Development of Zalo Vietnamese Text-to-Speech for VLSP 2019

Viet Lam Phung[1], Huy Kinh Phan[1], Anh Tuan Dinh[1], Khuong Duy Trieu[1] and Quoc Bao Nguyen[1,2]
Zalo Group - VNG Corporation
[2]*Information and Communication Technology University - Thai Nguyen University*
{lampv,baonq6}@vng.com.vn

***Abstract*— The paper describes the ZALO Text-to-Speech (TTS) system architecture and how the VLSP 2019 small and big corpus were integrated into our system. The core of ZALO system was based on (1) Tacotron-2 and WaveGlow neural vocoder, and (2) a generative adversarial model (GAN)-based TTS. The big corpus has inconsistent quality with noise and reverberation; which might reduce the quality of synthesized speech. Our key solution for the problem is to adopt sophisticated measurements to select the best portion of data from the given database. In the small corpus challenge, there is not enough data for training an end-to-end TTS system from scratch. We adapted a pre-trained end-to-end TTS using the small amount of data. We also trained a GAN-based TTS using the small corpus. Then we selected the best stimuli from the 2 systems. An analysis of mean opinion score for test sentences was conducted for the 2 challenges. For the big corpus challenge, our system achieved an average score of 4.10 (given an average score of 4.30 for natural speech.) For small corpus, our systems obtained an average score of 3.77 (given an average score of 4.58 for natural speech.)**

***Index Terms*— Tacotron-2, GAN, Text-to-speech, WaveGlow**

I. INTRODUCTION

Text-to-speech synthesis plays a key role in interactive, speech-based systems. In the International Workshop on Vietnamese Language and Speech Processing (VLSP) 2019, a TTS shared task was organized to evaluate Vietnamese TTS systems. This year evaluation provided a small corpus which simulated a low-resource condition, and a big corpus which simulated an imperfect, non-standard data condition. The two databases included raw text and corresponding audio files. The participants built voices using the 2 databases.

The big corpus features noisy, reverberant utterances, inconsistent recording conditions, and wrongly spoken utterances; which are very similar to randomly found data on Internet. Our solution emphasized the importance of transcription correction and utterances selection. In subsection II-A.1, we introduced an audio alignment system to obtain precise transcription of audio files. In II-A.2, we used different data selection metrics to select the best utterances from the given databases. For small corpus challenge, we adopted data augmentation techniques in subsection II-A.3 to generate an sufficient amount of data for training an end-to-end system. Moreover, we added prosodic pauses to the raw transcription as mentioned in subsection II-A.4 to assist our TTS systems to learn the prosody of the speakers.

II. Our system

A. Data Processing

The audio files were resampled to a sampling rate of 16KHz. For the big corpus, the audio files were denoised using a minimum-mean-square-error noise reduction algorithm [1]. For both corpus, we normalized the volume, then cut unvoiced parts of the audio files using a Gaussian Mixture Model (GMM) based Voice Activity Detection (VAD) algorithm [2].

Text normalization was also conducted. There are a number of English words in the two provided databases. Our solution is to borrow Vietnamese sounds to read the English words. Thus, the English words were splitted into possible Vietnamese syllables with voiceless sounds (for example, x sound) if necessary (for instance, "studio" becomes 'x- tiu-đi-ô'). We think this can make it easier for the model to learn the voiceless sounds. In addition, splitting English words introduced uncommon Vietnamese syllables which enriched the vocabulary of the training data set.

1) Audio Alignment System: An audio alignment system was trained to predict time boundaries of either words or syllables in an utterance. As a result, the system also recognized the "true" text transcription of an utterance. The system consisted of a VAD [2], a time delay neural network based acoustic model, and a character-level language model which is over-fitted on the normalized text transcription. In our experiments, we further used this information for data augmentation, data selection and automatic punctuation to improve the naturalness and prosody of synthesized voices.

2) Data Selection: Several metrics [3] were used to select the best portion of audio files from provided big corpus:

- **Misalignment errors**: We compared each utterance's original text transcription with its "true" text transcription coming from our audio alignment system. We removed any utterance with any mismatch between its original transcription and corresponding true transcription. In fact, the mismatch can due to bad transcription, poor acoustic condition, poor pronunciation or it may be due to poor acoustic models used for automatic alignment. But we are confident in the quality of our alignment system. For big corpus, we removed about 800 utterances.

- **Articulation**: The articulation was calculated from the power of speech parts P_{signal} and average syllable duration as follows:

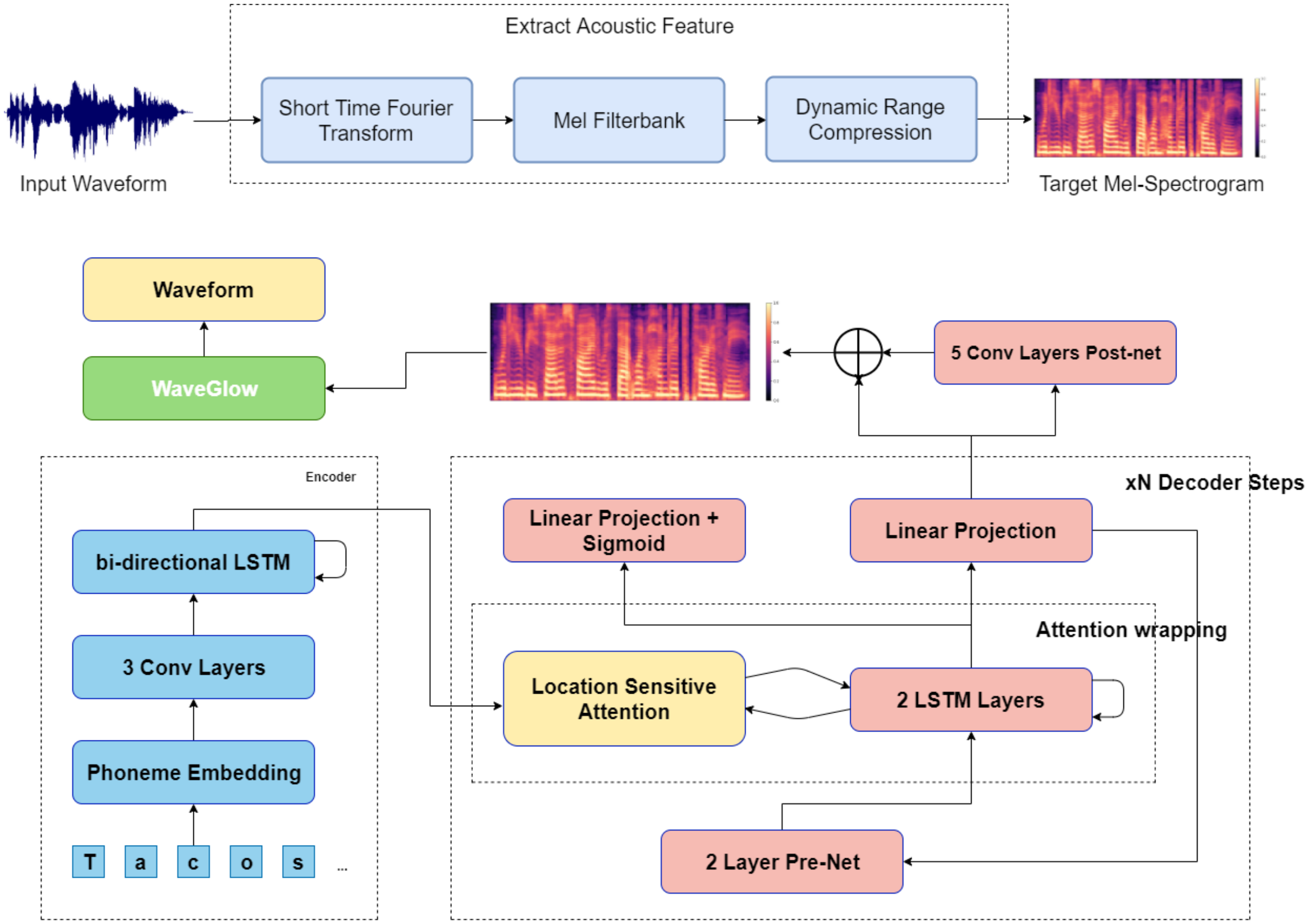

Fig. 1. End-to-end model architecture

$$\text{Articulation} = P_{signal} \times \text{Avg. syl. dur}$$

We filtered out the audio files with high articulation or hyper-articulation.

- **Standard deviation of syllable duration**: If standard deviation of syllable duration is high for a segment, it indicates that the narrator spoke sometimes fast and sometimes slow in that segment.

- **Non-fluency**: The value of non-fluency is high if there is a long pause within a sentence which reflects the non-fluency. Non-fluency was calculated as follows:

$$\text{Non-fluency} = \frac{\max(\text{internal silence duration})}{\text{Avg. syllable duration}}$$

- **Standard deviation of F0**: A high standard deviation of F0 might due to an expressive speech but can also be associated to poor performance of F0 estimator.

Based on these selection metrics, with the appropriate thresholds, we filtered out about 6 hours of audios with bad quality from big corpus.

3) Data Augmentation: For small corpus, we used an data augmentation method to make the model easier to converge and increase generalization. We first concatenated all the audio files in several different order to get the big audio files. Then we sliced them into short segments with a maximum length of 5s to obtain new training utterances. The text transcription of the new utterances were obtained using our audio alignment system. We have implemented this method to double number of samples in the small corpus.

4) Automatic Punctuation: Punctuation has been added to time-aligned text transcription at suitable silent intervals. We also marked bad silences (we found that speakers often read without fluency at silence intervals of about 0.05 or greater than 0.5 seconds) and syllables with either very short (speak very fast, which is common when pronouncing English words with many syllables) or very long duration (stressed). These helped the model learn the standard punctuation and greatly improved the prosody of the synthesized voices.

III. Introduction

A. *TTS Systems*

1) End-to-end System: Our system was composed of a recurrent sequence-to-sequence feature prediction network called Tacotron-2, which mapped text embedding to mel-spectrogram, followed by a WaveGlow vocoder - a Flow-based generative network for waveform generation from mel-spectrogram.

- Tacotron-2: Our network architecture was similar to [4], but with some changes. Phoneme embedding was used instead of character embedding. Stop tokens were not used. We used other parameters for mel-spectrogram feature extraction to better fit the data set (about sampling rate, pitch and loudness of the sound). For small corpus we reduced pre-net and encoder dimension to avoid over-fitting.

- WaveGlow: We also had a few changes compared to [5]. We used the same mel-spectrogram feature extraction as Tacotron-2. In addition, we created a learning rate schedule to optimize the training process and implemented ground-truth-aligned (GTA) training.

2) GAN System: The idea of using GAN [6] was inspired by the drawbacks of traditional mean-squared error loss function in many statistical parametric speech synthesis (SPSS) systems [7], [8]. The fact that errors over the whole sequence are averaged lead to over-smoothing drove us to pick up another loss function. Our preliminary results showed that an adversarial structure utilizing an adversarial-based loss outperformed a Merlin-based SPSS model [9] .

IV. Experiments

A. System Configuration

The big corpus and small corpus contain respectively 23 and 0.8 hours of utterances from two single female speakers. After data selection, The big corpus had 17 hours of utterances. While we increased the size of the small corpus to two hours using our data augmentation method.

For big corpus, the acoustic model Tacotron-2 was trained from scratch. For small corpus, we fined-tune a pre-trained Tacotron-2 model. We also trained our WaveGlow vocoder on the ground truth-aligned predictions of the acoustic model.

B. Results

The system was tested using a subjective MOS test. There are 24 participants listening to the stimuli of synthesized and natural speech. The participants gave each utterance a score on a 5-point scale with five is the highest score of naturalness. Details of the scores are given in Table I. In this table, NATU refers to natural speech in training data.

TABLE I

Average MOS results for small and big corpus

	Our system	NATU
Small corpus	3.77	4.58
Big corpus	4.10	4.30
Final score	**3.94**	4.44

V. Conclusion and Future works

In this report, we described our Vietnamese TTS system for VLSP 2019. In big corpus challenge, our approach yielded a very close MOS result to this of natural speech. We also have a good MOS score with the small corpus.

By testing many different solutions for these challenges, we find that data processing has a very important role in TTS systems, especially when the data is of poor quality. Data selection and automatic punctuation are the two main factors that help us achieve very good results on the big corpus.

As future works, we keep improving our TTS systems to get good results with small training corpus. The techniques, which are in our consideration, are GAN-TTS, and multi-speaker end-to-end TTS

References

[1] Dong Yu, Li Deng, Jasha Droppo, Jian Wu, Yifan Gong, Alex Acero, "A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition", IEEE, 3/2008, IEEE International Conference on Acoustics, Speech and Signal Processing

[2] Zhan Shen, Jianguo Wei, Wenhuan Lu, Jianwu Dang, "Voice activity detection based on sequential Gaussian mixture model with maximum likelihood criterion", IEEE, 10/2016, International Symposium on Chinese Spoken Language Processing (ISCSLP)

[3] F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, I. Ouyang, "Data Selection for Improving Naturalness of TTS Voices Trained on Small Found Corpuses", IEEE, 12/2018, IEEE Spoken Language Technology Workshop (SLT).

[4] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions", IEEE, 4/2018, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[5] Ryan Prenger, Rafael Valle, Bryan Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis", IEEE, 5/2019, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[6] Ian J. Goodfellow and Jean Pouget-Abadie and Mehdi Mirza and Bing Xu and David Warde-Farley and Sherjil Ozair and Aaron Courville and Yoshua Bengio, "Generative Adversarial Nets", 12/2014, NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems

[7] Y. Saito, S. Takamichi and H. Saruwatari, "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks," IEEE, 1/2018, IEEE/ACM Transactions on Audio, Speech and Language Processing

[8] Shan Yang and Lei Xie and Xiao Chen and Xiaoyan Lou and Xuan Zhu and Dongyan Huang and Haizhou Li, "Statistical Parametric Speech Synthesis Using Generative Adversarial Networks Under Multi-task Learning Framework", IEEE, 12/2017, IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

[9] Zhizheng Wu, Oliver Watts, Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System", 9/2016, ISCA Speech Synthesis Workshop