Exploring the Effect of Tones for Myanmar Language Speech Recognition Using Convolutional Neural Network (CNN)

Aye Nyein Mon[1(✉)], Win Pa Pa[1], and Ye Kyaw Thu[2]

[1] Natural Language Processing Laboratory, University of Computer Studies, Yangon, Myanmar
{ayenyeinmon,winpapa}@ucsy.edu.mm
[2] Artificial Intelligence Laboratory, Okayama Prefectural University, Okayama, Japan
ye@c.oka-pu.ac.jp

Abstract. Tone information is very helpful to improve automatic speech recognition (ASR) performance in tonal languages such as Mandarin, Thai, Vietnamese, etc. Since Myanmar language is being considered as a tonal language, the effect of tones on both syllable and word-based ASR performance has been explored. In this work, experiments are done based on the modeling of tones by integrating them into the phoneme set and incorporating them into the Convolutional Neural Network (CNN), state-of-the-art acoustic model. Moreover, to be more effective tone modeling, tonal questions are used to build the phonetic decision tree. With tone information, experiments show that compared with Deep Neural Network (DNN) baseline, the performance of CNN model achieves nearly 2% for word-based ASR or more than 2% for syllable-based ASR improvement over DNN model. As a result, the CNN model with tone information gets 2.43% word error rate (WER) or 2.26% syllable error rate (SER) reductions than without using it.

Keywords: Tone information
Automatic Speech Recognition (ASR) · Tonal language
Deep Neural Network (DNN) · Convolutional Neural Network (CNN)

1 Introduction

Nowadays, the performance of automatic speech recognition (ASR) systems is improved by exploring the new architecture and utilizing particular properties of the target language. Recently, Convolutional Neural Network (CNN) has shown significant performance in ASR tasks [1,2]. CNN has gained higher performance than Deep Neural Network (DNN) across different large vocabulary continuous speech recognition (LVCSR) tasks [3,4]. CNN has an ability to reduce the translational invariance and spectral correlations in the input signal. There are many

K. Hasida and W. P. Pa (Eds.): PACLING 2017, CCIS 781, pp. 314–326, 2018.
https://doi.org/10.1007/978-981-10-8438-6_25

ASR tasks that utilize the particular features of the target language. For tonal languages such as Mandarin, Thai, Vietnamese, etc., they used the particular features of their language which means the tonal information was augmented to acoustic modeling to improve their ASR performance [5,6].

Myanmar language is assumed as a tonal language. There are four different tones in Myanmar language. It has different meanings according to the different types of tones. The following example shows the four tones of the phoneme 'a' (အ) and its different meanings in Myanmar language. /a te/ (အာ တယ်) [to be wide open, to talk too much] with Tone1, /a: te/ (အား တယ်) [to be free] with Tone2, /a. te/ (အ တယ်) [to be dumb or dull or naive] with Tone3 and /a' te/ (အံ့ တယ်) [to place an order] with Tone4 [8]. Therefore, accurate tone recognition plays an important role in automatic Myanmar speech recognition. In this work, the effect of tones is explored using state-of-the-art acoustic modeling approach, CNN-based acoustic model for Myanmar language. In low-resourced condition, CNN is better than DNN because the fully connected nature of the DNN can cause overfitting and it degrades the ASR performance for low-resourced languages where there is a limited amount of training data [1]. CNN can alleviate these problems and it is very useful for a low-resourced language such as Myanmar. Moreover, CNN can model well tone patterns because it can reduce spectral variations and model spectral correlations existing in the signal. For tonal languages such as Vietnamese [9], the tonal information are incorporated into the phonetic decision tree and it showed promising result. Therefore, in this work, tonal questions are used to build the phonetic decision tree in order to get more sophisticated tone modeling.

This paper is organized as follows: In Sect. 2, Myanmar language and previous Myanmar ASR research works are discussed. Building the speech and text corpus is presented in Sect. 3. About pronunciation lexicon is in Sect. 4. In Sect. 5, CNN is introduced. The experimental setup is performed in Sect. 6. The evaluation result is discussed in Sect. 7. Error analysis is done in Sect. 8. The conclusion and future work are presented in Sect. 9.

2 Myanmar Language and Previous Myanmar ASR Research Works

In this section, we describe Myanmar language and its previous ASR research works.

2.1 Nature of Myanmar Language

Myanmar language is a tonal, syllable-timed language, largely monosyllabic and analytic language with a subject-object-verb word order. There are 12 basic vowels, 33 consonants, and 4 medials in Myanmar language. The basic consonants in Myanmar can be extended by medials. Syllables or words are formed by consonants combining with vowels. Generally, one syllable is formed by combining

Table 1. Characteristics of Myanmar tones

Register		Phonation	Length	Pitch
Low	a	Modal voice	Medium	Low
High	a:	Breathy voice	Long	High
Creaky	a.	Creaky voice	Medium	High
Checked	a?	Final glottal stop	Short	Varies

Table 2. Three example Myanmar sentences of training data

State Counsellor Daw Aung San Suu Kyi sends a message to the Myanmar Motion Picture Academy Awards Presentation Ceremony for 2016. နိုဝ် ထောင် ဆယ် ခြောက် ခုနှစ် မြန်မာ ရုပ်ရှင် အကယ်ဒမီ ထူးချွန်ဆု ပေးပွဲ အခမ်းအနား သို့ နိုင်ငံတော် ၏ အတိုင်ပင်ခံ ပုဂ္ဂိုလ် ဒေါ် အောင်ဆန်းစုကြည် မှ သတင်းလွှာ တစ် စောင် ပေးပို့ ခဲ့ ပါတယ် (nhi' htaun. hse. chau' ku. nhi' mjan ma. jou' shin a- ke da- mi htu: gyun hsu. pei: bwe: a- khan: a- na: dhou. nain gan do i. a- tain bin gan pou' gou do aun hsan: su. kyi mha. tha- wun lhwa ta- zaun pei: pou. ge. ba de)
The domestic gold price is an upward trend with the global price rising, according to the local gold market. ကမ္ဘာ ရွေ့ရေး မြင့်တက် လာ မှု ကြောင့် ပြည်တွင်း ရွှေ ဈေးကွက် အမြင့် ဘက် သို့ ဦးတည် နေ ကြောင်း ဒေသ ရွှေ ဈေးကွက် က ဆို ပါတယ် (ga- ba. shwei zei: mjin. te' la mhu. gyaun. pji dwin: shwei zei: gwe' a- mjin. be' thou. u: ti nei gyaun: dei tha. shwei zei: gwe' ka. hsou ba de)
The Lao Tennis Federation is selecting a squad to compete in the 29th Southeast Asian (SEA) Games to be held in Malaysia in August. လာအို တင်းနစ် အဖွဲချုပ် မှ လာမည့် သြဂုတ် လ မလေးရှား နိုင်ငံ တွင် ကျင်းပ မည့် နှစ် ဆယ် ကိုး ကြိမ်မြောက် အရှေ့တောင် အာရှ ဆီးဂိမ်း တွင် ပါဝင် ယှဉ်ပြိုင် ရန် အသင်း အဝှေက် အားကစားသမား များကို ရွေးချယ် နေ ပါတယ် (la ou tin: ni' a- hpwe. gyou' mha. la me. o: gou' la. ma- lei: sha: nain ngan dwin kyin: pa. mji. nha- hse. kou: gyein mjau' a- shei. taun a sha. hsi: gein: dwin pa win shin pjain jan a- thin: a- twe' a: ga- za: tha- ma: mja: gou jwei: che nei ba de)

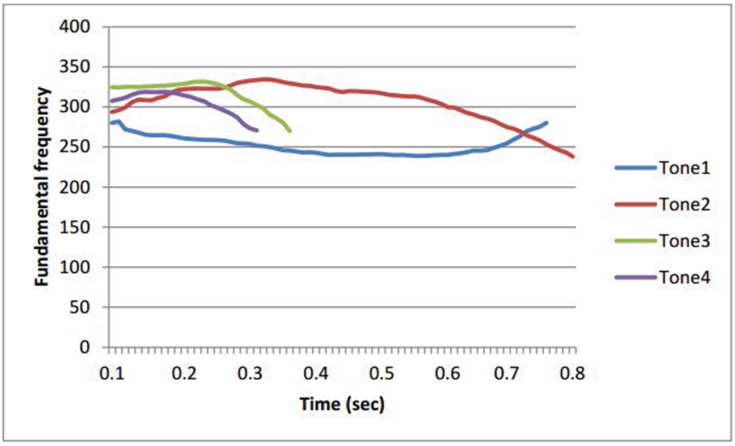

Fig. 1. Example of four tones of the Myanmar syllable 'a' (အ)

one consonant and vowel, and one word is composed of one or more than one syllable. There are 23 phonemes for 33 consonants since some consonant letters have the same phoneme. Myanmar tone is carried by syllable and is featured by both fundamental frequency and duration of syllable. There are four nominal tones transcribed in writing Myanmar: low, high, creaky and checked [7]. For example, the four types of tones of the phoneme 'a' (အ) are /a/ (အာ) , /a:/ (အား) , /a./ (အ) and /a'/ (အော်) and their fundamental frequency is shown in Fig. 1. Table 1 shows the characteristics of Myanmar tones.

2.2 Previous Myanmar ASR Research Works

This section discusses recent publications in Myanmar ASR. Hidden Markov Model with Subspace Gaussian Mixture Model (HMM-SGMM) continuous automatic speech recognition on Myanmar web news was developed by Mon et al. [10]. In this work, both Gaussian Mixture Model (GMM) and SGMM-based systems are compared using n-gram language model. Moreover, ASR performance was evaluated based on training data size, language model size and number of Gaussians. Soe and Thein [11] presented syllable-based speech recognition for Myanmar using GMM with Hidden Markov Model (GMM-HMM) approach. Syllable-based language model was used in this work. Myanmar language speech recognition with hybrid artificial neural network and hidden Markov model was demonstrated by Nwe and Myint [12]. This system used Artificial Neural Network and Hidden Markov Model (ANN-HMM) in acoustic model building and Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coding (LPCC) and Perceptual Linear Prediction (PLP) were used in feature extraction techniques. A Myanmar large vocabulary continuous speech recognition for travel domain was presented by Naing et al. [7]. In this work, DNN-based acoustic model was developed. Word error rate (WER) and syllable error rate (SER) were used as the evaluation criteria. The tonal phonemes and pitch features were applied in acoustic modeling and it showed that tones information is very helpful for Myanmar language. Myanmar continuous speech recognition based on Dynamic Time Wraping (DTW) and HMM was expressed by Khaing [13] using HMM approach. Combinations of LPC, MFCC and Gammatone Cepstral Coefficients (GTCC) techniques were used in feature extraction and DTW was applied in feature clustering.

3 Building the Speech and Text Corpus

We already built the 5-hour-data set for the broadcast news domain. It was used in my first paper [10] and in that paper, we extend our 5-hour-data set to 10-hour-data set. To build the speech corpus, there are many websites on the Internet that the Myanmar broadcast news is available. Among them, we collected the recorded speech with clear voice from Myanmar Radio and Television (MRTV), Voice of America (VOA), Eleven broadcasting, 7 days TV, and ForInfo news. Both local news and foreign news are included in the corpus. The audio files are cut from each news file. And then, the segmented files are set to a single channel (mono type) and the sampling rate is set to 16 kHz. The length of the segmented audio files is between 2 s and 25 s. Most of the audio files from the corpus are female voice, and only few audio files are male voice. The average number of words and syllables in one utterance is 33 words and 54 syllables. Most of the broadcast news at the online never transcribe into text. Therefore, the audio data is listened to and transcribed into text manually. Moreover, Myanmar words do not have boundary in writing like the other languages such as English. In order to define the word boundary, it needs to put a space between Myanmar words and this is done by using Myanmar word segmenting tool [14]. The segmentation and

spelling of the words are then manually checked. To increase the size of the text corpus, bootstrapping technique is used. The total time taken for preparing the corpus is about 5 months. Table 2 shows the three example Myanmar sentences from the training set.

4 Pronunciation Lexicon

Lexicon is a list of words, with a pronunciation for each word expressed as a phone sequence. To generate the pronunciation of the new words, grapheme-to-phoneme (g2p) converter [15] is used and then the new words are added to the lexicon. In this experiment, two types of dictionary are used: dictionary with tone and dictionary without tone.

4.1 Tonal Pronunciation Lexicon

A standard dictionary, Myanmar language commission (MLC) dictionary with tone, is adopted as baseline [16] and this dictionary is extended to words from the speech corpus. There are about 36,700 words in the lexicon. Table 3 shows a snippet of Myanmar phonetic dictionary with tone.

Table 3. Example of Myanmar phonetic dictionary with tone

Myanmar word	Phonetic
က	k a.
ကာ	k a
ကဇား	g a- z a:
ကသိ်	k a- th i'

4.2 Non-tonal Pronunciation Lexicon

Tone information is not included in this dictionary and Table 4 depicts a snippet of Myanmar phonetic dictionary without tone information.

Table 4. Example of Myanmar phonetic dictionary without tone

Myanmar word	Phonetic
က	k a
ကာ	k a
ကဇား	g a z a
ကသိ်	k a th i

5 Convolution Neural Network (CNN)

CNN has shown a greater success in ASR tasks than Deep Neural Network (DNN) recently [3,4] because of local filtering and max pooling layer of the CNN architecture where DNN has only the fully connected layers. Therefore, CNN has an advantage of handling in small frequency shifts that are common in speech signal and temporal variability due to varying speaking rate. CNN runs a small window over the input speech so that the weights of the network that looks through this window can learn from various features of the input.

Convolution layer - In this layer, each hidden activation h_i takes inputs from a small rectangular section of the previous layer, multiplying those local inputs (i.e., $[v_1, v_2, v_3]$) against the weight matrix W. The weight matrix, or the localized filter, will be replicated across the entire input space to detect a specific kind of local pattern. All neurons sharing the same weights compose a feature map. 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.

Pooling layer – The pooling layer similarly takes inputs from a local region of the previous convolutional layer and down-samples it to produce a single output from that region. One common pooling operator used for CNNs is max-pooling, which outputs the maximum value within each sub-region.

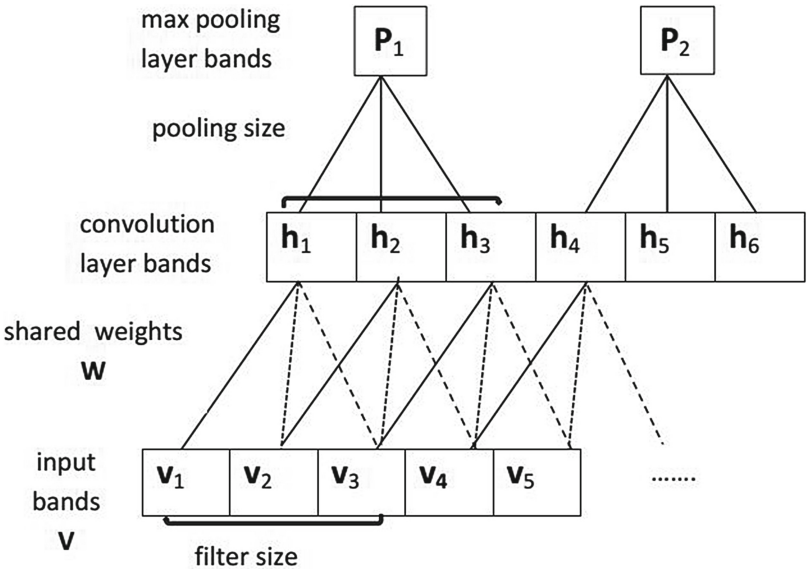

Fig. 2. A typical convolutional network architecture consisting of a convolutional and max-pooling layer. In this figure, weights with the same line style are shared across all convolutional layer bands with non-overlapping pooling.

One or more fully connected hidden layers are added on top of the final CNN layer in order to combine the features across all frequency bands. The Softmax layer interprets network output scores as conditional probabilities, for each class label. A typical convolutional neural network architecture [1] is shown in Fig. 2.

6 Experimental Setup

6.1 Data Sets

For training the acoustic model, a 10-hour-data set is used. There are 102 speakers with 3,530 utterances in the training set in total. For evaluating the performance, about 32-minute-test set is applied and there are 8 speakers (5 females and 3 males) with 193 utterances.

Table 5 shows the detailed information of the data sets.

Table 5. Train data and test data used in the experiment

Data	Size	Speakers			Utterance	UniqueWords
		Female	Male	Total		
TrainSet	10 h	79	23	102	3,530	2,590
TestSet	31 min 55 s	5	3	8	193	785

6.2 Word-Based and Syllable-Based ASR

Two ASR experiments are performed and they are word-based and syllable-based ASRs.

For the word-based ASR, training data and testing data are segmented by word units. The language model is trained from the word segmented training data and lexicon is also constructed at word level. The experiments show the ASR performance at the word level. WER is the typical ASR evaluation criterion and it is used to evaluate the word-based ASR performance.

In order to build the syllable-based ASR, all training data and testing data are segmented by the syllable units using the syllable breaking tool[1]. And then, the language model is built from syllable segmented training data. Lexicon is prepared with syllable units. For the syllable-based ASR, another alternative evaluation criterion, SER, is used to evaluate the ASR performance based on syllable units.

6.3 Language Model

The language model (LM) is built by using the transcription of all training speech with SRILM [17] language modeling toolkit. For both word-based and syllable-based ASR, 3-gram language model is used.

[1] https://github.com/ye-kyaw-thu/sylbreak.

6.4 Features and Acoustic Model

All the experiments are performed using Kaldi [18] speech recognition toolkit. In the experiments, three types of acoustic model are built and they are baseline (GMM-HMM), DNN and CNN. The Kaldi pitch tracker [19] is used to extract tone related pitch features which are augmented features to acoustic model. In the features, there are 3-dimensional pitch features (normalized pitch, delta-pitch, and voicing features).

GMM-HMM - As input features, mel frequency cepstral coefficients (MFCC) with delta and delta features with Cepstral Mean and Variance Normalization (CMVN). And then, Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) are applied to 9-splice frames and project to a 40-dimension feature. Then, Speaker Adaptive Training (SAT) is performed on LDA+MLLT model. It has 2,095 context dependent (CD) triphone states and an average of 14 Gaussian components per state.

DNN - 40-dimensional log mel-filter bank features with per-speaker mean and variance normalization are used as input features. No pre-training is used. There are 4 layers with 1,024 units per hidden layers.

CNN - The input features are the same with DNN. There are two convolution layers, one pooling layer and two fully connected layers with 1,024 hidden units. In the experiment, 1-Dimensional convolutional network is used. The size of the pooling is 3 and max pooling is used. There are 128 feature maps in the first convolutional layer and 256 feature maps in the second convolutional layer. 8×3 filter size is used and no pre-training is applied for the CNN experiment.

6.5 Tones Clustering Using Phonetic Decision Trees

The basic idea of using decision tree is to find similar triphones, and share the parameters between them. One of the biggest advantages of the decision trees is that they do not limit the scope of the questions in any way. This makes it possible to incorporate different sources of information in the decision process [20]. In Kaldi, the questions used to build the decision tree are generated automatically based on the tree clustering of the phones. Because of the flexible structure of Kaldi framework, extra questions for linguistic knowledge are allowed to add further tuning for a particular language. Therefore, two simple set of questions about tones are added for Myanmar language. They are questions to group phonemes with the same base vowel together and questions to group phonemes with the same tone together. Table 6 shows two examples of the tonal extra questions to incorporate into the decision tree.

Table 6. Example of the tonal extra questions to incorporate into decision tree

Same base vowel	Same tone
a. a a: a'	a. i. ei. an. e. in. o. ou. u. un.
ei. ei ei: ei'	a i ei an e in o ou u un

7 Evaluation Result

In this section, three experiments are used to evaluate the ASR performance.

Experiment1 (Exp1)

For the Experiment1, non-tonal dictionary, standard MFCC features for GMM model, and log Mel-filter bank (FBank) features for DNN and CNN are applied.

Experiment2 (Exp2)

The augmenting of pitch into acoustic features of each model and non-tonal dictionary are used in this setup.

Experiment3 (Exp3)

The last experiment, Experiment3 is with pitch features, tonal dictionary and the incorporation of tones into the phonetic decision tree.

Table 7. Evaluation of word-based ASR performance over tone and pitch features

Models	WER%		
	Exp1	Exp2	Exp3
GMM-HMM	46.08	44.92	42.33
DNN	45.04	42.14	39.89
CNN	**40.39**	**39.69**	**37.96**

In this experiment, the effect of pitch and tone features are explored for word-based ASR performance. Table 7 shows that the evaluation result over tone and pitch features. The lowest WER is shown in highlighted for each experiment. Without using tonal dictionary, Exp1 achieves WER about 46.08% for GMM model, 45.04% for DNN model and 40.39% for CNN model. As a result, CNN gains 5.69% and 4.65% absolute improvement over GMM and DNN. After augmenting the pitch features into each model, Exp2 gets WER reduction on 1.16%, 2.9% and 0.7% over GMM, DNN and CNN of Exp1. These results show that pitch features give better accuracy on all three models than without using it. Exp3 with pitch, tonal dictionary and the incorporation of tones to the decision tree achieves better results (3.75% and 2.59% absolute) on the conventional GMM model than Exp1 and Exp2. In comparison with Exp1 and Exp2, it notably improves the accuracy of 5.15% and 2.25% absolute in the DNN and 2.43% and 1.73% absolute in the CNN model. Using the tone and pitch features, it achieves the best WER of 37.96% with CNN.

Table 8 shows the syllable-based ASR performance over tones and pitch features. Exp1 without using tones and pitch information, it achieves SER about 34.96% on GMM, 34.63% on DNN, and 29.50% on CNN. For all models, it is observed that SER reduces greatly when the tones and pitch information is

Table 8. Evaluation of syllable-based ASR performance over tone and pitch features

Models	SER%		
	Exp1	Exp2	Exp3
GMM-HMM	34.96	33.98	31.44
DNN	34.63	32.79	29.47
CNN	**29.50**	**27.84**	**27.24**

used and the lowest SER is also highlighted. For the GMM-based model with Exp3, it achieves 3.52% and 2.54% absolute better than Exp1 and Exp2. Exp3 of DNN-based model gains better accuracy of 5.16% and 3.32% over the other two experiments. It shows that CNN-based model with tones and pitch information gets SER reduction of 2.26% and 0.6% than Exp1 and Exp2. Among the three different models, CNN is the best with 2.23% and 4.2% absolute better than DNN and GMM.

According to the Tables 7 and 8, it can be clearly seen that both tonal information and pitch features are important to improve ASR performance in Myanmar language.

8 Analysis of the Results of Tone Recognition

In this section, analysis results for four types of tones are discussed. The analysis is done to improve the future ASR performance. The tone error confusion matrices for both word and syllable-based ASR are shown in Tables 9 and 10. The highlighted values are the highest and the lowest accuracy of the tones. According to the Tables 9 and 10, it is observed that the overall accuracy, 88.71%, is achieved for word-based ASR and for syllable-based ASR, the average accuracy, 95.23%, is obtained. Tone1, which is the highest percentage of distribution among all tones, gets the best accuracy, 91.25% for the word-based model, and 97.25% for the syllable-based model. The lowest accuracy is got with Tone4, which is the least frequent tones. With respect to Tone4, there are about 87.28% accuracy in the word-based ASR and 93.90% in the syllable-based ASR.

Table 9. Confusion matrix of tone recognition for word-based ASR

	Tone1	Tone2	Tone3	Tone4	Accuracy %
Tone1	3,472	123	151	59	**91.25**
Tone2	141	1,786	82	36	87.33
Tone3	156	55	2,173	58	88.98
Tone4	40	43	78	1,105	**87.28**
	Overall accuracy %				88.71

Table 10. Confusion matrix of tone recognition for syllable-based ASR

	Tone1	Tone2	Tone3	Tone4	Accuracy%
Tone1	5,683	63	72	26	**97.25**
Tone2	83	2,566	42	12	94.93
Tone3	107	53	3,550	34	94.82
Tone4	31	31	37	1,524	**93.90**
	Overall accuracy %				95.23

Among the tone confusion pairs, the pairs of Tone1 with Tone3, Tone3 with Tone1, and Tone2 with Tone1 are most of the confused tone pairs. This is due to the speaking styles of news presenters. For example, the tone confusion pairs of Tone1 with Tone3 are ကာ (k a) is with က (k a.), တယ် (t e) is with တဲ့ (t e.), etc. Similarly, some of the example confusion pairs of Tone3 with Tone1 are အ (a.) is with အာ (a), အီ (i.) is with အီ (i), etc. The confusion tone pairs of Tone2 with Tone1 are အိုး (ou:) is with အို (ou), အား (a:) is with အာ (a), etc.

The analysis will be taken into account for the future research to improve the accuracy of all tones.

9 Conclusion

In this work, the effect of tone and pitch features was explored using state-of-the-art acoustic modeling technique, CNN, at both syllable and word levels of Myanmar ASR. It clearly shows that CNN notably outperforms DNN for tonal language like Myanmar. Moreover, the addition of tones into the phoneme set and acoustic features, and using the tonal extra questions into the building of phonetic decision tree are proved that they help to improve the ASR performance for Myanmar language. Using tone and pitch features, with GMM, DNN, and CNN-based models, better accuracy of 42.33%, 39.89% and 37.96% are achieved at the word level, and 31.44%, 29.47% and 27.24% are gained at the syllable level. As a result, in comparison with DNN, word-based CNN offers nearly 2% improvement and syllable-based CNN gains over 2% better accuracy with respect to the tone and pitch features. Furthermore, the CNN model with tone information gets WER of 2.43% or SER of 2.26% reductions than without using it.

For future work, more experiments will be done on CNN architecture using different parameters (number of feature maps, number of hidden units, number of convolutional layers, etc.) and pooling techniques. Moreover, more tonal features will be explored for Myanmar language.

References

1. Sainath, T.N., Mohamed, A., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for LVCSR. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, pp. 8614–8618, 26–31 May 2013
2. Sainath, T.N., Kingsbury, B., Mohamed, A., Dahl, G.E., Saon, G., Soltau, H., Beran, T., Aravkin, A.Y., Ramabhadran, B.: Improvements to deep convolutional neural networks for LVCSR, In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, pp. 315–320, 8–12 December 2013
3. Sainath, T.N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A., Dahl, G.E., Ramabhadran, B.: Deep convolutional neural networks for large-scale speech tasks. Neural Netw. **64**, 39–48 (2015)
4. Sercu, T., Puhrsch, C., Kingsbury, B., LeCun, Y.: Very deep multilingual convolutional neural networks for LVCSR. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, pp. 4955–4959, 20–25 March 2016
5. Hu, X., Saiko, M., Hori, C.: Incorporating tone features to convolutional neural network to improve Mandarin/Thai speech recognition. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014, Chiang Mai, pp. 1–5, 9–12 December 2014
6. Nguyen, V.H., Luong, C.M., Vu, T.T.: Tonal phoneme based model for Vietnamese LVCSR. In: 2015 International Conference Oriental COCOSDA Held Jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Shanghai, pp. 118–122, 28–30 October 2015
7. Naing, H.M.S., Hlaing, A.M., Pa, W.P., Hu, X., Thu, Y.K., Hori, C., Kawai, H.: A Myanmar large vocabulary continuous speech recognition system. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, (APSIPA 2015), Hong Kong, pp. 320–327, 16–19 December 2015
8. Htun, U.T.: Acoustic Phonetics and the Phonology of the Myanmar Language. School of Human Communication Sciences, La Trobe University, Melbourne (2007)
9. Luong, H.T., Vu, H.Q.: A non-expert Kaldi recipe for Vietnamese speech recognition system. In: Proceedings of WLSI/OIAF4HLT, Osaka, pp. 51–55, 12 December 2016
10. Mon, A.N., Pa, W.P., Thu, Y.K.: Building HMM-SGMM continuous automatic speech recognition on Myanmar web news. In: Proceedings of 15th International Conference on Computer Applications (ICCA 2017), pp. 446–453 (2017)
11. Soe, W., Thein, Y.: Syllable-based speech recognition system for Myanmar. Int. J. Comput. Sci. Eng. Inf. Technol. (IJCSEIT) 1–13 (2015)
12. Nwe, T.T., Myint, T.: Myanmar language speech recognition with hybrid artificial neural network and Hidden Markov Model. In: Proceedings of 2015 International Conference on Future Computational Technologies (ICFCT 2015), Singapore, pp. 116–122, 29–30 March 2015
13. Khaing, I.: Myanmar continuous speech recognition system based on DTW and HMM. Int. J. Innov. Eng. Technol. (IJIET) **2**(1), 78–83 (2013). ISSN 2319-1058
14. Pa, W.P., Thu, Y.K., Finch, A., Sumita, E.: Word boundary identification for Myanmar text using conditional random fields. In: Zin, T.T., Lin, J.C.-W., Pan, J.-S., Tin, P., Yokota, M. (eds.) GEC 2015. AISC, vol. 388, pp. 447–456. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-23207-2_46

15. Thu, Y.K., Pa, W.P., Finch, A., Ni, J., Sumita, E., Hori, C.: The application of phrase based statistical machine translation techniques to Myanmar grapheme to phoneme conversion. In: Computational Linguistics - 14th International Conference of the Pacific Association for Computaitonal Linguistics, PACLING 2015, Bali, Revised Selected Papers, pp. 238–250, 19–21 May 2015
16. Myanmar-English Dictionary: Department of the Myanmar Language Commission. Yangon, Ministry of Education, Myanmar (1993)
17. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit, pp. 901–904 (2002)
18. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB, December 2011
19. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, pp. 2494–2498, 4–9 May 2014
20. Hirsimäki, T.: A review: decision trees in speech recognition, 27 May 2003