The Fine-tuning based Speech Synthesis System entry in the VLSP Campaign 2019

Luong Tuan Dung
IOT department, VNG corporation
tuanluong04011996@gmail.com

***Abstract*—This paper describes my approach in text to speech shared-task in VLSP Campaign 2019. The system is based on Tacotron 2 architecture and modified to suit Vietnamese pronunciation. In addition, a preprocessing method for both audio and text training data is used, in order to increase the quality of the training corpus - an important condition that determines the quality of the output audio. I also present all my training experiments and fine-tuning models during the competition. My model achieves average MOS score 3.70 on Big Corpus and 3.99 on Small Corpus.**

***Index Terms*—Text-to-speech, Speech synthesis, Tacotron 2, End-to-end, Fine-tune**

I. Introduction

Nowadays, with the development of artificial intelligence, people are trying to build systems to imitate human behavior, and one of them is how human speak. The problem of generating speech from raw text has been researched and developed for decades and various techniques have been proposed. Concatenative speech synthesis was the state-of-the-art for many years which is based on the stringing together of of small segments of recorded speech [1], [2]. Statistical parametric speech synthesis (SPSS), which directly generates of speech features to be synthesized by a vocoder. Two commonly used acoustic features extractor are Hidden Markov Models (HMMs) [3]–[5] and Deep Neural Network (DNN) [5], [6]. SPSS solves many of the issues that Concatenate Speech Synthesis had, however, the output audio of these systems often sounds muffled and unnatural compared to human speech. Recently, end-to-end models have become a new trend with the introduction of models such as Tacotron [7], Tacotron2 [8], DeepVoice [9],... capable of generating voices perfectly human.

Several researches have been done to tackle TTS problem for Vietnamese. HMMs based methods achieve good intelligibility like T. T. Vu et al. system [10], or Q. S. Trinh system [11]. However, since HMMs bases on Markov assumption, it is not able to capture long-term dependencies, which can further improve naturalness. In VLSP Campaign 2018, DNN-based speech synthesis system of N. V. Thinh et al. system [12] surpassed many other systems to become champions with outperformed results in all threesubjects including naturalness, intelligibility, and MOS.

In this paper, I describe all my experiments to tackle Text to speech shared-task in VLSP Campaign 2019. The rest of the paper is organized as follows, section 2 presents my preprocess data phase and used model, section 3 describes our experimental setup and results, Section 4 concludes the paper.

II. Methodology

A. Audio Data Preprocessing

Text to speech shared-task in VLSP Campaign 2019 is composed of two training datasets: A big training dataset contains 15,000 utterances of a northern female speaker (about 23 hours) and A small training dataset with 1,000 utterances of a southern female speaker (about 45 minutes). Audio files in the big dataset was recorded in many different environments, mic holding distance, therefore there are different among them. Furthermore, some audio files contain loud white-noise and speaker speaks incoherently. In the other hand, audio files in small dataset are all clean.

After manual checking the big corpus, I discovered that the recording files recorded at the same time which share the same attributes of mic distance, background noise, voice quality, are arranged in sequence. I therefore grouped those files into 10 groups with the same time recorded before manually grouping them into 3 groups with different quality: good, medium, poor. The specific results are as follows:

- *Good files* (1825 files from 03187.wav to 05733.wav): files have almost no background noise.
- *Poor files* (116 files from 00984.wav to 01121.wav): files have background noise, muffled sound and mic set far away.
- *Medium files* (the rest): files have white noise

Then perform manual checks and corrections for files of good quality, including eliminating noise, cutting silence in sentences so that the voice naturally, eliminating recording errors.

For the rest of the big dataset and the whole small dataset, I removed the redundancy at the beginning and the end of audio files and normalized their amplitudes. Then resample all files to 22050Hz for training phase.

B. Text Data Preprocessing

For transcript files, we standardize syllables that do not belong to Vietnamese syllables such as numbers, foreign words and abbreviations by converting them to their standard phonetic forms using a handcraft dictionary and rules.

Vietnamese words are composed of syllables. The natural way to speak a compound word is speak all syllables quickly.

In the training dataset, the way the collaborator reads the text is often incoherent, pausing sometimes not right with natural speaking. Some words in training audio have the pause between 2 syllables of the same word. I do not want model learn this behavier, so, a Vietnamese Word Segmenter[1] is used to produce "_" symbols between syllables of the same words.

xa hoa → *xa_hoa*

The "_" character will be counted as a character in the character dictionary when training the model and the model will learn "_" as a "quick reading" symbol.

C. Model Architecture

The architecture is based on Tacotron 2 [8] model but changes the character dictionary set suitable for Vietnamese by replacing the character dictionary into the phonetic dictionary for Vietnamese [2] (including consonants, single vowels, double vowels, triple vowels in Vietnamese) and removed WaveNet layer of original architecture to reduce training time.

III. Experiments

All my experiments was trained on single GPU with batch-size 32. I use the Adam optimizer [13] with $\beta = 0.9$, $\beta = 0.999$, $\epsilon = 10^{-6}$ and a learning rate of 10^{-3} exponentially decaying to 10^5 starting after 50,000 iterations. We also apply L2 regularization with weight 10^{-6}.

In the first experiment, I trained the entire big dataset, the results obtained after 200000 steps generated sound files with lots of background noise, some generated audio has redundant noise in the end. Noise from training data affects significantly the model, namely white noise that is clearly visible in the spectrogram of generated audio like in some training audio. furthermore, the model has problem with short sentences which have only 1 or 2 words.

With 3 divided data groups, the good data group only has 1825 files, while the number of files has white noise takes 5 times bigger, so I decided to retrain a base model with a large amount of audio and then fine-tune the base model with a small audio set that has high quality. I also add 200 short audio which is cut from good audio files to handle the short sentences problems. The second experiment involved removing poor audio files from the training set and retraining the model. Stop the training process at step 50000 when the loss is about 0.56 (the model does not fit completely, at step 200000 in the first experiment, the training process converges and the loss is then 0.48).

Experiment three performed using a model from experiment 2 and continued fine-tune on a set of good quality audio files, continued training to step 200000, and this was the final model I use for a large dataset.

Test four continues to use a model from experiment 2 and fine-tune model on a small dataset, the training results up to step 200000 are the last models I use for small dataset.

IV. Result and Discussions

A. Result

The model was tested on private test set which contains 41 files for Big Corpus Voice and 83 files for Small Corpus Voice. Audio generated on private test set are sent to 24 people to rate from 1 to 5 with 0.5 point increments, from which a subjective mean opinion score (MOS) is calculated.

Table 1 shows a comparison of my model against a recorded private test set with the same speaker and quality with the training set.

TABLE I
MOS score of my model and natural voice

	Big Corpus	Small Corpus	Final Score
Natural voice	4.30	4.58	4.44
VNGGRD voice	3.70	3.99	3.85

My model achieves MOS score 3.70 and 3.99 on Big Corpus and Small Corpus and the overall score is 3.85. The results of the two generated voices are not much different (MOS score < 0.6 compared to the natural voice) while the Small Copus test set has many foreign words. But with the help of a quick reading symbol, those generated audio can be natural enough.

B. Discussion

The voice generated from the model is highly natural, however, the lack of processing of noise in the data, the lack of train in the WaveNet vocoder leads to the "mechanical" voice generated.

Using the word segmenter before training the model helps to learn the space character for quick reading "_". Because audio in big dataset often break unnaturally, leading to the model being able to learn this break pattern, the use of the "_" character will help the model never read two-syllable breaks. belongs to the same word. At the same time, this character also helps to link the syllables when transform a foreign word into Vietnamese syllables, helping to read foreign words more seamlessly and naturally.

Fine-tuning model is a good way to generate a new voice from a small amount of data. Big training set which is not required high quality can be easily crawled on the internet and the Small training set which is required high quality does not take long to prepare. In VLSP Campaign 2019, Small Corpus contains 45 minutes of audio, however, my experiment on less than 20 minutes of audio can still generate comparable voice as the 45-minutes voice.

V. Conclusion

This paper describes all my experiments to tackle text to speech shared-task in VLSP Campaign 2019. My model is base on the Tacotron architecture and fine-tune technique to generate voices that have quality close to that of natural human speech from a small amount of data. Experimental results showed that using cleaned data improves the quality of synthesized speech given by the Tacotron-based TTS system

[1] https://github.com/DungLuongTuan/Vietnamese-Word-Segmentation

[2] http://www.hieuthi.com/blog/2017/03/21/all-vietnamese-syllables.html

so that the preprocessing audio phase is truly important. My future work is to decrease the number of time need to synthesis natural voices.

References

[1] A. J. Hunt and A. W. Black, Unit selection in a concatenative speech synthesis system using a large speech database, in Proc. ICASSP, 1996, pp. 373376.

[2] A. W. Black and P. Taylor, Automatically clustering similar units for unit selection in speech synthesis, in Proc. Eurospeech, September 1997, pp. 601604.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, Speech parameter generation algorithms for HMMbased speech synthesis, in Proc. ICASSP, 2000, pp. 13151318.

[4] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, Speech synthesis based on hidden Markov models, Proc. IEEE, vol. 101, no. 5, pp. 12341252, 2013

[5] P. Taylor, Text-to-Speech Synthesis, Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[6] H. Zen, A. Senior, and M. Schuster, Statistical parametric speech synthesis using deep neural networks, in Proc. ICASSP, 2013, pp. 79627966.

[7] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, Tacotron: Towards end-to-end speech synthesis, in Proc. Interspeech, Aug. 2017, pp. 40064010

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, in Proc. ICASSP, 2018, pp. 4779-4783.

[9] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, Deep voice: Real-time neural text-to-speech, CoRR, vol. abs/1702.07825, 2017.

[10] T. T. Vu, M. C. Luong, and S. Nakamura, ”An hmm-based vietnamese speech synthesis system” in Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on.IEEE, 2009, pp. 116-121

[11] Q. S. Trinh, ”Hmm-based vietnamese speech synthesis,” in 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS). IEEE, 2015, pp. 349-353

[12] N. V. Thinh, N. Q. Bao, P. H. Kinh, D. V. Hai, ”DEVELOPMENT OF VIETNAMESE SPEECH SYNTHESIS SYSTEM USING DEEP NEURAL NETWORKS”, in Journal of Computer Science and Cybernetics, V.34, N.4 (2018), 349-363

[13] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in Proc. ICLR, 2015.