10th ISCA Workshop on Speech Synthesis (SSW 10)
20-22 September 2019, Vienna, Austria

Enhancing Myanmar Speech Synthesis with Linguistic Information and LSTM-RNN

Aye Mya Hlaing[1], *Win Pa Pa*[1], *Ye Kyaw Thu*[2,3]

[1]Natural Language Processing Lab., University of Computer Studies, Yangon, Yangon, Myanmar
[2]Language and Speech Science Research Lab., Waseda University, Tokyo, Japan
[3]Language and Semantic Technology Research Team, National Electronics and Computer Technology Center, PathumThani, Thailand

{ayemyahlaing, winpapa}@ucsy.edu.mm, wasedakuma@gmail.com

Abstract

Recently, Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) has become an attractive architecture in speech synthesis for its ability to learn long time-dependencies. Contextual linguistic information is an important feature for naturalness in speech synthesis and using that feature in various speech synthesis models improves the quality of the synthesized speeches for languages. In this paper, LSTM-RNN was applied in Myanmar speech synthesis, and the importance of contextual linguistic features and the effect of applying explicit tone information in different architectures of LSTM-RNN was examined using our proposed Myanmar question set. Experiments of LSTM-RNN, and a hybrid system of DNN and LSTM-RNN, i.e., four feedforward hidden layers followed by two LSTM-RNN layers, were done on Myanmar speech synthesis and compared with the baseline DNN. Both objective and subjective evaluations show that the hybrid of DNN and LSTM-RNN system gives more satisfiable synthesized speeches for Myanmar language than the LSTM-RNN and baseline DNN systems.

Index Terms: Long Short-Term Memory, LSTM, Myanmar speech synthesis, Myanmar Text to Speech, Linguistic feature, Question set

1. Introduction

The goal of text-to-speech (TTS) system is to generate a naturally sounding speech waveform for given input text. Recently, neural networks have been applied as acoustic models for statistical parametric speech synthesis (SPSS). Zen el at. proposed an approach which uses Deep Neural Network (DNN) to model the relationship between input features and their acoustic realizations [1]. The various training aspects of DNN as a generation model for TTS were investigated in [2]. However, one limitation of the feed-forward DNN-based acoustic modeling is that the sequential nature of speech is ignored [3]. Recurrent Neural Networks (RNNs) were applied for modeling sequential data that embodies correlations between consecutive frames in speech. However, the standard RNNs has the problem that the influence of a given input on the hidden layer either decays or blows up exponentially around the networks recurrent connections [4]. To overcome this vanishing gradient problem, the most effective solution so far is Long Short-Term Memory (LSTM) architecture [5]. LSTM is the most widely used RNN in speech processing because LSTM is capable of learning long time-dependencies [6].

In [7], RNNs with bidirectional Long Short-Term Memory (BLSTM) were adopted to capture the correlation information between any two frames in a speech utterance. The unidirectional LSTM RNNs with a recurrent output layer was proposed to apply acoustic modeling for SPSS to achieve low-latency speech synthesis in [3]. In [8], several variants of LSTM were examined and the forget gate and cell state of the LSTM were analyzed. Recent studies demonstrated that LSTMs can achieve significantly better performance on SPSS than DNN.

Little research has been performed for speech synthesis on Myanmar language former known as Burmese. Only three SPSS based papers on Myanmar speech synthesis are found publicly: HMM-based Myanmar TTS [9], CART-based Myanmar TTS [10], and DNN-based Myanmar speech synthesis [11]. In [9], the first HMM-based Myanmar TTS was operated at the syllable level and word information was used in CART-based Myanmar TTS in [10]. In [11], more contextual information was applied in Myanmar speech synthesis.

In this work, LSTM-RNN was applied in Myanmar speech synthesis to improve the naturalness of synthesized speech, and compared with DNN-based speech synthesis. The comparisons of LSTM-RNN architectures for Myanmar speech synthesis were experimented, and the detail analysis on the aspects of using linguistic features on LSTM-RNN based Myanmar speech synthesis was also conducted. As LSTM-RNN achieves better results on speech synthesis of other languages [3, 7, 12], we want to analyze whether it can get more natural synthesized speech for Myanmar speech synthesis. To the best of our knowledge, this is the first attempt to apply LSTM-RNN architecture in Myanmar speech synthesis.

The rest of this paper is organized as follows. Section 2 presents extracting linguistic information for Myanmar language and Section 3 describes LSTM-RNN based speech synthesis. Section 4 presents experimental setup of different network architectures for Myanmar speech synthesis and Section 5 reports the evaluation results on all experiments. Some issues of the performance of LSTM-RNN based speech synthesis for Myanmar language are discussed in Section 6 and Section 7 describes the conclusion.

2. Linguistic feature extraction for Myanmar language

The general speech synthesis architecture of Festival[1] was used for extracting contextual information from utterances. However, there is no phoneme features file and lexicon for Myanmar language in Festival. Therefore, we prepared phoneme features

[1]http://www.cstr.ed.ac.uk/projects/festival/

10.21437/SSW.2019-34

for consonants such as consonant type, place of articulation, consonant voicing, and lip rounding and phoneme features for vowels such as vowel frontness, vowel height, tone, and nasality [10]. Standard Myanmar phonemes and extended phonemes for foreign words [13] were used in this work. Myanmar pronunciation lexicon with syllable information was prepared because syllable is the basic sound unit bearing tone information in Myanmar language [10]. Many contextual labels reported in [11] formatted as HTS-style labels[2] have been extacted for Myanmar language.

A question set is used in linguistic feature extraction for DNN and LSTM-RNN based speech synthesis and it is also language dependent requirement. There is no publicly available Myanmar question set for linguistic feature extraction. We proposed Myanmar question set in [11] and it was used for extracting linguistic features. In Myanmar language, tone is the integral part of the pronunciation of syllable and can affect the meaning of that syllable. There are four types of tones in Myanmar language and prosodic features such as fundamental frequency and duration can be influenced by the tone type of the syllable. Therefore, questions about explicit tone information has been used in Myanmar question set although tone information is already included in the grapheme of the syllable. The updated Myanmar question set including 635 questions (622 phoneme questions and 13 related positional questions) was applied in linguistic feature extraction for Myanmar speech synthesis.

3. LSTM-RNN based speech synthesis

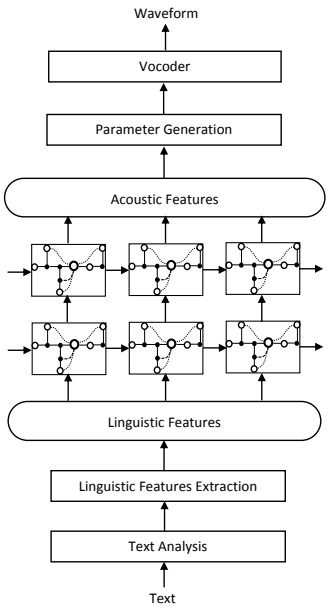

Figure 1: *A schematic diagram of LSTM-RNN based speech synthesis*

Figure 1 illustrates the schematic diagram of LSTM-RNN architecture for speech synthesis. In LSTM-RNN based speech synthesis, input features are extracted from contextual labels generated by text analysis phase. Input features includes binary features for categorical contexts (e.g. phoneme identity, tone type of the syllable) and numerical features for numerical contexts (e.g. the number of syllables in the word). The output features are acoustic features like spectral and excitation parameters, and their dynamic features. For training LSTM-RNNs, input features and output acoustic features can be force aligned frame-by-frame by HMMs in advance. The weights of LSTM-RNN are initialized randomly and then they are updated to minimize the mean squared error between the target features and predicted output features. At the synthesis time, the input feature vectors are extracted from the text analysis and then mapped to output acoustic vectors by a trained LSTM-RNN. The output acoustic features are used with the speech parameter generation algorithm. Finally, the vocoder outputs a synthesized waveform according to the given speech parameters.

[2] http://www.cs.columbia.edu/~ecooper/tts/lab_format.pdf

4. Experiments

4.1. Experimental setups

Myanmar phonetically balanced corpus (PBC) [9] built from Basic Travel Expression Corpus (BTEC) [14] was employed for building all speech synthesis for Myanmar language. The speech data was downsampled from 48kHz to 16kHz sampling. Myanmar PBC was divided into three subsets: 3,800 utterances for training, 100 utterances for development, and 100 utterances for testing. All sets are disjoint.

The proposed question set was used for extracting input linguistic features for Myanmar language. WORLD [15] vocoder was used to extract 60-dimensional Mel-Cepstral Coefficient (MCCs), 5-dimentional band aperiodicities (BAPs), and logarithmic fundamental frequencies (log F_0) at 5 msec frame step. A binary voiced/unvoiced feature was used for voicing information. Input linguistic features were min-max normalized to the range of [0.01, 0.99], and acoustic features were mean-variance normalized before training. Maximum likelihood parameter generation (MLPG) was applied to generate smooth parameter trajectories at generation time. Merlin speech synthesis toolkit [16] with Keras [17] python library was applied for modeling all systems on K80 GPU for training. DNN-based speech synthesis [11] was used as the baseline in this paper.

4.2. Network Architectures

The following network architectures of speech synthesis systems were used in our experiments:

1. DNN : a baseline system with six feedforward hidden layers of 1024 hyperbolic tangent units each
2. LSTM-1L : a single hidden layer with LSTM-RNN (512 units)
3. LSTM-2L : two hidden layers with LSTM-RNN (512 units each)
4. Hybrid-LSTM-1L : a hybrid of DNN and LSTM-RNN, five feedforward hidden layers of 1024 hyperbolic tangent units each, followed by a single LSTM-RNN layer with 512 units
5. Hybrid-LSTM-2L : a hybrid of DNN and LSTM-RNN, four feedforward hidden layers of 1024 hyperbolic tangent units each, followed by two LSTM-RNN layers with 512 units each

According to our preliminary results, we found that LSTM-RNN hidden layers with 512 units gave better objective results

than that with 256 and 1024 units. Therefore, LSTM-RNN hidden layers with 512 units have been used in all experiments. Silence frames were removed from the training data for avoiding overlearning silence labels in acoustic modeling. The weights of all LSTM-RNNs were initialized randomly and then they were updated to minimize mean squared error (mse) between target and predicted output features. Stochastic gradient descent (sgd) based learning rate scheduling was used for all hybrid systems and Adam optimizer [18] was used for LTSM-1L and LSTM-2L. Exact LSTM gradient with untrancated Back-propagation Through Time (BPTT) [4] was applied for training LSTM-RNNs. All systems were trained with batch size of 25 sentences. Hyperparameters for each system were optimized on the development set. Fixed momentum was used and learning rates were tuned in these systems. A linear activation function was used at the output layer for all systems.

5. Evaluation

The quality and naturalness of synthesized speeches generated by the systems described in Section 4.2 are evaluated in terms of objective and subjective measures.

5.1. Objective Evaluation

Objective results are used to measure the quality of synthesized speech in terms of distortions between the synthesized speech and natural speech of the original speaker. The objective measures are Mel-Ceptral Distortion (MCD) in dB, distortion of band aperiodicities (BAP) in dB, F_0 distortion in root mean square error (RMSE), and voiced/unvoiced error (V/UV) in percentage. The lower is the better.

5.1.1. Effect of contextual linguistic information

We analyzed the effect of contextual linguistic information on all LSTM-RNN architectures. As the LSTM-RNNs can access the past contextual information through their recurrent connections, the effect of preceding two contextual information on modeling all LSTM-RNN based speech synthesis systems was experimented. Figure 2 and 3 depict the comparisons of MCD and F_0 RMSE using C_635 and C_423 on all LSTM-RNN architectures for Myanmar speech synthesis respectively. C_635 refers 635 input linguistic features including current context, and preceding and succeeding two contexts at phoneme, syllable, word, and utterance levels. C_423 refers 423 input linguistic features including only current context, and succeeding two contexts at these levels. In this case, tone information is also included in contextual linguistic features of C_635 and C_423. 9 numeric features for frame related features are also used for all experiments. C_635 and C_423 are extracted by applying the proposed Myanmar question set. Figure 2 shows that applying C_635 on all architectures gets better prediction on Mel-Cepstrum than applying C_423. In Figure 3, all architectures applied C_635 except LSTM-1L get better F_0 RMSE than that applied C_423. These objective results confirm that preceding contextual information is still important for modeling LSTM-RNN based speech synthesis.

5.1.2. Effect of explicit tone questions in Myanmar question set

Though tone information is already included in the grapheme of vowels in Myanmar language, explicit tone information was added in the input linguistic features by applying questions about tone types of vowels in Myanmar question set. Comparisons of tone information and no tone information on modeling LSTM-RNN based speech synthesis were experimented. In this experiments, all the systems with tone information use C_635 input features. Using explicit tone information in modeling Myanmar speech synthesis give better MCD on all network architectures in the experiments according to Figure 4. As shown in Figure 5, all architectures modeling with explicit tone information except LSTM-1L get better F_0 RMSE than no explicit tone information. In general, we can conclude that explicit tone questions in Myanmar question set are useful for modeling Myanmar speech synthesis.

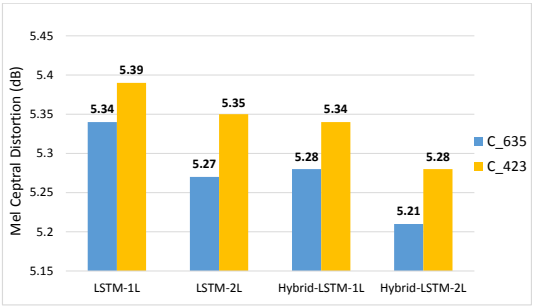

Figure 2: *Effect of left contextual information on MCD*

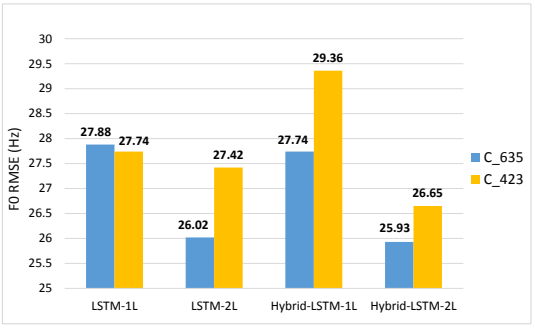

Figure 3: *Effect of left contextual information on F_0*

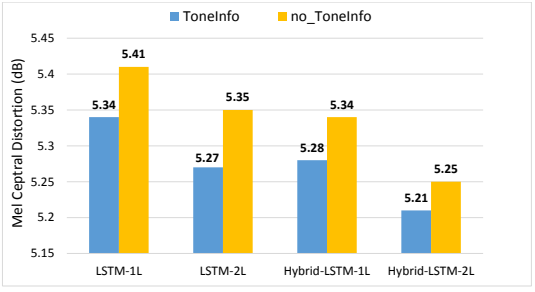

Figure 4: *Effect of explicit tone information on MCD*

5.1.3. Objective results of different network architectures

Table 1 presents the objective results of different network architectures for Myanmar speech synthesis. C_635 for contextual linguistic features and 9 numerical features for frame related features were applied in the experiments. It is observed that all LSTM-RNN based speech synthesis systems

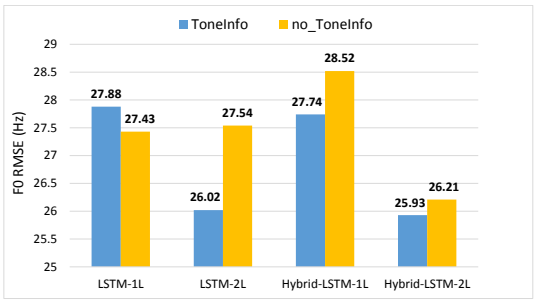

Figure 5: *Effect of explicit tone information on F_0*

achieve better objective results than the baseline DNN except BAP distortion of LSTM-1L and Hybrid-LSTM-1L. It shows that LSTM-2L objectively outperforms LSTM-1L across all objective measures, and Hybrid-LSTM-2L gets better objective results than Hybrid-LSTM-1L in terms of MCD, BAP, and F_0 RMSE. These results confirm that two hidden layers of LSTM-RNNs can give better performance over single hidden layer of LSTM-RNN. In particular, MCD of Hybrid-LSTM-2L architecture decreases 0.15(dB) from that of the baseline DNN, and F_0 RMSE of Hybrid-LSTM-2L 25.93(Hz) is significantly better than that of DNN 31.23(Hz). Hybrid-LSTM-2L is the best network architecture for Myanmar speech synthesis in our experiments.

Table 1: *Comparison of objective results for all network architectures for Myanmar speech synthesis*

	MCD (dB)	**BAP** (dB)	**F_0 RMSE** (Hz)	**V/UV** (%)
DNN (baseline)	5.36	0.21	31.23	5.47
LSTM-1L	5.34	0.21	27.88	5.31
LSTM-2L	5.27	0.20	26.02	5.26
Hybrid-LSTM-1L	5.28	0.21	27.74	**5.06**
Hybrid-LSTM-2L	**5.21**	**0.20**	**25.93**	5.16

5.2. Subjective Evaluation

The performance of DNN, LSTM-2L, and Hybrid-LSTM-2L systems was subjectively evaluated by perceptual tests. 30 utterances were randomly selected from the evaluation set and open domain, internet data. These utterances were synthesized by the baseline DNN, LSTM-2L, and Hybrid-LSTM-2L systems. Three AB preference tests (DNN vs. LSTM-2L, DNN vs. Hybrid-LSTM-2L, and LSTM-2L vs. Hybrid-LSTM-2L) were participated by 20 non-expert native speakers of age range from 20 to 40 years. The synthetic speeches were presented in random order in each pair of all three tests. Subjects were given 30 pairs of synthesized speeches and asked to choose the more natural one in each pair or "Neutral" if the difference between two speech samples cannot be perceived.

The scores of three AB preference tests with 95% confidence intervals are presented in Figure 6, 7, and 8. The higher preference scores on LSTM-2L and Hybrid-LSTM-2L over the baseline DNN can also be seen clearly in the figures 6 and 7. They confirm that LSTM-RNN based systems can generate more natural synthesized speech than DNN based system.

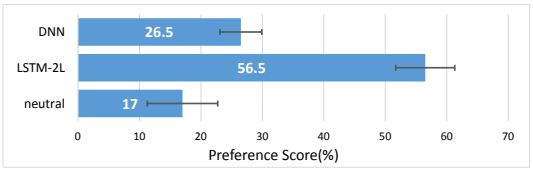

Figure 6: *Preference scores with 95% confidence intervals for DNN vs. LSTM-2L*

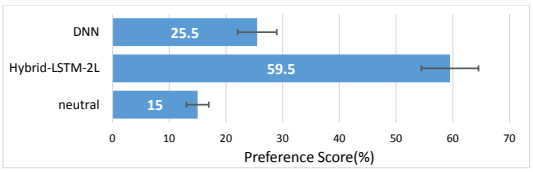

Figure 7: *Preference scores with 95% confidence intervals for DNN vs. Hybrid-LSTM-2L*

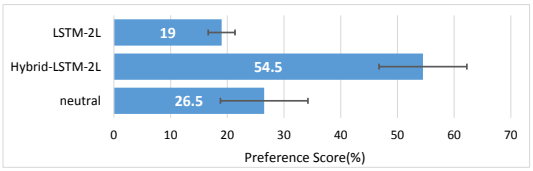

Figure 8: *Preference scores with 95% confidence intervals for LSTM-2L vs. Hybrid-LSTM-2L*

Again, the two LSTM-RNN based systems are compared in Figure 8 by the preference score and here, the performance of Hybrid-LSTM-2L is obviously preferred over LSTM-2L by the native listeners. According to the three preference tests, it can be concluded that the naturalness of Hybrid-LSTM-2L system is highly preferred than that of DNN and LSTM-2L.

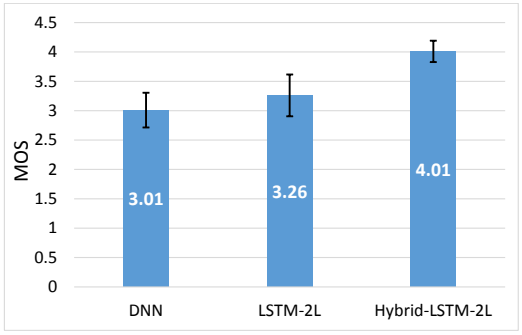

Figure 9: *Mean Opinion Scores (MOS) with 95% confidence intervals of DNN, LSTM-2L, and Hybrid-LSTM-2L*

The naturalness of the synthesized speeches generated by DNN, LSTM-2L, and Hybrid-LSTM-2L systems were further evaluated in terms of Mean Opinion Score (MOS), to confirm the results from the preference tests whether they give the same conclusion. The same 20 subjects from AB preference tests were also used in the MOS test. It is the subject to rate the naturalness of synthesized speeches on a scale from 1 to 5 where

1 is bad and 5 is excellent. The scores of DNN, LSTM-2L, and Hybrid-LSTM-2L are shown in Figure 9 with 95% confidence intervals of MOS results by the error bars. The LSTM-RNN based systems give higher MOS scores than the baseline DNN, and the Hybrid-LSTM-2L has the best result among all. Some samples of synthesized speeches generated by these systems are available for listening on here[3].

All AB preference tests and MOS test confirmed that LSTM-RNN based systems offer better performance than the baseline DNN, and furthermore, Hybrid-LSTM-2L outperform both DNN and LSTM-2L in terms of naturalness. It can be observed that the preference on Hybrid-LSTM-2L achieved the highest score not only in terms of objective but also subjective evaluation.

6. Discussion

It can be noticed that though LSTM-2L and Hybrid-LSTM-2L have only a slight difference in objective results, their subjective scores are notably different. In particular, the difference of MCD between two systems is only 0.06(dB) and the difference of F_0 RMSE is only 0.09(Hz). However, the difference of MOS results between two systems (0.75) is relatively high. The occurrence of breath pauses insertion in wrong places in LSTM-2L is more than that of DNN and Hybrid-LSTM-2L, made LSTM-2L to be less preferred by the listeners.

270 synthesized speeches (100 each from development and evaluation sets, and 70 from open internet data) were inspected on DNN, LSTM-2L, and Hybrid-LSTM-2L systems. It is found that LSTM-RNN based speech synthesis can reduce half of incorrect pronunciation of tones over DNN based speech synthesis. Better prediction of F_0 by LSTM-RNN contributed to the more natural synthesized speech of Myanmar speech synthesis in addition to better prediction of other factors (MCD, BAP, V/UV).

7. Conclusions

In this paper, the use of LSTM-RNN architecture for Myanmar speech synthesis has been investigated. The effect of contextual linguistic features extracted by using proposed Myanmar question set on LSTM-RNN based speech synthesis was explored and it shows that the preceding contextual information and explicit tone information are still important for modeling LSTM-RNN based speech synthesis though it has the ability of accessing past information through their recurrent connections. Both objective and subjective results confirm that LSTM-RNN based systems outperform DNN based system and the hybrid of DNN and LSTM-RNN offers more suitable network architecture for Myanmar speech synthesis in naturalness.

From this research work, it can be clearly concluded that the importance of correct phrase break makes the system to be more preferred. Therefore, using phrase break features in the network architecture would be our future work for better naturalness of Myanmar speech synthesis.

8. References

[1] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

[2] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3829–3833.

[3] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4470–4474.

[4] A. Graves, "Supervised sequence labelling," in *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 5–13.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[7] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Z. Wu and S. King, "Investigating gated recurrent neural networks for speech synthesis," *arXiv preprint arXiv:1601.02539*, 2016.

[9] Y. K. Thu, W. P. Pa, J. Ni, Y. Shiga, A. Finch, C. Hori, H. Kawai, and E. Sumita, "Hmm based myanmar text to speech system," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] A. M. Hlaing, W. P. Pa, and Y. K. Thu, "Word-based myanmar text-to-speech with clustergen," in *The 16th International Conference on Computer Applications (ICCA2018)*, 2018, pp. 203–208.

[11] A.-M. Hlaing, W.-P. Pa, and Y.-K. Thu, "Dnn based myanmar speech synthesis," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 142–146.

[12] X. Wang, S. Takaki, and J. Yamagishi, "A comparative study of the performance of hmm, dnn, and rnn based speech synthesis systems trained on very large speaker-dependent corpora," in *9th ISCA Speech Synthesis Workshop*, vol. 9, 2016, pp. 125–128.

[13] Y. K. Thu, W. P. Pa, F. Andrew, A. M. Hlaing, H. M. S. Naing, S. Eiichiro, and H. Chiori, "Syllable pronunciation features for myanmar grapheme to phoneme conversion," in *The 13th International Conference on Computer Applications (ICCA2015)*, 2015, pp. 161–167.

[14] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[15] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[16] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[17] F. Chollet *et al.*, "Keras: The python deep learning library," *Astrophysics Source Code Library*, 2018.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[3] `http://www.nlpresearch-ucsy.edu.mm/subeval.html`