The Future of Speaker Adaptation: Advancements in Text-to-Speech Synthesis Solutions

Ali Raheem Mandeel

Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
aliraheem.mandeel@edu.bme.hu

Abstract

Personalizing a text-to-speech (TTS) model is an admiringly advantageous application. The TTS model can create a speech for any target speaker using a limited dataset. However, many challenges are related to using small data, such as it is hard to generalize the model to unseen speakers, the similarity to the target speaker being too low, and low-quality synthesized speech in terms of prosody or naturalness, etc. Moreover, building a lightweight TTS model with a small dataset is necessary for low computational complexity. This research firstly exploits Continuous vocoder-based statistical parametric speech synthesis (SPSS) for speaker adaptation to reduce the computational complexity for real-time applications. Secondly, the lowest data and training time were measured to build an efficient end-to-end TTS model-based speaker adaptation. Finally, creaky voices and interrogative sentences' prosody were examined in an end-to-end TTS model to enhance the naturalness of the target speakers' synthesized speech.

Index Terms: speech synthesis, limited dataset, creaky voices, interrogative sentences

1. Introduction

1.1. Background

TTS models synthesize a human-like speech when trained on a large high-quality speech dataset. Customizing a TTS model for an independent speaker requires collecting sufficient data, which is a costly time and effort. One solution for this issue is speaker adaptation (sometimes called speaker cloning or customizing). The TTS model is trained on a large dataset (usually a multi-speaker or single-speaker dataset) and adapted/fine-tuned to a limited dataset of target speakers [1]. Another issue that TTS models suffer from is consuming or needing high computational resources for training/inference of the speech. This high-cost computational cost is due to using numerous parameters because of the largely used data. Speaker adaptation is also an essential solution for this problem by using a limited dataset.

1.2. Motivations and research goals

One crucial aspect of TTS is the robustness with which the model should be adapted to various speakers with various speech characteristics. Moreover, data availability is the main challenge for speakers with limited samples. Therefore, building a TTS model-based speaker adaptation with minimum data is demanding. Real-time inference speed using a lightweight computational resource model is preferred. In this paper, three directions or solutions for speaker adaptation with small data will be discussed:

- I implemented modifications of the conception of the average TTS model (Continuous vocoder [6]) to synthesize new target speakers and adapt to various domains. I used speaker adaptation to enhance Continuous vocoder (SPSS) to have a lightweight TTS model using limited data.
- I lowered target speakers' amount of adaptation data and parameters for efficient adaptation techniques. Then, I investigated a TTS model's end-to-end performance (Tacotron2) with the limited target speaker data adaptation.
- I enhanced the target speakers' synthesized speech naturalness. Specifically, interrogative sentences' prosody for spontaneous conversational speech was improved using the target speakers' limited datasets. Moreover, I investigated the impact of a creaky voice [7] within a synthesized speech in the state of arts TTS models to enhance the similarity of the synthesized speech to the original recording.

2. Related work

Limited data used to train TTS models has been a desirable topic recently. Numerous vocoders have been suggested over the past years, employing SPSS [8, 9] and neural vocoders [10, 11]. Neural vocoders, which can synthesize incredibly natural speech, frequently need to meet the requirements of real-time synthesis. In applications requiring only a modest computational complexity, conventional vocoder-based SPSS can be sufficient. The Continuous vocoder (SPSS vocoder) was utilized in the first objective of this research, and it was inspired by [6], which proposed a computationally applicable residual-based vocoder. In this vocoder, the excitation engages two one-dimensional parameters: continuous fundamental frequency (F0) and Maximum Voiced Frequency (MVF).
On the other hand, many studies tried to minimize the required datasets. End-to-end TTS model Tacotron2 was proved to synthesize one single sentence from a target speaker using a Romanian dataset [12]. Guided attention was implemented in Tacotron2 with limited Spanish and Basque data [13]. Guided attention has reduced the lost alignment (text/ phonemes) during the inference process.

3. Results

3.1. Speaker adaptation using Continuous vocoder

Continuous vocoder results were compared to the baseline vocoder (WORLD [14]). The open-source Merlin [15] framework was utilized to implement the experiment. An average voice model (AVM) was created (trained on nine speakers on the VCTK corpus [16]) and customized for the four target speakers (two females and two males) of roughly 14 minutes.

Feed-forward neural networks (FFNN) and two versions of the recurrent neural networks (LSTM and GRU) were utilized in this study.
According to objective testing, the Continuous vocoder could synthesize voices using RNN approximately equivalent to the baseline WORLD vocoder during speaker adaptation. In addition, the MUSHRA test [17] outcomes confirmed the effectiveness of the Continous vocoder's speaker adaptation method by receiving slightly lower rates than the WORLD vocoder in GRU and LSTM (see Figure 1). In contrast, the WORLD vocoder received higher scores than the Continous vocoder in FFNN.

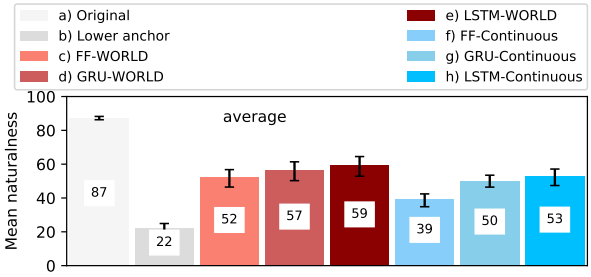

Figure 1: *MUSHRA scores for the naturalness/Continuous vocoder. Errorbars show the bootstrapped 95% confidence intervals.*

3.2. Speaker adaptation efficiency experiments

I investigated the dataset and training period required to construct a TTS model using end-to-end TTS (Tacotron2 [18]) and neural vocoder (WaveGlow [10]) with an unseen target speakers' datasets. A general model was trained on a multi-speaker dataset of 88.3 hours (Hi-Fi multi-speaker dataset [19]) and then adapted to four speakers (two females and two males). Also, two kinds of audio qualities (clean of signal-to-noise ratio (SNR) at least 40 dB and another data of SNR equal to 30 dB) were investigated in the experiment.
According to the findings (see Figure 2), the Tacotron2 model, which is trained on 100 sentences of data over a relatively short time (checkpoint 900), delivers acceptable synthesized speech quality. Moreover, there is no direct relation between the adaptation audio dataset's SNR and the synthesized speech quality in our experiment's results.

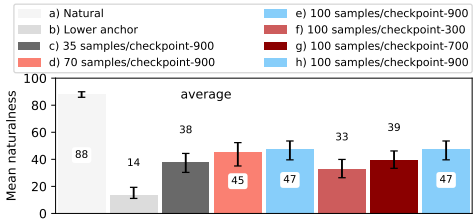

Figure 2: *Average naturalness ratings of the four speakers' speech/efficiency experiment.*

3.3. Speaker adaptation- creaky voice experiments

I investigated the impact of the creaky voice (irregular, glottalization, or vocal fry) in TTS with a limited target speaker's data (100 sentences). I adapted a pretrained FastSpeech 2 [20] (on LJSpeech dataset [21]) model to four target speakers. Three adaptation data scenarios were chosen (such as frequent irregular voice, few irregular voice, and randomly chosen sentences). The adaptation data was selected based on the creakiness percentage, measured automatically [22]. The results showed that the TTS model has successfully modeled the creakiness in the synthesized speech according to the objective evaluation. Moreover, the MUSHRA test showed that the creaky synthesized speech obtained a lower preference than the synthesized speech without creaky voices (see Figure 3).

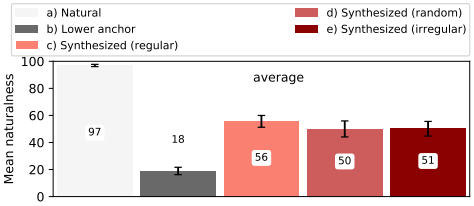

Figure 3: *Average naturalness ratings of the four speakers' speech/creaky voice experiment.*

3.4. Speaker adaptation prosody experiments-interrogative sentences

Using a small dataset (840 and 580 sentences for male and female speakers, respectively), I developed improved English intonation patterns of interrogative sentences for a TTS model (FastSpeech 2). I adapted FastSpeech 2 to interrogative sentences (first dataset) and frequently declarative sentences (second dataset). To find the adaptation data, I looked at how often interrogative sentences appeared. The objective and subjective evaluations revealed that the suggested model successfully created interrogative intonation prosody (see Figure 4). In the subjective evaluation, humming voices (wordless tone) were used to have accurate listeners' feedback. The proposed model (trained on interrogative sentences) obtained 62%, ground truth humming voices received 66%, and natural sentences with words obtained 94%.

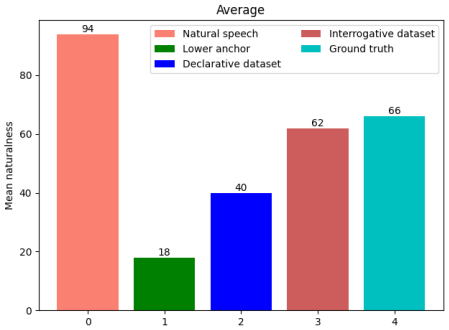

Figure 4: *Average naturalness ratings of the four speakers' speech/prosody experiment.*

4. Future works

I would like to explore Cross-lingual (from English to Arabic), i.e., adapting a pretrained Tacotron 2 model to Arabic target speakers with a minimum amount of data. Secondly, I plan to control the creaky voices percentage in the synthesized speech sentences using a limited dataset. Then, I will observe the impact of the diversity of these percentages on the similarity/naturalness of the out speech.

5. References

[1] A. R. Mandeel, M. S. Al-Radhi, and T. G. Csapó, "Investigations on speaker adaptation using a continuous vocoder within recurrent neural network based text-to-speech synthesis," Multimedia Tools and Applications, vol. 82, no. 10, pp. 15635–15649, Oct. 2022, doi: 10.1007/s11042-022-14005-5.

[2] E. Cooper et al., "Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, doi: 10.1109/icassp40776.2020.9054535.

[3] C.-M. Chien, J.-H. Lin, C. Huang, P. Hsu, and H. Lee, "Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2021, doi: 10.1109/icassp39728.2021.9413880.

[4] S.-F. Huang, C.-J. Lin, D.-R. Liu, Y.-C. Chen, and H. Lee, "Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1558–1571, 2022, doi: 10.1109/taslp.2022.3167258.

[5] J. Lee and J.-H. Chang, "One-Shot Speaker Adaptation Based on Initialization by Generative Adversarial Networks for TTS," Interspeech 2022, Sep. 2022, doi: 10.21437/interspeech.2022-934.

[6] T. G. Csapo, G. Nemeth, M. Cernak, and P. N. Garner, "Modeling unvoiced sounds in statistical parametric speech synthesis with a continuous vocoder," 2016 24th European Signal Processing Conference (EUSIPCO), Aug. 2016, doi: 10.1109/eusipco.2016.7760466.

[7] J. Laver, The Phonetic Description of Voice Quality., Cambridge Univ. Press, 1980.

[8] G. Degottex and Y. Stylianou, "A full-band adaptive harmonic representation of speech," Interspeech 2012, Sep. 2012, doi: 10.21437/interspeech.2012-138.

[9] T. Drugman and T. Dutoit, "The Deterministic Plus Stochastic Model of the Residual Signal and Its Applications," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 968–981, Mar. 2012, doi: 10.1109/tasl.2011.2169787.

[10] 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), May 2019, doi: 10.1109/icassp.2019.8683143.

[11] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks," Interspeech 2020, Oct. 2020, doi: 10.21437/interspeech.2020-2143.

[12] G. Saracu and A. Stan, "An analysis of the data efficiency in Tacotron2 speech synthesis system," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Oct. 2021, doi: 10.1109/sped53181.2021.9587411.

[13] V. García, I. Hernáez, and E. Navas, "Evaluation of Tacotron based Synthesizers for Spanish and Basque," Applied Sciences, vol. 12, no. 3, p. 1686, 2022, doi:10.3390/app12031686.

[14] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," IEICE Transactions on Information and Systems, vol. E99.D, no. 7, pp. 1877–1884, 2016, doi: 10.1587/transinf.2015edp7457.

[15] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), Sep. 2016, doi: 10.21437/ssw.2016-33.

[16] C. Veaux, J. Yamagishi and K. MacDonald, Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.

[17] "ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality," 2001.

[18] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in ICASSP, IEEE, 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[19] Bakhturina, E., Lavrukhin, V., Ginsburg, B., Zhang, Y. (2021) Hi-Fi Multi-Speaker English TTS Dataset. Proc. Interspeech 2021, pp. 2776-2780, doi: 10.21437/Interspeech.2021-1599.

[20] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," In International Conference on Learning Representations, 2020.

[21] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[22] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Comp. Speech & Lang.*, vol. 27, no. 4, pp. 1028-1047, 2013.