Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 2, April 2021, pp. 898~903
ISSN: 2302-9285, DOI: 10.11591/eei.v10i2.2539

The first FOSD-tacotron-2-based text-to-speech application for Vietnamese

Duc Chung Tran
Computing Fundamental Department, FPT University, Hanoi, Vietnam

Article Info

Article history:

Received Sep 11, 2020
Revised Nov 27, 2020
Accepted Dec 11, 2020

Keywords:

Application
Bot
Tacotron-2
Text-to-speech
Vietnamese

ABSTRACT

Recently, with the development and deployment of voicebots which help to minimize personnels at call centers, text-to-speech (TTS) systems supporting English and Chinese have attracted attentions of researchers and corporates worldwide. However, there is very limited published works in TTS developed for Vietnamese. Thus, this paper presents in detail the first Tacotron-2-based TTS application development for Vietnamese that utilizes the publicly available FPT open speech dataset (FOSD) containing approximately 30 hours of labeled audio files together with their transcripts. The dataset was made available by FPT Corporation with an open access license. A new cleaner was developed for supporting Vietnamese language rather than English which was provided by default in Mozilla TTS source code. After 225,000 training steps, the generated speeches have mean opinion score (MOS) well above the average value of 2.50 and center around 3.00 for both clearness and naturalness in a crowd-source survey.

Corresponding Author:

Duc Chung Tran
Computing Fundamental Department
FPT University, Hanoi, 155300, Vietnam
Email: chung.tranduc89@gmail.com

1. INTRODUCTION

Nowadays, the advances of technologies in artificial intelligence and machine learning have enabled wide development of automated tools for answering customers' queries, collecting surveys, addressing complaints without human involvements. These tools are usually chatbots [1-6], or more advanced, voicebots [3, 7-9]. For voicebots, it is essential to have engines called text-to-speech (TTS) for performing conversion of answering text to speech and playback to customer during a call. Usually, there are two steps in a TTS conversion: (i) converting text to melspectrogram; and (ii) synthesize melspectrogram to waveform [10].

The recently introduced end-to-end [11], neural network-based models for generating TTS [12, 13] are Tacotron [14], Tacotron-2 [15-17], Es-Tacotron-2 [15], WaveNet [18-20], WaveGlow [21]. In [14], a TTS based on Tacotron model was introduced to generate speech at frame level which enabled faster speech synthesis compared to approach using sample level approach. The training was based on a single professional female speaker with approximately 25 hours of recorded speeches; thus, input audios' quality can be guaranteed and variants are minimal. The input audios had sampling rate of 24 kHz and the training steps were up to 2 million. In order to reduce the training steps, it shall be possible for one to reduce the sampling rate. For synthesizing waveform, Griffin-Lim model (exisiting since 1984 [22] and catching attention to date [23]) was used [14]. For further improving Tacotron-2 model by addressing over-smoothness problem resulting in unnatural generated speeches, Y. Liu and J. Zheng [15] proposed adding an Es-Network into the existing model.

The idea was to make generated speeches more natural by employing the Es-Network for calculating the estimated melspectrogram residual and making this an additional task of Tacotron2 model. N. Li *et al.* [16] improved Tacotron2 model's speed during training by replacing its attention mechanism by a multi-head one. This was inspired by transformer network used in neural machine translation. However, the drawback of this approach is that it used text-to-phoneme conversion for processing data to learn English language which shall discard the meaning of the orginal end-to-end TTS engine proposed for Tacotron [19], Tacotron2. Although using WaveNet for synthesizing speech may improve speech quality [18, 24-27], its system will need to train two separate networks, one for converting speech to melspectrogram and the other for synthesizing the speech from the melspectrogram [20]. WaveNet variant such as WaveGlow [21], on similar dataset, also required training steps up to 580,000 with audio files sampled at 16 kHz. For synthesizing audio waveform from melspectrogram and for use in very large audio dataset (i.e., 960 hours from 2,484 speakers), multi-head convolutional neural network was proposed [28]. However, its performance for the case of low number of heads, i.e., 2, was just slightly above the average. Even though, [29, 30] also attempted to work on very large audio dataset using the proposed Deep Voice models, the results obtained were not as comparative as Tacotron2.

As seen from the above analysis, the developed engines mainly support English and Chinese, the most popular languages in the world. Meanwhile, Vietnamese is not supported yet. Although, the local TTS tools [31, 32] are supporting well Vietnamese language, there is little information about their back-end engines. In addition, among the developed models, Tacotron and Tacotron-2 are the most utilied end-to-end TTS. Eventhough, it lacks of support for Vietnamese. Therefore, this work presents the first open approach for tailoring a Tacotron-2-based TTS engine utilizing FPT open speech dataset (FOSD) [31, 33, 34]. To the best of author's knowledge, this work is the first that attempts to utilize the freely available to public dataset, FOSD. The main contributions of this work are:

- The newly developed cleaner for supporting Vietnamese speech generation using the TTS' back-end engine provided by Mozilla [35]
- The utilization of the publicly available dataset FOSD [34] for Vietnamese speech generation from text
- The method and analysis of a trained (up to 225,000 steps) TTS model for generating Vietnamse speech [9, 36]

The remaining of this paper is organized as follows; section 2 details the method; section 3 discusses results obtained; section 4 concludes this research.

2. RESEARCH METHOD

In this section, the overall research method is presented in Figure 1. At first, the approach for processing dataset is presented. Second, the core settings for Tacotron-2 engine to be trained and tested are outlined, this eases readers to further investigate the proposed approach. Third, the role of the developed Vietnamese cleaners, as part of the TTS engine is described to help readers better understanding the differences between English and Vietnamese texts. Next, the information of the trained model is presented to give readers how much effort was put to run the training model and at which conditions of the training model used in this work. Finally, the approach for creating input data (Vietnamese texts) is shown to provide various cases of the tests conducted in this work.

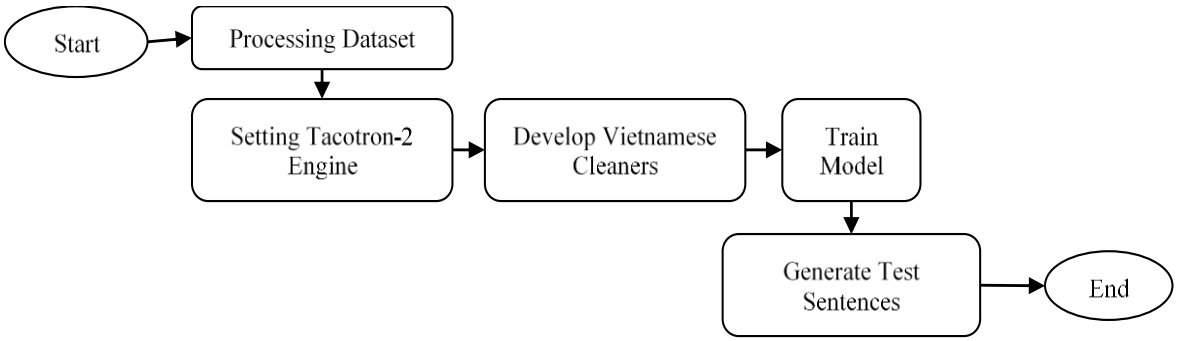

Figure 1. Overall research method

2.1. Dataset processing

The dataset contains over 25,000 audio files (approximately 30 hours of recording) in Vietnamese separated into two main subsets [33, 34]. All audio files are in compressed format (i.e., *.mp3) while their transcripts are stored in *.txt files within the same subfolders. The audio file bitrate is 64 kbps. In order to feed these audio files into the Mozilla-based TTS engine, by using SOX toolbox [37], they were all converted into *.wav format with bitrate of 352 kbps. In addition, all the audio files were placed together in one folder for training the model. The transcript files were also compiled into one file; each line follows the style: **audio_file_name**|**transcript**|**speech_start_time_1**-**speech_end_time_1** speech_start_time_2-end_time_2.

Here, the **audio_file_name** is the file name including the extension; the **transcript** is the text in the speech; speech durations are marked by two ends (i.e., **speech_start_time_1**-**speech_end_time_1**); if there are multiple speeches in one file, each duration is separated by a space character.

The transcript file was then separated into two *.csv files for training and testing the engine. The training file consisted of 23,000 transcript lines while the testing file consisted of 1,900 transcript lines. The detail step-by-step guidelines for this data processing can be found in [38].

2.2. Tacotron-2 architecture settings

In this work, Tacotron-2 architecture based on [19] was utilized since it provides better output quality compared to Tacotron architecture, recommended in Mozilla's notes to developer in [35]. Table 1 presents the typical configuration of the important parameters for training the model. In this table, the number of mel-spectrograms was 80, the number of short-time fourier transform (STFT) frequency levels (equals to size of linear spectrogram frame) was 1,025, same as the default value. The sampling rate was set to 22,050 Hz for faster training the Tacotron-2 architecture. Since the model used in this work was Tacotron-2, softmax function was used for calculating attention norm, suggested by Mozilla. The complete TTS' engine's configuration can be found in [9].

Table 1. TTS' Engine configuration parameters for training

Parameter	Value
num_mels	80
num_freq	1,025
sample_rate	22,050 Hz
Model	Tacotron2
attention_norm	softmax
min_seq_len	6 -> 10
max_seq_len	150 -> 100
use_phonemes	false
text_cleaner	Vietnamese_cleaners
datasets.name	fptopenspeechdata
datasets.path	/content/drive/My Drive/FPTOpenSpeechData/
datasets.meta_file_train	metadata_train.csv
datasets.meta_file_val	metadata_val.csv

In addition, the minimum and maximum sequence lengths were changed from 6 to 10 and 150 to 100 respectively after the first 100,000 training steps. This is to make the model faster to converge and more suitable with the existing dataset which has minimum sequence length of 2, maximum sequence length of 301, average sequence length of 52.43. As a result, 1,145 instances were discarded since they were out of the aforementioned sequence length range. In this work, using phoneme option was disabled since it was out of this research focus. Meanwhile, a new text cleaner namely "V**ietnamese_cleaners**" was newly developed for processing Vietnamese texts. The dataset path, meta file for training and validation were provided as well. It should be noted that, the model was trained completely on Google Colaboratory, a free TensorFlow-supported platform.

2.3. Vietnamese cleaners

The Vietnamese cleaners was developed to support Vietnamese language instead of English as in the original repository. The cleaner allows the special conversions of:

- symbols to words: e.g., "**+**" to "**cộng**" (English: plus)
- special characters to words: e.g., "**%**" to "**phần trăm**" (English: percent)
- special words to similar words with the same pronunciations: e.g., "**hỷ**" to "**hỉ**" (English: happy)
- number to words: e.g., "**11**" to "**mười một**" (English: eleven)

Here, it should be noted that all capitalized words were converted to lowercase to form uniform source texts before feeding to the network for training, validation and testing.

2.4. Training model

In order to prove that the developed Vietnamese cleaners are suitable for the model to generate clear Vietnamese speeches from random texts, the model was trained for 225,000 steps. As a result, the training loss was 0.10406 while the validation loss was 0.12349.

2.5. Random texts for speech generation

In Table 2, the uncorrelated random texts were selected for testing the trained TTS model. The first text was an unusual statement comparing sizes of "one" duck and a cow. In this text, the word "một" (English: "one") was used to test if the trained model could generate a speech containing a number. The second text was a statement describing a female having a name of "sơn", here, the letter "s" was not capitalized. The third text was a statement describing the event that two footballers were invited to Spain for career probation. The fourth text was a statement describing Hanoi streets during spring, near Vietnamese Lunar New Year. The fifth text was a statement describing how Vietnamese footballer stars spend money.

Table 2. Uncorrelated texts for testing model

No.	Input Text (Vietnamese)	Translated Version in English
1	một con vịt to như con bò	one duck is as big as a cow
2	chị sơn xinh gái nhỉ	sister son is beautiful
3	Không chỉ có Tuấn Anh, Văn Toàn cũng được mời sang thử việc tại Tây Ban Nha	Not only Tuan Anh, Van Toan (footballer) also was invited for career probation in Spain
4	Đào xuống phố sớm, nhiều tuyến đường Hà Nội đã rộn ràng sắc xuân	Peach blossom moves down early to street, many streets in Hanoi is already bustling the spring
5	Sao bóng đá Việt đua nhau tặng xế sang bạc tỷ cho người thân	Vietnamese star footballer following each other to buy billion-dong luxury cars for their relatives

3. RESULTS AND DISCUSSION

In this work, the results obtained from the trained Vietnamese TTS model is discussed. At first, the generated speeches are accessed based on its completeness. This indicates whether the model is able to generate complete speeches based on given texts. Second, the speeches are accessed based on its clearness and naturalness subject to MOS scores, the typical index for accessing the quality of generated speeches from TTS engine.

3.1. Completeness of the generated speeches

Out of the five generated speeches, three (the first, the second, and the fifth) were complete. The third speech missed 2/17 words while the fourth one missed 10/14 words (i.e., the second part of the sentence, after the comma). Further analyzing the missing words, Table 3 presents the frequencies of missing words in the training and validation sets which were used for training and validating the developed FOSD model. From the table, it could be seen that, the typical ratio of validation words over training words were from approximately 0.05 to 0.14

It is obvious that too little frequencies in the validation set could cause missing words in the generated speech, i.e., 2 times for the words "sắc" and "xuân". In addition, too many frequencies also could cause the same issue, i.e., from above 1,000 to over 2,000 or 3,000 times for the words "nhiều", "đã", and "Hà" respectively.

Table 3. Frequencies of missing words

No.	Word	Training	Validation	Ratio Validation/Training
1	Văn	167	20	0.1197
2	Toàn	267	23	0.0861
3	nhiều	1,038	81	0.0780
4	tuyến	57	8	0.1404
5	đường	395	31	0.0785
6	Hà	3,056	259	0.0848
7	Nội	166	9	0.0542
8	đã	1,829	125	0.0683
9	rộn	149	15	0.1007
10	ràng	49	7	0.1429
11	sắc	80	2	0.0250
12	xuân	35	2	0.0571

3.2. Clearness and naturalness of the generated speeches

A crowd-source survey was conducted on a set of 100 random participants who are students at FPT University to assess the clearness and naturalness of the generated speeches. Here, the naturalness refers to the state or quality of being natural (human-like) in the generated speeches while the clearness indicates the clarity (low noise) in the generated speeches. Based on the survey, 50% of the students used headphones while the other 50% used computer speakers for the test. In addition, all of the students had never heard about the sentences and speeches before. Their MOS were outlined in the Table 4. From the table, the MOS for clearness was ranging approximately from 2 to 4.5. Four out of five speeches were considered clear while the second speech was the least clear one. The clearest speech was the fifth, its MOS was 3.39 with standard deviation of 0.98 making it the best speech in the test set. Meanwhile, the MOS for the generated speeches' naturalness were typically slightly lower than those of clearness. Still, the fifth speech was the most natural speech in the test set. Here, three out of five speeches were above the average (about 2.50).

Table 4. Clearness and naturalness of generated speeches-MOS

No.	Clearness	Naturalness
1	2.95 ± 1.15	2.54 ± 1.12
2	2.62 ± 1.17	2.52 ± 1.07
3	2.94 ± 1.07	2.84 ± 1.00
4	2.97 ± 1.17	2.81 ± 1.02
5	3.39 ± 0.98	3.06 ± 1.07

4. CONCLUSION

This paper has presented the first approach for generating FOSD Tacotron-2-based TTS engine for Vietnamese. The work opens new insights into the generation of speeches from texts. To be particular, too little or excessively large frequencies of texts in training and validation sets could cause missing of the words in the generated speeches. Overall, all the generated speeches are above the average in terms of clearness and naturalness. Future works will explore more possibility of generating quality speeches from an optimal dataset.

REFERENCES

[1] G. Mao, J. Su, S. Yu and D. Luo, "Multi-Turn Response Selection for Chatbots With Hierarchical Aggregation Network of Multi-Representation," in *IEEE Access*, vol. 7, pp. 111736-111745, 2019

[2] B. Liu et al., "Content-Oriented User Modeling for Personalized Response Ranking in Chatbots," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 122-133, Jan 2018.

[3] H. Cuayáhuitl *et al.*, "Ensemble-based deep reinforcement learning for chatbots," *Neurocomputing*, vol. 366, pp. 118-130, Nov 2019.

[4] M. Chung, E. Ko, H. Joung, and S. J. Kim, "Chatbot e-service and customer satisfaction regarding luxury brands," *Journal of Business Research*, vol. 117, pp. 587-595, Sep 2018.

[5] S. Arsovski, H. Osipyan, M. I. Oladele, and A. D. Cheok, "Automatic knowledge extraction of any Chatbot from conversation," *Expert System with Application*, vol. 137, pp. 343-348, Dec 2019.

[6] Abhishek Das *et al.*, "Visual Dialog," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326-335, 2017.

[7] T. D. Chung, H. H. Son, and A. Khalyasmaa, "A Question Detection Algorithm for Text Analysis," in *Proceedings of the 2020 5th International Conference on Intelligent Information Technology*, pp. 61-65, Feb. 2020.

[8] T. D. Chung, M. Drieberg, M. F. Bin Hassan and A. Khalyasmaa, "End-to-end Conversion Speed Analysis of an FPT.AI-based Text-to-Speech Application," *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*, Kyoto, Japan, pp. 136-139, 2020.

[9] D. C. Tran, "The First Vietnamese FOSD-Tacotron-2-based Text-to-Speech Model Dataset," *Data Brief*, vol. 31, p. 105775, Aug 2020.

[10] Y. Shiga, J. Ni, K. Tachibana, and T. Okamoto, "Text-to-speech synthesis," in *SpringerBriefs in Computer Science*, pp. 39-52, 2020.

[11] Z. Jiang, F. Qin, and L. Zhao, "A Phoneme Sequence Driven Lightweight End-To-End Speech Synthesis Approach," *IOP Conf. Series: Journal of Physics: Conference Series*, vol. 1267, pp. 1-7, 2019.

[12] L. Luo, G. Li, C. Gong, and H. Ding, "End-to-end Speech Synthesis for Tibetan Lhasa Dialect," *OP Conf. Series: Journal of Physics: Conference Series*, vol. 1187, no. 5, pp. 1-7, 2019.

[13] P. Partila, J. Tovarek, G. H. Ilk, J. Rozhon and M. Voznak, "Deep Learning Serves Voice Cloning: How Vulnerable Are Automatic Speaker Verification Systems to Spoofing Trials?," in *IEEE Communications Magazine*, vol. 58, no. 2, pp. 100-105, February 2020.

[14] Y. Wang *et al.*, "Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model," *Arxiv*, pp. 1-10, 2017.

[15] Y. Liu and J. Zheng, "Es-Tacotron2: Multi-task Tacotron 2 with pre-trained estimated network for reducing the over-smoothness problem," *Information*, vol. 10, no. 4, pp. 1-13, 2019.
[16] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural Speech Synthesis with Transformer Network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 6706-6713, 2019.
[17] Y. Zheng, J. Tao, Z. Wen and J. Yi, "Forward–Backward Decoding Sequence for Regularizing End-to-End TTS," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2067-2079, 2019.
[18] J. Chorowski, R. J. Weiss, S. Bengio and A. van den Oord, "Unsupervised Speech Representation Learning Using WaveNet Autoencoders," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2041-2053, Dec. 2019.
[19] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, pp. 4779-4783, 2018.
[20] N. Adiga, V. Tsiaras and Y. Stylianou, "ON the Use of Wavenet as a Statistical Vocoder," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, pp. 5674-5678, 2018.
[21] R. Prenger, R. Valle and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, pp. 3617-3621, 2019.
[22] D. Griffin and Jae Lim, "Signal estimation from modified short-time Fourier transform," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236-243, April 1984.
[23] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, "Deep Griffin–Lim Iteration," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 61-65, 2019.
[24] S. Arik *et al.*, "Deep voice: Real-time neural text-to-speech," *arXiv*, 2017.
[25] A. Van Den Oord *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," *Proceedings of Machine Learning Research*, 2018.
[26] L. J. Liu, Z. H. Ling, Yuan-Jiang, Ming-Zhou, and L. R. Dai, "Wavenet vocoder with limited training data for voice conversion," *Interspeech*, pp. 1983-1987, 2018.
[27] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," *Interspeech*, pp. 1138-1142, 2017.
[28] S. Ö. Arık, H. Jun and G. Diamos, "Fast Spectrogram Inversion Using Multi-Head Convolutional Neural Networks," in *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94-98, Jan 2019.
[29] W. Ping *et al*., "Deep Voice 3: 2000-Speaker Neural Text-to-Speech," *conference paper at ICLR*, pp. 1-16, 2018.
[30] S. O. Arik *et al.*, "Deep voice 2: Multi-speaker neural text-to-speech," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 1-9, 2017.
[31] FPT, "FPT.AI-A comprehensive AI platform and integrated AI services," 2020.
[32] V. Group, "Viettel AI Open Platform," 2020.
[33] FPT Technology Innovation, "30 years–FPT shares 30 hours of recorded voice data," 2018.
[34] D. C. Tran, "FPT Open Speech Dataset (FOSD)-Vietnamese," 2020.
[35] Mozilla, "Deep Learning for Text to Speech." 2020.
[36] D. C. Tran, "The First FOSD-Tacotron-2-based Text-to-Speech Model Dataset," *Data in Brief*, vol. 31, p. 105775, 2020,
[37] SoX - Sound eXchange, "SoX - Sound eXchange." 2015.
[38] D. C. Tran, M. K. A. A. Khan, and S. S., "On The Training and Testing Data Preparation for End-to-end Text-to-speech Application," in *2020 11th IEEE Control & System Graduate Research Colloquium (ICSGRC 2020)*, pp. 1–4, 2020.

BIOGRAPHY OF AUTHOR

Tran Duc Chung completed bachelor and Ph.D. degrees, both in Electrical & Electronic Engineering, at PETRONAS University of Technology, Malaysia in 2014 and 2018 respectively. During the studies, his majors were Instrumentation & Control and WirelessHART networked control system correspondingly. He had worked for several companies namely Intel Technology Sdn. Bhd. Malaysia, R&D Centre - Sony EMCS Sdn. Bhd. Malaysia, Dasan Zhone Solutions Vietnam, FPT Technology Research Institute Vietnam. He has involved in authoring more than 50 indexed publications. Currently, he is with FPT University, Hanoi, Vietnam, researching in emerging technologies and topics including Natural Language Processing and Generation (NLP, NLG), and Applied Artificial Intelligence