Myanmar Text-to-Speech Synthesis Using End-to-End Model

Qinglai Qin
Yunnan University
Kunming,
Yunnan Province, China
+86 18687137636
qin_qinglai@163.com

Jian Yang*
Yunnan University
Kunming
Yunnan Province, China
+86 13078704171
Corresponding author
jianyang@ynu.edu.cn

Peiying Li
Yunnan University
Kunming
Yunnan Province, China
+86 15587175540
ieLiPeiYing@163.com

ABSTRACT
In this paper, we propose a Myanmar speech synthesis system based on an End-to-End neural network model, which integrates the Myanmar phone model into the Tacotron2 End-to-End model. Based on the Seq2seq model architecture, we use phone-level embedding to form a feature prediction network from phone sequences to Mel spectrum, and combine with a semi-supervised speech generation network to generate high-quality Myanmar synthesized speech. In addition, we introduced the BERT pre-training decoder module to assist the phone feature extraction, which reduces the system's dependence on the phone feature extraction network and improve the text feature richness. Compared with other Myanmar speech synthesis systems, this method effectively improves the naturalness and accuracy of synthesized speech under low resource conditions.

CCS Concepts
•**Computing methodologies→Natural language processing**

Keywords
Myanmar Speech Synthesis; Text-to-Speech; End-to-End; Pre-Trained Language Model;

1. INTRODUCTION
Synthesizing natural human speech has always been an important research topic. Text-to-speech [1] (TTS) is a research and development hotspot for speech synthesis at this stage. Traditional speech synthesis methods can be divided into waveform splicing synthesis methods and statistical parameter-based synthesis methods [2] [3]. The former selects the appropriate speech waveform units from the speech unit library according to the text and splices the waveform units to generate speech. The speech waveform synthesized by this method is discontinuous, and the scale of the speech unit library is large, so the application of the scene is limited. The speech synthesis method based on statistical parameters is mainly divided into two parts, the front-end text analysis and the back-end speech synthesis. The front-end text analysis is to get the syllable information from the text according to the rules, such as normalization, word segmentation, and phonetic pronunciation, while the back-end focuses on training the acoustic model according to the phoneme information, including the prediction of acoustic and prosodic parameters, and synthesizes speech according to the acoustic model. The synthesis method based on statistical parameters relies on a large amount of language knowledge and human work, and the errors accumulated in the processing of various parts will affect the final synthesis quality. In recent years, speech synthesis technology has developed rapidly. The End-to-End deep learning model brings significant performance improvements to speech synthesis technology [4] [5] [6] [7]. Sequence-to-sequence feature prediction networks can generate Mel spectrograms based on text, and then generate high-quality speech through a vocoder. The reference system Tacotron2 [7] used in this system is the outstanding method of this synthesis method.

Tacotron2 can be regarded as a conditional autoregressive model, in which the condition is summarized from the text by the attention mechanism, and then trained in an End-to-End structure. This approach simplifies the steps of text processing and requires less professional language knowledge and human work. However, there are great differences in the language characteristics of each language, and the End-to-End speech synthesis system needs to be optimized according to the language characteristics of different languages, so that it can obtain useful implicit speech feature representations from the input text.

The End-to-End speech synthesis system requires an encoder to convert the input text into a hidden speech feature representation. In theory, it is possible to model a large number of symbol sets implicitly with the support of a sufficient number of nonlinear conversion layers and a sufficient number of training data [8]. However, in the actual situation, the system needs to consider the scale of the training data and the calculation cost. Therefore, according to the language characteristics of different languages, the text input of Tacotron2 will be adapted. Take English and Chinese as examples, the original Tacotron2 was designed for English. It uses character input [5] [6] and can be easily applied to other Latin alphabet languages [9]. However, when the target language is Chinese, the large size of the Chinese character set makes it impractical to use character embedding. Some reports have used Pinyin syllables as input and achieved good results [10]. There are hundreds of combinations of Pinyin in Chinese. When the Pinyin set is combined with tone, the size of the Pinyin set with tone will be increased by four to five times. This is still a larger set compared with English character input. In recent reports, in order to reduce the size of the input label set, some recent

NLPIR 2020, December 18–20, 2020, Seoul, Republic of Korea

ACM ISBN 978-1-4503-7760-7/20/06…$15.00
DOI: https://doi.org/10.1145/3443279.3443295

studies in Chinese speech synthesis systems use the phone as the input [11], reducing the size of the symbol set. Both Myanmar and Chinese belong to the Sino-Tibetan language family [14], which is a pinyin language with a large number of syllables. Using syllables as encoder input will cause the problem of sparse data under the condition of low resources. If Myanmar bytes are used as encoder input, Myanmar characters belong to multi-byte characters. However, it has been reported that in multi-byte character languages, using bytes as input will require the model to learn short-term intra-character contextual relationships and long-term contextual relationships between characters and even word phrases, which is difficult to extract hidden features from Encoder [12]. Therefore, in this paper, Myanmar syllables are represented as a combination of initial consonants and tonal vowels with reference to Chinese consonant structure, and the system is modeled with phones as encoder inputs.

Myanmar belongs to the Myanmar branch of the Tibetan-Myanmar language family, which belongs to the Sino-Tibetan language family and is the official language of the Union of Myanmar [13] [14]. The research on Myanmar speech synthesis lags behind. At present, the synthesis methods still remain in the methods of waveform splicing [15] and synthesis based on statistical parameters [16] [17]. The waveform stitching method requires a large storage capacity, and the naturalness of the synthesis method based on statistical parameters is still low, so an End-to-End Myanmar speech synthesis method is proposed in this paper. This system adopts the method of phone sequence instead of the original text to translate the complex character structure into a one-dimensional Romanized phone sequence based on time order, which is convenient for the alignment of attention mechanism. At the same time, the standard Romanized sequence also makes the pronunciation of stacked words and pronunciation change in Myanmar explicit.

The End-to-End speech synthesis system needs pairs of < text, speech > data sets. In the case of lack of training data sets, the synthetic speech quality will decline. Myanmar is a low-resource language, and electronic resources are relatively scarce. Under the condition of low resources, the End-to-End model will have low richness of text features, which will affect the final quality of speech synthesis. In order to solve this problem, this paper introduces the pre-training language model BERT[21] as another representation of text information, and uses the auxiliary phone embedding module to better extract phone features, so as to improve the quality of Myanmar speech synthesis under the condition of low resources.

2. MYANMAR SPEECH SYNTHESIS

In this section, we will first introduce the benchmark End-to-End model Tacotron2, then introduce the phone embedding method for Myanmar and the text information enhancement based on BERT pre-training language model, and finally introduce the End-to-End Myanmar speech synthesis system from the overall structure.

2.1 Tacotron2

Tacotron2 model is a typical seq2seq model, and it is one of the End-to-End speech synthesis systems that can generate natural speech. Tacotron2 is mainly composed of two parts, an encoder and a decoder with attention mechanism. The encoder is to extract fixed-length feature vectors from text sequences. Firstly, the text sequence is embedded into the continuous vector with location information, and the embedded vector is obtained through the three-layer one-dimensional convolution layer to obtain the context information. Finally, the output of the convolution layer is sent to the bidirectional LSTM [18] network to generate a fixed-dimensional feature vector.

The generated fixed-dimensional encoder features are utilized in the decoder. The decoder of Tacotron2 is an autoregressive recurrent neural network with position-sensitive attention mechanism. The eigenvector generated by the encoder is used to predict the Mel spectrum coefficient one frame at a time. Based on the position-sensitive attention mechanism, the complete fixed-length text feature vector is summarized as the fixed-length feature vector of each decoder output. In this process, the encoder hidden state of the current step, the decoder hidden state and the position features obtained by the attention weight of the previous step are taken into account. In the decoder, the attention eigenvector and the previous prediction spectrum are spliced into the pre-net network to assist attention learning, and the output of pre-net and the attention eigenvector are spliced together to the two-layer unidirectional LSTM network. The output of LSTM is used for two purposes: one is to predict the target spectrum after linear transformation, and the residual is predicted through the five-layer convolution network to improve the effect of spectrum prediction. The second is to add sigmoid prediction stop token after linear transformation to determine when the reasoning process terminates speech generation. Finally, the predicted Mel spectrum coefficient is further generated by the vocoder WaveGlow [19].

2.2 Phone Embedding Based on Initials and Tonal Finals

Myanmar is a phonetic language with tones, which is divided into high falling tone, low flat tone, high flat tone and short promoting tone. Myanmar uses consonant letters as the basic characters, a total of 33, in addition to 10 numeric characters, 7 independent vowel characters, 7 non-independent vowel characters and 4 medial characters. Non-independent vowel characters can form compound vowels with consonants and tone symbols, and four medial characters can form consonant cluster with consonants. The syllable structure of Myanmar language mainly has two forms: CV and CVC (C is consonant, V is vowel), In addition, syllable can also be composed of multiple consonants, medials and vowels, that is, CCMV (M is medial), When writing, Myanmar characters are composed of one basic character and four directions of up, down, pre-addition and post-addition scripts, and its character composition is shown in Figure 1. Basic consonant characters and subscript represent compound consonants, and other characters represent vowels.

According to the phonetic characteristics of Myanmar and the small size of the existing data set, this paper adopts the combination of dictionaries and pronunciation rules, with reference to the Chinese vowel structure [20], and represents Myanmar syllables as a combination of initial consonants and tonal vowels. The initial consonant is composed of basic consonants and complex consonants, and the vowel is composed of vowels or vowels and subsequent consonants. For example, according to the structure of consonants, medials, vowels and tones, the composition of a syllable is CCMVCt, where t is the tone, and the composition form of the initials and tone finals is I-Ft, where I=CCM and Ft=VCt. The phonetic subset in this system is composed of 71 initials and 50 finals with tones.

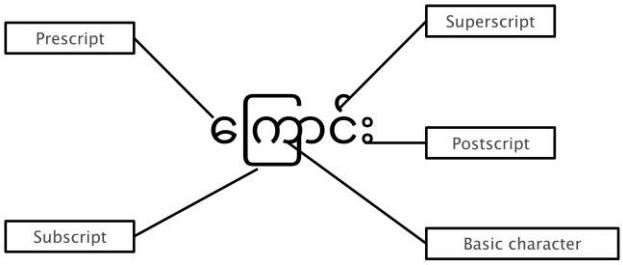


Figure 1. Structure Diagram of Myanmar Character

The phenomenon of stacked words and pronunciation change in Myanmar will lead to mispronunciation of synthetic speech if these problems are not specially dealt with. Pronunciation change means that the same Myanmar characters are pronounced differently in different contexts. The pronunciation change is mainly divided into two cases. One is changes in part of speech, The pronunciation of the same syllable changes when the part of speech is different, for example, ဆည် is pronounced as /shi^{22}/ when used as a noun and as /she^{22}/ when used as a verb; the second is to change according to the change of adjacent syllables, for example, စား is pronounced as /sa^{55}/ alone and /za^{55}/ when combined with က. Stacked words refers to the phenomenon that two consonant phonemes or vowel phonemes overlap on consonant phonemes, such as ကန္တာရ, the middle part is the overlapping part, which needs to be divided into ကန်/တာ/ရ according to the rule. In this system, the changed syllables are marked as different pronunciations according to the summary rule. The characters that have overlapped characters are first written as non-overlapping characters, and then the non-overlapping characters are marked as correct pronunciation. Through the above processing, the pronunciation ambiguity in the speech synthesis system is avoided.

2.3 Wordpiece Embedding Based on Bert Pre-Trained Model

The End-to-End speech synthesis system needs training data corresponding to text and speech. The research on English or Chinese speech synthesis is more mature, and the speech data is more abundant. The scale of a single speech database is more than 10 hours, or even tens of hours, such as LJspeech [23] and THCHS-30 [24]. Myanmar is a low-resources language, and the available voice materials are relatively small, for example, the data used in this system contains only 5.8 hours of recording. Under the condition of low resources, the introduction of pre-trained language model can enrich the text information of the encoder, accelerate the convergence speed of the model, and improve the prediction accuracy of stop tokens. And the rich context information in the pre-training language model also helps to improve the prosodic performance of synthetic speech. The purpose of this section is to assist Tacotron2 training through the rich text information in BERT's pre-training model, so as to enhance the richness of encoder text information and better guide the spectrum prediction of the decoder. We embed the original text in parallel through phone embedding and BERT word vector embedding to get two representations of text information, and provide both to Decoder. The BERT is first introduced below.

Bidirectional Encoder Representations from Transformers (BERT) is a word vector representation model that can describe character level, word level and sentence level. The word vector model containing rich text information can adapt to different task scenarios through downstream fine-tuning. BERT uses unmarked text to obtain word vectors with rich text information through random masking modeling (randomly masking tokens in a sentence) and next sentence prediction (predicting whether a sentence is adjacent or not). BERT consists of several Transformer [22] modules stacked together. Thanks to this multi-layer Transformer structure, the distance of any one or two token in a sentence can be converted into 1, so that the representation of the word vector of BERT will be generated based on the context of the left and right sides of all layers. The input of BERT is no longer characters or syllables, but the words are divided into a limited set of common sub-word units according to the BPE algorithm, called Wordpiece [25], which is a smaller word level than words. In the BERT pre-training model of Myanmar, because the Myanmar language uses UTF-8 encoding, the single syllable is based on the order of consonants-prefixes-vowels, so in the Wordpiece marking process of BERT. A syllable is divided into three tokens of [consonant] [medial] [vowel], and a Myanmar sentence is divided into a set of tokens with a syllable structure.

In the process of word embedding in the same Burmese sentence, there is no strict corresponding relationship between the Wordpiece sequence of BERT and the phoneme embedding sequence, and the length of the sequence is not the same. A syllable is divided into three sets [consonants] [intermediate sounds] [vowels with tones] in the Wordpiece sequence, and two sets [consonants] [vowels with tones] in the phoneme embedding sequence. There is a blank token in the phoneme embedding to represent the pause between syllables. Therefore, the corresponding relationship between the two embedded sequences of the same Burmese speech and the Mel frame is different. Figure 4 shows the result of phoneme embedding and Wordpiece embedding of a Burmese phrase. In order to achieve accurate alignment between the two sequences and Mel frames, the decoder provides independent position-sensitive attention modules for the two kinds of embedding.

In the process of word embedding in the same Myanmar sentence, there is no strict corresponding relationship between the Wordpiece sequence of BERT and the phone embedding sequence, and the length of the sequence is not the same. A syllable is divided into three sets [consonants] [intermediate sounds] [vowels with tones] in the Wordpiece sequence, and two sets [consonants] [vowels with tones] in the phone embedding sequence. There is a blank token in the phone embedding to represent the pause between syllables. Therefore, the corresponding relationship between the two embedded sequences of the same Myanmar speech and the Mel frame is different. Figure 2 shows the result of phone embedding and Wordpiece embedding of a Myanmar phrase. In order to achieve accurate alignment between the two sequences and Mel frames, the decoder provides independent position-sensitive attention modules for the two kinds of embedding. The word vector of BERT can represent context information based on location. On the one hand, the context information of BERT model improves the richness of text features of Tacotron2 model encoder, on the other hand, it improves the problem of incorrect pronunciation caused by stacked words and pronunciation change. The [CLS] token is included at the

beginning of the Wordpiece sequence of the BERT, and the [CLS] token contains the characteristic information of the entire sentence. In the experimental part, this paper will conduct separate experiments on BERT word embedding with [CLS] tokens and without [CLS] tokens to discuss the influence of [CLS] tokens on guiding speech synthesis.

2.4 Construction of Myanmar Speech Synthesis System

In the Myanmar speech synthesis model, 121 phones are used to represent Myanmar pronunciation information, and these phones are embedded into a series of phone index tokens. The labeling of the phone set refers to the pronunciation guidance of the dictionary, and the phonetic variation and reduplication can be explicitly expressed in the phone sequence, which is convenient for the encoder to partially extract the text feature information. The process from the original text to the phone sequence and then to the phone index tokens has been fully integrated into the system.

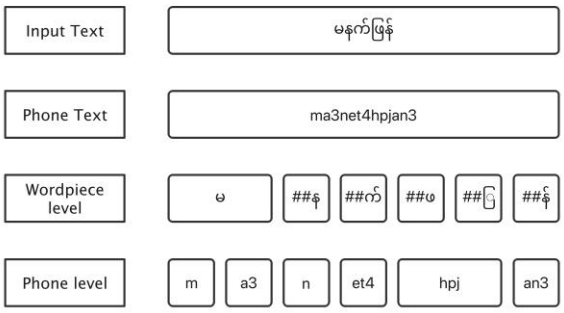


Figure 2. An Example of Myanmar Phone Level and Wordpiece Level

The original text is marked as a Wordpiece sequence and input to BERT, to provide the representation of the last Transformer module through the linear layer to the decoder. The decoder provides a separate position-sensitive attention module for the output of the BERT and generates a location-based attention context vector of the same size as the taoctron2 encoder. The context vector of the BERT is connected to the context vector of the Tacotron2 and then exposed to the autoregressive recursive network of the decoder part. The overall structure is shown in Figure 3.

3. EXPERIMENTS

3.1 Experimental Data and Settings

In this work, we prepared a Myanmar <text, voice> data set, recorded by a single female announcer, a total of 4890 sentences, audio lengths ranging from 2 seconds to 8 seconds, the total voice duration is about 5.8 hours, 16-bit PCM is used for encoding, the sampling rate is 16000Hz, and the text is encoded in UTF-8. In each audio sample, the beginning and end contain 50ms of silence. The data set is divided into 4665 sentence as training set, 188 sentence as test set and 37 sentence as verification set.

In the encoder part of the model, the system uses the structure of CNN + Bi-LSTM to obtain a context vector of fixed dimensions, while in the decoder part, we retain the default hyperparameters to train the model. In the Wordpiece embedding part of BERT, the model uses multilingual case model based on BERT pre-trained parameters. The BERT embedding dimension is 768, and the linear layer output dimension of the BERT encoder is 512.

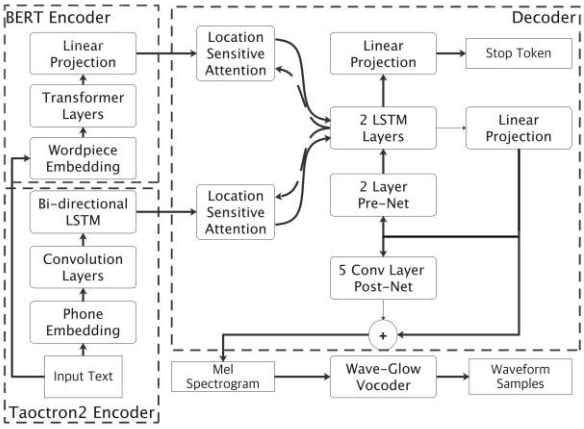
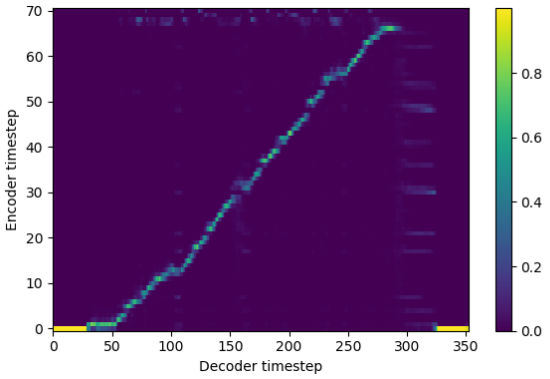


Figure 3. Block Diagram of Our Phone Embedding Model with Bert Pre-Trained Encoder

3.2 Evaluation of Experimental Results

In order to accurately evaluate the experimental results, this paper uses objective evaluation and subjective evaluation to evaluate the experimental results. First of all, we visualize the alignment results of the two attention layers in the decoder, and take the alignment results as the objective evaluation basis of the experiment. In the experimental results, we show the BERT attention alignment map with [CLS] token and without [CLS] token respectively. In the alignment graph with [CLS] token, we can see that in the synthesis of speech, the synthesis of each frame refers to the [CLS] token vector, because the [CLS] token contains the information of the whole sentence. In the experiment, we find that the learning speed of the model with [CLS] token is faster, but when predicting the stop token, the performance of the model without [CLS] token is better. The model with [CLS] token is usually difficult to end its decoding process. Models without [CLS] tokens learn slowly at the initial stage of training, but with the progress of training, the advantages of models without [CLS] gradually become apparent. Figure 4 and 5 show the visual attention alignment results of the two models when the number of training steps is 100k. Both models can focus on the correct phones when generating audio sequences. Figure 6 shows the speech spectrogram of the speech generated by two methods, both of which can generate a complete Myanmar speech.



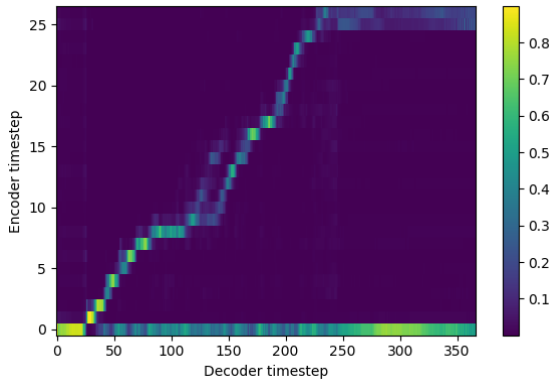


Figure 4. Visual Alignment Results for Models with [CLS] Token

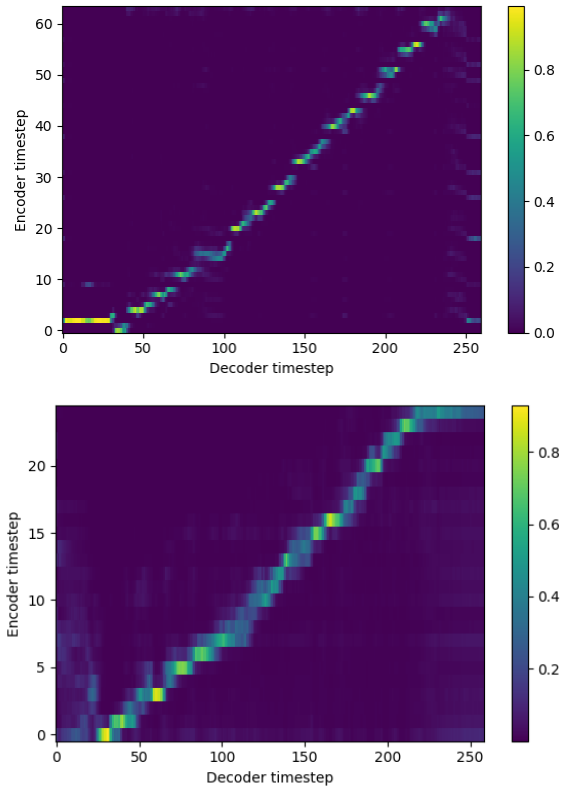


Figure 5. Visual Alignment Results For Models Without [CLS] Token

In Figure 4 and 5, the pre-50ms and post-50ms of the training corpus are silent segments, which affects the alignment results of the first 25 inference steps and the last 25 inference steps in the figure. In the encoder section, the number of character embedding steps is greater than that of the BERT encoder step, because character embedding contains not only the phone structure, but also white space tokens between syllables. In the BERT encoder, multiple vowels are represented by a single token, and there is no whitespace token between syllables, which results in different encoder inference steps between the two methods.

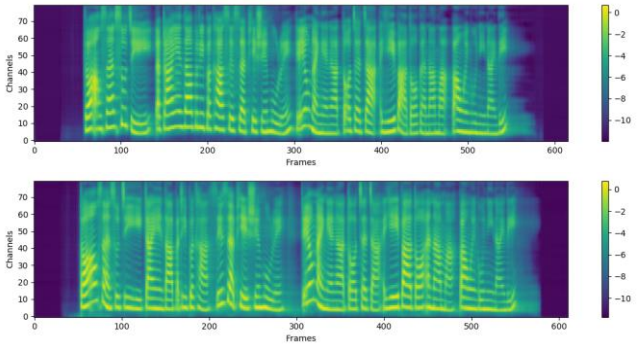


Figure 6 Two Methods of Synthesizing Speech Spectrum. Wordpiece Embedding with [CLS] Token (up), Wordpiece Embedding without [CLS] Token (down)

We use the mean opinion score (MOS) [26] to evaluate the comprehensive speech quality of the model. In the 100k-based training step, 20 samples were selected from the test set, and 20 listeners were asked to score the overall impression of these samples. Table 1 shows the scoring results of the synthetic speech of the two models.

Table 1. MOS Test Result of Our Model

Model	MOS
Tacotron2+BERT with [CLS]	3.5
Tacotron2+BERT without[CLS]	3.7

From the MOS score, it can be seen that the performance of the speech synthesis model without [CLS] token is better than the speech synthesis model with [CLS] token. Both models can usually synthesize high-quality Myanmar speech, indicating the effectiveness of Myanmar phone structure embedding. Due to the lack of Myanmar < text, speech > data, the experiment can synthesize high-quality Myanmar speech using only 5.8 hours of Myanmar recording corpus, which shows that the phone structure instead of character structure used in this paper is effective. For End-to-End speech synthesis in Myanmar, the replacement of phone structure effectively reduces the demand for training data in Myanmar speech synthesis. At the same time, the introduction of BERT pre-training model to enhance text information can effectively accelerate the convergence speed of the model and reduce the training steps.

4. CONCLUSION

In this paper, we implement a Myanmar speech synthesis system based on End-to-End model, which uses phone embedding to reduce the amount of training data needed by the speech synthesis system, and introduces the BERT pre-trained model as other inputs of the encoder to enhance text feature extraction. The experimental evaluation shows that the method adopted by the system is effective. The system can synthesize Myanmar speech with less training data, which improves the accuracy and naturalness of the synthesized speech.

5. ACKNOWLEDGMENTS

This research is supported by the Natural Science Foundation of China (No.61961043).

6. REFERENCES

[1] P. Taylor, Text-to-Speech Synthesis, Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp.1039–1064, 2009.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proc. ICASSP, 2013, pp. 7962–7966.

[4] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep Voice: Real-time neural text-to-speech," CoRR, vol. abs/1702.07825, 2017.

[5] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. "Char2wav: End-to-end speech synthesis". 2017.

[6] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J.Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A.Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," CoRR, vol. abs/1703.10135, 2017.

[7] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. Saurous, Y. Agiomyrgiannakis, and Y.Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in ICASSP, 2018.

[8] Yasuda Y, Wang X , Yamagishi J . "Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis [J]". arXiv, 2020.

[9] K. Park, T. Mulc, CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages, in: Proc. Interspeech 2019, 2019, pp. 1566–1570. URL: http://dx:doi:org/10:21437/ Interspeech:2019-1500. doi:10:21437/Interspeech:2019-1500

[10] J. Li, Z. Wu, R. Li, P. Zhi, S. Yang, H. Meng, "Knowledge-Based Linguistic Encoding for End-to-End Mandarin Text-to-Speech Synthesis", in: Proc. Interspeech 2019, pp. 4494-4498.URL:http://dx:doi:org/10:21437/Interspeech:2019-1118.doi:10:21437/Interspeech:2019-1118

[11] Y. Lu, M. Dong, Y. Chen, "Implementing prosodic phrasing in Chinese end-to-end speech synthesis", in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7050–7054. doi:10:1109/ICASSP:2019:8682368.

[12] B. Li, Y. Zhang, T. Sainath, Y. Wu, W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes", in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5621–5625.doi:10:1109/ICASSP:2019:8682674.

[13] Wang.D,"Burmese Language tutorial(NO,1),"Beijing University Publishing House,2012.(in Chinese).

[14] Wang D. "A comparative study of Burmese and Sino-Tibetan languages [M]". Beijing: Kunlun Publishing House, 2012: 1-14.

[15] Kyawt. Y, and Tomio. T, Myanmar text-to-speech system with rule-based tone synthesis, Acoustical Science and Technology, vol. 32, no. 5, 2011, pp. 174-181.

[16] Hlaing A M , Pa W P , Thu Y K . Myanmar Number Normalization for Text-to-Speech[C]// International Conference of the Pacific Association for Computational Linguistics. Springer, Singapore, 2017.

[17] Ye K, Win P, Jinfu N, Yoshinori S,Andrew F, Chiori H, Hisashi K, Eiichiro S HMM Based Myanmar Text to Speech System. Interspeech 2015.

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

[19] R. Prenger, R. Valle and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 3617-3621, doi: 10.1109/ICASSP.2019.8683143

[20] Wang D. A comparative study of Burmese and Sino-Tibetan languages [M]. Beijing: Kunlun Publishing House, 2012: 1-14.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in NAACL-HLT, 2019.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[23] K. Ito, "The LJ speech dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[24] D. Wang and X. Zhang, "THCHS-30: A free chinese speech corpus," arXiv preprint arXiv:1512.01882, 2015.

[25] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[26] R. C. Streijl, S. Winkler, and D. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," Multimedia Systems, vol. 22, pp.213–227, 03 2016.