Improving prosodic phrasing of Vietnamese text-to-speech systems

Pham Ngoc Phuong
Thai Nguyen University
phuongpn@tnu.edu.vn

Chung Tran Quang
Hanoi University of Science and Technology
chungtran@vais.vn

Quang Minh Nguyen
Vietnam Artificial Intelligence Solution
minhnq@vais.vn

Quoc Truong Do
Vietnam Artificial Intelligence Solution
truongdo@vais.vn

Abstract

End-to-end TTS architecture which is based on Tacotron2 is the state-of-art system. It breaks the traditional system framework to directly converts text input to speech output. Although it is shown that Tacotron2 is superior to traditional piping systems in terms of speech naturalness, it still has many defects in building Vietnamese TTS: 1) Not good at prosodic phrasing for long sentences, 2) Not good at expression for foreign words. In this paper, we used 2 methods to solve these defects: 1) Pause detection system for predicting and inserting punctuation into long sentences to improve speech naturalness. 2) Translation system for transcribing foreign words to Vietnamese words. In the VLSP 2020 evaluation campaign, our model achieved a mean opinion score (MOS) of 3.31/5 compared to 4.22/5 of humans.

Index Terms— Text-to-speech, TTS, Vietnamese TTS, end-to-end speech synthesis

1 Introduction

Text-to-Speech (TTS) study is widely applied in real-life but it is still a challenge in the field of speech processing. Many techniques have been proposed such as concatenative synthesis (Hunt and Black, 1996), statistical parametric speech synthesis (SPSS). Although concatenative synthesis can reach highly natural synthesized speech, the approach is inherently limited by properties of the speech corpus used for the unit selection process. Meanwhile, SPSS allows product direct speech smoothly and intelligibly by a vocoder. A full SPSS system consists of text analysis, feature generation, and waveform generation modules a, some SPSS techniques are used for Vietnamese TTS: Hidden Markov model (HMM) (Tokuda et al., 2000), Deep neural networks(DNN) (Ze et al., 2013), generative adversarial networks (GAN)(Saito et al., 2017) and End-to-end architectures(Wang et al., 2017). Currently, DNN approaches have gradually replaced HMM models for the duration model and acoustics model. However, the generated voice is often muffled and becomes unnatural. Wavenet (Oord et al., 2016), Wave RNN (Kalchbrenner et al., 2018), GAN (Saito et al., 2017) produces audio with significantly improved naturalness but requirements deep experience and voices that are not as realistic as they are in reality. An end-to-end architecture (Tacontron 2 and WaveGlow vocoder) include five components: linguistic analysis, acoustic model, duration model, parameter generation, and post-filtering are replaced by encoder-attention-decoder networks (Wang et al., 2017; Shen et al., 2018), to be able to effectively optimize the mapping from input text to acoustic features. Finally, a neural vocoder such as Waveglow generated a waveform from the generated mel-spectrogram.

However, in a long sentences or long phrases, speech synthesis results will not be natural. This comes from the fact that human speakers usually break phrases by inserting word transitions instead of punctuation for the sake of expressivity, better comprehension or only taking a breath. The term phrasing is used to describe the phenomenon of grouping words into phrases and separating these phrases with pauses or punctuation inserts. In addition, there are many foreign words in the sentences that are not in the Vietnamese phonetic dictionary. If only replacing foreign words with International Phonetic Alphabet (IPA), the synthesized sentence will not be pronounced in Vietnamese standard. In this paper, 2 methods are applied to synthesize sen-

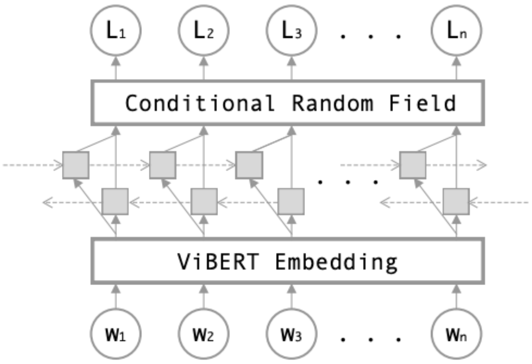

Figure 1: The CaPu model insert the punctuation into the sentences.

tences more naturally: 1) Pause detection module will insert punctuation into sentences to improve prosody of the TTS system, 2) Translation module will transforms foreign words into the Vietnamese standard pronunciation word.

2 Prosodic and pronunciation modeling

2.1 Prosodic modeling

When reading long sentences, the reader always stops at the punctuation or at the position of two or more words of equal syntactic importance (such as noun, verb, etc). So, pause prosodic detection is extremely important affecting the prosody of the TTS system. However, the provided data from the VLSP organization (Trang et al., 2020) was the result of the ASR system, so it had the text only. The synthetic sound quality of the deep neural network depends on the input data. Thus, adding the punctuation at a suitable position can enhance the prosody of our system. To solve this challenging problem, we integrate the Capitalization and Punctuation (CaPu) model (Nguyen et al., 2020) to recover the punctuation of the sentences. The CaPu model not only inserts the punctuation automatically to correct the text format but also places the punctuation at the location relating to breathing.

The CaPu model includes three components that is the embedding layer, the recurrent layer, and the classification layer. More specifically, the embedding layers is ViBERT model that embedded the input sentences to the fixed vectors. The fixed vectors passed through the bidirectional GRU layers. followed by the conditional random field layer to classify the punctuation-tag of each input word. ViBERT is a variation of RoBERTa$_{base}$ model with fewer layers than the original model, it contains 4 encoder layers, the number of heads is 4 and the hidden dimension size is 512. The model has 4 bidirectional GRU layers, the hidden size of GRU cell is 512. The figure 1 depicts CaPu architecture.

To train CaPu model, we collected a huge of text from many domains on the internet including wikipedia, law, politics, etc. This document has the punctuation in accordance with Vietnamese standard style. To mimic the pause of the reader, we use word time-stamp of the ASR system. If the silent time is more than 0.3 second, we put the commas at this silent position. Finally, we trained the CaPu model with the processed data. As a result, CaPu model can insert the punctuation at the proper location by 2 strategy, Vietnamese standard and reader style. Besides, we also added a dot at the end of transcript text to present the end of audio. The result of the CaPu model:

Raw transcript:

cảm giác đó đến một cách đột ngột nhưng mụ xua đuổi nó đi không cho nó chạm tới mụ cũng như không để cho nó chạm tới nền cộng hòa

After add commas to transcript:

cảm giác đó đến một cách đột ngột **,** nhưng mụ xua đuổi nó đi **,** không cho nó chạm tới mụ **,** cũng như không để cho nó chạm tới nền cộng hòa **.**

2.2 Pronunciation modeling

One of the biggest challenges for the VLSP Text-To-Speech (Trang et al., 2020) is that the transcript text has many foreign words. Because foreign words are out of the Vietnamese vocabulary and can not convert to the phoneme directly. This leads to trouble for the participants when joining and building the Vietnamese TTS system. To handle and tackle this problem, we used Vietnamese sound to pronounce these English words. For example, “kuttner” will be pronounced by “cắt nơ”, seeing more examples in Table 1. In order to transform from foreign words to Vietnamese words, we used the popular translation model-Transformer$_{base}$ (Vaswani et al., 2017) model.

The Transformer architecture has two modules, the encoder, and the decoder, and 2 component is connected through an attention mechanism. The Transformer model that we used for this challenge is composed of a stack of N=6 identical layers for both the encoder and decoder.

To train this translation model, we must create a large number of pair of English-Vietnamese words. The total dataset that we produced is more than 1

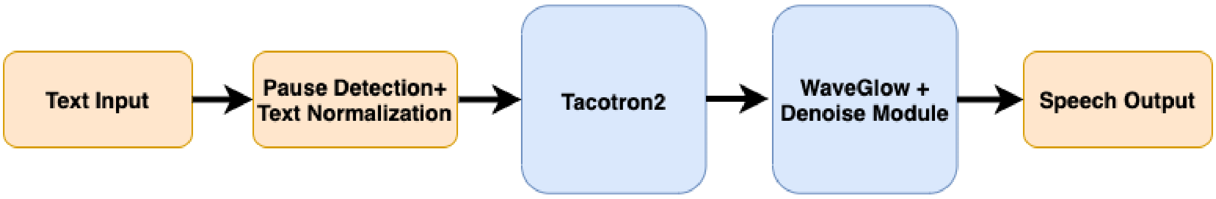
Text Input
Pause Detection+ Text Normalization
Tacotron2
WaveGlow + Denoise Module
Speech Output

Figure 2: Our TTS pipeline, the input text passes to the pause detection and text normalization module. Subsequently, the processed data passes to Tacotron2 and WaveGlow to generate speech synthesis

English word	Vietnamese word
kuttner	cắt nơ
Anderson	an đơ sơn
vera	vê ra
reme	rê mi

Table 1: Convert English words to Vietnamese words

hundred million pairs. The result of the translation model was displayed in Table 1.

3 Text-to-Speech System

Nowadays, for the TTS task, the end-to-end speech synthesis pipeline consists of two phases, 1. converting text to Mel-spectrogram and 2. converting Mel-spectrogram to waveform synthesis. The model Tacotron2 combining with WaveGlow vocoder is still state-of-the-art for the TTS task. Tacotron2 is a deep neural network receiving a text to predict Mel-spectrogram signal. Then Mel-spectrogram will be converted to waveform thanks to WaveGlow. However, we realized that synthetic speech was noisy. Therefore, we used a denoiser model, attaching at the end of the WaveGlow model.

- *Tacotron2*: The network has two components an encoder and a decoder. We had a small change comparing with the original model. To adapt to the characteristic of the Vietnamese language, the input model was phoneme level instead of character level. Phoneme character passed to the embedding layer, which represented by 512-dimensional. Afterward, these vectors passed through a stack of 3 convolutional layers, followed by single bi-directional LSTM layers to generate the encoded features. The encoder output was consumed by an attention network which yielded a fixed-dimensional vector. Finally, the decoder had the mission of converting this vector to a Mel-spectrogram. To train the Tacotron2 model, we minimized the output of the model with ground trust using mean squared error(MSE).

- *WaveGlow*: The network that we used for the TTS challenge was similar to the original model. The model transformed the output of the Tacotron2 model to the waveform signals. WaveGlow is deployed using only a single network and single cost function, so it is fast, efficient and can produce high quality audio synthesis. The network has 12 coupling layers and 12 invertible 1 x 1 convolutions. In coupling module has 8 layers of dilated convolutions with 512 channels used as residual connections and 256 channels in the skip connection. For the challenge, we used the pre-trained model provided by the author to synthesize the audio.

- *Denoise Module*: This module will reduce the noise of synthetic audio generated from WaveGlow. Firstly, we produced bias audio by using WaveGlow infer a zero Mel-spectrogram with shape 1x80x88. Then both synthetic audio and bias audio will be transformed to Mel-spectrogram by the short-time Fourier transform method. Next, we used the synthetic Mel-spectrogram minus the bias Mel-spectrogram. As a result, we received the final Mel-spectrogram and we used the inverse Fourier transform function to convert it back to audio.

4 Experimental Setup

4.1 Dataset

The duration of the training dataset is about 5-6 hours of a single female speaker and has 7770 audio files. The duration of each file is from 2s to 11s. The sample rate is 44100Hz, 2 channels. To train the model, we resampled to a sample rate of 20500Hz and also convert it to mono channel (1 channel). Besides, we decreased the volume of each file audio by 50%. To reduce noise for the training data, audio in training dataset will be trimmed the silence at start and end position. All transcript text in the dataset is spelled out, for example, “30” is written as “ba mươi”.

Data Processing	Evaluation
No	Speech synthesis can not read the foreign words, the pause in the sentences is unnatural
Pause detection	Speech synthesis can pause at the punctuation correctly, prosody seem naturally
Pause detection + Text Normalization	Speech synthesis can pronounce foreign words.

Table 2: Data processing and evaluate the system

4.2 Experimental Setup

Both CaPu and translation model were implemented by Fairseq (Ott et al., 2019) framework. We used Adam optimizer with beta factor (0.9, 0.98), the learning rate of 0.0005. Conditional Random Field (CRF) loss was applied to train the model and the learning rate scheduler was the inverse square root. The warm-up initial learning rate is 1e-7, and the batch size is 64.

To train the Tacotron2 model, we use GeForce RTX 2080 Ti, 11GB, the learning rate is 1e-3, the weight decay is 1e-6 , the batch size is 64. Adam optimizer with β_1=0.9 and β_2=0.999, $epsilon$=1e-6.

5 Result

We used Tacotron2+Waveglow to evaluate the TTS system. We conducted many experiments relating to data processing, see Table 2 for more detail. Finally, when we combined 2 methods processing pause detection and text normalization, the TTS system yielded speech synthesis naturally. Not only prosody seem natural, but also our system can pronounce foreign words similar to Vietnamese people.

MOS was applied to evaluate the system. The speech synthesis was evaluated by three groups of listeners: speech experts, volunteers, and undergraduates. The listeners will have 5 options to give a score from 1-5: excellent(5), good(4), fair(3), poor(2), 1(bad).

In the VLSP 2020's challenge, as shown in Table 3, our architecture achieved a MOS of 3.31 for the naturalness. For intelligibility, the rate of hearing correct words is 83.10% and the rate of listening to correct syllabi's is 82.90%

	MOS
Our system	3.31
Human	4.22

Table 3: MOS Result for the VLSP Dataset

6 Conclusion and future works

In this paper, we describe our architecture for the Vietnamese Text-to-speech system. For the data from an organization, our approach yielded a MOS of 3.31. By conducting many experiments, we realized that data processing is very important in this challenge. By converting English words to Vietnamese words, also add commas to transcript text, these techniques assist model producing utterance synthesis very naturally.

In the future, we can experiment with more state-of-the-art architecture such as Hifi-Gan, Mel-Gan, Glow-TTS. Also, exploring many challenges of TTS such as how to training TTS with small data, TTS adaptation, etc.

References

Andrew J Hunt and Alan W Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 373–376. IEEE.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. *arXiv preprint arXiv:1802.08435*.

Thai Binh Nguyen, Quang Minh Nguyen, Thi Thu Hien Nguyen, Quoc Truong Do, and Chi Mai Luong. 2020. Improving vietnamese named entity recognition from speech using word capitalization and punctuation recovery models. *Proc. Interspeech 2020*, pages 4263–4267.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2017. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):84–96.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 2000. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE.

Nguyen Thi Thu Trang, Nguyen Hoang Ky, Pham Quang Minh, and Vu Duy Manh. 2020. Remaining problems with state-of-the-art techniques in proceedings of the seventh international workshop on vietnamese language and speech processing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Afully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*.

Heiga Ze, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE.