Interspeech 2022
18-22 September 2022, Incheon, Korea

A Unified System for Voice Cloning and Voice Conversion through Diffusion Probabilistic Modeling

Tasnima Sadekova, Vladimir Gogoryan, Ivan Vovk, Vadim Popov, Mikhail Kudinov, Jiansheng Wei

Huawei Noah's Ark Lab

sadekova.tasnima@huawei.com, vladimir.gogoryan@huawei.com

Abstract

Text-to-speech and voice conversion are two common speech generation tasks typically solved using different models. In this paper, we present a novel approach to voice cloning and any-to-any voice conversion relying on a single diffusion probabilistic model with two encoders each operating on its input domain and a shared decoder. Extensive human evaluation shows that the proposed model can copy a target speaker's voice by means of speaker adaptation better than other known multimodal systems of such kind and the quality of the speech synthesized by our system in both voice cloning and voice conversion modes is comparable with that of recently proposed algorithms for the corresponding single tasks. Besides, it takes as few as 3 minutes of GPU time to adapt our model to a new speaker with only 15 seconds of untranscribed audio which makes it attractive for practical applications.

Index Terms: speech synthesis, voice cloning, voice conversion, diffusion probabilistic modeling

1. Introduction

Voice cloning is the task usually formulated as adding a new voice to a multi-speaker text-to-speech (TTS) system [1, 2]. Although recent advances in deep learning enable almost natural speech synthesis [3, 4, 5, 6], neural TTS models normally require a large amount of transcribed data for training whereas common voice cloning applications put serious constraints on the amount of data available for a target voice. One of the most popular ways to obtain well-performing TTS for the target voice under these constraints is to fine-tune a multi-speaker TTS system on a small amount of the target speaker data. Such approach is usually referred to as *speaker adaptation*, and many modern voice cloning systems employing it achieve rather good quality both in terms of speech naturalness and speaker similarity even when adapted using a few recordings with a total duration of several minutes [7, 8].

While voice cloning is essentially a text-to-speech technology allowing to copy the voice of the target speaker, another approach based on voice conversion can serve this purpose: in this approach linguistic information about the source utterance is extracted from speech rather than from text. In practice, it is preferable to have an *any-to-any* voice conversion model, i.e., the one capable of copying a target voice while preserving source speech content when both source and target speakers do not necessarily belong to the training dataset. Lately, several successful models of this kind have been described [9, 10, 11, 12].

Despite the fact that TTS and voice conversion have much in common, not so many multimodal systems designed to solve both voice cloning and voice conversion tasks have been known so far [13, 14]. In this paper, we present a novel hybrid system consisting of three separately trained modules: the text encoder, the mel encoder and the shared decoder. The whole system is essentially a diffusion probabilistic model (DPM) [15] trying to convert speaker-independent acoustic features extracted either from text by means of the text encoder or from source spoken utterance by means of the mel encoder to the target mel-spectrogram by employing speaker-dependent score matching network which we call the decoder. DPMs have shown good performance in various speech-related tasks [6, 16] including voice conversion [17], so we extend the capabilities of the model described in the latter paper to voice cloning and test its performance on both voice cloning and voice conversion tasks. Moreover, we show that due to the hybrid nature of our model speaker adaptation can now be performed on untranscribed data. Enabling speech synthesis under such a constraint on adaptation data can be helpful in practical applications and attracts interest from researchers [14, 18].

The rest of this paper is organized as follows: Section 2 describes the model we propose and the way we train it; the results of the performance evaluation of our model are given in Section 3; Section 4 concludes with our findings.

2. Multimodal System Description

We follow the same diffusion modeling framework as in [17]. The forward diffusion transforms any mel-spectrogram X_0 into a normal random variable $X_1 \sim \mathcal{N}(\bar{X}, \mathrm{I})$ where I is identity matrix and $\bar{X} = \varphi(X_0)$ is the "average voice" mel-spectrogram predicted by the mel encoder φ so that the prior $\mathcal{N}(\varphi(X_0), \mathrm{I})$ is speaker-independent and preserves the linguistic content of the encoded speech X_0. The reverse diffusion parameterized by the speaker-conditional decoder is trained to approximate the forward diffusion trajectories backwards in continuous time variable $t \in [0, 1]$. As a result, a well-trained decoder enables generative modeling by sampling $\hat{X}_1$ from the prior $\mathcal{N}(\bar{X}, \mathrm{I})$ and simulating paths of the reverse diffusion parameterized with this decoder on the unit time interval $[0, 1]$ with any appropriate numerical scheme. The resulting sample $\hat{X}_0$ at initial time point is the output of the voice conversion model proposed in [17].

The multimodal system we propose in this paper is an extension of the voice conversion system described in [17]: we keep the latter system unchanged and add the text encoder ψ trained to convert text input T into the same "average voice" mel-spectrograms $\bar{X}$ which serve as the training targets for the mel encoder φ. The whole hybrid system consisting of three separately trained modules is illustrated in Figure 1.

2.1. Mel encoder

The mel encoder φ is trained to minimize mean square error between output mel-spectrograms $\bar{X}_\varphi = \varphi(X_0)$ and "average voice" mel-spectrograms $\bar{X}_{GT}$. Ground truth mel features $\bar{X}_{GT}$ are obtained by running Montreal Forced Aligner [19] on LibriTTS and replacing features corresponding to each phoneme

10.21437/Interspeech.2022-10879

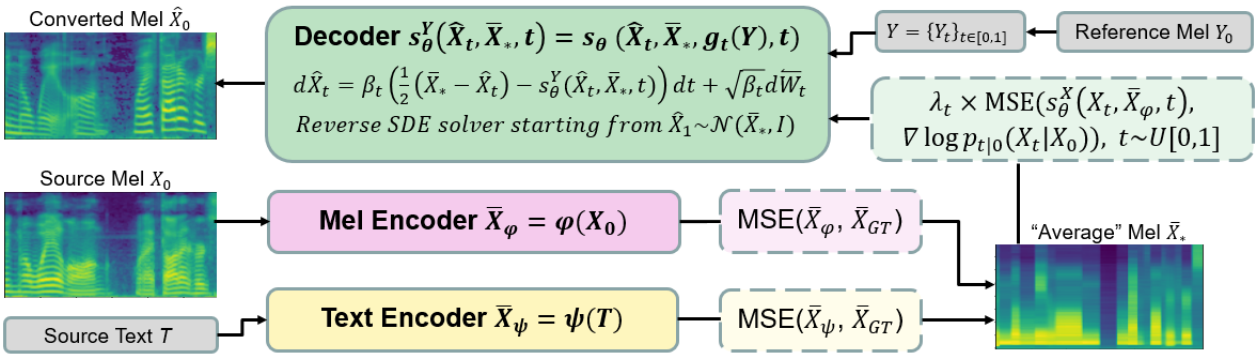

Figure 1: *The proposed multimodal system. The mean $\bar{X}_*$ of the prior in this DPM is either $\bar{X}_\varphi = \varphi(X_0)$ in voice conversion mode when the source mel-spectrogram X_0 is the input or $\bar{X}_\psi = \psi(T)$ in voice cloning mode when the source text T is the input. The decoder conditioned on the trajectory Y of the reference mel-spectrogram Y_0 under the forward diffusion is trained and fine-tuned for the prior whose mean is $\bar{X}_\varphi$.*

in the input mel-spectrogram X_0 with the ones corresponding to this particular phoneme averaged across the whole LibriTTS dataset. Since a phoneme can last for a varying number of acoustic frames, every phoneme is represented by mean mel features corresponding to a single frame, and all frames aligned with a particular phoneme in X_0 are replaced with the same single-frame mean mel features in $\bar{X}_{GT}$.

The mel encoder is composed of a pre-net, 6 Transformer blocks with multi-head self-attention (the number of heads is 2) followed by the final linear projection layer. The pre-net is composed of 3 layers of $1D$ convolutions with kernel size 5 and 192 channels followed by a fully-connected layer.

2.2. Decoder

Once the mel encoder φ parameterizing the DPM prior $\mathcal{N}(\bar{X}_\varphi, \mathrm{I})$ is trained, we fix its parameters and train the decoder corresponding to the reverse diffusion. As in [15], we formalize our DPM employing Stochastic Differential Equations [20] rather than discrete-time Markov chains. The forward X and reverse $\hat{X}$ diffusion processes are given by the following SDEs:

$$dX_t = \frac{1}{2}\beta_t(\bar{X}_\varphi - X_t)dt + \sqrt{\beta_t}d\overrightarrow{W}_t\,, \tag{1}$$

$$d\hat{X}_t = \left(\frac{1}{2}(\bar{X}_\varphi - \hat{X}_t) - s_\theta^Y(\hat{X}_t, \bar{X}_\varphi, t)\right)\beta_t dt + \sqrt{\beta_t}d\overleftarrow{W}_t\,, \tag{2}$$

where $t \in [0,1]$, $\overrightarrow{W}$ and $\overleftarrow{W}$ are forward and backward standard Brownian Motions independent of each other and β_t is a non-negative noise schedule. The reverse SDE (2) is parameterized with the score matching network s_θ and describes the process of transforming random variables $\hat{X}_1$ from the prior distribution $\mathcal{N}(\bar{X}_\varphi, \mathrm{I})$ centered at "average voice" mel-spectrograms $\bar{X}_\varphi$ into target mel-spectrograms $\hat{X}_0$. To capture target speaker identity we integrate the speaker encoding network $g_t(\cdot)$ into the score matching network s_θ and train these two networks jointly:

$$s_\theta^Y(\hat{X}_t, \bar{X}_\varphi, t) = s_\theta(\hat{X}_t, \bar{X}_\varphi, g_t(Y), t), \tag{3}$$

where the decoder parameters are denoted by θ and $Y = \{Y_s\}_{s\in[0,1]}$ is the whole trajectory of the reference mel-spectrogram computed for the target speaker (or just the training mel-spectrogram at training) under the forward diffusion (1).

The decoder is trained to maximize weighted variational lower bound on data log-likelihood which, as shown in [15, 6], can be formalized in terms of the following score matching loss:

$$\mathcal{L} = \int_0^1 \mathbb{E}_{X_0,\xi}\left[\|\sqrt{\lambda_t}s_\theta^X(X_t, \bar{X}_\varphi, t) + \xi\|_2^2\right]dt, \tag{4}$$

where $\lambda_t = 1 - e^{-\int_0^t \beta_s ds}$ and ξ is a sample from $\mathcal{N}(0, \mathrm{I})$ such that X_t is obtained by the formula

$$X_t = X_0 e^{-\frac{1}{2}\int_0^t \beta_s ds} + \bar{X}_\varphi(1 - e^{-\frac{1}{2}\int_0^t \beta_s ds}) + \sqrt{\lambda_t}\xi. \tag{5}$$

The decoder consists of UNet-based score matching network s_θ processing mel-spectrograms as $2D$ images at 3 different resolutions by 3×3 convolutions with channel numbers 256, 512 and 1024 and the speaker encoder $g_t(Y)$ processing target speaker embedding described in [21] broadcast-concatenated with noisy mel-spectrogram Y_t by a stack of six 3×3 convolutions with channel numbers increasing from 128 to 512 followed by average pooling. At generation stage, we use the maximum likelihood reverse SDE (2) solver proposed in [17] with 30 iterations.

2.3. Text encoder

We employ the text encoder designed as a modification of an autoregressive TTS model called Non-Attentive Tacotron [22] to enable voice cloning mode. Non-Attentive Tacotron (NAT) is a variant of a popular Tacotron2 [3] acoustic feature generator in which the attention module is replaced with the explicit duration predictor. In our model, we dropped the decoder part from the original NAT described in [22]. The text encoder is trained as a usual TTS model with the exception that ground truth acoustic features $\bar{X}_{GT}$ are artificial "average voice" mel-spectrograms obtained according to the procedure described in Section 2.1. Due to the nature of our target acoustic features we applied simple upsampling procedure through repetition in accordance with the predicted durations rather than Gaussian upsampling as in [22].

The encoder in NAT modification we utilize consists of a stack of three convolutional layers with kernel size 5 and 512 channels followed by a bidirectional LSTM with 256 units and trained with Mean Square Error (MSE) loss. Its outputs

are passed to the duration predictor – a two-layer bidirectional LSTM with 256 units trained in a supervised manner with MSE loss. Ground truth phoneme durations are obtained with Montreal Forced Aligner.

2.4. Operating modes

The proposed model can perform both voice cloning and voice conversion: the mel encoder followed by the decoder is used to perform voice conversion whereas the text encoder combined with the decoder corresponds to voice cloning task. Speaker adaptation consists in fine-tuning the decoder of our multimodal system on target speaker's data while both encoders remain speaker-independent. Thus, speaker adaptation is possible on untranscribed data because the decoder training requires only target mel-spectrograms X_0 and "average voice" mel-spectrograms $\bar{X}$ approximated by the mel encoder output $\bar{X}_\varphi = \varphi(X_0)$. Note that although in principle it is possible to replace $\bar{X}_\varphi$ with either $\bar{X}_{GT}$ or $\bar{X}_\psi = \psi(T)$, such methods require either alignment between speech frames and phonemes or text transcription thus putting additional constraints to speaker adaptation scenario.

2.5. Related models

Apart from two encoders and the decoder producing acoustic features, the first successful neural model capable of both voice cloning and voice conversion called NAUTILUS [13] has an additional decoder converting encoder outputs to text. This decoder is not used at generation stage and serves only as a regularizer during training. In contrast with our model, all four modules in NAUTILUS are trained jointly with a mixed loss function consisting of four terms which can potentially lead to unstable training.

The hybrid system which resembles ours to the most extent is described in [14]. The main conceptual difference is that this system utilizes the same approach as NAUTILUS to make both encoders output the same distribution – KL divergence between two encoder output distributions is added to the loss function. Our model uses more reliable approach training both encoders separately with the same targets in a supervised manner.

3. Experiments

To train a base multi-speaker model, we used LibriTTS dataset containing 1150 speakers. To train the mel encoder, the decoder and the text encoder, we used Adam optimizer with initial learning rates 0.0005, 0.0001 and 0.0005 and batch sizes 128, 32 and 128 correspondingly. These three modules were trained for 300, 110 and 300 epochs, respectively. For fair comparison of our model with multimodal systems [14] and [13] mentioned in Section 2.5 we fine-tuned the decoder on 10 and 60 (approximately 5 minutes) untranscribed adaptation utterances as in the corresponding papers for 150 iterations of Adam with an initial learning rate of 0.0001. For comparison with single-task models (i.e., the ones capable of either voice cloning or voice conversion but not both) we fine-tuned our model on 1 minute of untranscribed adaptation data (equivalent to $12 - 13$ utterances).

To test our model against multimodal systems NAUTILUS [13] and the best performing model $\mathbf{BaB}^{\mathbf{all}}$ from [14], we chose the source/target pairs from the corresponding demo pages since these models are not open-sourced. Thus, to compare our model with NAUTILUS we synthesized 8 utterances in voice conversion mode and 4 utterances in voice cloning mode. Since the authors of [14] test their model only on voice cloning task, for comparison with their model we just synthesized 4 utterances in voice cloning mode. We performed A/B testing in both speaker similarity and speech naturalness on Amazon Mechanical Turk. Each synthesized utterance was evaluated by 30 assessors for voice cloning and 15 assessors for voice conversion. To ensure the high quality of the results, only Master workers were allowed to do all subjective evaluation tasks described in this section.

Table 1: *Our model (A) compared with [13] (B) for voice cloning/voice conversion on untranscribed data*

Which sounds more similar?	A	Same	B
Voice cloning	**76.5**%	5.0%	18.5%
Voice conversion	**70.0**%	8.3%	21.7%
Which sounds more natural?	A	Same	B
Voice cloning	**80.7**%	6.7%	12.6%
Voice conversion	**79.2**%	7.5%	13.3%

Table 2: *Our model (A) compared with [14] (B) for voice cloning on untranscribed data*

Which sounds more similar?	A	Same	B
Voice cloning	**85.6**%	4.2%	10.2%
Which sounds more natural?	A	Same	B
Voice cloning	**82.2**%	2.5%	15.3%

As for comparison with single-task models, we chose 25 target speakers from VCTK. Thus, like in A/B tests, our model performed both voice cloning and voice conversion for unseen speakers only. For voice cloning tests we synthesized 5 sentences for each of 25 target speakers unless stated otherwise. For voice conversion tests source speakers were chosen from the whole VCTK dataset at random. Every synthesized utterance was evaluated by 5 assessors. The overall number of unique workers who participated in these tests was 53. We assessed both speaker similarity and speech naturalness on a 5-point scale with step size 0.5, the lowest and the highest scores being 1 and 5 correspondingly. In addition to Mean Opinion Scores (MOS), in Tables 3, 4 and 5 we report the overall GPU fine-tuning time (in minutes).

The pre-trained universal HiFi-GAN [23] vocoder was used to convert mel-spectrograms (we utilized conventional 80 mel features as in [17]) into raw waveforms. A subset of speech samples used in the subjective human evaluation is available at our demo page `https://diff-vc-vcl.github.io`.

3.1. Comparison with multimodal systems

The results of A/B tests are given in Tables 1 and 2. These results allow us to conclude that the model we propose is significantly better ($p < 10^{-5}$ in sign test) than two other multimodal systems when adapted on a small amount of untranscribed data both in terms of speech naturalness and speaker similarity. We believe this is due to more reliable speaker-independent acoustic features ("average voice" mel features instead of some latent representation) and the fact that diffusion models are trained on noisy data X_t which can act as data augmentation and reduce

Table 3: *Voice conversion comparison on VCTK.*

Model	Similarity	Naturalness	Time
FS-PPG-VC	3.67 ± 0.10	3.38 ± 0.09	$10'$
BNE-PPG-VC	$\mathbf{4.29 \pm 0.07}$	$\mathbf{4.24 \pm 0.07}$	$6'$
Ours	4.24 ± 0.07	4.20 ± 0.07	$\mathbf{3'}$
Ground Truth	4.58 ± 0.05	4.68 ± 0.05	—

Table 4: *Voice cloning comparison on VCTK.*

Model	Similarity	Naturalness	Time
Tacotron-SMA	$\mathbf{4.31 \pm 0.08}$	$\mathbf{4.23 \pm 0.06}$	$20'$
FastSpeech	4.02 ± 0.10	3.98 ± 0.07	$6'$
StyleSpeech	3.96 ± 0.09	3.77 ± 0.07	$10'$
Grad-TTS	4.19 ± 0.09	3.94 ± 0.07	$\mathbf{2'}$
Ours	4.15 ± 0.08	4.18 ± 0.06	$3'$

Table 5: *Voice cloning comparison on the internal dataset.*

Model	Similarity	Naturalness	Time
Tacotron-SMA	3.75 ± 0.12	3.31 ± 0.13	$20'$
FastSpeech	3.26 ± 0.13	3.35 ± 0.12	$6'$
StyleSpeech	2.89 ± 0.13	2.75 ± 0.11	$10'$
Grad-TTS	3.66 ± 0.14	3.21 ± 0.13	$\mathbf{2'}$
Ours	$\mathbf{3.94 \pm 0.12}$	$\mathbf{4.01 \pm 0.10}$	$3'$

Table 6: *MOS depending on the amount of adaptation data.*

Adapted on	Similarity	Naturalness	Steps	Time
5 seconds	4.30 ± 0.09	3.91 ± 0.11	150	$1'10''$
15 seconds	4.40 ± 0.09	4.02 ± 0.09	300	$3'$
30 seconds	4.40 ± 0.09	4.02 ± 0.09	300	$3'30''$
60 seconds	4.43 ± 0.09	4.05 ± 0.09	150	$3'$

the amount of data necessary to train or fine-tune such models.

3.2. Comparison with single-task systems

It is shown in [24] that voice conversion models based on Phonetic Posteriorgrams (PPGs) provide high synthesis quality and compete even with the models taking text transcription corresponding to the source utterance as input. So, we chose two PPG-based baselines: *BNE-PPG-VC* [12] that showed the best voice conversion quality in [17] among non-diffusion models and *FS-PPG-VC* similar to the model introduced in [25], but with Tacotron2 [3] and HiFi-GAN instead of FastSpeech [4] and LPCNet [26] used in the original paper. As far as voice cloning is concerned, we considered four baselines: *Tacotron-SMA* – multi-speaker Tacotron2 with step-wise monotonic attention described in detail in [27]; *FastSpeech* – multi-speaker FastSpeech2; *StyleSpeech* – multi-speaker StyleSpeech [28] without meta-learning improvement; and *Grad-TTS* - multi-speaker Grad-TTS [6].

For mel-spectrogram inversion, all models used a universal HiFi-GAN vocoder. Every model considered in this section was trained on LibriTTS. Also, every baseline model used the same speaker encoding approach as the one described in [21] with the same pre-trained speaker verification network used there. All the baselines were fine-tuned on 1 minute of transcribed adaptation data and only their decoders were adapted except for *Tacotron-SMA* in which we also fine-tuned the attention module.

The performance of voice conversion and voice cloning models is given in Tables 3 and 4 respectively. *BNE-PPG-VC* and the hybrid model we propose show almost the same performance in voice conversion both in terms of speaker similarity and speech naturalness and their adaptation to unseen VCTK speakers is unproblematic and fast on GPU. As for voice cloning, although our hybrid model is competitive with *FastSpeech*, *StyleSpeech* and *Grad-TTS*, its performance on VCTK speakers is inferior to that of the best performing *Tacotron-SMA* which achieves good synthesis quality when fine-tuned on a target speaker's voice particularly because attention fine-tuning allows to capture the target speaker's manner of speech. However, this feature may become a drawback of *Tacotron-SMA* and result in attention failure when this model is adapted to an unseen voice from unseen domain. As for the model we propose, the text encoder responsible for prosody is fixed during fine-tuning, so the model is more robust to domain shift, thus being potentially more stable in realistic conditions. This statement is supported by Table 5 which reports the results of speaker adaptation experiments on the internal dataset of 10 speakers recorded in more realistic acoustic conditions worse than those of VCTK dataset.

3.3. Adaptation data requirements

We also studied the influence of the amount of adaptation data on the quality of the speech synthesized by our system to understand the minimum adaptation data requirements sufficient to reach the best possible performance. Table 6 demonstrates the results of the analysis carried out on 10 VCTK and 5 internal speakers for voice cloning task. For each speaker, 5 sentences were synthesized and each of them was evaluated by 5 assessors. The model we propose proved to be extremely data-efficient achieving its nearly best performance when adapted on as few as 15 seconds of target speaker's voice which is equivalent to 3 short sentences. Also, it takes only 3 minutes to adapt our model on GPU with 300 SGD steps.

4. Conclusion

In this paper, we have described the novel multimodal system capable of voice cloning and voice conversion. It is designed as a diffusion probabilistic model whose prior is parameterized either with the encoder operating on mel input domain for voice conversion task or the one operating on text input domain for voice cloning. During speaker adaptation, our system makes use of the mel encoder to avoid the need for transcription of the adaptation audio. For both voice cloning and voice conversion, the proposed system shows performance superior to that of other known multimodal systems and comparable to that of the systems designed for the corresponding single tasks. Moreover, it takes only 3 minutes to fine-tune our model on GPU which, combined with relatively low requirements for adaptation data, makes the proposed system promising from the perspective of practical applications.

5. References

[1] Q. Xie, X. Tian, G. Liu *et al.*, "The Multi-Speaker Multi-Style Voice Cloning Challenge 2021," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8613–8617.

[2] D. Tan, H. Huang, G. Zhang *et al.*, "CUHK-EE voice cloning system for ICASSP 2021 m2voc challenge," *CoRR*, vol. abs/2103.04699, 2021. [Online]. Available: https://arxiv.org/abs/2103.04699

[3] J. Shen, R. Pang, R. Weiss *et al.*, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4779–4783.

[4] Y. Ren, Y. Ruan, X. Tan *et al.*, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 3171–3180.

[5] Y. Ren, C. Hu, X. Tan *et al.*, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations*, 2021.

[6] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139. PMLR, 2021, pp. 8599–8608.

[7] S. Arik, J. Chen, K. Peng *et al.*, "Neural Voice Cloning with a Few Samples," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 10 019–10 029.

[8] M. Chen, X. Tan, B. Li *et al.*, "AdaSpeech: Adaptive Text to Speech for Custom Voice," in *International Conference on Learning Representations*, 2021.

[9] K. Qian, Y. Zhang, S. Chang *et al.*, "AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. PMLR, 09–15 Jun 2019, pp. 5210–5219.

[10] J. Chou and H. Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*. ISCA, 2019, pp. 664–668.

[11] D. Wang, L. Deng, Y. Yeung *et al.*, "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 1344–1348.

[12] S. Liu, Y. Cao, D. Wang *et al.*, "Any-to-Many Voice Conversion With Location-Relative Sequence-to-Sequence Modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[13] H. Luong and J. Yamagishi, "NAUTILUS: A Versatile Voice Cloning System," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2967–2981, 2020.

[14] ——, "A Unified Speaker Adaptation Method for Speech Synthesis using Transcribed and Untranscribed Speech with Backpropagation," *ArXiv*, vol. abs/1906.07414, 2019. [Online]. Available: https://arxiv.org/abs/1906.07414

[15] Y. Song, J. Sohl-Dickstein, D. P. Kingma *et al.*, "Score-Based Generative Modeling through Stochastic Differential Equations," in *International Conference on Learning Representations*, 2021.

[16] N. Chen, Y. Zhang, H. Zen *et al.*, "WaveGrad: Estimating Gradients for Waveform Generation," in *International Conference on Learning Representations*, 2021.

[17] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme," in *International Conference on Learning Representations*, 2022.

[18] Y. Yan, X. Tan, B. Li *et al.*, "Adaspeech 2: Adaptive text to speech with untranscribed data," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6613–6617.

[19] M. McAuliffe, M. Socolof, S. Mihuc *et al.*, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[20] R. Liptser and A. Shiryaev, *Statistics of Random Processes*, ser. Stochastic Modelling and Applied Probability. Springer-Verlag, 1978, vol. 5.

[21] Y. Jia, Y. Zhang, R. Weiss *et al.*, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 4480–4490.

[22] J. Shen, Y. Jia, M. Chrzanowski *et al.*, "Non-Attentive Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modeling," *ArXiv*, vol. abs/2010.04301, 2020. [Online]. Available: https://arxiv.org/abs/2010.04301

[23] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, virtual*, 2020.

[24] K. Kim, S. Park, J. Lee *et al.*, "Assem-VC: Realistic Voice Conversion by Assembling Modern Speech Synthesis Techniques," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[25] S. Zhao, H. Wang, T. Nguyen *et al.*, "Towards Natural and Controllable Cross-Lingual Voice Conversion Based on Neural TTS Model and Phonetic Posteriorgram," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5969–5973.

[26] J. Valin and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis through Linear Prediction," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5891–5895.

[27] V. Popov, S. Kamenev, M. Kudinov, S. Repyevsky, T. Sadekova, V. Bushaev, V. Kryzhanovskiy, and D. Parkhomenko, "Fast and lightweight on-device TTS with Tacotron2 and LPCNet," in *Interspeech 2020*, 2020.

[28] D. Min, D. Lee, E. Yang *et al.*, "Meta-StyleSpeech: Multi-Speaker Adaptive Text-to-Speech Generation," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139. PMLR, 2021, pp. 7748–7759.