PVAE-TTS: ADAPTIVE TEXT-TO-SPEECH VIA PROGRESSIVE STYLE ADAPTATION

Ji-Hyun Lee[1], *Sang-Hoon Lee*[2], *Ji-Hoon Kim*[1], *Seong-Whan Lee*[1,2]

[1]Department of Artificial Intelligence, Korea University, Seoul, Korea
[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

ABSTRACT

Adaptive text-to-speech (TTS) has attracted increasing interests for the purpose of training TTS systems without tons of high quality data. Nevertheless, existing adaptive TTS systems still show low adaptation quality for novel speakers, since it is hard to learn an extensive speaking style with limited data. To address this issue, we propose progressive variational autoencoder (PVAE) which generates data with adapting to style gradually. PVAE learns a progressively style-normalized representation, which is a key component of progressive style adaptation. We extend PVAE to PVAE-TTS, a multi-speaker adaptive TTS model which generates natural speech with high adaptation quality for novel speakers. To further improve the adaptation quality, we also propose dynamic style layer normalization (DSLN) which utilizes a convolution operation. The experimental results demonstrate the superiority of PVAE-TTS in terms of both subjective and objective evaluations.

Index Terms— text-to-speech, speech synthesis, adaptive TTS, speaker adaptation

1. INTRODUCTION

Recent text-to-speech (TTS) [1, 2, 3, 4, 5, 6] has made dramatic improvement in speech quality. However, training the TTS system requires tons of high quality audio with the corresponding text, which is laborious and time-consuming to collect. This leads to an increasing demand for adaptive TTS, which aims to synthesize high-quality speech with a small amount of training data.

One approach for adaptive TTS is a *zero-shot adaptation* in which the TTS model generates speech conditioned on speaker embedding extracted from the audio of novel speakers [7, 8]. For example, Meta-StyleSpeech [9] used speaker embedding from mel encoder, leveraging an episodic meta-learning with additional discriminators. The other approach is a *few-shot adaptation* which fine-tunes a pre-trained TTS model with limited data [10, 11, 12]. AdaSpeech [13] proposed acoustic condition modeling and conditional layer normalization (CLN) to reduce the number of adaptation parameters while maintaining speech quality. Despite recent improvements, the synthesized audio is still not enough compared to human speech in terms of naturalness and similarity.

One major reason for low adaptation quality is the extensive speaking style information. Since speaking style includes various information such as tone and stress, it is hard to learn the speaking style with a few samples. To alleviate this problem, in this paper, we propose progressive variational autoencoder (PVAE) which enables progressive style adaptation. PVAE extracts progressively style-normalized representation; then gradually adapts to the style. By leveraging PVAE, we present PVAE-TTS which generates high-quality audio even from limited data by gradually learning various speaking style. Moreover, we introduce dynamic style layer normalization (DSLN) to enhance the adaptation quality. This makes the model adapt to the style effectively via a convolution operation. We assessed PVAE-TTS on subjective and objective evaluations. The experimental results demonstrate that PVAE-TTS synthesizes high-quality speech, outperforming existing methods in all metrics. The audio samples are available at `https://prml-lab-speech-team.github.io/demo/PVAE-TTS`

2. PROGRESSIVE VARIATIONAL AUTOENCODER

Since style contains various factors, it is a challenge to adapt to the style with a small amount of data. To overcome this difficulty, we propose PVAE which leverages bidirectional-inference variational autoencoder [14, 15]. Specifically, the prior and the approximate posterior are defined as $p_\theta(\mathbf{x},\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}_1)[\prod_{l=1}^{L-1} p_\theta(\mathbf{z}_l|\mathbf{z}_{l+1})]p_\theta(\mathbf{z}_L)$ and $q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x})[\prod_{l=1}^{L-1} q_\phi(\mathbf{z}_{l+1}|\mathbf{z}_l)]$, respectively. Here, $\mathbf{z} = \mathbf{z}_1, \ldots, \mathbf{z}_L$ refers to the hierarchy of latent variables where L denotes the number of hierarchy. PVAE consists of a deterministic bottom-up path and top-down path, sharing parameters each other. Along the bottom-up path, PVAE extracts progressively style-normalized features from the data $\mathbf{x}$ and keeps them inside. Then, along the top-down path, PVAE produces style-adapted features using stored style-normalized features, while maximizing evidence lower bound:

$$\mathcal{L}(\theta,\phi) \equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}_1|\mathbf{x})||p(\mathbf{z}_1)) - \sum_{l=2}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_{<l}|\mathbf{x})}[D_{KL}(q_\phi(\mathbf{z}_l|\mathbf{x},\mathbf{z}_{<l})||p_\theta(\mathbf{z}_l|\mathbf{z}_{<l})] \quad (1)$$

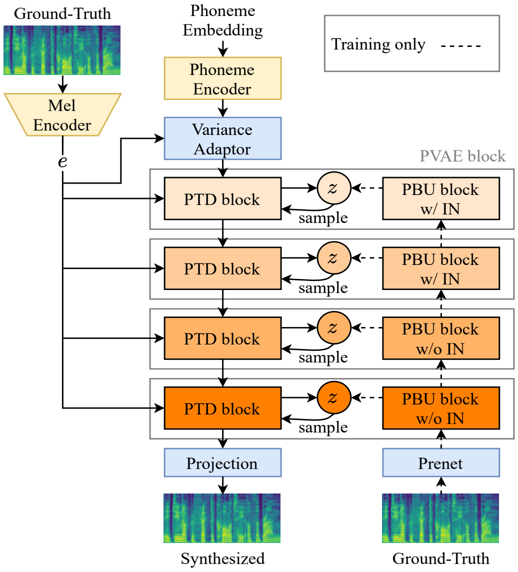


Fig. 1. The architecture of PVAE-TTS. The PBU blocks for bottom-up path are only used in training.

In addition, since PVAE removes speaker information progressively along the bottom-up path, it can be formulated as $P(\hat{s} = s_i|\mathbf{z}_{1,i}) \geq \cdots \geq P(\hat{s} = s_i|\mathbf{z}_{L,i})$, where $\hat{s}$ denotes the predicted style. $\mathbf{z}_l$ is a progressively style-normalized representation from the data $\mathbf{x}$ having a style $s_i \in S$, where S denotes a set of styles.

3. PVAE-TTS

In this section, we present PVAE-TTS extended from PVAE. PVAE-TTS is composed of a phoneme encoder, mel encoder, variance adaptor and 4 stacks of PVAE blocks. In particular, each PVAE block is composed of progressive top-down (PTD) block and progressive bottom-up (PBU) block, which are based on the FFT block [1]. We follow the same architecture of phoneme encoder and variance adaptor in FastSpeech2 [1]. To extend the model to multi-speaker TTS, we extract the style embedding e through mel encoder in Meta-StyleSpeech [9]. We concatenate style embedding e with the input of each variance predictor which predicts duration, pitch, and energy in phoneme-level. The model architecture is shown in Fig. 1.

3.1. Progressive Style Normalization

The goal of the bottom-up path is to extract hierarchical features with removing style information progressively. To remove style information progressively, we apply instance normalization (IN) in the third and fourth PBU blocks, which helps remove speaker information while preserving the linguistic information [16]. We also add information bottleneck layer in the same PBU blocks. As discussed in Sec. 4.5, both method contribute to remove style information gradually.

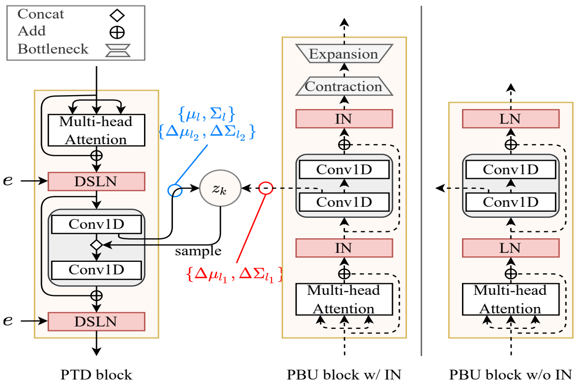


Fig. 2. Details of PVAE block.

Due to this capacity, each PBU block generates progressively style-normalized representations.

3.2. Information Sharing

At each PBU and PTD block, the parameters of the prior and approximate posterior distributions $\{\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\}$, $\{\Delta\boldsymbol{\mu}_{l_1}, \Delta\boldsymbol{\Sigma}_{l_1}\}$, and $\{\Delta\boldsymbol{\mu}_{l_2}, \Delta\boldsymbol{\Sigma}_{l_2}\}$ are acquired from the preceding 1D convolution network of each block, described in Fig. 2. The prior distribution $p_\theta(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{y})$ and the approximate posterior distribution $q_\phi(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x}, \mathbf{y})$ are defined as:

$$p_\theta(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{y}) := \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \tag{2}$$

$$q_\phi(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x}, \mathbf{y}) := \mathcal{N}(\boldsymbol{\mu}_l + \Delta\boldsymbol{\mu}_{l_1} + \Delta\boldsymbol{\mu}_{l_2}, \boldsymbol{\Sigma}_l \cdot \Delta\boldsymbol{\Sigma}_{l_1} \cdot \Delta\boldsymbol{\Sigma}_{l_2}) \tag{3}$$

where $\boldsymbol{\Sigma}$ denotes the covariance matrix, and $\mathbf{x}$ and $\mathbf{y}$ refer to the mel-spectrogram and text, respectively. Since variance cannot be negative, we apply a softplus function to $\boldsymbol{\Sigma}$ to ensure positive variance. The approximate posterior distribution can be interpreted as overall information containing approximate likelihood of the bottom-up path and generative distribution from top-down prior information [17]. During training, the latent variable $\mathbf{z}_l$ is sampled from the approximate posterior $q_\phi(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x}, \mathbf{y})$, while sampled from prior $p_\theta(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{y})$ during inference. This architecture allows the bottom-up path to share style-normalized information with the top-down path, and thus it helps the model adapt to the style progressively.

3.3. Progressive Style Adaptation

Along the top-down path, the text information is converted into mel-spectrogram with progressive style adaptation. Each PTD block generates a progressively style-adapted feature using progressively style-normalized features shared by the corresponding PBU block.

Instead of adding or concatenating style embedding e with encoder output, CLN [13] and SALN [9] use an element-wise product and a matrix addition. However, using such simple operation is hard to reflect style containing various information. To achieve high-quality and dynamic adaptation, we propose dynamic style layer normalization (DSLN) utilizing a convo- lution operation. Conditioned on style embedding e, DSLN takes the hidden vector h as an input. We add a single linear layer to predict the filter weight $\mathbf{W}_e$ and bias $\mathbf{b}_e$ from e. DSLN produces a style-adapted feature by performing 1D group convolution operation.

$$DSLN(h, e) = \mathbf{W}_e \otimes LN(h) + \mathbf{b}_e \tag{4}$$

where $\otimes$ denotes convolution operation on normalized h given weight $\mathbf{W}_e$ and bias $\mathbf{b}_e$ of convolution filter. Note that $\{\mathbf{W}_e, \mathbf{b}_e\}$ is not a learnable parameter but adaptively predicted according to the given style e. We add DSLN in all PTD blocks as shown in Fig. 2. Through DSLN, PVAE-TTS can generate natural speech with high adaptation quality.

3.4. Training Objectives

PVAE-TTS performs variational inference and generates mel-spectrogram along the top-down path using text information.

$$\mathcal{L}_{recon} = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y})] \tag{5}$$

$$\begin{aligned}\mathcal{L}_{KL} &= D_{KL}(q_\phi(\mathbf{z}_1|\mathbf{x},\mathbf{y})||p(\mathbf{z}_1|\mathbf{y})) \\ &+ \sum_{l=2}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_{<l}|\mathbf{x},\mathbf{y})}[D_{KL}(q_\phi(\mathbf{z}_l|\mathbf{x},\mathbf{y},\mathbf{z}_{<l})||p_\theta(\mathbf{z}_l|\mathbf{y},\mathbf{z}_{<l})]\end{aligned} \tag{6}$$

The overall loss function for PVAE-TTS becomes:

$$\mathcal{L}_{Total} = \mathcal{L}_{recon} + \beta\mathcal{L}_{KL} + \lambda\mathcal{L}_{var} \tag{7}$$

where β increased from 0 to 1 during the first 20% of training steps and $\lambda = 1$. $\mathcal{L}_{var}$ denotes the summation of L_2 distance of variance predictor [1]. L_1 distance is used for $\mathcal{L}_{recon}$.

4. EXPERIMENTS

4.1. Experimental Setup and Details

We pre-trained our model on LibriTTS (base speaker) and adapted on the VCTK (novel speaker). We used *train-clean-100* and *train-clean-360* of LibriTTS dataset which contain approximately 244 hours of audio with 1,151 speakers, while VCTK dataset contains 44 hours of audio with 108 speakers. For pre-training, 80% of each base speaker's utterance was used as a training set, and the rest was used as a test set. For few-shot adaptation, we randomly selected 20 samples of each novel speaker for adaptation and constructed a test set which is not used for the adaptation.

The audio was downsampled at 22,050Hz and converted into a mel-spectrogram with a 1,024 FFT size, 1,024 window size and 256 hop size. The mel-spectrogram was converted to a waveform by pre-trained HiFi-GAN [18]. To extract the phoneme duration, we used a Montreal forced alignment [19].

The hidden dimension, number of attention heads, and FFT filter size of the PTD blocks are 256, 2, and 1,024, respectively, while those of PBU blocks are set to 16, 2, and 64. This is because during training, the model naively reconstructs the mel-spectrogram with only using shared information from PBU blocks not using phoneme information when the aforementioned numbers are too large. The dimension of the latent variable is set to 8 for the same reason as above.

Table 1. 5-scale MOS, 4-scale SMOS with 95% CI and SV-EER results of multi-speaker TTS on base speakers.

Model	**MOS(↑)**	**SMOS(↑)**	**EER(↓)**	**WER(↓)**
GT	3.85±0.05	3.57±0.04	11.43	7.17
GT *mel*+Vocoder	3.77±0.05	3.49±0.04	11.88	7.72
FastSpeech2	3.55±0.06	3.16±0.05	18.39	20.46
Meta-StyleSpeech	3.60±0.05	3.22±0.05	17.83	18.99
PVAE-TTS	**3.68±0.05**	**3.32±0.04**	**17.12**	**18.55**

We pre-trained PVAE-TTS for 100K steps and fine-tuned for 3000 steps. Furthermore, we trained our model with 64 batch size on a single NVIDIA RTX A6000 GPU. We fine-tuned only 4 stacks of PVAE block and variance adaptor to avoid overfitting. For optimizer, we used Adamax optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ using the same learning rate schedule in [1], using initial learning rate of 10^{-3} and warm-up steps of 10K.

4.2. Evaluation Metrics

We conducted both subjective and objective evaluations on the audio synthesized by PVAE-TTS. For subjective evaluation, we conducted 5-scale mean opinion scores (MOS) for naturalness and 4-scale similarity MOS (SMOS) for speaker similarity via Amazon MTurk where at least 20 raters were asked to rate the naturalness and similarity. We report mean and confidence intervals (CI) of MOS and SMOS. For objective evaluation, we computed speaker verification equal error rate (SV-EER) by pre-trained speaker verification model [20] and word error rate (WER) by fine-tuned wav2vec 2.0 [21].

4.3. Evaluation on Multi-speaker TTS

Before evaluating the adaptation quality of PVAE-TTS, we evaluated synthesized speech for base speakers. We compared PVAE-TTS with other settings including: 1) GT, ground-truth audio; 2) GT *mel*+Vocoder, audio synthesized by HiFi-GAN using ground-truth mel-spectrogram; 3) FastSpeech2, multi-speaker FastSpeech2, which added the style embedding to the encoder output; 4) Meta-StyleSpeech.

With the results shown in Table 1, we have several observations: 1) LibriTTS has low MOS score and high SV-EER score compared to VCTK in Table 2. This is because VCTK has high audio quality compared to LibriTTS in general. 2) PVAE-TTS outperforms other models in all metrics. The results indicate that PVAE-TTS improves the speech quality for base speakers.

Table 2. 5-scale MOS, 4-scale SMOS with 95% CI and SV-EER results of zero-shot$^\diamond$ and few-shot* adaptation on novel speakers. We use 20 samples per speaker for fine-tuning.

Model	MOS(↑)	SMOS(↑)	EER(↓)	WER(↓)
GT	4.00±0.04	3.59±0.04	6.48	15.02
GT *mel*+Vocoder	3.90±0.05	3.55±0.04	6.85	15.95
FastSpeech2$^\diamond$	3.72±0.05	3.24±0.05	16.37	20.69
Meta-StyleSpeech$^\diamond$	3.74±0.05	3.27±0.05	15.63	23.27
PVAE-TTS$^\diamond$	**3.75±0.05**	**3.31±0.05**	**14.07**	**20.14**
FastSpeech2*	3.75±0.05	3.30±0.05	13.89	21.73
Meta-StyleSpeech*	3.77±0.05	3.33±0.05	12.96	16.83
PVAE-TTS*	**3.83±0.04**	**3.38±0.05**	**10.18**	**16.25**

4.4. Evaluation on Adaptive TTS

We evaluated PVAE-TTS in both zero-shot and few-shot adaptation. For few-shot adaptation, we fine-tuned the variance adaptor and decoder of 3) FastSpeech2 and 4) Meta-StyleSpeech with 20 samples per speaker, similar to ours.

As Table 2 shows, PVAE-TTS outperforms the other models in zero-shot adaptation, since the quality of similairy is the main evaluation metric in adaptive TTS [10]. Furthermore, PVAE-TTS shows superior quality compared to the other models in few-shot adaptation, showing the highest MOS, SMOS, SV-EER and WER. This implies that the quality of PVAE-TTS is better than the others in terms of both naturalness and similarity.

Moreover, we investigated the quality of synthesized speech according to different numbers of adaptation samples (1, 5, 20, 50) used for fine-tuning. Fig. 3(a) shows the results. As the number of adaptation samples used for fine-tuning increases, the adaptation quality is improved, showing higher MOS and lower SV-EER score.

4.5. Progressive Style Normalization

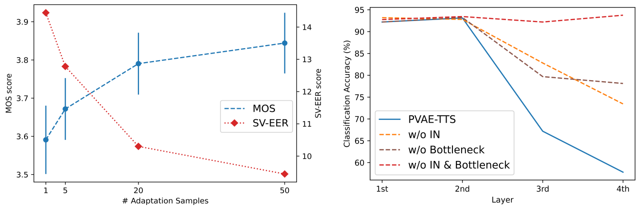
To verify that PBU blocks remove style information progressively, we trained another classifier to classify speaker identity on each PBU blocks. As Fig. 3(b) shows, the first and second layers show high classification accuracy since IN and information bottleneck were not used. When neither was used, the speaker information remains in all PBU blocks. We found that using IN only or information bottleneck only removed speaker information slightly because of the residual connections in the PBU blocks. Notably, the speaker classification accuracy decreased progressively when using both IN and information bottleneck; this implies that both help the model remove speaker information progressively.

4.6. Ablation Study

We conducted ablation studies of progressive style adaptaction and DSLN. Specifically, we compared DSLN with SALN [9], and progressive style adaptation with non-progressive adaptation. As shown in Table 3, DSLN helps the model adapt to the style more effectively, as all metrics decreased when using SALN instead of DSLN. Moreover, when adapting to the style non-progressively, naturalness and similarity decreased significantly compared to progressive style adaptation. This demonstrates that both DSLN and progressive style adaptation contribute to the naturalness and adaptation performance.

Table 3. 5-scale MOS, 4-scale SMOS with 95% CI and SV-EER results of ablation studies on few-shot adaptation.

Method	MOS(↑)	SMOS(↑)	EER(↓)	WER(↓)
PVAE-TTS	**3.83±0.04**	**3.38±0.05**	**10.18**	**16.25**
w/o DSLN	3.76±0.05	3.36±0.05	12.47	17.90
w/o Progressive	3.71±0.05	3.25±0.05	13.52	18.84



(a) MOS and SV-EER results.

(b) Classification results.

Fig. 3. (a) MOS with 95% confidence intervals and SV-EER results according to the different numbers of adaptation samples. (b) Speaker classification accuracy on each PBU block.

5. CONCLUSION

We presented PVAE-TTS which can generate intelligible speech with high adaptation quality by leveraging PVAE. Since it is hard to learn the extensive speaking style information, PVAE learns progressively style-normalized information and adapts to the style progressively. We also proposed DSLN to effectively generate style-adapted feature. We assessed PVAE-TTS on various metrics. From the results, we demonstrated that both DSLN and progressive style adaptation contribute to quality of synthesized audio. Additionally, PVAE-TTS outperforms existing methods for base speakers as well as novel speakers in terms of naturalness and speaker similarity. For future work, we will extend our PVAE to image generation and PVAE-TTS to expressive TTS.

6. ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079 , Artificial Intelligence Graduate School Program (Korea University)) and Netmarble AI Center

7. REFERENCES

[1] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations (ICLR)*, 2021.

[2] Sang-Hoon Lee, Hyun-Wook Yoon, Hyeong-Rae Noh, Ji-Hoon Kim, and Seong-Whan Lee, "Multi-SpectroGAN: High-Diversity and High-Fidelity Spectrogram Generation with Adversarial Style Combination for Speech Synthesis," in *AAAI Conference on Artificial Intelligence*, 2021.

[3] Hyun-Wook Yoon, Sang-Hoon Lee, Hyeong-Rae Noh, and Seong-Whan Lee, "Audio Dequantization for High Fidelity Audio Generation in Flow-Based Neural Vocoder," in *Interspeech*, 2020, pp. 3545–3549.

[4] Yoonhyung Lee, Joongbo Shin, and Kyomin Jung, "Bidirectional Variational Inference for Non-autoregressive Text-to-Speech," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee, "Fre-GAN: Adversarial Frequency-consistent Audio Synthesis," in *Interspeech*, 2021, pp. 2197–2201.

[6] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A Survey on Neural Speech Synthesis," *arXiv:2106.15561*, 2021.

[7] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al., "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[8] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi, "Zero-shot Multi-speaker Text-to-Speech with State-of-the-art Neural Speaker Embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6184–6188.

[9] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang, "Meta-StyleSpeech: Multi-speaker Adaptive Text-to-Speech Generation," in *International Conference on Machine Learning (ICML)*, 2021.

[10] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al., "Sample Efficient Adaptive Text-to-Speech," in *International Conference on Learning Representations (ICLR)*, 2018.

[11] Henry B Moss, Vatsal Aggarwal, Nishant Prateek, Javier González, and Roberto Barra-Chicote, "Boffin TTS: Few-shot Speaker Adaptation by Bayesian Optimization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7639–7643.

[12] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, Hong-Gyu Jung, and Seong-Whan Lee, "GC-TTS: Few-shot Speaker Adaptation with Geometric Constraints," in *International Conference on Systems, Man, and Cybernetics (SMC)*, 2021.

[13] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu, "AdaSpeech: Adaptive Text to Speech for Custom Voice," in *International Conference on Learning Representations (ICLR)*, 2021.

[14] Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther, "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[15] Arash Vahdat and Jan Kautz, "NVAE: A Deep Hierarchical Variational Autoencoder," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee, "One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Interspeech*, 2019, pp. 664–668.

[17] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther, "Ladder Variational Autoencoders," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 3738–3746.

[18] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[19] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment using Kaldi.," in *Interspeech*, 2017, pp. 498–502.

[20] Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung, "Clova Baseline System for the Voxceleb Speaker Recognition Challenge 2020," *arXiv:2009.14153*, 2020.

[21] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.