IMPROVING MYANMAR AUTOMATIC SPEECH RECOGNITION WITH OPTIMIZATION OF CONVOLUTIONAL NEURAL NETWORK PARAMETERS

Aye Nyein Mon[1], Win Pa Pa[2] and Ye Kyaw Thu[3]

[1,2]Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar
[3] Language and Speech Science Research Lab., Waseda University, Japan

ABSTRACT

Researchers of many nations have developed automatic speech recognition (ASR) to show their national improvement in information and communication technology for their languages. This work intends to improve the ASR performance for Myanmar language by changing different Convolutional Neural Network (CNN) hyperparameters such as number of feature maps and pooling size. CNN has the abilities of reducing in spectral variations and modeling spectral correlations that exist in the signal due to the locality and pooling operation. Therefore, the impact of the hyperparameters on CNN accuracy in ASR tasks is investigated. A 42-hr-data set is used as training data and the ASR performance was evaluated on two open test sets: web news and recorded data. As Myanmar language is a syllable-timed language, ASR based on syllable was built and compared with ASR based on word. As the result, it gained 16.7% word error rate (WER) and 11.5% syllable error rate (SER) on TestSet1. And it also achieved 21.83% WER and 15.76% SER on TestSet2.

KEYWORDS

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

1. INTRODUCTION

Automatic speech recognition research has been doing for more than four decades. The purpose of speech recognition system is to be convenient for humans with a computer, robot or any machine in interaction via speech. Speech recognition is the converting of speech signal to text. Speech recognition systems can be characterized by a number of parameters. These parameters specify the performance of speech recognition systems such as utterances (isolated words vs. continuous speech), speaking style (read speech formal style vs. spontaneous and conversational speech with causal style), speaking situation (human-to-machine speech vs. human-to-human speech), speaker dependency (speaker dependent, speaker adaptive, vs. speaker independent), vocabulary size (small vocabulary of fewer than 100 words, medium vocabulary of fewer than 10,000 words, vs. large vocabulary of more than 10,000 words.), language model (N-gram statistical model, finite state model, context free model, vs. context sensitive model), environmental condition (high signal-to-noise of greater than 30 dB, medium SNR (signal to noise ratio), vs. low SNR of lower of 10dB) and transducer (close-talking microphone, far-field microphone, microphone array, vs. telephone) [1].

At present and later decade, speech recognition systems have been become in many other nations for their owned languages. They have applied statistical based pattern recognition to develop

DOI: 10.5121/ijnlc.2018.7601

ASR for their languages. Furthermore, they used the particular features of their languages such as tones, pitch, etc. For example, Mandarin, Vietnamese, Thai, etc., they appended tones related information to acoustic modelling to increase their ASR accuracy [2] [3] [4].

For low-resourced languages, the researchers have developed their speech recognition systems by building speech corpus from scratch. For instance, AGH corpus of Polish speech [5], Agglutinative Hungarian spontaneous speech database [6], Bengali speech corpus [7], Bulgarian speech corpus [8], European Portuguese corpus [9], etc. were developed for ASR technology.
There are previous works for Myanmar language using deep learning techniques. Thandar Soe [et.al,] [10] built a robust automatic speech recognizer by using deep convolutional neural networks (CNNs). The multiple acoustic models were trained with different acoustic feature scales. The multi-CNN acoustic models were combined based on a Recognizer Output Voting Error Reduction (ROVER) algorithm for final speech recognition experiments. They showed that integration of temporal multi-scale features in model training achieved the lower error rate over the best individual system on one temporal scale feature. Aye Nyein Mon [et.al,] [11] has designed and developed a speech corpus from web news for Myanmar ASR. The speech corpus consists of 15 hours of read speech data collected from online web news in which there are 155 speakers (109 females and 46 males). Moreover, this corpus was evaluated by using state-of-the-art acoustic modeling approach, CNN. The experiments were conducted by means of the default parameters of CNN. It showed that corpora built on Internet resources get promising results. In this paper [12], Aye Nyein Mon [et.al,] explored the effect of tones for Myanmar language speech recognition using Convolutional Neural Network (CNN). Experiments are conducted based on the modeling of tones by integrating them into the phoneme set and incorporating them into convolutional neural network, state-of-the-art acoustic model. Moreover, tonal questions are utilized to build phonetic decision tree. In this work, the hyperparameters of CNN architecture are not changed and they are used as default settings for all experiments. It showed that in comparison with Deep Neural Network (DNN) baseline, the CNN model achieves a better ASR performance over DNN. It was investigated that the CNN model with tone information got the better accuracy than that of without tone information. Hay Mar Soe Naing [et.al,] [13] presented a Myanmar large vocabulary continuous speech recognition system for Travel domain. Phonemically-balanced corpus consisting of 4000 sentences and 40 hours of speech was used for training data. In this work, three kinds of acoustic models were developed: Gaussian mixture model (GMM), DNN (Cross Entropy), and DNN (state-level minimum Bayes risk (sMBR) and compared their performance. The tonal and pitch features were incorporated to acoustic modeling and experiments were conducted with and without those features. The differences between the word-based language model (LM) and syllable-based LM were also investigated. The best WER of and SER were achieved with sequence discriminative training DNN (sMBR).

In this paper, the Myanmar ASR performance is improved by altering hyperparameters of CNN for Myanmar ASR. The CNN has locality property that can reduce the number of neural network weights to be learned and hence decreases overfitting. Moreover, pooling operation is very useful in handling small frequency shifts that are common in speech signals. Therefore, the different parameters of CNN such as feature map numbers and pooling sizes have been changed to get the better ASR accuracy for Myanmar language. In addition, as Myanmar is a tonal and syllable-timed language, the syllable-based Myanmar ASR is developed by using a 42-hr-data set. And then, the performance of word-based and syllable-based ASRs is compared by using two open test sets: web data and recorded data.

This paper is formed as follows. In Section 2, about Myanmar language is presented. Convolutional Neural Network (CNN) is described in Section 3. Experiments such as improving CNN architecture and comparison of word-based and syllable-based ASRs are done in Section 4. Conclusion and future work are recapped in Section 5.

2. MYANMAR LANGUAGE

The Myanmar language is Sino-Tibetan language spoken in Myanmar. Most of the Myanmar people about 33 million have spoken the official language of Myanmar. Ethnic groups related to Myanmar have used as a first language and 10 million of ethnic minorities in Myanmar as a second language.

Myanmar is a tonal, largely monosyllabic and analytic language. Myanmar language has 4 contrasts tones and syllables can have different meanings on the basis of tone. A tone is denoted by a diacritic mark. A simple syllable structure of the Myanmar language is composed of either a vowel by itself or a consonant combined with a vowel. Unlike the other language such as English, it has subject-object-verb order and there are no spaces between words in writing. Therefore, it needs to do segmentation to determine the word boundaries. Myanmar characters has 3 groups: consonants (known as “Byee”), medials (known as “Byee Twe”), and vowels (known as “Thara”). There are 33 basic consonants, 4 basic medials, and 12 basic vowels in Myanmar script. Myanmar numerals are decimal-based for counting [14].

3. CONVOLUTIONAL NEURAL NETWORK (CNN)

The Convolutional Neural Network is a feed-forward artificial neural network, which includes convolutional and pooling layers. The CNN needs the input data to be organized in a certain way. For processing speech signals, it requires to use features which are organized along frequency or time (or both) so that the convolution operation can be correctly applied. In this work, a number of one-dimensional (1-D) feature maps are used to consider frequency variations only which normally occur in speech signals due to speaker differences [15].

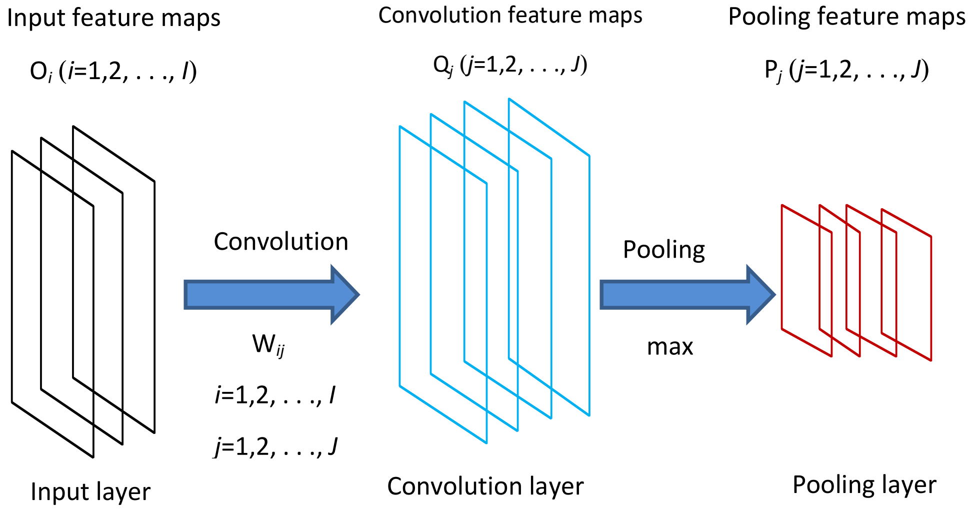

Figure 1: An illustration of one CNN “layer” consisting of a pair of a convolution ply and a pooling ply in succession, where mapping from either the input layer or a pooling ply to a convolution ply is based on eq. (2) and mapping from a convolution ply to a pooling ply is based on eq. (3).

Once the input feature maps are formed, the convolution and pooling layers apply their respective operations to generate the activations of the units in those layers, in sequence, as depicted in Figure 1.

3.1 Convolution Ply

As shown in Figure 1, all input feature maps (assume I in total), O_i (i =1,…, I) are mapped into a number of feature maps (assume J in total),Q_j (j = 1, …, J) in the convolution ply based on a number of local filters ($I \times J$ in total), $w_{i,j}$ (i = 1, . . ., I; j = 1, . . ., J). The mapping can be represented as the well-known convolution operation in signal processing [15]. Assuming input feature maps are all one dimensional, each unit of one feature map in the convolution layer can be computed as:

$$q_{j,m} = \sigma\left(\sum_{i=1}^{I} \sum_{n=1}^{F} o_{i,n+m-1} w_{i,j,n} + w_{0,j}\right), \qquad (j = 1, \dots, J) \tag{1}$$

where,

$o_{i,m}$ = the m-th unit of the i-th input feature map O_i,

$q_{j,m}$ = the m-th unit of the j-th feature map Qj in the convolution ply,

$w_{i,j,n}$ = the nth element of the weight vector, $w_{i,j}$, connecting the ith feature map of the input to the jth feature map of the convolution layer,

F = the filter size which is the number of input bands that each unit of the convolution layer receives

When, the equation (1) is written as a more concise matrix form using the convolution operator ∗,

$$Q_j = \sigma\left(\sum_{i=1}^{I} O_i * w_{i,j}\right)(j = 1, \dots, J), \tag{2}$$

where,

O_i = the i-th input feature map,

wi,j = each local filter with the weights flipped to adhere to the convolution operation definition.

A complete convolutional layer is composed of many feature maps, generated with different localized filters, to extract multiple kinds of local patterns at every location.

3.2 Pooling Ply

Each neuron of the pooling layer receives activations from a local region of the previous convolutional layer and reduces the resolution to produce a single output from that region. The pooling function computes some overall property of a local region by using a simple function like maximization or averaging. In this case, max pooling function is applied. The max pooling layer performs down-sampling by dividing the input into rectangular pooling regions, and computing the maximum of each region. The max pooling layer is defined as:

$$p_{i,m} = max_{n=1}^{G}\ q_{i,(m-1)\times s+n} \tag{3}$$

where,

G = the pooling size,

s = the *shift size*, determines the overlap of adjacent pooling windows.

4. EXPERIMENTS

Experiments are done by using our developed speech corpus that has 42-hr-data size. 3-gram language model with Kneser-Ney discounting is built by using SRILM language modeling toolkit [16]. Two open test sets are used and the detailed statistics on the training and testing sets are depicted in Table 1.

The input features are 40-dimensional log mel-filter bank features padded with 11 frames (5 frames left and right frame contexts). The targets are 2,052 context dependent (CD) triphone states that are generated from GMM-HMM model.

For CNN training, 0.008 of constant learning rate is diminished by half based on cross-validation error decreasing. When the error rate has no more noticeably reducing or the error rate started to increase, training process stopped. Stochastic gradient descent with a mini-batch of 256 training examples is utilized for backpropagation.

TESLA K80 GPU is used for all the neural networks training. Kaldi [17] automatic speech recognition toolkit is applied to do the experiments.

Table 1: Train and test sets statistics

Data	**Size**	**Domain**	**Speakers**			**Utterance**
			Female	**Male**	**Total**	
TrainSet	42 Hrs 39 Mins	Web News + Recorded Data	219	88	307	31,114
TestSet1	31 Mins 55 Sec	Web News	5	3	8	193
TestSet2	32 Mins 40 Sec	Recorded Data	3	2	5	887

4.1. Optimization of CNN Parameters

In this experiment, the effect of changing various CNN parameters is investigated on Myanmar ASR performance. For this work, one dimensional convolution across frequency domain is applied and there are two convolutional layers, one pooling layer and two fully connected hidden layers with 300 units per layer. Pooling size and feature map numbers have a great impact to improve the ASR performance [18] [19]. Hence, the ASR accuracy is evaluated on two open test sets, web news and recorded data, by varying the different feature map numbers and pooling sizes.

4.1.1 Number of Feature Maps of First Convolutional Layer

Firstly, different numbers of feature maps in the first convolutional layer are changed and compared on the evaluation results. The filter size of the first layer is set to 8. The number of feature maps is altered to 32, 64, 128, 256, 512, and 1,024 respectively. Table 2 depicts the WER% on two open test sets based on changing the feature map numbers of first convolutional layer.

Table 2: Evaluation results on number of feature maps of the first convolutional layer

Number of Feature Maps	**WER%**	
	TestSet1	**TestSet2**
32	18.97	24.01
64	18.59	23.80
128	**18.18**	**23.50**
256	18.52	23.75
512	18.31	23.65
1,024	18.53	23.85

It is observed that the lowest WERs on both test sets are obtained with feature map numbers of 128. According to the Table 2, the WERs reduce slightly when the number of feature maps is increased up to 128. But, if the number of feature maps exceeds 128, it did not get the lower WERs and there is a small increment of WER. As the result, the lowest WERs are attained with 128 feature maps.

4.1.2 Pooling Size

Pooling has ability in handling the small frequency shifts in speech signal. Max pooling has achieved a greater accuracy than the other pooling types in ASR tasks [18]. Thus, for this experiment, max pooling is used and the best pooling size is investigated. The pooling layer is added above the first convolution layer. The experiments are done by varying max pooling size with a shift size of 1. The feature maps of first convolution layer are fixed at 128. Table 3 gives that the evaluation results based on the different pooling sizes.

Table 3: Evaluation results on pooling size

	WER%	
	TestSet1	**TestSet2**
No pooling	18.18	23.50
pool size=2	18.05	23.00
pool size=3	17.22	22.56
pool size=4	18.06	23.33

From the Table 3, it can be found that the first convolution layer using 128 feature maps, followed by the max pooling size 3 has the lowest WERs, 17.22% on TestSet1 and 22.56% on TestSet2. It indicates that after the pooling layer is appended, the WER diminishes significantly than without using it. Hence, pooling has an impact on the ASR performance and the lowest error rates are achieved with the pooling size of 3.

4.1.3 Number of Feature Maps of Second Convolutional Layer

The second convolutional layer is added on top of the pooling layer and the experiments are further done to investigate the best number of feature maps for the second convolutional layer. The filter size of this layer is set to 4. The best result of 128 feature maps is fixed in the first convolution layer. And, the max pooling layer with pool size 3 is also fixed in this experiment.

From Table 4, when the number of feature maps in the second convolutional layer is altered to 32, 64 and 128 respectively, it did not get the lower error rates on test sets. But, when the feature map numbers are set above 128, it slightly decreases the WERs on both test sets. Hence, the more number of filters the convolutional layers have, the more acoustic features get extracted and the better the network becomes better at recognizing acoustic patterns of unseen speakers.

Table 4: Evaluation results on number of feature maps in the second convolutional layer

Number of Feature Maps	**WER%**	
	TestSet1	**TestSet2**
32	18.14	23.40
64	18.40	23.64
128	17.86	22.65
256	17.03	22.20
512	17. 02	21.89
1,024	**16.70**	**21.83**

Therefore, it is analyzed that 128/1,024 feature maps in first and second convolutional layers with max pooling size 3 gives the best accuracy for Myanmar ASR.

4.2 Comparison of Syllable vs. Word-Based ASR

Syllable is a basic unit of Myanmar language and every syllable has a meaning in Myanmar language. Therefore, in this experiment, the syllable-based ASR is developed and the evaluation results of the syllable-based ASR and word-based ASR are compared. The above CNN architecture, 128/1,024 feature maps in first and second convolutional layers with max pooling size 3 are applied to develop the ASR with syllable units. As in word-based ASR, 3-gram language model with Knesey-Ney discounting is used.

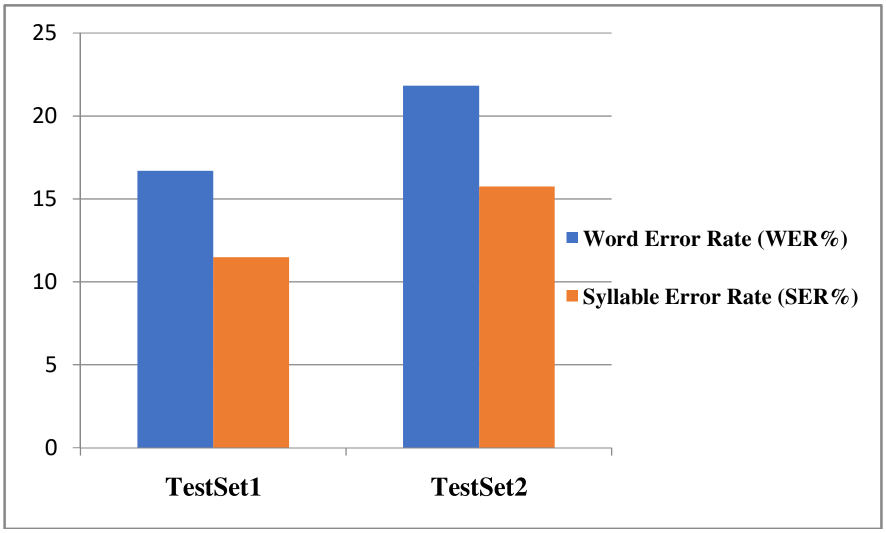

Figure 2: Comparison of word and syllable-based ASRs

Figure 2 shows the comparison of word-based and syllable-based ASRs in terms of word error rate (WER%) and syllable error rate (SER%). It can be clearly seen that SER% are significantly less than the WER% for both test sets. Therefore, it can be assumed that the accuracy of ASR using syllable units is better than that of the ASR using word units. Nevertheless, for fair comparison, the syllables of the hypothesis text of the syllable-based ASR are changed to words levels and the WERs of that words are calculated again. Then, the error rates of the syllable-based and word-based ASRs are compared again. Consequently, it is observed that the error rates of the syllable-based ASR are greater than that of the word-based ASR. Therefore, it can be said that the accuracy of the word-based ASR is better than that of the syllable-based ASR.

Moreover, the hypothesis texts of word-based and syllable-based ASRs are manually evaluated by one native Myanmar people. It is observed that the recognition output of word-based ASR is more effective, meaningful and accurate than the syllable-based ASR's output. Therefore, it can be summarized that the word-based ASR can give more reasonable output text than syllable-based ASR.

The example reference and hypothesis sentences of word-based ASR are expressed as follows. The reference text consists of 41 correct words. In the hypothesis text, the total count of insertion, substitution and deletion words are 6. So, the WER obtains 14.63%. The error words in the hypothesis text are expressed using numbers.

Reference Text of Word-based ASR:

ဒီ ပြပွဲ မှာ ဆိုလို့ရှိရင် တိုရစ်စင် အောက် **မှု**[1] **အကြိုးဝင်**[2] တဲ့ ဟိုတယ် တွေ ခရီးသွား အေးဂျင့် တွေ **အဲလိုင်း**[3] တွေ အာမခံ အကြောင်းအရာ တွေ အပြင် ခရီးသွား လုပ်ငန်း နဲ့ ဆက်စပ် တဲ့ တခြား အဖွဲ့အစည်း တွေ ရဲ့ ပြခန်း တွေ ကို **ပါ**[4] ပါဝင် ခင်းကျင်း ပြသ သွား မှာ ဖြစ် ပါတယ် ရှင်

Hypothesis Text of Word-based ASR:

ဒီ ပြပွဲ မှာ ဆိုလို့ရှိရင် တိုရစ်စင် အောက် **မှာ**[1] **အကြော ဝင်**[2] တဲ့ ဟိုတယ် တွေ ခရီးသွား အေးဂျင့် တွေ **အိမ် လိုင်း**[3] တွေ အာမခံ အကြောင်းအရာ တွေ အပြင် ခရီးသွား လုပ်ငန်း နဲ့ ဆက်စပ် တဲ့ တခြား အဖွဲ့အစည်း တွေ ရဲ့ ပြခန်း တွေ ကို **က**[4] ပါဝင် ခင်းကျင်း ပြသ သွား မှာ ဖြစ် ပါတယ် ရှင်

The example reference and hypothesis sentences of syllable-based ASR are shown as follows. The reference sentences of word-based and syllable-based ASRs are the same. The total number of syllables in the reference texts is 73. The total insertion, substitution and deletion syllables in the hypothesis text are 9 and so, 12.33% SER is attained. The highlighted words that describing using numbers in the hypothesis are the wrong words.

Reference Text of Syllable-based ASR:

ဒီ ပြ ပွဲ မှာ ဆို လို့ ရှိ ရင် **တို**[1] **ရစ်**[2] **စင်**[3] အောက် **မှု**[4] အ **ကြိုး**[5] ဝင် တဲ့ ဟို တယ် တွေ ခ ရီး သွား **အေး**[6] **ဂျင့်**[7] တွေ **အဲ**[8] လိုင်း တွေ အာ မ ခံ အ ကြောင်း အ ရာ တွေ အ ပြင် ခ ရီး သွား လုပ် ငန်း နဲ့ ဆက် စပ် တဲ့ တ ခြား အ ဖွဲ့ အ စည်း တွေ ရဲ့ ပြ ခန်း တွေ ကို **ပါ**[9] ပါ ဝင် ခင်း ကျင်း ပြ သ သွား မှာ ဖြစ် ပါ တယ် ရှင်

Hypothesis Text of Syllable-based ASR:

ဒီ ပြ ပွဲ မှာ ဆို လို့ ရှိ ရင် **ထိုး**[1] **ရေး**[2] **ဆယ်**[3] အောက် **မှာ**[4] အ **ကြို**[5] ဝင် တဲ့ ဟို တယ် တွေ ခ ရီး သွား **အေ**[6] **ဂျင်**[7] တွေ **အိမ်**[8] လိုင်း တွေ အာ မ ခံ အ ကြောင်း အ ရာ တွေ အ ပြင် ခ ရီး သွား လုပ် ငန်း နဲ့ ဆက် စပ် တဲ့ တ ခြား အ ဖွဲ့ အ စည်း တွေ ရဲ့ ပြ ခန်း တွေ ကို **က**[9] ပါ ဝင် ခင်း ကျင်း ပြ သ သွား မှာ ဖြစ် ပါ တယ် ရှင်

5. CONCLUSION

In this work, the better accuracy of automatic speech recognition for Myanmar language is investigated by changing the hyperparameters of CNN. It is found that feature map numbers and pooling sizes of CNN have a great impact on ASR performance. By varying CNN hyperparameters, the lower error rates are achieved. In addition, as syllable is a basic unit of Myanmar language, syllable-based ASR is also created and the performance of syllable-based and word-based ASRs is compared. It can be concluded that although the SER% of syllable-based ASR is less than the WER% of word-based ASR, the recognition output of word-based ASR is more effective, meaningful and accurate than that of syllable-based ASR. And thus, word-based ASR can give better recognizable results than syllable-based ASR for Myanmar language.

In the future, Myanmar ASR will be developed by using state-of-the-art end-to-end learning approach.

REFERENCES

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

2. X.Hu, M.Saiko, and C.Hori, "Incorporating Tone Features to Convolutional Neural Network to Improve Mandarin/Thai Speech Recognition", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014, Chiang Mai, Thailand, December 9-12, 2014, pp.1-5

3. V.H.Nguyen, C.M.Luong, and T.T.Vu, "Tonal Phoneme Based Model for Vietnamese LVCSR", 2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Shanghai, China, October 28-30, 2015, pp.118-122.

4. X. Lei, M. Siu, M. Hwang, M. Ostendorf, T. Lee, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition", INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006, 2006.

5. P.Zelasko, B. Ziolko, T.Jadczyk, and D.Skurzok, "AGH Corpus of Polish Speech", Language Resources and Evaluation, vol. 50, no. 3, pp. 585--601, 2016.

6. T.Neuberger, D.Gyarmathy, T.E.Graczi, V.Horvath, M.Gosy, and A.Beke, "Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language", in Text, Speech and Dialogue - 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings, 2014, pp. 424--431.

7. S.Mandal, B.Das, P.Mitra, and A.Basu, "Developing Bengali Speech Corpus for Phone Recognizer Using Optimum Text Selection Technique", in International Conference on Asian Language Processing, IALP 2011, Penang, Malaysia, 15-17 November, 2011, 2011, pp. 268--271.

8. N.Hateva, P.Mitankin, and S.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, Portoroz, Slovenia, May 23-28, 2016., 2016.

9. F.Santos and T.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.

10. T.Soe, S.S.Maung, N.N.Oo, "Combination of Multiple Acoustic Models with Multi-scale Features for Myanmar Speech Recognition", International Journal of Computer (IJC), [S.l.], Volume 28, No 01, pp.112-121, February. 2018. ISSN 2307-4523. Available at: <http://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/1150>. Date accessed: 03 Sep. 2018.

11. A.N.Mon, W.P.Pa, Y. K.Thu and Y. Sagisakaa, "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), Seoul, 2017, pp. 1-6.

12. A.N.Mon, W.P.Pa and Y.K.Thu, "Exploring the Effect of Tones for Myanmar Language Speech Recognition Using Convolutional Neural Network (CNN)", In: Hasida K., Pa W. (eds) Computational Linguistics. PACLING 2017. Communications in Computer and Information Science, vol 781. Springer, Singapore.

13. H.M.S.Naing, A.M.Hlaing, W.P.Pa, X.Hu, Y.K.Thu, C.Hori, and H.Kawai, "A Myanmar Large Vocabulary Continuous Speech Recognition System", In Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2015, Hong Kong, December 16-19, 2015, pages 320–327, 2015.

14. Myanmar Language Committee, "Myanmar Grammar", Myanmar Language Committee, Ministry of Education, Myanmar, 2005.

15. O.A-Hamid, A-R.Mohamed, H.Jiang, L.Deng, G.Penn, D.Yu, " Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 10, October 2014

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

17. D.Povey, A.Ghoshal, G.Boulianne, L.Burget, O.Glembek, N.Goel, M.Hannemann, P.Motlicek, Y.Qian, P.Schwarz, J.Silovsky, G.Stemmer, and K.Vesely, "The Kaldi Speech Recognition Toolkit", IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, Dec 2011, IEEE Catalog No.: CFP11SRW-USB.

18. T.N.Sainath, B.Kingsbury, A.Mohamed, G.E.Dahl, G.Saon, H.Soltau, T.Beran, A.Y.Aravkin, and B.Ramabhadran, "Improvements to Deep Convolutional Neural Networks for LVCSR", 2013 IEEE Workshop on Automatic Speech Recognition and Understanding,Olomouc, Czech Republic, December 8-12, 2013, pp.315-320.

19. W. Chan, 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, South Brisbane, Queensland, Australia, April 19-24, 2015, p. 2056-2060.

AUTHORS

Aye Nyein Mon received M.C.Sc. (Credit), in Computer Science, from University of Computer Studies, Yangon (UCSY), Myanmar. She is currently pursuing Ph.D in University of Computer Studies, Yangon. Her research interests are Natural Language Processing, Speech Processing, and Machine Learning.

Dr. Win Pa Pa is now working as a Professor and has been doing research at Natural Language Processing lab of UCSY. She has been supervising Master and Ph.D thesis on Natural language processing such as Information Retrieval, Morphological Analysis, Part of Speech Tagging, Parsing, and Automatic Speech Recognition and Speech Synthesis. She took part in the project of ASEAN MT, the machine translation project for South East Asian languages. She also participated in the projects of Myanmar Automatic Speech Recognition and HMM Based Myanmar Speech Synthesis (Text to Speech) that were the research collaboration between NICT, Japan and UCSY.

Dr. Ye Kyaw Thu is a Visiting Researcher of Language and Speech Science Research Lab., Waseda University, Japan and also a co-supervisor of masters' and doctoral students of several universities. His research interests lie in the fields of AI, natural language processing (NLP) and human-computer interaction (HCI). His experience includes research and development of text input methods, corpus building, statistical machine translation (SMT), automatic speech recognition (ASR) and text to speech (TTS) engines for his native language Myanmar. He is currently, working on robot language acquisition research and sign language processing.