ၎င်း တို့ ၏ စိုးရိမ်ရေမှတ် အသီးသီး ထက် ကျော်လွန် နေ မှု သည် နောက် သုံး ရက် အတွင်း လျော့ကျ သွား နိုင် တယ် လို့ မိုးလေဝသ နဲ့ ဇလဗေဒ ဦးစီးဌာန မှ ခန့်မှန်း ထား ပါတယ်
[The water levels of Ayeyawady, Sittoung, Ngawun and Toe rivers which have exceeded their respective danger levels are expected to decrease in the next three days, according to the Meteorology and Hydrology Department.]

Fig. 1. Example Myanmar sentences from the corpus

3.4. Speech Utterance

Some of the broadcast news from online already has transcription but, some does not have. Therefore, we manually transcribe them into text as transcription of the speech if it is not available. Myanmar language needs to segment the text because there is no space between words when writing. So, the transcribed texts are segmented into words using [8]. In addition, the segmented texts are checked by hand again to get the correct segmentation. Finally, the spelling of the words is manually checked. There are 33 words and 54 syllables on average in one utterance. Bootstrapping technique is used to increase the text corpus size. Myanmar 3 Unicode font is used in building the text corpus. Example sentences from the text corpus are as shown in Figure 1.

4. STATISTICS OF CORPUS

The corpus has 20 hours speech data and 178 speakers (126 females and 52 males) with 7,332 utterances.

The detailed information of the corpus is shown in Table 2.

Table 2. Corpus data set size

Data	Size	Speakers			Utterance	UniqueWords
		Female	Male	Total		
TrainSet	20 hr	126	52	178	7,332	9,560

The percentage of the news collection from various news sites are presented in Table 3. Most of the broadcast news is the data from the websites of MRTV, Eleven, forInfoNews, and 7daysTV. Totally, there are about 96% gathered from them. Some of the broadcast news is from British Broadcasting Corporation (BBC) Burmese news[5] and Frequency Modulation (FM) radio. Moreover, Hluttaw speech data also includes in the corpus. The data collected from them are about 4%. The most type of news in the corpus is business, politics, and crime.

[5]https://www.facebook.com/bbcburmese/

Table 3. % of collection of the news from different sites

News Sites	% of Collection
MRTV	20.43
Eleven	23.31
ForInfoNews	24.82
7daysTV	27.73

Table 4 gives the basic statistics such as the number of words, and n-gram counts of the text corpus. 3-gram language model is used and it is trained by using SRILM [9] language modeling toolkit.

Table 4. Statistic on n-grams count

Attribute	Value
Number of words	232,603
Uni-grams	9,674
Bi-grams	72,439
Tri-grams	28,875

The perplexity of a language model is a measure which indicates how good the language model is. The results of the perplexity measures on both test sets using the trained language model are shown in Table 5.

Table 5. Perplexity measures of the language model for both TestSet1 and TestSet2

Attribute	Value of TestSet1	Value of TestSet2
Number of sentences	193	154
Number of words	6,741	4,815
OOV Words	422	181
Perplexity	119.58	91.85

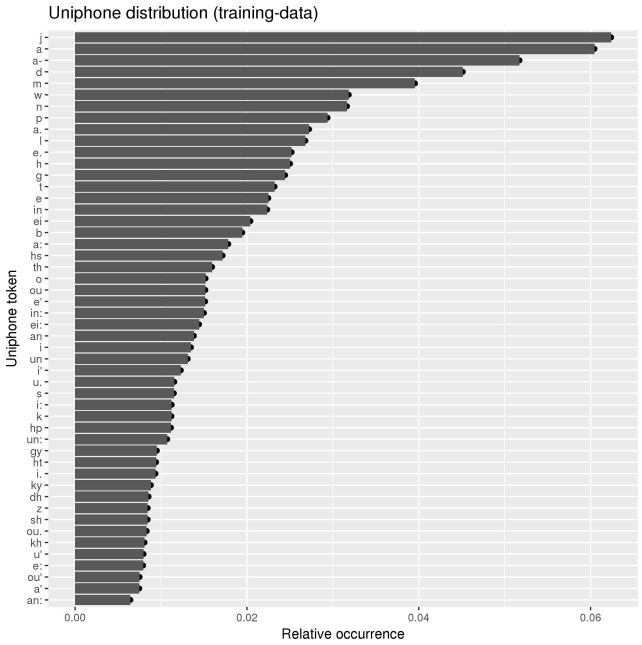

Fig. 2. The top 50 most frequent uniphones distribution

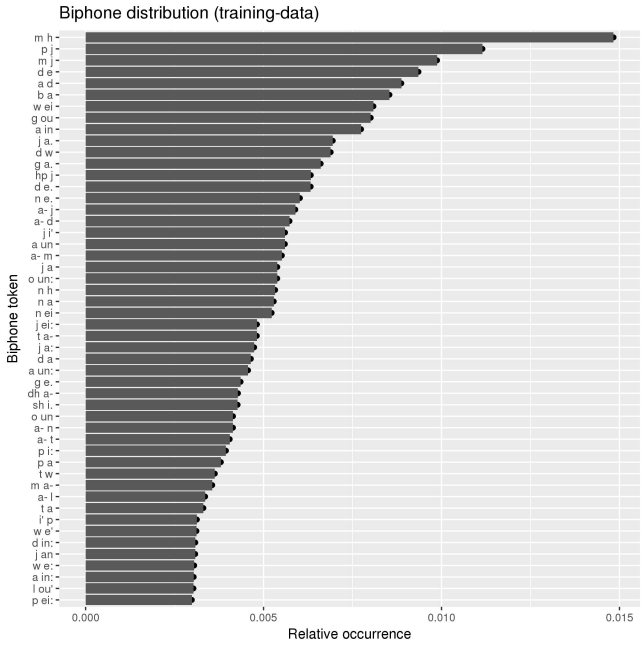

Fig. 3. The top 50 most frequent biphones distribution

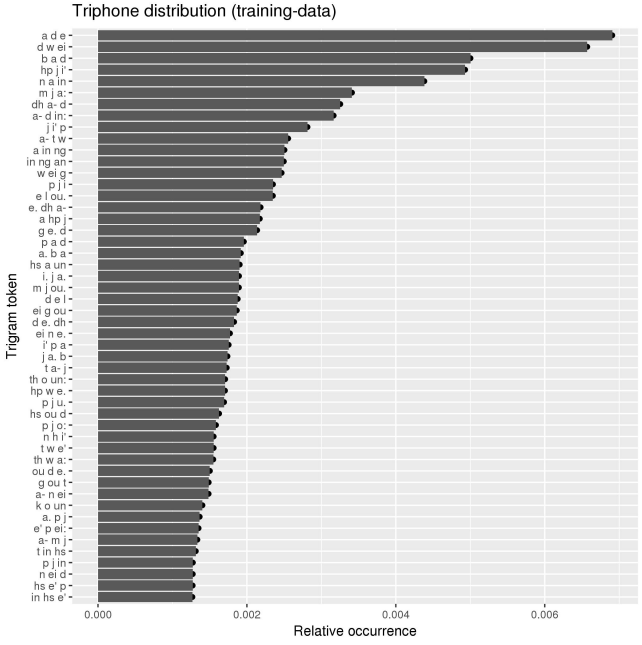

Fig. 4. The top 50 most frequent triphones distribution

4.1. Phones coverage in speech corpus

The number of monophones, diphones, and triphones of the training text is evaluated and their coverage is calculated based on Myanmar-English Dictionary which was published by Myanmar Language Commission (MLC). The MLC dictionary is a standard dictionary of Myanmar language. This dictionary consists of 26.6K unique words and 105 phonemes with tone information are identified [10]. In the training set, there are 65 syllables and it covers 89.71% of syllables. The number of diphones and triphones is 2,731 and 30,706 respectively. It covers 82.91% of diphones and 98.85% of triphones.

Figure 2 depicts the 50 most frequent uniphones occurrences based on the statistical analysis of the training corpus. Among the other types of uniphones, the phone 'j' is the most commonly found in news corpus. The total number of this phone is 53,460. The phone 'tr' is the lowest distribution of uniphones with 14 occurrences.

The top 50 most diphones appearings are shown in Figure 3. The most commonly distribution of diphone is the pair of 'Consonant+Consonant' [/m/+/h/] with 12,352 occurrences. The least distribution pairs are 'Vowels+Vowels' combination. Example pairs are /i/+/ou/, /o:/+ /o/, /e'/+/in:/, etc,. The count of these pairs is one in this corpus.

Figure 4 displays the top 50 most triphone occurrences. In triphone distribution case, the most distribution of triphone is the 'Consonant+Consonant+vowel' pair [/a/+/d/+/e/]. The total count of this pair is 5,585 and most of the combination of 'Vowels+Consonants+Vowels' is the smallest amount of triphone distribution. Those triphone pairs are /in./+/kh/+/an:/, /ei/+/w/+/ei/, /o./+/sh/+/o/, etc.

5. EXPERIMENTAL SETUP

In this section, the experimental setup for two test sets, and acoustic models are discussed.

5.1. Setup for Two Different Test Sets

For these experiments, two different test sets are used to evaluate the ASR performance on the created speech corpus.

5.1.1. TestSet1

For TestSet1, broadcast news data from web is used. The texts are randomly selected and there are 193 utterances in this set. It has 785 unique words. The number of females and males are 5 and 3, and the total is 8. The speakers from the test sets are different from those of training sets. The duration of this corpus is about 32 min.

5.1.2. TestSet2

Recorded data via mobile phone is used as TestSet2. Galaxy Grand Prime model of Samsung brand android phone was used in recording the voice from natives via built-in voice recorder and microphone. The age of the speakers is between 24 and 35. There are 154 utterances in this set. This includes about 30 min voice data from 5 males and 5 females. Each person records about 3 minutes. Most of the speakers are Bamar and some of the speakers are Kayin and Rakhine races of Myanmar major national ethnic races. All recorded data is set to single channel (mono) type and 16 kHz is used as sampling rate. The detailed information is shown in Table 6.

Table 6. Test set size

Data	Size	Speakers			Utterance	UniqueWords
		Female	Male	Total		
TestSet1	31 min 55 sec	5	3	8	193	785
TestSet2	30 min 7 sec	5	5	10	154	1,065

5.2. Acoustic model

All experiments are done using Kaldi [11] speech recognition toolkit. In this experiment, three different acoustic models (Gaussian Mixture Model (GMM), Deep Neural Network (DNN), and CNN) are developed.

Baseline GMM-HMM: Input features for GMM are Mel Frequency Cepstral Coefficients (MFCC) with delta and delta features after applying linear discriminated analysis (LDA) and maximum likelihood linear transform (MLLT). Moreover, speaker adaptive training is performed on top of the LDA+MLLT model. There is an average of 44 Gaussian components per state.

DNN-based acoustic model: There are 4 layers with 300 units per hidden layers. Input features for the DNN consist of 40-dimensional log mel-filter bank features. The DNN is trained without using pre-training.

CNN-based acoustic model: Input features of CNN are the same with DNN input features. It has two convolution (conv) layers, one pooling (pool) layer and two fully connected (fc) layers with 300 hidden units. The filter size is 8×4 and no pre-training is done. Max pooling is used and pooling size is 2 with pool step 1. There are 128 and 256 feature maps in the first and second convolutional layers.

6. EVALUATION ON THE SPEECH CORPUS

In this section, evaluation on the speech corpus is performed using three different models: GMM, DNN, and CNN. Word error rate (WER) is used as evaluation criteria. From the Table 7, it can be clearly seen that using

Table 7. Evaluation on the speech corpus with different models and test sets

Models	WER%	
	TestSet1	**TestSet2**
GMM-HMM	29.03	28.66
DNN	27.08	24.76
CNN	**24.73**	**22.95**

the best acoustic model, CNN, the performance on both TestSet1 and TestSet2 has shown the slight difference although they are not the same nature, data from web and data recorded by ourselves. Therefore, it can be said that the model trained from the web data makes the accuracy on both test sets effective. In comparison with the baseline model, GMM, DNN-based model achieves better performance of 1.95% on TestSet1 and 3.90% on TestSet2 over the GMM. CNN gains absolute 4.30% and 2.35% on TestSet1, and 5.71% and 1.81% on TestSet2 over the GMM and DNN. The lowest WER of ASR performance, 24.73% on TestSet1 and 22.95% on TestSet2, are obtained by using CNN.

7. CONCLUSION

This paper has designed and developed news speech corpus for Myanmar ASR. Moreover, this corpus was evaluated by using state-of-the-art acoustic modeling approach, CNN. Furthermore, two different test sets are applied to evaluate our training model. One is the web

data and it is the same nature of the training data. The other is the different nature of the training data and it is the voices that are recorded via mobile phone by reading broadcast news text. With CNN-based model, it is found that the accuracy on both test sets achieves satisfactory results and it showed that the corpus built on Internet resources gives promising results.

Currently, in our corpus, there are 7,332 utterances in 20 hours length. The corpus will be growing with recorded speech other than news domain.

Acknowledgement

We would like to thank Ms. Myat Aye Aye Aung, University of Computer Studies, Yangon, participating in speech corpus creation.

8. REFERENCES

[1] Piotr Zelasko, Bartosz Ziólko, Tomasz Jadczyk, and Dawid Skurzok, "AGH Corpus of Polish Speech," *Language Resources and Evaluation*, vol. 50, no. 3, pp. 585--601, 2016.

[2] Tilda Neuberger, Dorottya Gyarmathy, Tekla Etelka Gráczi, Viktória Horváth, Mária Gósy, and András 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, September 8-12, 2014. Proceedings*, 2014, pp. 424--431.

[3] Sandipan Mandal, Biswajit Das, Pabitra Mitra, and Anupam Basu, "Developing Bengali Speech Corpus for Phone Recognizer Using Optimum Text Selection Technique," in *International Conference on Asian Language Processing, IALP 2011, Penang, Malaysia, 15-17 November, 2011*, 2011, pp. 268--271.

[4] Neli Hateva, Petar Mitankin, and Stoyan Mihov, "Bulphonc: Bulgarian Speech Corpus for the Development of ASR Technology," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 2016.

[5] Fabíola Santos and Tiago Freitas, "CORP-ORAL: Spontaneous Speech Corpus for European Portuguese," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.

[6] U Thein Htun, "Some Acoustic Properties of Tones in Burmese," in *South-East Asian Linguistics 8: Tonation, D. Bradley, Ed. (Australian National University, Canberra, 1982)*, p. 77–116.

[7] "Praat:doing phonetics by www.fon.hum.uva.nl/praat," accessed June 2010.

[8] Win Pa Pa, Ye Kyaw Thu, Andrew M. Finch, and Eiichiro Sumita, "Word Boundary Identification for Myanmar Text Using Conditional Random Fields," in *Genetic and Evolutionary Computing - Proceedings of the Ninth International Conference on Genetic and Evolutionary Computing, ICGEC 2015, August 26-28, 2015, Yangon, Myanmar - Volume II*, 2015, pp. 447--456.

[9] Andreas Stolcke, "Srilm - An Extensible Language Modeling Toolkit," 2002, pp. 901--904.

[10] "Myanmar-English Dictionary," Department of the Myanmar Language Commission, Yangon, Ministry of Education, Myanmar, 1993.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society.