INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

HierVST: Hierarchical Adaptive Zero-shot Voice Style Transfer

Sang-Hoon Lee, Ha-Yeong Choi*, Hyung-Seok Oh, Seong-Whan Lee*†

Department of Artificial Intelligence, Korea University, Seoul, Korea

{sh_lee, hayeong, hs_oh, sw.lee}@korea.ac.kr

Abstract

Despite rapid progress in the voice style transfer (VST) field, recent zero-shot VST systems still lack the ability to transfer the voice style of a novel speaker. In this paper, we present HierVST, a hierarchical adaptive end-to-end zero-shot VST model. Without any text transcripts, we only use the speech dataset to train the model by utilizing hierarchical variational inference and self-supervised representation. In addition, we adopt a hierarchical adaptive generator that generates the pitch representation and waveform audio sequentially. Moreover, we utilize unconditional generation to improve the speaker-relative acoustic capacity in the acoustic representation. With a hierarchical adaptive structure, the model can adapt to a novel voice style and convert speech progressively. The experimental results demonstrate that our method outperforms other VST models in zero-shot VST scenarios. Audio samples are available at https://hiervst.github.io/.

Index Terms: voice conversion, voice style transfer, zero-shot voice conversion, self-supervised speech representation

1. Introduction

Recently, voice conversion (VC) systems [1, 2, 3, 4, 5] have shown rapid progress, with significant performance in voice style transfer (VST). Concurrently, progress in neural vocoder models [6, 7, 8, 9, 10] has accelerated the development of VC systems because of their ability to generate high-fidelity waveform audio, and the end-to-end VC systems [11, 12, 13, 14] have garnered significant interest by generating high-quality converted waveform audio by combining the VC and neural vocoder. However, end-to-end models still have low speaker adaptation performance and require text transcripts to disentangle linguistic representations from speech. Hence, there is a limitation where a paired text-audio dataset is required.

To utilize a non-parallel speech dataset, AutoVC [1] introduces an information bottleneck on content representation to disentangle the content and style information, and train the model with only self-reconstruction loss. However, there is a trade-off between audio quality and VST performance according to the information bottleneck size and there is a difficulty in choosing the appropriate information bottleneck size. F0-AutoVC [15] extends AutoVC to use an additional pitch contour from the source speech, and transforms the normalized pitch contour to the target pitch contour using the target speech statistics. Despite these pitch contour guides, most F0 extraction algorithms have a problem of extracting inaccurate F0 causing unnaturalness by generating a noisy sound and a voice style different from target speaker.

[11, 16] utilizes a discrete unit of self-supervised speech representation and quantized representation of normalized F0 to reconstruct speech, and convert the speech only by replacing the speaker representation. NANSY [17] utilizes continuous self-supervised speech representation, and introduces a speech perturbation to acquire only the linguistic representation from speech. HierSpeech [18] also uses self-supervised speech representation to extract the linguistic representation from speech, but text transcripts are required to regularize the linguistic representation to contain only linguistic information. Diffusion-based VC systems [19, 20] also show an improvement in generative performance. However, they also require text transcripts to train the average-Mel encoder from the extracted phoneme alignment [21]. In addition, most models still have limitations in zero-shot VC, resulting from a lack of ability in VST.

To address the above problems, we propose HierVST, a hierarchical adaptive end-to-end VST system. We adopt a multi-path self-supervised speech representation from a single speech by restoring the speaker-agnostic linguistic representation from perturbed speech and extracting the speaker-related linguistic representation from the original speech. We also introduce a hierarchical adaptive generator (HAG) with source modeling, and connect multiple representations through hierarchical variational inference. We found that hierarchical adaptation is the key to the success of zero-shot VC. Moreover, we present prosody distillation for enhanced linguistic representation and unconditional generation on the HAG to improve the acoustic capacity on acoustic representation for better speaker adaptation. The experimental results demonstrate that our model outperforms the others in terms of audio quality and speaker similarity on the zero-shot VST without any text transcripts.

2. HierVST

We present a hierarchical adaptive end-to-end VST system, HierVST. For untranscribed voice conversion, we introduce a multi-path self-supervised speech representation, and adopt hierarchical variational inference to connect the speech representations. Furthermore, we introduce a HAG, prosody distillation, and unconditional generation for better speaker adaptation. The details are described in the following subsections.

2.1. Speech representation

For voice conversion, we first decompose the speech into perturbed linguistic representation, linguistic representation, and acoustic representation and resynthesize the speech from disentangled representations. Following [12, 18], we use a high-resolution linear spectrogram to extract the acoustic representation. For speaker adaptation, we also extract the style representation from the Mel-spectrogram.

*Equal contribution
†Corresponding author

10.21437/Interspeech.2023-1608

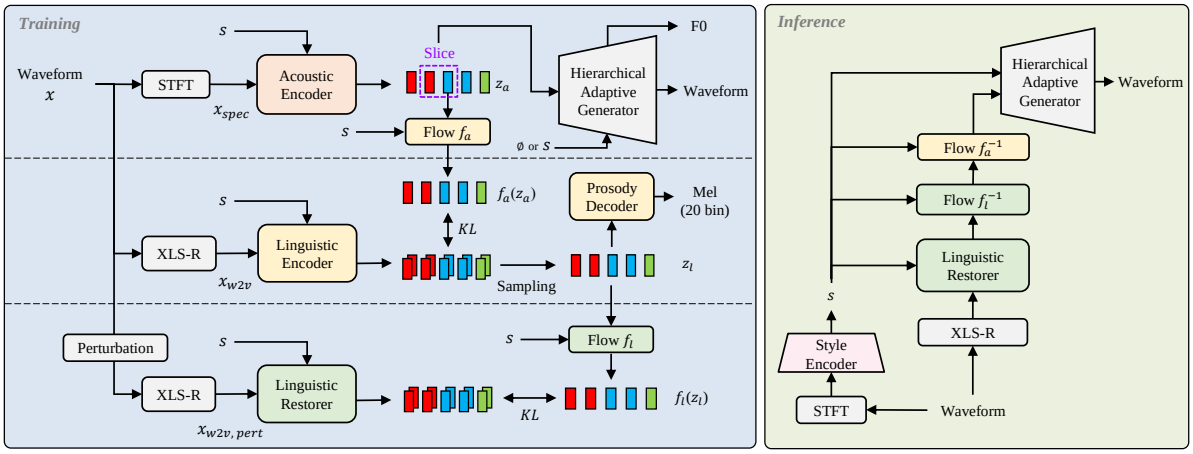

Figure 1: *Overall framework of HierVST*

2.1.1. Linguistic representation

Following [18], the wav2vec feature x_{w2v} is extracted from the representation from the middle layer of XLS-R, which is a pre-trained self-supervised model with a large-scale cross-lingual speech dataset. For speech disentanglement, we introduce a multi-path self-supervised speech representation by utilizing data perturbation [17] to reduce the content-irrelevant features from the same self-supervised speech model. The extracted $x_{w2v,pert}$ from the perturbed speech is fed to the linguistic restorer to restore the linguistic representation. The extracted x_{w2v} is fed to the linguistic encoder to extract an enhanced linguistic representation.

2.1.2. Style representation

For global voice style representation (timbre information), we extract the style representation from the Mel-spectrogram. The style encoder [22] is utilized to extract the style representation which is an averaged style vector of the single sentence, and this encoder is jointly trained with the model in an end-to-end manner. For hierarchical style adaptation, this style representation is fed to all networks including the linguistic restorer, linguistic encoder, acoustic encoder, normalizing flow modules, and HAG. For the zero-shot VST scenario, we do not use the speaker ID information, and we only extract the style representation from the speech.

2.2. Hierarchical variational autoencoder

We adopt the structure of HierSpeech [18] for an end-to-end VST system replacing the text encoder with a linguistic restorer. We utilize the perturbed linguistic representation $x_{w2v,pert}$ as conditional information c to hierarchically generate waveform audio. We additionally use the enhanced linguistic representation from the self-supervised representation of the original waveform, which is not perturbed. Moreover, the raw waveform audio is reconstructed from the acoustic representation which is extracted using a linear spectrogram during the training. To connect acoustic and multi-path linguistic representations, we utilize hierarchical variation inference, and the optimization objective of HierVST can be defined as follows:

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x)}\Big[\log p_{\theta_d}(x|z_a) - \log \frac{q_{\phi_a}(z_a|x_{spec})}{p_{\theta_a}(z_a|z_l)} - \log \frac{q_{\phi_l}(z_l|x_{w2v})}{p_{\theta_l}(z_l|c)}\Big] \quad (1)$$

where $q_{\phi_a}(z_a|x_{spec})$ and $q_{\phi_l}(z_l|x_{w2v})$ are the approximate posteriors for the acoustic and linguistic representations respectively. $p_{\theta_l}(z_l|c)$ represents a prior distribution of linguistic latent variables z_l given condition c, $p_{\theta_a}(z_a|z_l)$ denotes a prior distribution on acoustic latent variables z_a, and $p_{\theta_d}(x|z_a)$ is the likelihood function represented by a HAG that produces data x given latent variables z_a. In addition, we use the normalizing flow to improve the expressiveness of each linguistic representation. For the reconstruction loss, we use multiple reconstruction terms of a HAG as described in the following subsection.

2.3. Hierarchical adaptive generator

For end-to-end VC, we additionally introduce the HAG G which consists of the source generator G_s and waveform generator G_w as illustrated in Figure 2. The generated representations including acoustic representation z_a, style representation s are fed to G_s, and G_s generates the refined pitch representation p_h and auxiliary F0 predictor is used to enforce the F0 information on p_h as follows:

$$L_{pitch} = \|p_x - G_s(z_a, s)\|_1, \quad (2)$$

where p_x is the ground-truth (GT) log-scale F0. Subsequently, G_w synthesizes the waveform audio from z_a, p_h, s hierarchically, and we use the reconstruction loss between the GT and generated Mel-spectrogram transformed from waveform audio using STFT with Mel-filter ψ as follows:

$$L_{STFT} = \|\psi(x) - \psi(G_w(z_a, p_h, s))\|_1. \quad (3)$$

In addition, we utilize adversarial training [23, 24] to improve audio quality. We adopt the multi-period discriminator (MPD) [6][1] and the multi-scale STFT discriminator (MS-STFTD) [25] which can reflect the characteristic of real and imaginary components from a complex-valued STFT as:

$$\mathcal{L}_{adv}(D) = \mathbb{E}_{(x,z_a)}\Big[(D(x)-1)^2 + D(G(z_a,s))^2\Big], \quad (4)$$

$$\mathcal{L}_{adv}(\phi_a, \theta_d) = \mathbb{E}_{(z_a)}\Big[(D(G(z_a,s))-1)^2\Big] \quad (5)$$

2.4. Prosody distillation

We introduce prosody distillation to extract the enhanced linguistic representation z_l from the linguistic encoder. z_l is fed

[1]When we remove MPD for fast training, we observed that audio quality perceptually decreases.

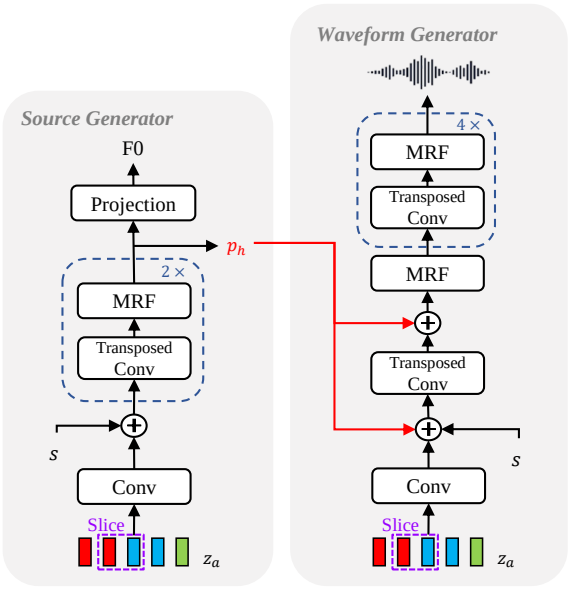

Figure 2: *Hierarchical adaptive generator*

to the prosody decoder which generates the first 20 bins of Mel-spectrogram containing the prosody representation. Unlike ProsoSpeech [26] which restricts the speaker information from the prosody vector, we make z_l acquire speaker-related prosody information for enhanced linguistic information. We use the prosody loss $\mathcal{L}_{prosody}$ which minimizes the $l1$ distance between the 20 bin of GT and reconstructed Mel-spectrogram.

2.5. Unconditional generation

For speaker adaptation, we use style representation as a condition for the network within the entire framework as mentioned in Section 2.1.2. We observed that the speaker adaptation is performed mainly in the HAG. Hence, we introduce an unconditional generation in the hierarchical generator to increase the speaker characteristic on the acoustic representation for progressive speaker adaptation. Following [27], we simply replace the style representation s with the null speaker embedding $\varnothing$ by a 10% chance, so we can treat the model as a conditional and unconditional model in the single model.

3. Experiment and result

3.1. Dataset and preprocessing

We use the large-scale multi-speaker dataset, LibriTTS [28] to train the model (*train-clean-360* and *train-clean-100*), which consists of about 300-hours of speech for 1,151 speakers. We use *dev-clean* subset for validation. To evaluate the zero-shot VST task, we utilize the VCTK dataset [29]. For both datasets, we downsample audio to 16 kHz. For self-supervised speech representation, the downsampled audio is fed to the XLS-R model to extract the linguistics-related representation from the middle layer of XLS-R, and this representation is a sequence of 1024-dimensional vectors downsampled from 16 kHz audio (320× downsampled scale). We also utilize high-resolution F0 which is a sequence of F0 extracted from audio (80× downsampled scale). For the Mel-spectrogram, we transform audio using the short-time Fourier transform (STFT) with a hop size of 320, a window size of 1,280, an FFT size of 1,280, and 80 bins of Mel-filter.

Table 1: *Many-to-many VST results from LibriTTS dataset*

Method	nMOS	sMOS	CER	WER	EER	SECS
GT	4.55±0.04	3.97±0.01	0.54	1.84	-	-
HiFi-GAN [6]	4.17±0.04	3.86±0.03	0.60	2.19	-	0.986
AutoVC [1]	2.57±0.06	2.21±0.05	5.34	8.53	33.30	0.703
VoiceMixer [31]	2.84±0.06	2.49±0.05	2.39	4.20	16.00	0.779
DiffVC [19]	3.50±0.06	3.02±0.05	7.99	13.92	11.00	0.817
SR [11]	2.75±0.06	2.32±0.05	6.63	11.72	33.30	0.693
YourTTS [32]	2.83±0.06	2.35±0.04	5.43	8.79	8.00	0.769
HierVST (Ours)	**4.06±0.05**	**3.29±0.04**	**0.84**	**2.22**	**5.25**	**0.827**

3.2. Training

We use the AdamW optimizer [30] with the same setting of [18]. We train HierVST with a batch size of 128 for 600k steps on four NVIDIA A100 GPUs (six days). For one-shot VST, we fine-tune the model with only a single sample of novel speakers for 1,000 steps and we initialize the same AdamW optimizer but a lower learning rate of 1×10^{-4}. We train the model for ablation study with a batch size of 64 on two A100 GPUs for 300k steps. For efficient training, we use a segment audio of 61,440 frames for input audio and utilize the windowed generator training with additional sliced audio of 9,600 frames.

3.3. Implementation details

The linguistic restorer, linguistic encoder, and acoustic encoder consist of 16 layers of non-causal WaveNet with 192 hidden dimensions. The flow modules including f_l and f_a consist of four affine coupling layers with four layers of WaveNet. For the HAG, the source generator consists of two upsampling layers of [2,2] and two multi-receptive field fusion (MRF) blocks, and the waveform generator consists of HiFi-GAN [6] and a conditional layer from the representation of the source generator. We use the upsampling rate of [4,5,4,2,2] and an initial channel of 512. For the discriminator, we use the MPD [6] and the MS-STFTD [25] with five different sizes of window([2048,1024,512,256,128]). We use a shallow feed-forward transformer network with two layers and 768 hidden dimensions for prosody distillation. For the unconditional generation, we set the ratio of unconditional generation p_{uncond} to 0.1. We fine-tune the model only with the conditional generation. The number of entire model parameter for inference is 45M.

3.4. Many-to-many VST

We compared our model with five baseline models: (1) AutoVC [1], information bottleneck based VC model. (2) VoiceMixer [31], similarity-based information bottleneck and adversarial training based VC model. (3) DiffVC [19], diffusion-based VC model. (4) Speech Resynthesis (SR) [11], an end-to-end model using discrete speech units. (5) YourTTS[2] [32], an end-to-end speech synthesis model, based on VITS [12]. Following [18], we conduct naturalness mean opinion score (nMOS) and similarity MOS (sMOS) for subjective evaluation metrics. To evaluate linguistic consistency, we also calculate the character error rate (CER) and word error rate (WER) by Whisper large model [33]. For the speaker similarity measurements, we calculate the equal error rate (EER) of the automatic speech recognition model [34] and speaker embedding cosine similarity (SECS) of Resemblyzer[3] between the target and converted speech.

Table 1 shows that our model achieves a significant improvement in all evaluation metrics. Specifically, audio quality improved and the speaker adaptation quality increased in terms

[2]We used an official pre-trained model. However, this model was trained with LibriTTS, VCTK, and an additional dataset. In addition, YourTTS utilizes text transcripts for training.

[3]https://github.com/resemble-ai/Resemblyzer

Table 2: *Zero-shot VST results on unseen speakers from VCTK*

Method	nMOS	sMOS	CER	WER	EER	SECS
GT	4.42±0.04	3.98±0.01	0.21	2.17	-	-
HiFi-GAN [6]	4.15±0.05	3.91±0.02	0.21	2.17	-	0.989
AutoVC [1]	2.47±0.05	1.79±0.05	5.14	10.55	37.32	0.715
VoiceMixer [31]	2.79±0.05	2.28±0.06	**1.08**	**3.31**	20.75	0.797
DiffVC [19]	3.51±0.07	2.44±0.05	6.92	13.19	24.01	0.785
SR [11]	2.27±0.05	2.15±0.06	2.12	6.18	27.24	0.750
YourTTS [32]	2.69±0.05	2.31±0.06	2.42	6.08	**4.02**	0.848
HierVST (Ours)	**4.12±0.05**	**2.70±0.06**	1.14	3.46	**5.06**	**0.850**

of nMOS and sMOS, respectively. Also, our model can convert the speech with a small loss of content information, where the CER and WER are much lower than others even though our model is trained without text transcripts. The objective metrics for speaker similarity also show that our model can adapt well to target voice style. Although HierVST has a similar structure using variational inference augmented with the normalizing flow [32], our hierarchical structure has better speaker adaptation and audio quality including naturalness and pronunciation.

3.5. Zero-shot VST

We compared the performance of zero-shot VST on the VCTK dataset. Table 2 shows that only our model can adapt to novel speakers in terms of EER and SECS. Note that YourTTS is trained with the VCTK dataset so the VST scenario of YourTTS is not the zero-shot VST. Nonetheless, the zero-shot speaker adaptation results of our model show a speaker adaptation quality similar to that of YourTTS in terms of EER and SECS. Furthermore, our model also achieves much better performance on both subjective metrics than others, and this means our model robustly converts speech even in the zero-shot VST scenario with a hierarchical adaptive structure.

Table 3: *One-shot VST results on VCTK dataset according to the number of fine-tuning steps*

Metric \ Step	0 (zero-shot)	100	300	500	1000	1500
CER (↓)	1.14	0.74	0.76	**0.66**	0.79	1.13
WER (↓)	3.46	2.77	2.85	**2.72**	3.05	3.63
EER (↓)	5.06	2.67	2.25	1.56	0.80	**0.50**
SECS (↑)	0.85	0.87	0.89	0.90	0.91	**0.92**

3.6. One-shot VST

We compared the performance with zero-shot and one-shot VST with different numbers of fine-tuning steps. Table 3 demonstrated that fine-tuning with one sample can improve the VST performance in terms of EER and SECS. However, the linguistic consistency decreased after overfitting to the small training samples so we only fine-tune the model with 1,000 steps.

3.7. Ablation study

3.7.1. Hierarchical VAE

We adopt the hierarchical VAE (HVAE) to restore the perturbed linguistic representation and to increase the speaker adaptation quality. Table 4 shows that removing the HVAE significantly decreases the performance of speaker adaptation. However, we found that the hierarchical structure requires more training steps to achieve the lower CER and WER for proper pronunciation in that the HierVST trained with 600k steps has a lower CER and WER. Also, the model has better naturalness which means that the hierarchical structure reduces the degradation of audio quality by regularizing an acoustic representation with a speaker-related linguistic representation. Note that it is necessary to perturb the waveform audio to remove the speaker-relevant information in the linguistic representation, so the model trained without audio perturbation is not able to convert the voice style.

Table 4: *Results of ablation study on zero-shot VST scenario*

Method	p_{uncond}	CER	WER	EER	SECS
HierVST (Ours)	0	2.56	5.86	6.73	0.843
	0.1	2.12	4.95	**6.25**	**0.847**
	0.2	2.04	4.79	7.77	0.838
	0.5	1.69	4.13	8.25	0.836
− PD	0	5.48	11.81	8.5	0.835
− PD − HVAE	0	1.05	3.66	11.75	0.816
− PD − HVAE − HAG	0	**0.78**	**3.09**	13.75	0.816

3.7.2. Hierarchical adaptive generator

We modify the HiFi-GAN by combining it with the source generator. With the distillation of the source-related representation, the model with a HAG synthesizes audio with better quality as indicated in Table 4 and the adaptation performance also increased regarding EER.

3.7.3. Prosody distillation

Although hierarchical VAE can improve the VST quality, the model has a higher CER and WER. Therefore, we additionally introduce prosody distillation (PD) for enhanced linguistic representation. Adding prosody distillation improves the overall performance regarding all metrics with an enhanced linguistic representation. We also compared the 20 bins of Mel-spectrogram with the full-band of the Mel-spectrogram, and the model trained with 20 bins of Mel-spectrogram has a lower F0 $l1$ distance in the source generator, therefore, we used only 20 bins for the prosody distillation.

3.7.4. Unconditional generation

We train the model with unconditional generation on the HAG with different unconditional ratios. We found that increasing the unconditional ratio improved the pronunciation of the converted speech. However, a model with a small ratio could generate converted speech with better speaker adaptation. Table 4 shows that adopting an unconditional generation with a proper ratio simply improved the model capacity for generation tasks.

4. Conclusion

We present HierVST, which can convert speech by hierarchically transferring the voice style. With only a speech dataset, we restored the linguistic representation from the disentangled representation, reproduced the enhanced linguistic and rich acoustic representation, and generated high-quality converted speech. Furthermore, we improve the capacity of the entire model using prosody distillation and unconditional generation. The experimental results demonstrated that our model can generate converted speech with high-fidelity audio and high-quality speaker adaptation. We see that our hierarchical adaptive structure can be adopted in unit-based speech-to-speech translation systems to generate an expressive voice style of translated speech. Although our model generates high-quality converted speech, our model has little controllability without converting the timbre. In future works, we will utilize pitch and duration to directly control the intonation and rhythm of speech.

5. Acknowledgements

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 No. 2021-0-02068, Artificial Intelligence Innovation Hub) and Clova Voice, NAVER Corp., Seongnam, Korea.

6. References

[1] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot Voice Style Transfer with Only Autoencoder Loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[2] S. Yang, M. Tantrawenith, H. Zhuang, Z. Wu, A. Sun, J. Wang, N. Cheng, H. Tang, X. Zhao, J. Wang, and H. Meng, "Speech Representation Disentanglement with Adversarial Mutual Information Learning for One-shot Voice Conversion," in *Proc. Interspeech 2022*, 2022, pp. 2553–2557.

[3] S.-H. Lee, H.-R. Noh, W.-J. Nam, and S.-W. Lee, "Duration Controllable Voice Conversion via Phoneme-Based Information Bottleneck," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1173–1183, 2022.

[4] P. Bilinski, T. Merritt, A. Ezzerg, K. Pokora, S. Cygert, K. Yanagisawa, R. Barra-Chicote, and D. Korzekwa, "Creating New Voices using Normalizing Flows," in *Proc. Interspeech 2022*, 2022, pp. 2958–2962.

[5] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "DDDM-VC: Decoupled Denoising Diffusion Models with Disentangled Representation and Prior Mixup for Verified Robust Voice Conversion," *arXiv preprint arXiv:2305.15816*, 2023.

[6] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[7] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-GAN: Adversarial Frequency-Consistent Audio Synthesis," in *Interspeech*, 2021.

[8] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Proc. Interspeech 2021*, 2021, pp. 2207–2211.

[9] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion Probabilistic Model Based Neural Vocoder with Adaptive Noise Spectral Shaping," in *Interspeech*, 2022.

[10] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Interspeech*, 2021.

[12] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[13] Y. Liu, R. Xue, L. He, X. Tan, and S. Zhao, "DelightfulTTS 2: End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders," in *Proc. Interspeech 2022*, 2022, pp. 1581–1585.

[14] H.-S. Choi, J. Yang, J. Lee, and H. Kim, "NANSY++: Unified voice synthesis with neural analysis and synthesis," in *The Eleventh International Conference on Learning Representations*, 2023.

[15] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6284–6288.

[16] G. Maimon and Y. Adi, "Speaking Style Conversion With Discrete Self-Supervised Units," *arXiv preprint arXiv:2212.09730*, 2022.

[17] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural Analysis and Synthesis: Reconstructing Speech from Self-Supervised Representations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 251–16 265, 2021.

[18] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, "HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis," in *Advances in Neural Information Processing Systems*, 2022.

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

[20] T. Sadekova, V. Gogoryan, I. Vovk, V. Popov, M. Kudinov, and J. Wei, "A Unified System for Voice Cloning and Voice Conversion through Diffusion Probabilistic Modeling," in *Proc. Interspeech 2022*, 2022, pp. 3003–3007.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[22] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-StyleSpeech: Multi-Speaker Adaptive Text-to-Speech Generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7748–7759.

[23] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee, "Multi-SpectroGAN: High-Diversity and High-Fidelity Spectrogram Generation with Adversarial Style Combination for Speech Synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 13 198–13 206.

[24] H. Chung, S.-H. Lee, and S.-W. Lee, "Reinforce-Aligner: Reinforcement Alignment Search for Robust End-to-End Text-to-Speech," in *Proc. Interspeech 2021*, 2021, pp. 3635–3639.

[25] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

[26] Y. Ren, M. Lei, Z. Huang, S. Zhang, Q. Chen, Z. Yan, and Z. Zhao, "ProsoSpeech: Enhancing Prosody With Quantized Vector Pre-training in Text-to-Speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7577–7581.

[27] S. Kim, H. Kim, and S. Yoon, "Guided-TTS 2: A Diffusion Model for High-quality Adaptive Text-to-Speech with Untranscribed Data," *arXiv preprint arXiv:2205.15370*, 2022.

[28] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," 2019, pp. 1526–1530.

[29] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[30] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[31] S.-H. Lee, J.-H. Kim, H. Chung, and S.-W. Lee, "VoiceMixer: Adversarial Voice Style Mixup," *Advances in Neural Information Processing Systems*, vol. 34, pp. 294–308, 2021.

[32] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-scale Weak Supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[34] Y. Kwon, H. S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: lessons from VoxSRC 2020," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.