Development of Smartcall Vietnamese Text-to-Speech for VLSP 2020

Manh Cuong Nguyen
Smartcall JSC
dodo.proptit.99@gmail.com

Khuong Duy Trieu
Smartcall JSC
duytrkh@gmail.com

Ba Quyen Dam
Smartcall JSC
dambaquyen.ptit@gmail.com

Thu Phuong Nguyen
ICTU, Thai Nguyen University
ntphuong@itcu.edu.vn

Quoc Bao Nguyen
Smartcall JSC and ICTU, Thai Nguyen University

Abstract

An end-to-end text-to-speech (TTS) system (e.g. consisting of Tacotron-2 and WaveGlow vocoder) can achieve the state-of-the art quality in the presence of a large, professionally-recorded training database. However, the drawbacks of using neural vocoders such as WaveGlow include 1) a time-consuming training process, 2) a slow inference speed, and 3) resource hunger when synthesizing waveform from spectral features. Moreover, the synthesized waveform from the neural vocoder can inherit the noise from an imperfect training data. This paper deals with the task of building Vietnamese TTS systems from moderate quality training data with noise. Our system utilizes an end-to-end TTS system that takes advantage of the Tacotron-2 acoustic model, and a custom vocoder combining a High Fidelity Generative Adversarial Networks (HiFiGAN)-based vocoder and a WaveGlow denoiser. Specifically, we used the HiFiGAN vocoder to achieve a better performance in terms of inference efficiency, and speech quality. Unlike previous works, we used WaveGlow as an effective denoiser to address the noisy synthesized speech. Moreover, the provided training data was thoroughly preprocessed using voice activity detection, automatic speech recognition and prosodic punctuation insertion. Our experiment showed that the proposed TTS system (as a combination of Tacotron-2, HiFiGAN-based vocoder, and WaveGlow denoiser) trained on the preprocessed data achieved a mean opinion score (MOS) of 3.77 compared to 4.22 for natural speech, which is the best result among participating systems of VLSP 2020's TTS evaluation.

Index Terms— End-to-end TTS, Tacotron-2, HiFi-GAN, WaveGlow, vocoder

1 Introduction

Text-to-speech synthesis plays a crucial role in speech-based interaction systems. In the last two decades, there have been many attempts to build high quality Vietnamese TTS systems. A data processing scheme proved its efficacy in optimizing naturalness of end-to-end TTS systems trained on Vietnamese found data (Phung et al., 2020). Text normalization methods were explored; utilizing regular expressions and language model (Tuan et al., 2012). New prosodic features (e.g. phrase breaks) were investigated, which showed their efficacy in improving naturalness of Vietnamese hidden Markov models (HMM)-based TTS systems (Dinh et al., 2013; Trang et al., 2013; Phan et al., 2013). Different types of acoustic models were investigated such as HMM (Dinh et al., 2013), deep neural networks (DNN) (Nguyen et al., 2019), and sequence-to-sequence models (Phung et al., 2020). For postfiltering, it was shown that a global variance scaling method may destroy the tonal information; therefore, exemplar-based voice conversion methods were utilized in postfiltering to preserve the tonal information (Tuan et al., 2016). To our knowledge, there is little to none research on vocoders for Vietnamese TTS systems, especially when the training data is moderately noisy.

In the International Workshop on Vietnamese Language and Speech Processing (VLSP) 2020, a TTS challenge (Trang et al., 2020) required participants to build Vietnamese TTS systems from a provided moderately noisy corpus. The corpus included raw text and corresponding audio files. However, the corpus has incorrect pronunciation of a foreign language, the slight buzzer sounds in audio data, and many incorrectly labeled words, which pose significant challenges to participants. For example, a general neural vocoder will learn the buzzer sounds from the corpus, and introduce

it to the synthesized speech.

In previous VLSP 2019's TTS evaluation, Tacotron-2 and WaveGlow neural vocoder were combined to achieve the best speech quality in Vietnamese speech synthesis (Lam et al.). However, HiFiGAN vocoder significantly outperformed WaveGlow vocoder in term of vocoding quality and efficiency (Kong et al., 2020). In the paper, we present the complete steps of building our end-to-end TTS system combining data preprocessing (Phung et al., 2020) and end-to-end modeling which showed that the system addressed the data problems and achieved high performance and high efficiency.

In particular, we introduced a solution that combines HiFiGAN and WaveGlow denoiser as a custom vocoder to enhance the quality of the final synthesized sound. Specifically, in Section II, we present the TTS system architecture consisting of a Tacotron-2 network followed by the HiFiGAN model as a vocoder and the WaveGlow model as a denoiser. The use of HiFiGAN has both improved aggregation speed and reduced resource size, and utilizing WaveGlow denoiser significantly reduces unexpected noise of synthesized speech. The challenges of naturalness, background noise and buzzer noises in the artificial sound were also overcome by combining Tacotron-2, a HiFiGAN-based vocoder and a WaveGlow denoiser.

2 SYSTEM ARCHITECTURE

2.1 Data Preprocessing

We inherited the data processing method (as shown in Figure 1) proposed in (Phung et al., 2020). We remove non-speech segments from the audio files using Voice Activity Detection (VAD) model (Kim and Hahn, 2018). As for textual data, we normalized the original text to lower case without punctuation, then use the results from an Automatic Speech Recognition (ASR) (Peddinti et al., 2015) model to define unvoiced intervals to automatic punctuation to improve the naturalness and prosody of synthesized voices (Phung et al., 2020). Moreover, there is an enormous number of English words in the provided databases, so our solution is to borrow Vietnamese sounds to read the English words. Even, the English words can consist of Vietnamese syllables and English fricative sounds (for example, x sound) if necessary (for instance, "study" becomes 'x-ta-đi'), which can make it easier for the model to learn the fricative sounds. Also, by selecting the pronunciation of English words, we introduced uncommon Vietnamese syllables, which enriched the vocabulary of the training data set. The overall text normalization was carried out using regular expressions and a dictionary. Finally, we manually reviewed and corrected the transcription. The data processing scheme is shown in Figure 1

2.1.1 Voice Activity Detection

We used the Voice Activity Detection (VAD) module to split long audio files of many sentences into short speech segments corresponding to many new sentences. Additionally, large silences at the beginning and the end of each audio were removed. We utilized the a VAD model (Kim and Hahn, 2018) including a Long Short Term Memory Recurrent Neural Network (LSTM-RNN)-based classification.

2.1.2 Automatic Speech Recognition and Speech Punctuation

We utilized a Automatic Speech Recognition (ASR) system to obtain the time stamps of each word or each sound in each sentence. Moreover, the within-sentence pauses were identified and considered as potential punctuation. We marked a pause as a punctuation when its duration is greater than a threshold of 0.12 seconds. Then, the punctuation was added to input text. Without the added punctuation, the Tacotron-2 may align short pauses to any word or phoneme; which significantly reduce the quality of the synthesized voice.

The ASR acoustic model is the state-of-the art Time Delay Neural Network (Peddinti et al., 2015). To achieve the best performance on provided VLSP data, the language model is trained to over-fit the provided data.

2.2 Proposed text-to-speech systems

We proposed a text-to-speech system which is robust to noisy training data. Our system (as shown in Figure 2) was composed of a recurrent sequence-to-sequence feature prediction network called Tacotron-2, which mapped text embedding to acoustic features, followed by a Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis (HiFiGAN)-based vocoder. When using the HiFiGAN-based vocoder alone, we realized that the synthesized speech was noisy. As a result, we utilized the WaveGlow model to denoise the synthesized sound. Therefore, our proposed speech synthesis system includes a Tacotron-2 as a

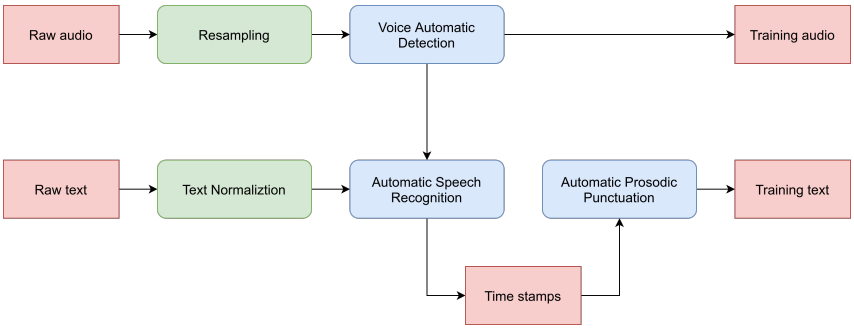

Figure 1: Data Processing Scheme

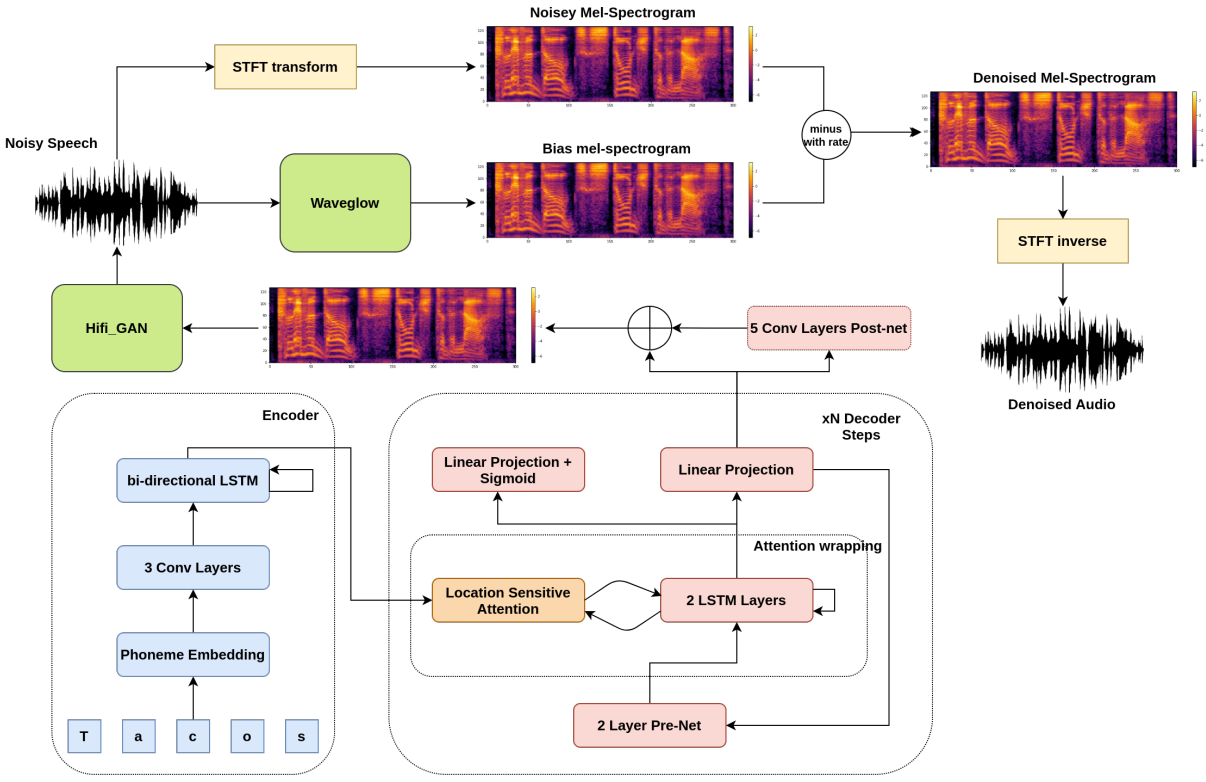

Figure 2: End-to-end system architecture

acoustic model, a HiFiGAN-based vocoder, and a WaveGlow denoiser.

- *Tacotron-2*: In previous VLSP 2019's TTS evaluation, Tacotron-2 was utilized in Vietnamese speech synthesis to achieve the best speech quality (Lam et al.). Therefore, we utilized Tacotron-2 as our TTS acoustic model. Our network architecture was almost similar to (Shen et al., 2017), with some modifications. Firstly, character embedding was used instead of phoneme embedding, which can take advantage of a more flexible and diverse pronunciation dictionary for the Vietnamese dataset. Lastly, we changed some parameters to better fit the data set which has a sampling rate of 22050 Hz, a minimum frequency of 75 Hz, and a amaximum frequency of 7600 Hz.

- *HiFiGAN*: To achieve better vocoding quality and higher efficiency, we utilized a HiFiGAN-based vocoder instead of WaveGlow vocoder. Our network architecture was similar to config V1 (Kong et al., 2020). A mel-spectrogram was used as input of generator and upsamples it through transposed convolutions until the length of the output sequence matches the temporal resolution of a raw waveform.

- *WaveGlow*: Our network architecture was similar to (Prenger et al., 2019). However, we only use WaveGlow for audio's noise reduction. First, we generate bias audio with mel-spectrogram from Tacotron-2 (sigma=0.0). And then we transform bias audio to bias mel-spectrogram. Next, for audio's noise reduction, we took the converted mel-spectrogram from the HiFi-GAN output minus the mel-spectrogram bias by a "denoiser strength" of 0.15. Finally, we obtained the last mel-spectrogram and converted it back to sound.

3 Experiments

The goal of the subjective experiments is to show the efficacy of our proposed method when the training data is noisy. We used the Tacotron-2 acoustic model in combination with different vocoders including 1) WaveGlow vocoder (denoted as WaveGlow), 2) HiFiGAN vocoder (denoted as HiFiGAN), and 3) our proposed method combining HiFiGAN-based vocoder and WaveGlow denoiser (denoted as HiFiGAN+Denoiser). The target natural speech is denoted as NAT.

3.1 Network Training

The original corpus contained 9 hours and 23 minutes of speaking from a female speaker. And after removing the unvoiced parts, the corpus had 8 hours and 21 minutes of speech. All data has been entered to train from scratch for the Tacotron-2 model. We also trained our HiFiGAN and WaveGlow model on the ground truth-aligned predictions.

3.2 Experimental Results

We submitted our proposed system (described in Section 2) to the VLSP 2020's TTS evaluation. The system was evaluated using the VLSP organizer's subjective MOS test. There were 24 participants listening to the stimuli of synthesized and natural speech. The participants gave each utterance a score on a 5-point scale including "very bad", "bad", "fair", "good", and "very good". Details of the results of the second MOS test are given in Table 1.

Our system	NAT
3.77	4.22

Table 1: Average MOS of our proposed system (described in Section 2) from VLSP's TTS evaluation

We conducted the second Mean Opinion Score (MOS) test to evaluate the performance of four vocoders (WaveGlow, HiFiGAN, and HiFiGAN+WaveGlow) in speech synthesis. Each listener listened to 20 test sentences and rate the quality of each sentence in a 5-point scale including "very bad", "bad", "fair", "good", and "very good". In total, there are 20 (sentences) $\times$ 4 (systems) = 80 (trials) [1] in a Latin-square design. We need 80 $\div$ 20 = 4 listeners to cover all the trials.There were 12 participants in the the test.

We summarize the perceptual characteristics of each speech synthesis systems in Table 2. The Figure 3 showed that our proposed system (denoted as HiFiGAN+Denoiser) has a highest MOS. The proposed system is better than natural speech (NAT) due to the fact that the target natural speech is noisy. The results showed that HiFiGAN vocoder outperformed WaveGlow vocoder when the training data is noisy.

We also ran the benchmarks for three models on the same Nvidia GTX 1080 Ti GPU hardware,

[1] Samples are available at: https://proptitclub.github.io/paper/index.html

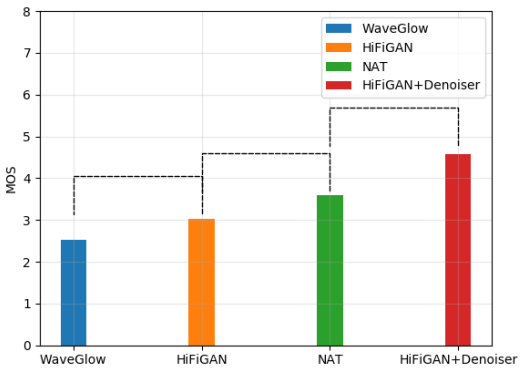

Figure 3: Average MOS of four systems. Dashed lines show statistically significant differences with p-value < 10^{-8}

Systems	Evaluate
WaveGlow	Each pronouncing word has a buzzer, however, the background noise is noticeable
HiFiGAN	The sound quality of each word has been improved, the background noise is moderate
HiFiGAN+Denoiser	The sound is clean

Table 2: Experimental reviews

with the same set of samples to show the inference efficicenty of using HiFiGAN-based vocoder. Statistics of real-time factor (RTF) values, which tells how many seconds of speech are generated in 1 second of wall time, are shown in Table 3. The results show that the speech synthesis rate of the model with HiFiGAN vocoder compared to the model with WaveGlow vocoder is 1.8 times, which hugely improves the speed performance of the system. For the system with both HiFiGAN and WaveGlow, the speed performance is approximate to the model using only HiFiGAN, because the denoising process of WaveGlow is not computationally exhausting. The results indicate that the HiFiGAN-based vocoder has better inference efficiency than the WaveGlow vocoder.

On the other hand, the resource consumption of our proposed model increases due to the use of both HiFiGAN and WaveGlow denoiser. While the number of HiFiGAN's parameters is 13.92 million, the WaveGlow has six times more parameters than HiFiGAN (as shown in Table 4). And the total number of parameters using both models is 101.65 million.

Systems	RTF
WaveGlow	4.00
HiFiGAN	7.37
HiFiGAN+Denoiser	7.25

Table 3: RTF results

Models	Param (M)
WaveGlow	87.73
HiFiGAN	13.92
HiFiGAN and WaveGlow	101.65

Table 4: Number of parameters

4 CONCLUSION AND FUTURE WORKS

In this report, we have presented our Vietnam TTS system for VLSP 2020. As for the challenge, our approach yields MOS result pretty close to this of natural speech. By testing various solutions to these challenges, we found that combining the methods to develop a custom vocoder played a significant role in the quality of synthesized speech. And the system efficiency was also significantly improved. As a result, the challenges of naturalness, background noise and buzzer noises in the artificial sound have been overcome. We plan to investigate other types of neural vocoders for improving the quality of speech synthesis.

References

Anh Tuan Dinh, Thanh Son Phan, Tat Thang Vu, and Chi Mai Luong. 2013. Vietnamese hmm-based speech synthesis with prosody information. In *Eighth ISCA Workshop on Speech Synthesis, Barcelona, Spain*.

J. Kim and M. Hahn. 2018. Voice activity detection using an adaptive context attention model. *IEEE Signal Processing Letters*, 25(8):1181–1185.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *ArXiv*, abs/2010.05646.

Phung Viet Lam, Phan Huy Kinh, Dinh Anh Tuan, Trieu Khuong Duy, and Nguyen Quoc

Bao. Development of zalo vietnamese text-to-speech for vlsp 2019. http://vlsp.org.vn/sites/default/files/2019-10/VLSP2019-TTS-PhungVietLam.pdf. Accessed: Oct, 2019.

Thinh Van Nguyen, Bao Quoc Nguyen, Kinh Huy Phan, and Hai Van Do. 2019. Development of vietnamese speech synthesis system using deep neural networks. In *Journal of Computer Science and Cybernetics*, volume 34, pages 349–363.

V. Peddinti, D. Povey, and S. Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

T. Phan, T. Duong, A. Dinh, T. Vu, and C. Luong. 2013. Improvement of naturalness for an hmm-based vietnamese speech synthesis using the prosodic information. In *The 2013 RIVF International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 276–281.

Viet Lam Phung, Phan Huy Kinh, Anh Tuan Dinh, and Quoc Bao Nguyen. 2020. Data processing for optimizing naturalness of vietnamese text-to-speech system. *arXiv:2004.09607*.

R. Prenger, R. Valle, and B. Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884.

Nguyen Thi Thu Trang, Nguyen Hoang Ky, Pham Quang Minh, and Vu Duy Manh. 2020. Vietnamese text-to-speech shared task vlsp 2020: Remaining problems with state-of-the-art techniques in proceedings of the seventh international workshop on vietnamese language and speech processing (vlsp 2020). In *International workshop on Vietnamese Language and Speech Processing (VLSP 2020)*.

Nguyen Thi Thu Trang, Albert Rilliard, Tran Do Dat, and Christophe d'Alessandro. 2013. Prosodic phrasing modeling for vietnamese tts using syntactic information. In *Proceedings of Interspeech, Lyon, France*.

Dinh Anh Tuan, Phi Tung Lam, and Phan Dang Hung. 2012. A study of text normalization in vietnamese for text-to-speech system. In *Proceedings of Oriental COCOSDA Conference, Macau, China*.

Dinh Anh Tuan, Phan Thanh Son, and Masato Akagi. 2016. Quality improvement of vietnamese hmm-based speech synthesis system based on decomposition of naturalness and intelligibility using non-negative matrix factorization. In *Advances in Information and Communication Technology. ICTA 2016. Advances in Intelligent Systems and Computing, vol 538. Springer, Cham.*