DNN-based Myanmar Speech Synthesis

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

[1]Natural Language Processing Lab., University of Computer Studies, Yangon, Yangon, Myanmar
[2]Language and Speech Science Research Lab., Waseda University, Tokyo, Japan
{ayemyahlaing, winpapa}@ucsy.edu.mm, wasedakuma@gmail.com

Abstract

This paper presents Deep Neural Network (DNN) as the generative model for Myanmar Speech Synthesis. A question set for Myanmar language is proposed and used in context clustering of HMM-based speech synthesis and extracting input features for DNN-based speech synthesis. We investigated the effectiveness of precise state boundaries and coarse phone boundaries on aligning input linguistic features and output acoustic features for training DNN. The experimental results in objective evaluation show that the state boundary information give better result than phone boundary information in training DNN in terms of acoustic features, MCD, F0 and V/U, and the subjective listening tests confirm that DNN-based speech synthesis get a significant improvement over a conventional HMM-based speech synthesis in naturalness.

Index Terms: statistical parametric speech synthesis; hidden Markov model; deep neural network; Myanmar speech synthesis, Myanmar Text to Speech

1. Introduction

Hidden Markov model (HMM) based speech synthesis was popular in the last decade because of its flexibility in changing speaker identities, emotions, and speaking styles [1]. However, some limitations of decision tree-clustered context-dependent HMMs are highlighted by Zen, et al. [2]. One of the major factors of degrading the quality of synthesized speech is the accuracy of acoustic models and Deep Neural Network (DNN) has applied for modeling the relationship between linguistic features and acoustic features [2].

In recent years, artificial neural network-based acoustic models have become the state-of-the-art acoustic modeling in speech synthesis area. DNN can yield better synthesized speech than HMM [3]. Multi-task learning and stacked bottleneck features have employed on DNN [4]. Fan, et al. applied Recurrent Neural Networks (RNNs) with Bidirectional Long Short Term Memory (BLSTM) that can capture deep information in a sentence, to acoustic modeling for speech synthesis. Zen, et al. proposed LSTM-RNN which can access input-features up to current frame for low-latency speech synthesis [5].

Little research has been performed for speech synthesis on Myanmar language. Rule-based Myanmar Text to Speech (TTS) [6], diphone-concatenation based speech synthesis [7], Phoneme based Myanmar TTS [8], HMM-based Myanmar TTS [9], and CART-based Myanmar TTS [10] are found publicly. That HMM-based Myanmar TTS [9] operates at the syllable level. Word information is applied on building CART-based Myanmar TTS [10]. However, in these two statistical TTS systems [9, 10], a question set for context clustering has not been used. There is no publicly available question set for Myanmar language.

In this study, DNN was applied as a generative model for Myanmar speech synthesis and an investigation on the state-level and phone-level alignment between input linguistic features and acoustic features for training DNN was presented. We proposed a question set used in context clustering of HMM-based speech synthesis and extracting input features for DNN-based speech synthesis. HMM-based speech synthesis for Myanmar language was also conducted and taken as the baseline to compare with DNN-based speech synthesis.

This paper is organized as follows: Section 2 describes the introduction of Myanmar language and Section 3 outlines the data preparation for speech corpus and input text. Section 4 describes the overview of HMM-based speech synthesis and Section 5 presents the overview of DNN-based speech synthesis. Section 6 reports details of experimental setups and results. Some issues of performance analysis are discussed in Section 7 and Section 8 concludes the paper.

2. Myanmar language

Myanmar language is the official language of Myanmar, and it is spoken as the first language by 32 million people and as the second language by another 10 million people[1]. Myanmar script has 33 basic consonants, 4 basic medials, 12 basic vowels, other symbols and special characters. The consonants have only 23 distinct pronunciation because some consonants have the same pronunciation in Myanmar language.

A syllable is composed of one or more characters and one or more syllables can be formed as the word in Myanmar language. If the syllable final glottal stop is regarded as a tonal feature and the non-final neural vowel as an atonic vowel, there are four phonological tones in Myanmar [11]. Acoustic properties such as fundamental frequency (F_0) and duration can be greatly influenced by the tone type of the syllable. In Table 1, four tones are described marking on the Myanmar phoneme “ka”. The meaning of syllable is changed depending on the tone type. Figure 2 shows the fundamental frequency and duration of Myanmar phoneme “a”.

[1]https://en.wikipedia.org/wiki/Languages_of_Myanmar

Table 1: *An example of four Myanmar tones*

Tone	IPA	Phonation	Myanmar	Meaning
Tone 1	kà	Normal voice	ကာ	cover
Tone 2	ká	Breathy voice	ကား	car
Tone 3	ka̰	Creaky voice	က	dance
Tone 4	kaʔ	Final Glotal stop	ကတ်	stick

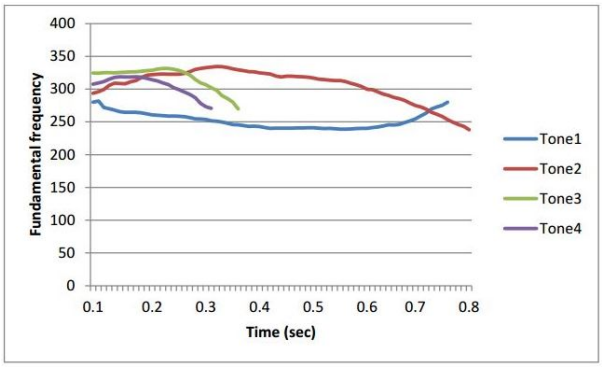

Figure 1: *An example of four tones in Myanmar phoneme "a" [12]*

3. Data preparation

3.1. Speech corpus

Myanmar phonetically balanced corpus (PBC) was built from Basic Travel Expression Corpus (BTEC) [13] by applying a greedy algorithm [14] and contained 5,276 sentences. Among them, 4,000 utterances (3.59 hrs) recorded by female speaker were used for building both HMM-based and DNN-based speech synthesis for Myanmar language. The speech data was downsampled from 48kHz to 16kHz sampling.

3.2. Question set for Myanmar language

Contextual labels and question set for context clustering are language-dependent requirements for HMM-based speech synthesis. Since Myanmar is a tonal language, tone-dependent questions such as tone types of two preceding, current and two succeeding vowels have been considered in the question set. A question set for Myanmar language was built manually by referencing Acoustic phonetics and phonology of the Myanmar language book [15] and English question set. However, Part of Speech (POS) information and intonation information such as tones and break indices (ToBI) are not included in this question set. The Myanmar question set has 810 questions including 799 phoneme questions and 11 related positional questions.

3.3. Linguistic features extraction

Festival[2] was used for extracting linguistic features for Myanmar language. Myanmar pronunciation lexicon with syllable information was prepared for extracting syllable information from the input text. Myanmar Language Commission (MLC) [16] dictionary and words from selected sentences were included in Myanmar pronunciation lexicon. Phrase Based Statistical Machine Translation (PBSMT) based grapheme to phoneme conversion [17] was applied to get the pronunciation of words from sentences. Phoneme symbols defined in [18] and phoneme features defined in [10] were used in this work.

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

3.4. Contextual information for Myanmar language

There are many contextual factors such as phone identity factors, locational factors that affect spectrum, F_0 pattern and duration [19]. For Myanmar language, the following contextual factors are taken into account in training both HMM-based and DNN-based speech synthesis:

phoneme level

- two preceding, current, two succeeding phonemes
- position of current phoneme in current syllable (forward, backward)

syllable level

- number of phonemes in preceding, current and succeeding syllable
- position of current syllable in the current word (forward, backward)
- position of current syllable in the utterance (forward, backward)
- the number of syllables before and after the current syllable in the utterance
- vowel within current syllable

word level

- number of syllables in the preceding, current and succeeding word
- position of current word in the utterance (forward, backward)
- number of words before and after the current word in the utterance

utterance level

- number of syllables in the utterance
- number of words in the utterance

4. HMM-based speech synthesis

Statistical parametric speech synthesis which uses a hidden Markov model (HMM) as its generative model is typically called HMM-based speech synthesis [1]. In the training part, spectral parameters, excitation parameters and duration are modeled in a unified framework of HMM. This part performs the maximum-likelihood estimation of the HMM parameters by using the Baum-Welch algorithm. HMM-based speech synthesis uses various linguistic contexts for the context-dependent modeling of HMMs. A decision-tree based context clustering technique is applied to distributions for spectrum, F_0 and state duration.

In the synthesis part, a given input text is first converted into a sequence of context-dependent labels. These contextual labels are used to access the decision

tree to get the context-dependent HMMs and a sentence-level HMM is constructed by concatenating context-dependent HMMs. A sequence of speech parameters including Mel-Cepstral Coefficients (MCCs) and log F_0 values including voiced/unvoiced decisions is determined so as to maximize the output probability using the speech parameter generation algorithm. Finally, a speech waveform is resynthesized directly from the generated spectral and excitation parameters by using the mel-log spectral approximation (MLSA) filter.

5. DNN-based speech synthesis

Figure 2 illustrates DNN-based speech synthesis framework with 3 hidden layers [2]. The input text is converted to a sequence of input features x_t at frame t which contain the binary features for categorical contexts (*e.g.*, is-current-phoneme-kh?) and numerical features (*e.g.*, the number of words in the utterance, the relative position of current frame in the current phoneme). The output features y_t at frame t are spectral and excitation parameters, and their dynamic features. The weights of DNN are trained by using pairs of input and output features extracted from training data.

In synthesis time, input features are extracted from the input text and then these are mapped to output features (mean and variances of speech parameter vector sequence) by the trained DNN. The speech parameter generation algorithm can generate smooth trajectories of speech parameter features which satisfy the statistics of static and dynamic features. Finally, the vocoder outputs a synthesized waveform given the speech parameters.

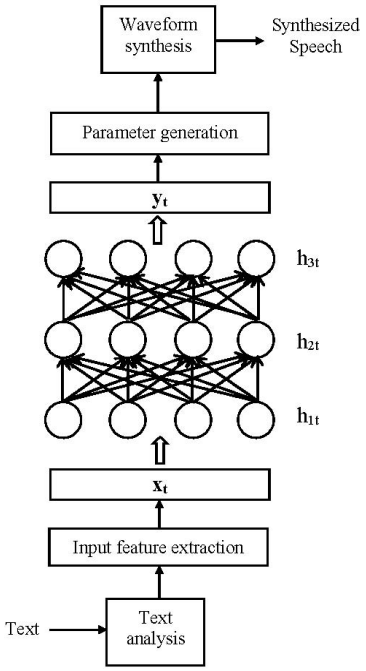

Figure 2: *A DNN-based speech synthesis framework with three hidden layers, h_{ij} denotes activation at i^{th} layer at j^{th} frame*

6. Experiments

6.1. Experimental setups

The architecture of the DNNs was 6-hidden layers, 1024 units per layer. The tangent or tanh function was used as the hidden activation function, and a linear activation function was used at the output layer. 3,800 utterances from Myanmar speech corpus were used for training, 100 utterances as development set, and 100 utterances as evaluation set. These all sets are disjoint.

For training DNNs, input linguistic features and output acoustic features are generally needed to be force aligned by HMMs in advance. DNN using state level alignment ($\mathrm{DNN_{st}}$) and DNN using phone level alignment ($\mathrm{DNN_{ph}}$) were experimented in this work. In $\mathrm{DNN_{st}}$, the input linguistic features and the output acoustic features were aligned at the precise state level. For training $\mathrm{DNN_{st}}$, HVite from HTK tools[3] was used to do forced alignment. For training $\mathrm{DNN_{ph}}$, the input and output features were aligned at phone level and used input features to indicate the coarse boundaries in a given phone and Ergodic Hidden Markov Model (EHMM) in CLUSTERGEN [20] setup was applied for doing this forced alignment.

The questions used for extracting input features from linguistic contexts were manually selected from the proposed question set of decision tree system. The input features for all DNN-based systems consisted of 645 features including 622 binary features for categorical linguistic contexts (*e.g.*, phoneme identities, tone types) and 25 numeric features for numerical linguistic contexts (*e.g.* the number of syllables in a word, the number of frames in the current phoneme). WORLD [21] was used to extract 60-dimensional MCCs, 5-dimensional band aperiodicities (BAPs) and logarithmic fundamental frequency ($\log F_0$) at 5 msec frame intervals. Input features were normalized using min-max to the range of [0.11, 0.99] and output features were normalized to zero mean and unit variance. Maximum likelihood parameter generation (MLPG) was applied to generate smooth parameter trajectories from DNN outputs and spectral enhancement post-filtering was applied to MCCs. Merlin speech synthesis toolkit [22] was used for modeling DNNs and training was done on GPU.

HMM-based system was trained on the same speech data and used standard five-state left-to-right Hidden Semi-Markov Models (HSMM) with no skip. The proposed question set for Myanmar language was used for decision tree based context clustering. Spectral envelope, fundamental frequency, and duration were modeled simultaneously by the corresponding HMMs. Decision tree state clustering used a minimum description length (MDL) factor of 1.0. Global variance (GV) enhancement and modulation spectrum-based postfilter were applied on training HMM-based system. The publicly available HTS toolkit [4] was used to implement this HMM-based speech synthesis for Myanmar language.

6.2. Objective evaluation

The quality of synthesized speech is measured objectively between the speech of the original speaker and the syn-

[3] `http://htk.eng.cam.ac.uk/download.shtml`
[4] `http://hts.sp.nitech.ac.jp`

thesized speech. Mel-cepstral distortion (MCD) in dB, F_0 distortion in root mean squared error (RMSE) and voiced/unvoiced (V/U) swapping error in percentage are used as the objective measures. Table 2 shows the results of objective measures of HMM-based system and DNN-based ones. By comparing the objective results of DNN-based systems with those of HMM-based system, the DNN-based systems outperform the HMM-based one in log F_0 prediction and V/U swapping. In particular, the RMSE of F_0 is reduced from 39.693 Hz to 31.233 Hz and V/U error rate is also reduced from 8.316% to 5.470%. On the other hand, the HMM-based system achieves a better performance in Mel-cepstrum prediction. As the comparison of DNN_{st} and DNN_{ph}, the DNN_{st} has better prediction than DNN_{ph} across all acoustic parameters. It shows that training DNN with aligned state boundaries is more efficient for generating better synthesized speech than training DNN with aligned coarse boundaries.

Table 2: *Comparison of objective results on HMM-based system and DNN-based systems*

	MCD(dB)	$\boldsymbol{F_0}$ **RMSE(Hz)**	**V/U(%)**
HMM	5.015	39.693	8.316
DNN_{st}	5.355	31.233	5.470
DNN_{ph}	5.564	32.472	6.548

6.3. Subjective evaluation

The performance of HMM-based system and DNN-based ones are further evaluated by subjective listening tests. Two AB preference tests were conducted to compare the performance of these systems. 20 utterances were randomly selected for these tests and synthesized by all systems. 24 native Myanmar people were participated in these preference tests. Each subject evaluated 20 pairs. The subjects can choose one of their preferences from three options: (1) speech sample generated by the first system is better than that of the second system, (2) speech sample generated by the second system is better than that of the first system, and (3) neural which means the difference between speeches generated by both systems cannot be perceived or difficult to judge which one is better.

Figure 3 shows the preference scores of first AB listening test. This shows the speech synthesized by the DNN-based system is significantly preferred than the HMM-based system. The preference score (87%) of the DNN-based system is higher than the HMM-based system (4%). According to the second AB listening test shown in Figure 4, the perception difference between the DNN_{st} and DNN_{ph} is not significant. Their preference scores are 19% and 21% respectively. Some samples of synthesized speech are given on the web link[5].

7. Discussion

The preference score of DNN-based synthesized speech was 83% more than that of HMM-based speech in contrary to mel-cepstral distortion of DNN-based system was 0.34 higher than that of HMM-based one as the scores

[5] `http://www.nlpresearch-ucsy.edu.mm/subtest.html`

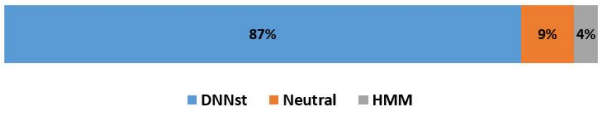

Figure 3: *The preference scores of HMM-based system and DNN-based system*

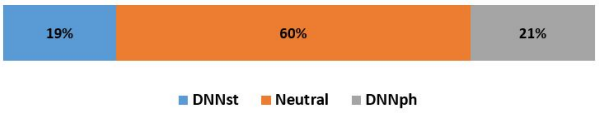

Figure 4: *The preference scores of two DNN-based systems (DNN_{st} and DNN_{ph})*

can be seen in Table 2 and Figure 3. There are some small noises in HMM-based synthesized speech and it might be the cause of listeners did not prefer HMM-based synthesized speech. 270 synthesized speeches of 100 from development set, 100 from test set and 70 from open internet data were inspected on all three systems, HMM, DNN_{st} and DNN_{ph}. Types of tone and vowel errors found in syllable-based HMM are discussed in [9], but we found there are only about 0.45% incorrect pronunciation of normal voice (Tone 1) to breathy voice (Tone 2). And there are also 3 skipped (missing) phonemes they are occurred in the synthesized speech of DNN_{ph} but it did not occur in DNN_{st}. The listening tests in the DNN-based synthesized speech did not find incorrect vowels pronunciation, while they are occurred in HMM-based synthesized speech. From the listening tests, it can be concluded that the naturalness of the DNN-based system is better than the HMM-based system.

8. Conclusion

This paper presented the analysis of HMM, DNN_{st} and DNN_{ph}. The question set for Myanmar language is proposed and used to train both HMM and DNNs. According to the listening tests, DNN-based speech synthesis is more preferable for its naturalness than HMM-based speech synthesis and more effective on the small dataset. According to the experiments, DNN_{st} can generate better Myanmar synthesized speech than DNN_{ph}. In future work, more linguistic features will be applied for DNN expecting to reduce the tone errors and for better naturalness. More experiments will be done on DNNs and LSTM-RNNs for Myanmar speech synthesis.

9. References

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, ``Speech synthesis based on hidden markov models,'' *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234--1252, 2013.

[2] 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.

[3] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, ``On the training aspects of deep neural network (dnn) for para-

metric tts synthesis,'' in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3829--3833.

[4] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, ``Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis,'' in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4460--4464.

[5] 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.

[6] K. Y. Win and T. Takara, ``Myanmar text-to-speech system with rule-based tone synthesis,'' *Acoustical science and technology*, vol. 32, no. 5, pp. 174--181, 2011.

[7] E. P. P. Soe and A. Thida, ``Diphone-concatenation speech synthesis for myanmar language,'' *International Journal of Science, Engineering and Technology Research*, vol. 2, no. 5, pp. pp--1078, 2013.

[8] C. S. Hlaing and A. Thida, ``Phoneme based myanmar text to speech system,'' *International Journal of Advanced Computer Research*, vol. 8, no. 34, pp. 47--58, 2018.

[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] U. Thein-Tun, ``The domain of tones in burmese,'' *SST 1990 Proceedings*, pp. 406--411, 1990.

[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 *International Conference of the Pacific Association for Computational Linguistics*. Springer, 2017, pp. 314--326.

[13] 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.

[14] J. Ni, T. Hirai, and H. Kawai, ``Constructing a phonetic-rich speech corpus while controlling time-dependent voice quality variability for english speech synthesis,'' in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I--I.

[15] D. T. Tun, ``Acoustic phonetics and the phonology of the myanmar language,'' *School of Human Communication Sciences, La Trobe University, Melbourne, Australia*, 2007.

[16] M. L. Commission, *Myanmar-English Dictionary*. Dunwoody Pr, 1996.

[17] Y. K. Thu, W. P. Pa, A. Finch, J. Ni, E. Sumita, and C. Hori, ``The application of phrase based tatistical machine translation techniques to myanmar grapheme to phoneme conversion,'' in *International Conference of the Pacific Association for Computational Linguistics*. Springer, 2015, pp. 238--250.

[18] 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.

[19] K. Tokuda, H. Zen, and A. W. Black, ``An hmm-based speech synthesis system applied to english,'' in *IEEE Speech Synthesis Workshop*, 2002, pp. 227--230.

[20] A. W. Black, ``Clustergen: A statistical parametric synthesizer using trajectory modeling,'' in *Ninth International Conference on Spoken Language Processing*, 2006.

[21] 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.

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