Myanmar Text to Speech

Badounmar, Hnin Yu Hlaing, Hlaing May Tin,
Nan Yu Hlaing, Thida San, Zun Hlaing Moe

***Abstract*— Text-to-speech system typically consists of a text analysis front-end, an acoustic model and a speech synthesizer. Since these components are trained independently and rely on extensive domain expertise. In this project to address these problems, we apply Tacotron2 through tensorflow, which is an end-to-end generated text-to-speech (TTS) model with syllable and word-level. Given <text, audio> pairs, the model can be trained from scratch with random initialization. End-to-end system can be trained on a small number of manually labeled text and audio paired data sets brings many advantages. In our experiment, according to the difficulties of the use of GPU, we trained 10 text and speech on step 10k for a week and investigated on closed test. Tacotron2 achieves a 3.8 subjective 5-scale mean opinion score on closed tests with listeners, outperforming a production parametric system in terms of naturalness. The clarity performance of the syllable-level is better than the word-level. In addition, since Tacotron2 generates speech at the frame level, it's substantially faster than sample-level autoregressive methods.**

***Index Terms*—TTS, Tacotron, Tensorflow, MOS, Encoder, Decoder**

I. Introduction

THE main motivation for this project is to investigate the end-to-end text to speech synthesis with small corpus. TTS is a common research area of digital signal processing (DSP) and natural language processing (NLP). It is intended to generate human-like speech from the input text or sentences, a natural-sounding in terms of intelligibility and quality. The main task of TTS is to convert any text information into standard and smooth speech in real time. The speech synthesis is not a new problem, but it is still one of the challenges for organizations and businesses. The modern TTS trend is more complex. Deep learning (DL) is a new research direction in the machine learning area in recent years. In this project, we experimented with TTS by Tacotron, one of deep learning methods which is to synthesize speech directly from the characters. It does not need phoneme-level alignment and can be trained on completely from scratch given <text, audio> pairs. There are a number of generative models already exist for this purpose, but some of them are not necessarily end-to-end as they usually have models developed and trained separately. Among these models, we used Tacotron as it is truly an end-to-end generative model that can fulfill our goal.

The remainder of this paper is organized as follows. In Section 2, we describe the related work. Section 3 briefly introduces Myanmar Language. Section 4 describes methodology, Section 5 presents the overview of experimental setup, results and discussion. Lastly, we conclude in section 6.

II. Related Work

In recent years, DNN based generative models for Myanmar Speech synthesis can yield better synthesized speech than HMM [1]. In [2], Quang Pham Huu proposed a deep learning architecture to the problem of speech synthesis, Tacotron model. The output of Tactron on both BigCorpus and SmallCorpus achieved high-quality speech audio. Lwin et al., proposed Tacotron-2 model synthesize speech directly from the characters and they experimented on Myanmar text and audio pairs [3]. Kim et al., researched a generative flow of speech synthesis with monotonic alignment search without any external aligner and they obtained the comparable speech quality to Tacotron-2 [4]. Chuxiong Zhang et al., introduced Tacotron for Mandrain Chinese TTS with prosodic features to generate more natural speech and obtained better by adding the prosodic system as the front-endsystem for Tacotron [6].

III. 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. Myanmar script has 33 basic consonants, 4 basic medials, 12 basic vowels, other symbols and special characters. The consonants have only 23 distinct pronunciations because some consonants have the same pronunciation in the 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 anatonic vowel, there are four phonological tones in Myanmar [5].

Badounmar is with the Faculty of Computer Science, Computer University (Thaton), Thaton, Myanmar.

Hnin Yu Hlaing is with the Faculty of Information Science, University of Computer Studies(Meiktila), Meiktila, Myanmar.

Hlaing May Tin is with the Faculty of Computer System and Technology, Myanmar Institute of Information Technology (MIIT), Mandalay, Myanmar.

Nan Yu Hlaing, Zun Hlaing Moe and Thida San are with the Faculty of Information Science, Myanmar Institute of Information Technology (MIIT), Mandalay, Myanmar.

IV. Methodology

In this section, we describe the methodology used in this end-to-end text to speech synthesis. In the experiment of this project, used Tacotron2 was used with tensorflow.

A. Tacotron

End-to-end speech synthesis system combines text analysis front-end, acoustic model and speech synthesizer into a unified framework without requiring phoneme level alignment and it can be trained on large scale of text and audio pairs with minimum human annotation. In this project, we used Tacotron-2, a fully end-to-end speech synthesis model that directly maps the input text to mel-spectrogram. The backbone of Tacotron is a seq2seq model with attention. Tacotron takes character sequence as inputs and outputs raw spectrograms which are later reconstructed using an algorithm called Griffin-Lim. It operates at frame-level which is why it is faster than sample-level auto-regressive models like WaveNet and SampleRNN. It consists of an encoder, an attention-based decoder, and a post-processing net. The encoder is responsible for building up a well-defined summarized representation of the input character sequence. The decoder has to learn about the alignment between the text representations and the output audio frames based on the context [3]. Figure 1 depicts the model architecture, which includes an encoder, an attention-based decoder, and a post-processing network. On the high level, our model takes characters as input and the resulting spectral frame data is then converted to a waveform.

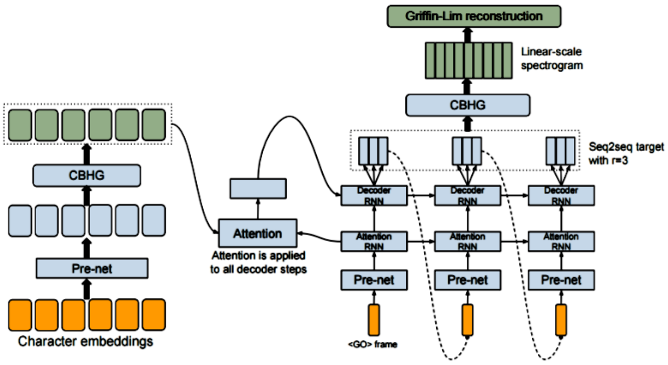


Fig. 1: Model architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Graiffin-Lim reconstruction algorithm to synthesize speech

B. CBHG Module

CBHG consists of a one-dimensional convolution filter bank followed by a highway network and a two-way gated loop unit cyclic neural network (RNNCBHG is a powerful module to extract feature representations of sequences. A CNN works well for identifying simple patterns within your data which will then be used to form more complex patterns within higher layers. A 1D CNN is very effective when you expect to derive interesting features from shorter (fixed-length) segments of the overall data set and where the location of the feature within the segment is not of high relevance.

C. Encoder

The purpose of the encoder is to extract the robust sequence representation of the text. The input to the encoder is a sequence of characters, each character entered is a one-hot vector and embedded in a continuous vector. A set of non-linear transformations is then applied to each character vector, collectively referred to as "pre-net." The CBHG module transforms the output of the prenet into the final representation of the encoder and passes it to the subsequent attention module.

D. Decoder

Decoder turns the internal representation of the input signal into a Mel-spectrogram. A very important element of the network is the PostNet, designed to improve the spectrogram generated by the decoder.

E. Post-Processing Network and Waveform Synthesis

The post-processing network's task is to convert the output of seq2seq into a target representation that can be synthesized into a waveform. The CBHG module is used as a post-processing network. Waveform is synthesized from the predicted spectrogram using the Griffin-Lim algorithm. Griffin-Lim algorithm enables a partial restore of the signal after fast Fourier transforms. It can reduce artifacts, probably due to its harmonic enhancement.

V. Results and Discussion

A. Corpus Statistics

For this experiment, we used sentences from the "Grade 2" Myanmar textbook published by the Ministry of Education. Firstly we prepared myanmar sentences and audio files. Myanmar sentences are segmented into syllable and word-level. For the syllable segmentation, we used a regular expression perl script developed by Saya Ye Kyaw Thu [7].To prepare the speech corpus, we used audio that has been recorded for Myanmar Braille TTS.

B. Implementation

To implement this project, we installed the following requirements:

- Python3
- tensorflow framework 14.1
- numpy 1.19.2
- scikit-learn 0.20.3
- librosa 3.0.3
- falcon 1.2.0
- tqdm 4.31.1
- matploatlib 3.0.3

In the implementation, we prepared 10 sentences totally, this is for about 15 minutes in the format of text and audio parallel pair for training and two sentences for testing on closed test. Training time is a week for ten sentences. The maximum input text length is 117 and the maximum number of frames in the input audio is 6.

C. Evaluation

We used Mean Opinion Score (MOS) for the evaluation of the model output. Testing the outputs contains eleven listeners. We collected the evaluation rate from the listeners based on three conditions: clarity, naturalness and accurate over 1-5 rating. Rating 5 is the best condition. Listeners answer the questions. For testing the clarity and naturalness they answer these questions “How could you hear the sound clarity? ” and “ How does the sound is natural?”. For the accurate of the output sentence, we evaluate the condition based on this question “How many words can her the listener?” Calculate the accurate words based on how the listener hear the sentence accurately. For instance , there are six words in test sentence “ဘယ် ကို သွား ခဲ့ သ လဲ ။ ”. If the user hear only 4 words, we denote an accurate rate of 3.33 for that sentence . If the user hear all words, the accurate rate is 5. The MOS scores from the listeners are shown in figure 6. The genrerated output of syllable and word-level waveforms are shown in figure 2 to figure 5. According to figure 5, there is noise in the word-level generated waveform. Therefore, the intelligibility of syllable-level is better than the word-level.

Input test sentence: အ ခန်း (၁) ။
Waveform for test sentence:

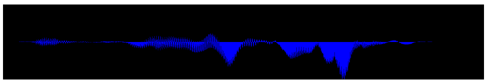


Fig. 2: Syllable-Level Generated Waveform

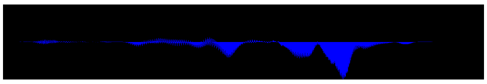


Fig. 3: Word-Level Generated Waveform

Input test sentence: ဘယ် ကို သွား ခဲ့ သ လဲ ။
Waveform for test sentence:

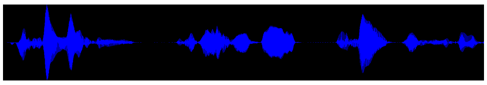


Fig. 4: Syllable-Level Generated waveform

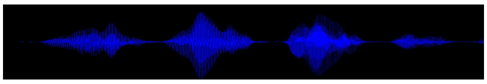
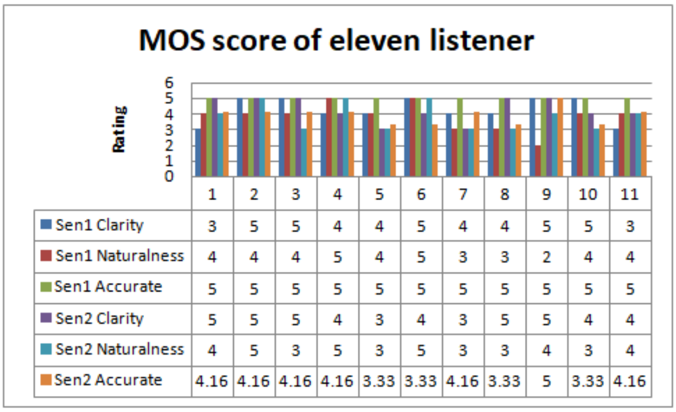


Fig. 5: Word-Level Generated waveform

According to figure 7: clarity and naturalness of both test sentences are not significantly. However the first sentence is more accurate than the second one. According to our investigation, text length of the first sentence is shorter than the second and tonal significance is also better than the second one. Therefore the result of the first sentence is more better than the second one. The tone is also important for the text to speech.



	1	2	3	4	5	6	7	8	9	10	11
Sen1 Clarity	3	5	5	4	4	5	4	4	5	5	3
Sen1 Naturalness	4	4	4	5	4	5	3	3	2	4	4
Sen1 Accurate	5	5	5	5	5	5	5	5	5	5	5
Sen2 Clarity	5	5	5	4	3	4	3	5	5	4	4
Sen2 Naturalness	4	5	3	5	3	5	3	3	4	3	4
Sen2 Accurate	4.16	4.16	4.16	4.16	3.33	3.33	4.16	3.33	5	3.33	4.16

Fig. 6: Rating Scores from the Listeners

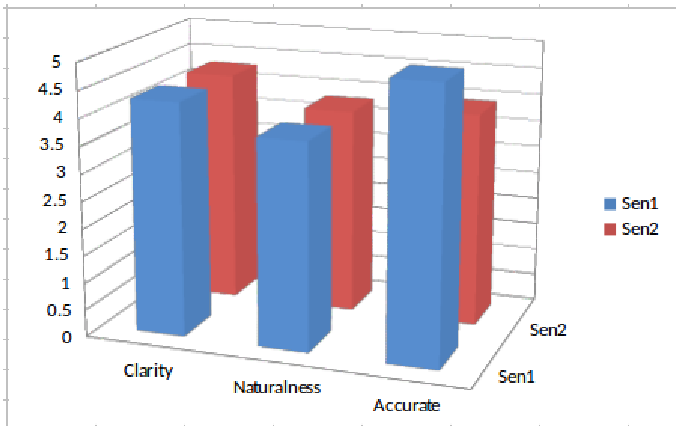


Fig. 7: Speech Quality of the Output Model

VI. Conclusion

This mini project was contributed the evaluation of the end-to-end speech synthesis with tacorton model. This is the study of end-to-end speech synthesis with small corpus. We evaluated end-to-end TTS with the small corpus by using ten sentences from “Grade 2” Myanmar Basic Education Textbook. Although the syllable-level achieved 3.8 MOS score, the word-level achieved 3.4 on closed tests with listeners. syllable-level also obtained more clearance result than word-level. Good speech output depends not only on recording condition but also on tone signature. By experimenting this mini project, we experienced that tacotron2 can work well even in a small corpus. In the near future, we will plan to test large corpus on tacotron2.

References

[1] Aye Mya Hlaing, Win Pa Pa and Ye Kyaw Thu, “DNN based Myanmar Speech Synthesis”, The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages 29-31 August 2018, pp, 142-146..

[2] Quang Pham Huu, “The End-to-End Speech Synthesis System for the VLSP Campaign 2019”.

[3] Htoo Pyae Lwin, Yuzana Win and Tomonari Masada, “Myanmar Text-to-Speech with End-to-End Speech Synthesis”.

[4] Jaehyeon Kim, Sungwon Kim, Jungil Kong and Sungroh Yoon, “ooo”, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, pp, 1-11.

[5] U. Thein-Tun, “The domain of tones in burmese”, SST 1990 Proceedings, pp. 406–411, 1990.

[6] Chuxiong Zhang, Sheng Zhang and Haibing Zhong, “A Prosodic Mandarin Text-to-Speech System Basedon Tacotron”, Proceedings of APSIPA Annual Summit and Conference, 2019, pp, 165-169.

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