International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 4, August 2019, pp. 3194~ 3202
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i4.pp3194-3202

UCSY-SC1: A Myanmar Speech Corpus for Automatic Speech Recognition

Aye Nyein Mon[1], Win Pa Pa[2], Ye Kyaw Thu[3]
[1,2]Natural Language Processing Lab, University of Computer Studies, Yangon
[3]Language and Speech Science Research Lab., Waseda University, Japan

Article Info

Article history:

Received Dec 18, 2018
Revised Feb 15, 2019
Accepted Mar 21, 2019

Keywords:

Automatic Speech Recognition
Myanmar Language
Speech Corpus
Convolutional Neural Network (CNN)

ABSTRACT

This paper introduces a speech corpus which is developed for Myanmar Automatic Speech Recognition (ASR) research. Automatic Speech Recognition (ASR) research has been conducted by the researchers around the world to improve their language technologies. Speech corpora are important in developing the ASR and the creation of the corpora is necessary especially for low-resourced languages. Myanmar language can be regarded as a low-resourced language because of lack of pre-created resources for speech processing research. In this work, 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. The total size of the speech corpus is over 42 hrs. There are 25 hrs of web news and 17 hrs of conversational recorded data. The corpus was collected from 177 females and 84 males for the news data and 42 females and 4 males for conversational domain. This corpus was used as training data for developing Myanmar ASR. Three different types of acoustic models such as Gaussian Mixture Model (GMM) - Hidden Markov Model (HMM), Deep Neural Network (DNN), and Convolutional Neural Network (CNN) models were built and compared their results. Experiments were conducted on different data sizes and evaluation is done by two test sets: TestSet1, web news and TestSet2, recorded conversational data. It showed that the performance of Myanmar ASRs using this corpus gave satisfiable results on both test sets. The Myanmar ASR using this corpus leading to word error rates of 15.61% on TestSet1 and 24.43% on TestSet2.

Corresponding Author:

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

1. INTRODUCTION

Speech is the most natural form of communication among humans. Numerous spoken languages are employed throughout the world. As communication among human beings is mostly done vocally, it is natural for people to expect speech interfaces with the computer. Automatic speech recognition (ASR) means the conversion of spoken words into computer text. A lot of automatic speech recognition research is currently being conducted by the researchers around the world for their languages [1] [2]. Current ASR system use statistical models constructed on speech data. Therefore, speech corpus is important for statistical model based automatic speech recognition and it affects the performance of a speech recognizer. For well-resourced languages, speech researchers have used publicly available resources from online. However, for low-resourced languages, they have to build

the corpora by themselves for developing the ASR systems [3] [4]. They used online resources such as broadcast news, daily conversational data, etc., or the data are recorded by themselves for developing the ASR.

Myanmar language can be considered as a low-resourced language regarding the linguistic resources available for NLP. Lack of proper data is the main problem when it comes to speech recognition research in the Myanmar Language. Therefore, speech corpus is needed to build for developing Myanmar ASR. This gives impetus to build the speech corpus for Myanmar language. There was our work done in developing a speech corpus for Myanmar language [5]. In the work, speech data was collected only from specific domain, web news. The total length of the speech corpus is 20 hrs. It involves 126 females and 52 males, totalling 178 speakers. There are 7,332 utterances in the corpus, which are collected from local and foreign news. The corpus was used in developing Myanmar continuous speech recognition. It was evaluated on two test sets: web data and news recorded by 10 natives. It yielded 24.73% WER on TestSet1 and 22.59% WER on TestSet2.

In this task, the domain of the speech corpus is extended. The speech corpus is named as "UCSY-SC1" and it is constructed by using two types of domain: web news and daily conversations. The web news data size is increased to 25 hrs.The web news is the already recorded data collected from the web. Its sentences are longer than the conversation sentences. This corpus consists of daily conversational data. The shorter daily conversation sentences are obtained from ASEAN language speech translation thru' U-Star[1] and the web. They are recorded by ourselves using the recording device. They are very short sentences. There are 17 hrs of conversational data. Thus, the total speech corpus size is over 42 hrs. This corpus is used as training data and the experimental results of GMM, DNN and CNN with sequence discriminative training approaches are presented. This is a milestone for Myanmar ASR development.

This paper is organized as follows. The introduction to Myanmar language is presented in Section 2. A speech corpus developed for Myanmar language is explained in Section 3. Evaluation on the corpus is done in Section 4. Conclusion and future work are summarized in Section 5.

2. NATURE OF MYANMAR LANGUAGE

Myanmar Language (formerly known as Burmese) is the official language of Myanmar. The Myanmar script derives from Brahmi script of South India. Myanmar text is a string of characters without any word boundary markup. There are no spaces between words in Myanmar language. Myanmar language has 33 basic consonants, 12 vowels, and 4 medials. Phonology is a system of the combination of vowels and consonants. Myanmar phonology is structured by just one vowel, or one vowel and consonant, consonant combination symbols. The vowels have their own sounds in Myanmar language. Therefore, just only one vowel can produce clear sound such as အ /a̱/၊ အာ /à/၊ အား /á/. Myanmar consonants have no clear own sound and if it is combined with a vowel, it can produce the clear sound. Example is 'က' /k/ + 'အား' /á/ = 'ကား' /ká/. The 12 basic vowels are အ /a̱/, အာ /à/, အိ /i̱/, အီ /ì/, အု /u̱/, အူ /ù/, ဧအ /èi/, အဲ /é/ ဧအာ /ɔ́/, ဧအော် /ɔ̀/, အံ /àɴ/, အို် /ò/. Myanmar syllables are basically formed by the consonant and vowel combination [6]. As an example, consider the combination of 'အူ' /ù/ vowel and 'က' /k/ consonant makes one syllable 'ကူ' /kù/ as 'က' /k/ + 'အူ' /ù/ = 'ကူ' /kù/.

3. UCSY-SC1 SPEECH CORPUS BUILDING

Building a speech corpus is the first step for developing any automatic speech recognition (ASR) system, especially for low-resourced languages, and it is crucial for the statistical ASR system. Moreover, the accuracy of a speech recognizer depends on the speech corpora. Speech corpora for well-resourced languages such as English are publicly available for ASR research. However, being a low-resourced language, Myanmar language has no existing speech corpora. A speech corpus can be built mainly in two methods. The first method is to gather 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 [7].

[1] http://www.ustar-consortium.com/qws/slot/u50227/index.html

3.1. Collecting the Data from the Web News

The first approach is used to collect the web news data. Today, the internet has various resource types, for example, social media, blogs, twitter, and new portals, which offer a lot of speech data and which can be freely downloaded. Moreover, it has been proved that the corpora created on internet resources yielded promising results [8] [9]. Therefore, speech data was collected first from the web news. The duration of the web data collecting process lasted one year and it involved two persons including the author.

3.1.1. Speech Corpus Preparation

The web news is downloaded from the sites of Myanmar Radio and Television (MRTV),Voice of America (VOA), facebook pages of Eleven broadcasting, 7days TV, ForInfo news, GoodMorningMyanmar, British Broadcasting Corporation (BBC) Burmese news and breakfast news. Both local and foreign news are contained in the corpus. The web news videos are converted to wave file format. After that, the audio files are segmented with Praat[2]. All the audio files are formatted with sample frequency 16,000 Hz and mono channel. The length of each audio file is between 2 seconds and 30 seconds.

3.1.2. Speaker Information

The news presenters are professional, well-experienced and well-trained. Therefore, they have clear voice in news broadcasting. Female news presenters are dominant in the web news. Hence, in this corpus, fewer male speakers are involved than females. The ages of the speakers are under 35.

3.1.3. Text Corpus Preparation

Most of the broadcast news items from the web have transcriptions. However, the transcriptions are manually done if they are not available and Myanmar3 Unicode is used for that purpose. Word segmentation is done by hand as Myanmar language has no word boundary. This is performed based on Myanmar-English dictionary [10] and this dictionary is also applied to check the spelling of the words. The average lengths of the utterances in this corpus are 33 in words and 54 in syllables. Web news data has 8,973 unique sentences and 11,040 unique words. The example news sentences from the corpus are shown in Figure 1. The format of each sentence is the utterance-id followed by the transcription of each sentence.

ucsy-mrtv-aungmyothu_1000 ဒီ သတင်း ကို တော့ ပုသိမ် သတင်း ဌာနခွဲ မှ ပေးပို့ ထား တာ ဖြစ် ပါတယ် ရှင်
[This news is sent from the Pathein news station.]

ucsy-eleven-chosetpaing_2002 သမ္မတ ရဲ့ ပထမဆုံး မိနှစ်ခွန်း ဟာ ခုနစ် မိနစ် နီးပါး ကြာမြင့် ခဲ့ ပါတယ်
[The President's first speech lasted nearly seven minutes.]

Figure 1. Example sentences of the corpus on news

3.2. Recording Daily Conversations

The second approach (designing the text corpus first and recording the speech by reading the collected text) was used for collecting the conversational data. It took 3 months for data recording and 11 people were involved in the speech and text segmentation.

3.2.1. Text Corpus Preparation

The daily English conversations from ASEAN language speech translation thru' U-Star are translated into Myanmar for text corpus building. The conversational data contains 2,156 unique sentences and 1,740 unique words. There are 2,000 sentences in the ASEAN language speech translation and they are the conversations in hotels, restaurants, streets, telephones, etc. The rest 156 daily conversational sentences are collected from the web. The spelling of the text is manually checked and the words are segmented as the news data. The sentences contained in the corpus are shorter than those of the news domain. The average lengths of the conversational sentences are 11 in words and 15 in syllables. The example sentences for the daily conversational domain are shown in Figure 2.

[2]http://www.fon.hum.uva.nl/praat/

The format of each sentence is similar to that of the news domain (utterance id followed by each utterance).

ucsy-record-ayechanmay_16201 အဝတ် တွေ က ဈေးပေါ် တယ် လို့ ထင် တယ်
[I think the clothes are cheap.]
ucsy-record-ayechanmay_16202 သူငယ်ချင်း တွေ ရောက် လာ ချိန် မှ စားသောက်ဆိုင် က ထွက် တော့ မယ်
[When the friends arrive, I am going to leave from the restaurant.]
ucsy-record-ayechanmay_16203 ကျေးဇူး ပြု၍ အကူအညီ လို ရင် ပြော ပါ
[Please tell me if you need help.]

Figure 2. Example sentences of the conversational data

3.2.2. Speaker Information

The sentences are recorded by 4 male speakers and 42 female speakers, who are the faculty members and students of the University of Computer Studies, Yangon, Myanmar. Since the number of females exceeds that of males in our university, many female speakers are represented in the corpus. The ages of the speakers are between 19 and 40.

3.2.3. Speech Recording and Segmentation

The recording work was done in a laboratory of our university. It is a very quiet place with no external effects from the room like echo and background noises. It is also a healthy place to work in because people can breathe well and feel relaxed. Tascam DR-100MKIII[3] was used for speech recording. It is intended to be used for audio designers and engineers and it has an easy-to-use interface with robust reliability. The audio files are formatted with sample frequency 16,000 Hz and mono channel with 16 bits encoding. The recorded files are segmented with the audacity tool[4]. Moreover, the silent portion of each utterance is discarded. In a speech corpus, audio and text data should be aligned. So each recorded sentence is listened to and checked with their corresponding text transcription and made necessary corrections. If the speakers do not have clear voices, the recordings are done repeatedly until they are satisfactory and smooth. All speakers read at normal pace.

3.2.4. Normalization to Transcription

Some of the transcriptions of broadcast news and daily conversions obtained from online consists of non-standard words. They are numbers, dates, abbreviations acronyms, symbols, and English names such as names of organization, things, persons, animals, social media, etc. The pronunciations of these words cannot be found in the dictionary. Therefore, it is necessary to do text normalization and transliteration into Myanmar language. In this work, those words are manually transcribed into Myanmar words as the transcribers listen to their corresponding audios. Table 1 shows the example words that need to be normalized.

Table 1. Example of text normalization

Description	Example	Normalization
Date	**၂၀၁၆-၂၀၁၇** (2016-2017)	နှစ် ထောင့် ဆယ့် ခြောက် နှစ် ထောင့် ဆယ့် ခုနစ်
Time	**၃ နာရီ ၅၅ မိနစ်** (3 Hours 55 Minutes)	သုံး နာရီ ငါး ဆယ့် ငါး မိနစ်
Number	**၁၁၄ ဦး** (114 persons)	တစ် ရာ တစ် ဆယ့် လေး ဦး
Digit	**09-448045577**	သုည ကိုး လေး လေး ရှစ် သုည လေး ငါး ငါး ခုနစ် ခုနစ်
Acronyms	**FDA**	အက်ဖ် ဒီ အေ
Person Name	**Mr. Filippno Grandi**	မစ္စတာ ဖီလစ်နို ဂရမ်းဒီ

[3]https://tascam.com/us/product/dr-100mkiii/top
[4]https://www.audacityteam.org/

3.3. Phone Coverage in the Speech Corpus

Phone coverage is vital for improving the ASR accuracy. Myanmar-English dictionary developed by Myanmar Language Commission (MLC) [10] is used as the baseline and this dictionary is extended with the vocabularies of the speech corpus. There are 38,376 words in the lexicon. In the training set, there are 67 phonemes and it covers 94.37% of phonemes. Table 2 describes an example of Myanmar lexicon.

Table 2. Example of Myanmar lexicon

Myanmar Word	Phoneme
အ (Dump)	/a̱/
အားကစား (Sport)	/á gəzá/
အာကာသ (Space)	/à kà θa̱/

The distributions of phonemes for both consonant and vowel phonemes occurring in the speech corpus are analyzed. The frequency data on consonant distribution of the corpus are given in Figure 3.

The phoneme /j/ has the most occurrences in the corpus. This is because the phoneme represents some medials such as '◌ျ'/ya̱ pi̱/, '◌ြ'/ya̱ yi?/ and the consonants 'ရ' and 'ယ' are defined as /j/ phoneme. The second most frequent occurrence is the phoneme /d/ because the consonants 'ဒ' and 'ဓ' are represented by the same phoneme /d/. The Myanmar word '◌ြ့' /tri̱/ rarely appears in Myanmar language. Therefore, the pronunciation phoneme of the word, /tr/ phoneme, is found only 1 time in the texts. A few nasal phonemes, /ng/ and /nj/, are found.

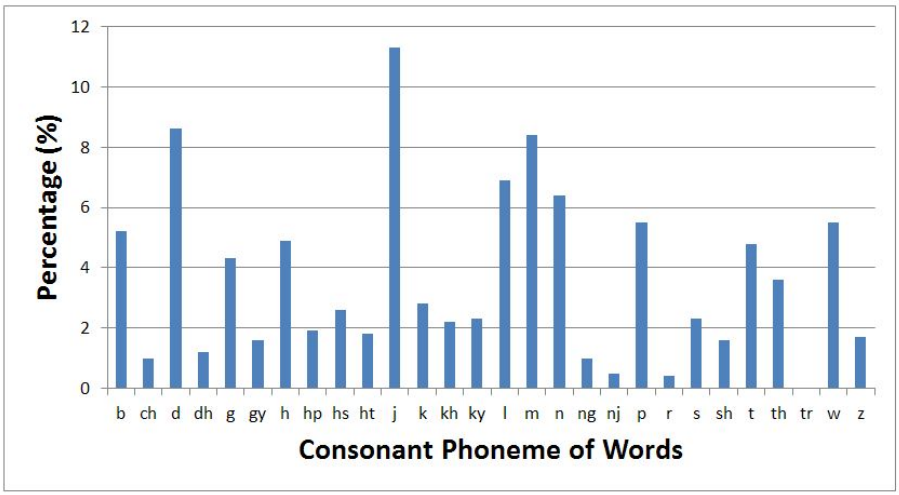

Figure 3. Consonant phonemes distribution of UCSY-SC1 corpus

The frequency of the vowel distribution of the corpus is shown in Figure 4. All vowel phonemes appear in the corpus. The most frequent phoneme is the phoneme /a/ with tone1 and most of the pronunciation of the words is formed with the vowel phoneme. For example, the words 'ခေါင်း'/káuɴ/ is composed of the phonemes of /k/ + /a/+/un:/ and 'ကိုင်း'/káɪɴ/ is formed by the combination of the phonemes /k/+/a/+/in:/.

The second most frequent phoneme is the /a-/ with neural tone. In Myanmar language, the basic vowels (/i/ /ì/, /ei/ /èi/, /e/ /è/, /a/ /à/, /o/ /ɔ̀/, /ou/ /ò/, /u/ /ù/) have their own properties. While these vowels are influenced by the contextual sounds, they change to neutralized vowels when their own properties decrease. Therefore, most of the Myanmar words are found with neutral tone in the corpus.

For example,

/ná/ + /jwɛʔ/ ==> /nə/ + /jwɛʔ/

Most of the nasalized vowels such as /ai'/ /aiʔ/, /an./ /a̱ɴ/, /ei'/ /eɪʔ/, /in./ /ɪ̱ɴ/, /u'/ʊ/ and /un./ /ũ/ are the least frequent phonemes in the corpus.

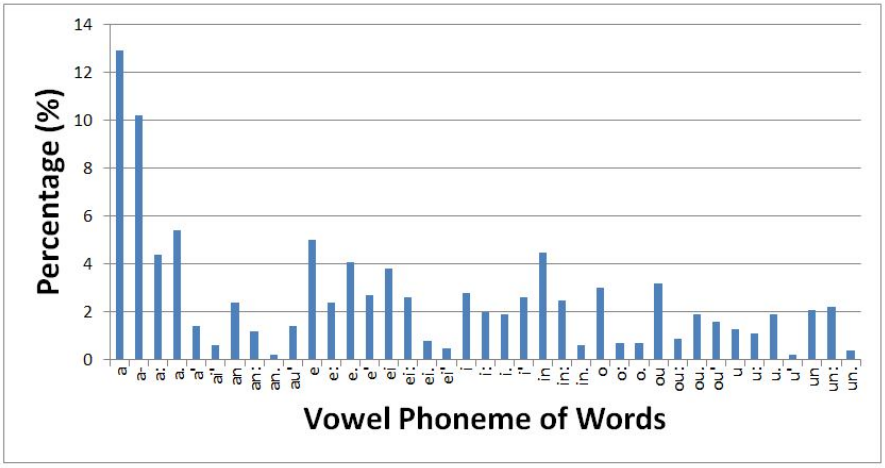

Figure 4. Vowel phonemes distribution of UCSY-SC1 corpus

3.4. Statistics of the Corpus

The speech corpus consists of two types of domain: web news and conversational data. The detailed statistics of the corpus is shown in Table 3. The corpus consists of 306,088 words. 11,696 words are unique and nearly 37% occurs only once. About 5% of unique words appear between 100 and 1,000 times. Only nearly 1% is found more than 1,000 times in the unique words.

Table 3. UCSY-SC1 corpus statistics

Data	Size	Speakers			Utterance	UniqueWord
		Female	Male	Total		
Web News	25 Hrs 20 Mins	177	84	261	9,066	9,956
Daily Conversations	17 Hrs 19 Mins	42	4	46	22,048	1,740
Total	42 Hrs 39 Mins	219	88	307	31,114	11,696

4. EVALUATION ON THE CORPUS

In this work, experiments are done to evaluate the quality of the speech corpus on Myanmar ASR.

4.1. Experimental Setup

The details of the experimental setup for data sets, acoustic and language models are dealt with in this section. The impact of training data sizes on the ASR performance is investigated in this experiment. Four different data sizes -10 hrs, 20 hrs, 30 hrs, and 42 hrs - are used for incremental training. The detailed statistics on the train and test sets are displayed at Table 4. TestSet1 is the open test data, which is web news data. TestSet2 is also open test data and it is the conversational data from natives recorded with voice recorders and microphones.

Table 4. Statistics on train and test sets

Data	Size	Speakers			Utterance
		Female	Male	Total	
TrainSet	10 Hrs 5 Mins	79	23	102	3,530
	20 Hrs 2 Mins	126	52	178	7,332
	30 Hrs 3 Mins	174	86	260	15,556
	42 Hrs 39 Mins	219	88	307	31,114
TestSet1	31 Mins 55 Sec	5	3	8	193
TestSet2	32 Mins 40 Sec	3	2	5	887

4.2. GMM-based Acoustic Model

Kaldi speech recognition toolkit [11] is adopted to develop the experiments. For GMM-based acoustic model training, the standard 13-dimensional Mel-Frequency Cesptral Coefficients (MFCC) features and its first and second derivatives without energy features are applied. After that, cepstral mean and variance normalized (CMVN) is performed on the features. Linear discriminant analysis (LDA)is used to splice 9 frames together and project down to 40 dimensions. A maximum likelihood linear transform (MLLT) is estimated on the LDA features. The feature-space Maximum Likelihood Linear Regression (fMLLR) is used for speaker adaptive training (SAT). The baseline GMM model has 2,052 context dependent (CD) triphone states and an average of 34 Gaussian components per state.

4.3. CNN and DNN Acoustic Model

As input features, 40-dimensional log mel-filter bank features are applied for CNN and DNN acoustic models. For DNN, 4 hidden layers with 300 units per hidden layers are used. For CNN, 256 and 128 feature maps in first and second convolutional layers are set respectively with 8 and 4 filter sizes. The pooling size is set to 3 with pool step 1. The fully connected network has 2 hidden layers with 300 units per hidden layers. Cross-entropy training is performed on CNN and DNN acoustic models. Restricted Boltzmann machines (RBMs) are built on top of the CNN training. Additionally, a 6-layer DNN with cross-entropy training is done and 6 iterations of state-level minimum Bayes risk (sMBR) for discriminative training are performed [12]. The training procedure of the CNN (sMBR) is depicted in Figure 5. A constant learning rate of 0.008 is used to train the neural networks. Next, the learning rate is decreased by half through cross-validation error reduction. When the error rate stops decreasing or starts increasing, the training procedure is stopped. Stochastic gradient descent is applied with a mini-batch of 256 training examples for backpropagation. TESLA K80 GPU is used for all the neural network training.

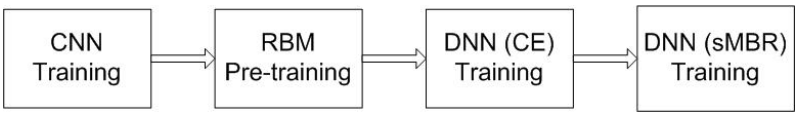

Figure 5. Training flow of CNN (sMBR)

4.4. Experimental Result

In this experiment, the ASR performance is evaluated on different corpus sizes. The three different acoustic models such as GMM, DNN, and CNN models are developed and compared their results. Convolutional Neural Network (CNN) has achieved a better performance than Deep Neural Network (DNN) and Gaussian Mixture Model (GMM) in different large vocabulary continuous speech recognition (LVCSR) tasks [13] [14] [15] because the fully connected nature of DNN can cause overfitting and it decreases the ASR performance for low-resourced languages. CNN can model well tone patterns because it has an ability to reduce the translational invariance and spectral correlations in the input signal. Furthermore, as a sequence discriminative training can minimize the error on the state labels in a sentence, the DNN with sequence training is done on top of the CNN training. It is obvious in this work that CNN (sMBR) significantly outperforms the GMM and DNN acoustic models for a low-resourced and tonal language, Myanmar language.

Figures 6 and 7 show word error rates (WERs) of TestSet1 and TestSet2 based on training data sizes. According to the Figures 6, when the training data set size is increased from 10 hrs to 20 hrs, the WERs of TestSet1 decrease considerably because it is the same domain with the training sets. However, the error rates of TestSet1 are not reduced notably even when the training data size is increased from 30 hrs to 42 hrs because the augmented data is from the different domain. In Figure 7, the word error rates of TestSet2 obviously decrease over the increasing training data size. This is because the augmented data of the training sets of 30 hrs and 42 hrs are the same domain with the TestSet2, which results in diminishing the word error rates of TestSet2. It can be clearly observed that when the amount of training data is increased, WERs are decreased. The largest amount of training data, 42-hr-data set, has the lowest WERs on both test sets. Thus, the training data size has a great impact on the ASR performance.

Figure 6. Word error rate on TestSet1 versus training data

Figure 7. Word error rate on TestSet2 versus training data

According to the evaluation results, the error rates of TestSet1 are lower than that of TestSet2. This is because the news presenters have clear and sharp voices than the voices in the recorded conversational data. Moreover, the total length of the web news data is longer than that of the recorded conversational data. It is found that CNN outperformed DNN and GMM on both test sets. As the result, using CNN (sMBR) leads to the lowest WERs of 15.61% on TestSet1 and 24.43% on TestSet2.

5. CONCLUSION

This paper introduces a UCSY-SC1 corpus for Myanmar speech processing research. The corpus consists of two domains: web news and daily conversational data recorded by ourselves. A detailed description of the collection of text and speech corpus for each domain is presented. The total duration of the UCSY-SC1 corpus is 42 hrs and 39 mins. The corpus consists of 261 speakers for the web news and 46 speakers for conversational domain. Moreover, the phone coverage of the corpus is analyzed. The speech corpus is used as training data for building Myanmar ASR. This is a milestone for Myanmar ASR development. The effect of the training data sizes on recognition accuracy is also analyzed by means of GMM, DNN, and CNN acoustic models. Two test sets, web news and recorded conversational data, are used to evaluate the ASR accuracy. It is found that the accuracy on web news data is better than that of the recorded conversational data. The CNN (sMBR)-based model

outperforms the GMM and DNN models. It leads to the lowest error rates of 15.61% WER on TestSet1 and 24.43% WER on TestSet2 by using this corpus.

As Myanmar is a low-resourced language, creating the speech corpora is essential and it is believed that this corpus will be of some use for future Myanmar speech processing research. The corpus will be further expanded by more speech data and Myanmar ASR will hopefully be developed by means of the end-to-end learning approach.

ACKNOWLEDGMENTS

We would like to thank all faculty members and students of University of Computer Studies, Yangon, for participating in the data collection task.

REFERENCES

[1] J.Xu, *et al.,* "Agricultural Price Information Acquisition Using Noise-Robust Mandarin Auto Speech Recognition," *International Journal of Speech Technology*, vol/issue:21(3), pp.681-688, 2018.

[2] M.O.M.Khelifa, *et al.,* "Constructing Accurate and Robust HMM/GMM Models for an Arabic Speech Recognition System," *International Journal of Speech Technology*, vol/issue:20(4), pp.937-949, 2017.

[3] N.D.Londhe, *et al.,* "Chhattisgarhi speech corpus for research and development in automatic speech recognition," *International Journal of Speech Technology*, vol/issue:21(2), pp.193-210, 2018.

[4] P.Zelasko, *et al.,* "AGH corpus of Polish speech," *Language Resources and Evaluation Journal*, vol. 50, 2015.

[5] A.N.Mon, *et al.,* "Developing a Speech Corpus from Web News for Myanmar (Burmese) Language," *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1-6, 2017.

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

[7] T.Nadungodage, *et al.,* "Developing a Speech Corpus for Sinhala Speech Recognition," *ICON-2013: 10th International Conference on Natural Language Processing,CDAC Noida, India*, 2013.

[8] J.Staš, *et al.,* "TEDxSK and JumpSK: A New Slovak Speech Recognition Dedicated Corpus," *Journal of Linguistics/Jazykovedný casopis*, vol/issue:68(2), pp.346 - 354, 2017.

[9] M.Ziołko, *et al.,* "Automatic Speech Recognition System Dedicated for Polish," *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy*, pp.3315-3316, 2011.

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

[11] D.Povey, *et al.,* "The Kaldi Speech Recognition Toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[12] K.Vesely, *et al.,* "Sequence-discriminative Training of Deep Neural Networks," *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, pp. 2345-2349, August 25-29, 2013.

[13] W.Chan and I.Lane, "Deep Convolutional Neural Networks for Acoustic Modeling in Low Resource Languages," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015*, pp. 2056-2060, 2015.

[14] T.N.Sainath, *et al.,* "Improvements to Deep Convolutional Neural Networks for LVCSR," *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 315-320, 2013.

[15] T.Sercu, *et al.,* "Very Deep Multilingual Convolutional Neural Networks for LVCSR," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, pp. 4955-4959, 2016.