VNU Journal of Science:
Computer Science and Communication Engineering
Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

TTS - VLSP 2021: Development of Smartcall Vietnamese Text-to-Speech

Le Ba Hoai[1,*], Nguyen Van Hoc[1,*], Dam Ba Quyen[1],
Nguyen Thu Phuong[2], Nguyen Quoc Bao[1,2]

[1]Smartcall JSC. Hanoi, Vietnam
[2]*Information and Communication Technology University, ThaiNguyen, Vietnam*

Received 27 December 2021
Revised 5 April 2022; Accepted 5 May 2022

Abstract: Recent advances in deep learning facilitate the development of end-to-end Vietnamese text-to-speech (TTS) systems with high intelligibility and naturalness in the presence of a clean training corpus. Given a rich source of audio recording data on the Internet, TTS has excellent potential for growth if it can take advantage of this data source. However, the quality of these data is often not sufficient for training TTS systems, e.g., noisy audio. In this paper, we propose an approach that preprocesses noisy found data on the Internet and trains a high-quality TTS model on the processed data. The VLSP-provided training data was thoroughly preprocessed using 1) voice activity detection, 2) automatic speech recognition-based prosodic punctuation insertion, and 3) Spleeter, source separation tool, for separating voice from background music. Moreover, we utilize a state-of-the-art TTS system that takes advantage of the Conditional Variational Autoencoder with the Adversarial Learning model. Our experiment showed that the proposed TTS system trained on the preprocessed data achieved a good result on the provided noisy dataset.

Keywords: Text-to-speech, Spontaneous Speech, Vietnamese.

1. Introduction

Text-to-Speech (TTS) is a technology that converts any text into a speech signal. With TTS, human-machine communication is easier and more natural than ever. As a result, it has great potential and can be applied to several applications (e.g., audiobooks, movie narrations, response services in telecommunications, and virtual assistants). Through decades of research and development, end-to-end speech synthesis systems for a single language have achieved outstanding results and produced natural human-like voices even in real-time. Based on these advances, recent end-to-end neural TTS models have been extended to enable control of speaker identity, controllability, or multilingual.

* Corresponding author.
E-mail address: lebahoaidongson@gmail.com

https://doi.org/10.25073/2588-1086/vnucsce.348

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 the naturalness of end-to-end TTS systems trained on Vietnamese found data [1]. Text normalization methods were explored, utilizing regular expressions and language model [2]. New prosodic features (e.g., phrase breaks) were investigated, which showed their efficacy in improving the naturalness of Vietnamese hidden Markov models (HMM)-based TTS systems [3, 4]. The pronunciation of foreign words is also improved [5]. For post-filtering, it was shown that a global variance scaling method might destroy the tonal information; therefore, exemplar-based voice conversion methods were utilized in post-filtering to preserve the tonal information [6]. With clean datasets, existing approaches give good results and produce quality sound. At the VLSP 2020 competition, [7] proposed an approach to utilize an end-to-end TTS system that takes advantage of the Tacotron-2 [8] acoustic model, and a custom vocoder combining with a High Fidelity Generative Adversarial Networks [9] (HiFiGAN)-based vocoder and a WaveGlow [10] denoiser. The use of sophisticated models (e.g., HiFiGAN) is probably a key role of the success of the TTS systems. Recently, the English TTS systems using the Conditional Variational Autoencoder with the Adversarial Learning model (VITS have outperformed HiFiGAN-based systems) [11]. Therefore, we examine the efficacy of the state-of-the art VITS for building Vietnamese TTS systems.

Developing the Vietnamese TTS systems with sophisticated models requires a large amount of good quality data. However, building such dataset is costly because it requires professional speakers and dedicated recording equipment. Moreover, according to our observations, potential data for training the TTS model is available on many websites such as Youtube, Facebook,... If we can take advantage of this rich data source, we will save the cost of building datasets and making TTS technology accessible to more people. The use of the found data for building TTS systems has received more and more attention in recent years [1, 12, 13].

Dataset of the competition [14] exploited the voice source from a female youtuber. The challenges of using spontaneous speech are

i) poor quality (e.g., inconsistent speaking rate)

ii) background noise such as music,

iii) sometimes mixed with other voices,

iv) differences in intensity, stress, prosody and voice styles (voice with diverse rhythms) across the dataset.

v) wrong transcripts (although validated manually), making it difficult for current TTS models to learn to produce good quality voices.

Spontaneous speech in TTS is especially important if you want to apply TTS in natural conversations, instead of just using TTS for readable text. With the training data being spontaneous speech, the quality of TTS will be more natural in human-machine communication applications. To build a spontaneous speech training dataset in ideal studio conditions is not easy, expensive and time consuming when the speakers do not have a prepared script in advance. Therefore, it is very necessary to exploit the available spontaneous speech data sources.

Table 1. Average MOS of our proposed system from VLSP's TTS evaluation (shown in final report [14]):

Test	MOS
in-domain	3.8
out-domain	3.5

In this paper, we propose an approach that uses spontaneous speech datasets to build a TTS model that produces high-quality voices. Our approach is based on a rigorous preprocessing pipeline [1] and the Conditional Variational Autoencoder with the VITS. Preprocessing pipelines include:

1) Spleeter-based noise processing [15], 2) Sentence splitting by Voice Activity Detection, and 3) Automatic Speech Recognition-based text and punctuation normalization. Our experiment shows that the proposed TTS system trained on preprocessed data achieved good results on a

non-clean dataset with a Mean Opinion Score (MOS) of 3.8 on the test in domain and 3.5 on the test out domain (as shown in Table 1).

2. Data Pre-processing

In this section, we present our approach to build high-quality TTS systems on noisy training corpus.

2.1. Data Preprocessing

Our dataset consists of many problems: i) background music, ii) other people's voices, iii) fast speaking rate, iv) slow speaking rate, v) unavailability of speech, vi) labeled words not correct. Our processing steps are: remove the abnormal speaking rate, merge them to reduce noise, then trim the audio again based on Voice Activity Detection (VAD), relabel and add punctuation using the Automatic Speech Recognition (ASR) system. We address the problems with a data processing scheme (as shown in Figure 1).

2.1.1. Data Selection

The representation of each word will vary depending on the pronunciation speed of the speaker. We calculate the speaking rate (words per second - WPS). Then, we exclude audio with a slower or faster rate than usual. After statistics and listening to many different levels, we select a speaking rate from 2.9 to 5.4. In addition, we also only used audio with word count >= 4. Other male voices usually have a reasonably fast speaking speed WPS > 5, so we check the audios with WPS > 5 and listen to the audios immediately at the end. Together we sorted out about 65 audios containing male voices. Audios containing many strange noises such as crowd sounds, prolonged laughter, and music usually have WPS < 2, so they are eliminated. Finally, after this WPS step, we removed 1249 audios from the raw dataset and left 4029 audios to do the following steps.

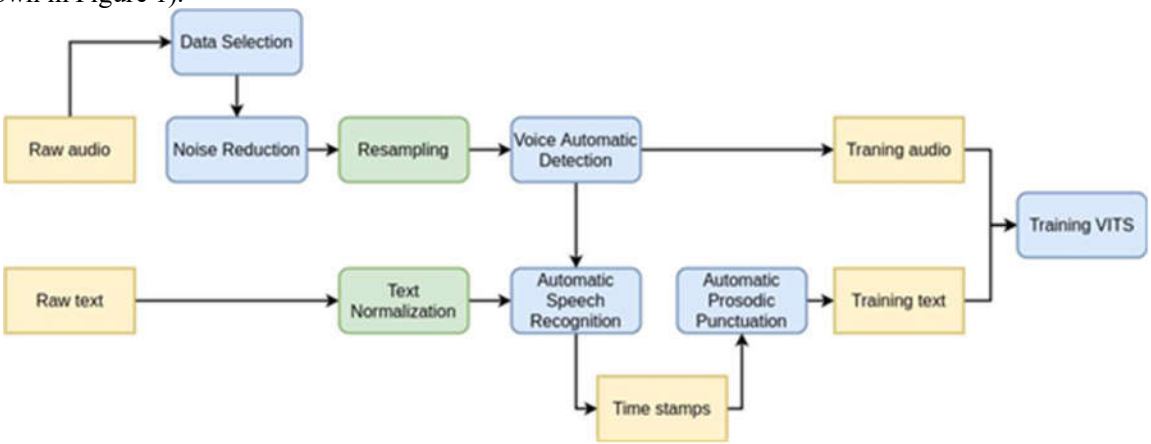

Figure 1. Data Processing Scheme.

2.1.2. Noise Reduction

The biggest challenge for provided (Youtube) data is background music. We consider background music as another sound source and refer to the problem of sound source separation for processing. Source separation can be thought of as speaker dimerization but for music. The speaker dimerization models have to differentiate between the voices of different speakers and then split the original audio into multiple tracks corresponding to each speaker. We apply Spleeter, which is a source separation Python library created by the Deezer R&D team [15], to address the problem. It provides pre-trained state-of-the-art models built using Tensorflow for various source separation tasks. Our original audio is short, mostly three to five seconds. We realize that Spleeter does not give good results with the short audios. Therefore, we

concatenate every 50 audios into a longer one with a duration of three to five minutes. Finally, we apply Spleeter for denoising. We perceive that the noise is drastically reduced.

2.1.3. Voice Activity Detection

We use the Voice Activity Detection (VAD) module to split long audio files of many sentences into short speech segments corresponding to many new sentences. Additionally, significant silences at the beginning and the end of each audio were removed. We utilized the VAD model [16] including a Long Short

Term Memory Recurrent Neural Network (LSTM-RNN)-based classification.

2.1.4. Automatic Speech Recognition and Speech Punctuation Insertion

We use the Automatic Speech Recognition (ASR) system to get the timestamp of each word or sound in a given case and a given sentence. Furthermore, punctuation marks in sentences are identified and considered potential punctuation marks. With timestamps defined, we mark the pause as a punctuation mark when its duration is more significant than the 0.12-second threshold. Then punctuation is added to the input text. Suppose the input text to the VITS model does not contain punctuation. In that case, the model will encounter audio with no pauses in the sentence, significantly reducing the synthesized voice quality. Adding punctuation helps the model understand that the sound is a pauses audio instead of understand that the sound is still from the previous word. The inflection and tone that it uses will vary depending on the punctuation you use. It makes the TTS voice achieve the intended meaning and clarity as well as enlivening it. The ASR audio model is a modern time-delay neural network [17]. The language model is trained to get the best performance on fed VLSP data.

3. System Architecture

We propose a text-to-speech system that can synthesize speech from the text in the most natural way when training on a dataset of multiple speech styles. We use Variational Inference with adversarial learning for end-to-end Text-to-Speech [11] or VITS for short for our system (shown in Figure 2). Our system was composed of a end-to-end model for speech synthesis called VITS and the input to the system is processed using character embedding before going into the model for speech synthesis. The model can process the audio data of voices with different speech styles and aims to reproduce the most natural sound possible.

VITS: We compare our system VITS to baseline systems such as Tacotron2+Waveglow, Tacotron2+HifiGAN, Speedyspeech+HifiGAN, but these models are all weak in handling inconsistent data in voice quality, reading speed, and background noise,... According to the evaluation results published in the VITS report, when applying this method to the English language, the results show that VITS's MOS score is the highest when compared to previous methods like Tacotron2+Hifigan, and the results are also close to reality.

Our model architecture is almost the same as [11], with a bit of change to work with Vietnamese. First, we change the embedding way to fit the Vietnamese sentence dataset,instead of using the text-to-phonemes conversion as suggested in the original source-code, we use the easier way of character embedding instead, specifically we split the input sentence into characters and numbering. Those characters follow the Vietnamese alphabet, including the accents in Vietnamese. After embedding, the data will be put into VITS to proceed to create the sound. Lastly, we have a few changes to some of the parameters, including changing the sampling rate to be 22.05kHz.

4. Experiments

In this section, we show the effectiveness of the method we used when training the data for various intonation and speaking styles. As we said, the model we use is the end-to-end VITS model for training.

4.1. Data

The data provided by the organizers is the data of one speaker, there is no consistency in speaking style between audio files, and the audio contains much noise in which typical forms such as background music, laughter, the sound of objects, noise from the environment. Some audio has more than one speaker voice. As for the text, there are many non-Vietnamese words, and the pronunciation of the labeled words is not correct. The punctuation also disappears. In addition, there are misspelled Vietnamese words such as "s" and "x", "tr", and "ch". 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-di'), which can make it easier for the model to learn the fricative sounds. In general, the provided data has a lot of noise and difficulty training the model. We manually reprocess the data to make sure the labeled words are correct. Our processed data is available here.

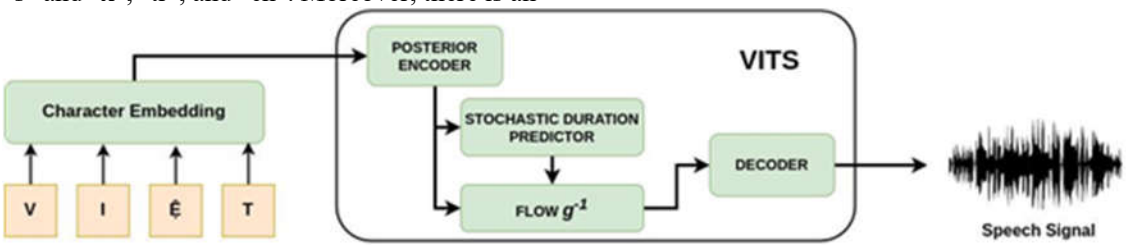

Figure 2. End-to-end system architecture.

4.2. Network Training

The VLSP-provided corpus contains 7 hours and 12 minutes of speech from a female speaker. Moreover, after preprocessing, the corpus had 5 hours and 21 minutes of speech. We train the VITS model with a batch size of 16, using the Adam optimizer configured with an initial learning rate of 0.0002 from scratch. It took six days on one Nvidia RTX 3090 to create a usable model.

4.3. Experimental Results

We submitted our proposed system (described in Section) to the VLSP 2021's TTS evaluation [18]. The system was evaluated using the VLSP organizer's subjective MOS test. Twenty-four participants were listening to the stimuli of synthesized and natural speech. there are both 100 indomain and 100 outdomain sentences synthesized from the model. The participants gave each utterance a score on a 5-point scale, including "excellent" (5), "good" (4),"fair" (3), "poor" (2), "bad" (1). Details of the second MOS test results are given in Table 1. The results are calculated based on two sets of in-domain and out-domain, where in-domain are texts with the same topic and style as the training text, out-domain are sentences with different topics and styles.The above difference occurs because the text of the indomain is spontaneous sentences and the speed of the voice is fast, the way to express emotions through sentences is very natural when compared to the text of out domain which are sentences and dialogues in stories, in plays. The in-domain MOS score is higher than the out-domain one as expected. Our intelligibility result is not good (78,2%) because we keep the original speaking speed which is already faster than normal. Speak speed is too fast can distorted word or miss word so it's harder to hear and understand than usual. The samples can be found here.

5. Conclusion

In this paper, we present our Vietnam TTS system for VLSP 2021, a method to build a TTS model that generates a good voice from data collected on the internet. Our approach yields

positive results with a preprocessing pipeline and a model capable of learning from data of different voice styles. The challenges of naturalness, background noise, voice style, and background music in synthesized voice have been overcome. With this approach, we take advantage of the rich source of recorded data on the internet, making TTS accessible to many people who cannot afford to build their datasets. We plan to research and apply a denoiser model to the post-processing step to make the voice smoother for improving the quality of speech synthesis.

Acknowledgments

This work has been supported by Vietnam National University, Hanoi (VNU), under Project No. QG.14.04.

References

[1] V. L. Phung, P. H. Kinh, A. T. Dinh, Q. B. Nguyen, Data processing for optimizing naturalness of vietnamese text-to-speech system (2020). arXiv:2004.09607.

[2] D. A. Tuan, P. T. Lam, P. D. Hung, A study of text normalization in vietnamese for text-to-speech system, in: Proceedings of Oriental COCOSDA Conference, Macau, China, 2012.

[3] A. T. Dinh, T. S. Phan, T. T. Vu, C. M. Luong, 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), 2013, pp. 276–281.

[4] T. T. T. Nguyen, A. Rilliard, D. D. Tran, D'Alessandro, Prosodic phrasing modeling for Vietnamese TTS using syntactic information, in: Annual Conference of the International Speech Communication Association (INTERSPEECH 2014), Singapore, Singapore, 2014, pp. 2332–2336.

[5] C. M. Nguyen, L. V. Phung, C. T. Bui, T. V. Truong, H. T. Nguyen, Learning vietnamese-english code-switching speech synthesis model under limited code-switched data scenario, in: D. N. Pham, T. Theeramunkong, Governatori, F. Liu (Eds.), PRICAI 2021: Trends in Artificial Intelligence, Springer International Publishing, Cham, 2021, pp. 153–163.

[6] D. A. Tuan, P. T. Son, M. Akagi, 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, 2016.

[7] M. C. Nguyen, K. D. Trieu, B. Q. Dam, Q. B. Nguyen, Development of smartcall Vietnamese text-to-speech for VLSP 2020, in: Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, Association for Computational Lingustics, Hanoi, Vietnam, 2020, pp. 24–29.

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu, Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions, CoRR abs/1712.05884. arXiv:1712.05884.

[9] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, CoRR abs/2010.05646. arXiv:2010.05646.

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

[11] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, CoRR abs/2106.06103.

[12] E. Cooper, A. Chang, Y. Levitan, J. Hirschberg, Data Selection and Adaptation for Naturalness in HMM-Based Speech Synthesis, in: Proc. Interspeech 2016, 2016, pp. 357–361.

[13] F. . Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, I. Ouyang, Data selection for improving naturalness of tts voices trained on small found corpuses, in: 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 319–324. doi:10.1109/SLT.2018.8639642.

[14] N. T. T. Trang, N. H. Ky, Vlsp 2021 - tts challenge: Vietnamese spontaneous speech synthesis, VNU Journal of Science: Computer Science and Communication Engineering 38 (1).

[15] R. Hennequin, A. Khlif, F. Voituret, M.

Moussallam, Spleeter: a fast and efficient music source separation tool with pre-trained models, Journal of Open Source Software 5 (50) (2020) 2154, deezer Research.

[16] J. Kim, M. Hahn, Voice activity detection using an adaptive context attention model, IEEE Signal Processing Letters 25 (8) (2018) 1181–1185. doi:10.1109/LSP.2018.2811740.

[17] V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[18] A. for Vietnamese Language, S. Processing, Vlsp 2021 - vietnamese text-to-speech, https://vlsp.org.vn/vlsp2021/eval/tts (2021).