INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation

Ha-Yeong Choi, Sang-Hoon Lee, Seong-Whan Lee†

Department of Artificial Intelligence, Korea University, Seoul, Korea
{hayeong, sh_lee, sw.lee}@korea.ac.kr

Abstract

Although voice conversion (VC) systems have shown a remarkable ability to transfer voice style, existing methods still have an inaccurate pitch and low speaker adaptation quality. To address these challenges, we introduce Diff-HierVC, a hierarchical VC system based on two diffusion models. We first introduce DiffPitch, which can effectively generate F_0 with the target voice style. Subsequently, the generated F_0 is fed to DiffVoice to convert the speech with a target voice style. Furthermore, using the source-filter encoder, we disentangle the speech and use the converted Mel-spectrogram as a data-driven prior in DiffVoice to improve the voice style transfer capacity. Finally, by using the masked prior in diffusion models, our model can improve the speaker adaptation quality. Experimental results verify the superiority of our model in pitch generation and voice style transfer performance, and our model also achieves a CER of 0.83% and EER of 3.29% in zero-shot VC scenarios.

Index Terms: voice conversion, diffusion models, pitch generation, speech restoration, zero-shot style transfer

1. Introduction

Voice conversion (VC) tasks typically convert the voice of a source speaker into the voice of a specific target speaker, and the linguistic information of the converted target speaker must be consistent with the source speech. The primary concept of VC is to disentangle the individual components of speech so that each component can be controlled and transformed to the target speaker voice. Recently, the VC system has significantly advanced with deep learning approaches, allowing clarity and naturalness of the converted voice [1, 2, 3, 4, 5, 6]. Moreover, the expansion of the conversion system has enabled effective applications in a variety of fields, such as cross-lingual [7, 8] and emotional VC [9, 10]. Despite these advancements, converted voice is still perceived as unnatural owing to mispronunciation in converted speech, and low speaker adaptation performance is still a challenge that requires addressing [11, 12].

Pitch modeling is essential to achieving speech intelligibility and naturalness in VC and text-to-speech (TTS) tasks. Pitch characteristics are crucial in speaker identity and correct pronunciation [13]. [14, 15] train the model using normalized fundamental frequency (F_0) to obtain the same mean and variance for all speakers. This approach contributes to expressiveness by considering the pitch information. However, F_0 is not entirely separated from the speaker style, so that it still causes perceptual unnaturalness in the conversion. SR [3] uses a vector quantized variational auto-encoder (VQ-VAE) to learn a speaker-irrelevant pitch representation. Although speaker-irrelevant pitch can be extracted, mispronunciation occurs due to the loss of pitch information during vector quantization. In addition, it is difficult to precisely predict the pitch of a voice with a high degree of expressiveness.

To address the above problem, we propose Diff-HierVC, a novel diffusion-based hierarchical VC system. Diff-HierVC consists of DiffPitch and DiffVoice, which hierarchically convert the voice style from disentangled speech representations. DiffPitch generates the pitch information of the target speaker during the inference step, and DiffVoice constructs a high-quality Mel-spectrogram utilizing the generated pitch information and the source-filter representation according to the source-filter theory. We found that a hierarchical VC architecture is an effective structure for decoupling speech components and generating the converted speech. Moreover, using the data-driven prior, we improved the conversion performance by regulating the inception of the denoising process of the diffusion model. Furthermore, we introduce a masked prior that allows the diffusion model to consider the context and condition for better generalization ability and robust training. The experimental results denote that pitch and voice modeling have considerable effects. Our main contributions are summarized as follows:

- We propose Diff-HierVC, a diffusion-based hierarchical VC system with robust pitch generation and masked prior for expressive zero-shot voice style transfer.
- To the best of our knowledge, this is the first study to utilize the diffusion process to generated F_0. We demonstrated that using the generated F_0 by the denoising diffusion process rather than conventional pitch modeling methods resulted in more accurate pronunciation and natural intonation of the converted voice.
- The experimental results reveal that Diff-HierVC achieves a significantly improved zero-shot style transfer in various conversion scenarios with cross-lingual and expressive real-world speech dataset. Our demos are available at `https://diff-hiervc.github.io/`.

2. Background: diffusion models

Diffusion models have shown extraordinary performance in generative tasks in various domains, such as images, videos, and audio, and have recently achieved considerable success in multi-modal tasks [16, 17]. Specifically, in speech, the diffusion model is utilized in applications such as audio generation [18, 19], speech enhancement [20], and TTS synthesis [21, 22]. The fundamental concept underlying the stochastic differential equation (SDE)-based continuous-time diffusion process [23] is to train an estimator that repeatedly removes noise by estimating log-density gradient of data and generates samples with an iterative denoising process via SDE. The SDE-based diffusion model was also applied to VC task using a maximum likelihood (ML)-SDE solver [5] for fast sampling.

†Corresponding author

 10.21437/Interspeech.2023-817

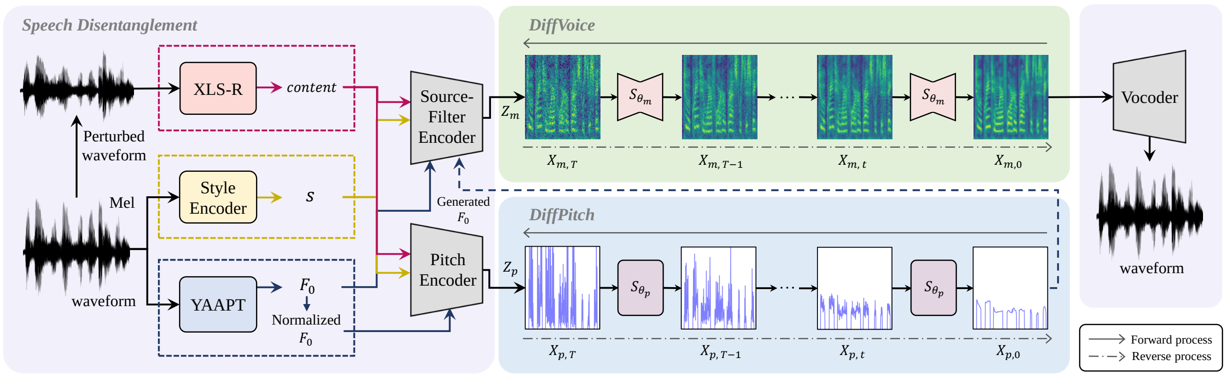

Figure 1: *Overall framework*

3. Diff-HierVC

3.1. Speech disentanglement

As illustrated in Figure 1, we first analyze speech into representations of content, pitch, and style: (1) Data perturbation [4] is applied to the input waveform to eliminate content-irrelevant information. Subsequently, we extract the content features from the intermediate layer representation of XLS-R [24], a pre-trained self-supervised model using a large-scale cross-lingual speech dataset. (2) We utilize a style encoder [25] to extract the voice style, which is the speaker style representation from the Mel-spectrogram. The style embedding serves as a guide for both the content encoder and pitch encoder. (3) We extract a fundamental frequency (F_0) using the YAAPT algorithm [26] with a $4\times$ high-resolution higher than Mel-spectrogram for precise pitch extraction. The content encoder receives $\log(F_0+1)$, and the pitch encoder takes the normalized F_0 as the mean and variance of the source speaker's F_0.

3.2. Hierarchical VC

For hierarchical VC, we introduce a two-stage diffusion models, DiffPitch and DiffVoice. DiffPitch initially converts the F_0 with the target voice style, and the converted F_0 is fed to DiffVoice to convert the speech with the target voice style hierarchically. The details of each diffusion models are described as follows.

3.2.1. DiffPitch

We introduce DiffPitch, a pitch generator based on the diffusion process. To consider continuous pitch information, we adopt a WaveNet [27] based conditional diffusion model [18], which can iteratively obtain a significant receptive field with a single denoiser. The pitch encoder transforms the normalized F_0 of the source speech to the pitch representation Z_p. We regularized the pitch representation by pitch reconstruction loss to utilize Z_p as a data-driven prior of DiffPitch as follows:

$$\mathcal{L}_{pitch} = \|X_p - Z_p\|_1. \quad (1)$$

The diffusion process of DiffPitch uses the log-scale F_0 extracted using the YAAPT algorithm as a target ground-truth X_p. The forward process of the DiffPitch is defined as follows:

$$dX_{p,t} = \frac{1}{2}\beta_t(Z_p - X_{p,t})dt + \sqrt{\beta_t}d\mathbf{w}_t\ , \quad (2)$$

where $t \in [0,1]$, β_t regulates the amount of stochastic noise injected in the process, and $\mathbf{w}_t$ is the forward standard Wiener process. DiffPitch executes denoising to recover the original pitch contour in the reverse process. The reverse process of the pitch denoiser is defined as follows:

$$d\hat{X}_{p,t} = \left(\frac{1}{2}(Z_p - \hat{X}_{p,t}) - s_{\theta_p}(\hat{X}_{p,t}, Z_p, t)\right)\beta_t dt + \sqrt{\beta_t}d\bar{\mathbf{w}}_t, \quad (3)$$

where $\bar{\mathbf{w}}_t$ denote the backward standard Wiener process. According to [5], in the forward process, a sample of noisy pitch is drawn from the following distribution:

$$p_{t|0}(X_{p,t}|X_{p,0}) = \mathcal{N}\left(e^{-\frac{1}{2}\int_0^t \beta_s ds}X_{p,0} + \left(1 - e^{-\frac{1}{2}\int_0^t \beta_s ds}\right)Z_p\right.$$
$$\left., \left(1 - e^{-\int_0^t \beta_s ds}\right)\mathrm{I}\right), \quad (4)$$

where I is the identity matrix. Distribution (4) is Gaussian, thus we obtain the following equation:

$$\nabla \log p_{t|0}(X_{p,t}|X_{p,0}) = -\frac{X_{p,t} - X_{p,0}(e^{-\frac{1}{2}\int_0^t \beta_s ds}) - Z_p(1 - e^{-\frac{1}{2}\int_0^t \beta_s ds})}{1 - e^{-\int_0^t \beta_s ds}}. \quad (5)$$

Therefore, DiffPitch approximates the score function with the following denoising objective:

$$\mathcal{L}_p = \mathbb{E}_{X_0, X_t}\left[\lambda_t \left\|\left(s_{\theta_p}(X_{p,t}, Z_p, s, t)\right) - \nabla \log p_{t|0}(X_{p,t}|X_{p,0})\right\|_2^2\right], \quad (6)$$

where $s_{\theta p}$ is the pitch score estimator and $\lambda_t = 1 - e^{-\int_0^t \beta_s ds}$. Furthermore, we derive fast sampling using the ML-SDE solver [5], which maximizes the log-likelihood of forward diffusion with the reverse SDE solver. During inference, the converted F_0 from the pitch encoder is utilized as a prior of DiffPitch, and DiffPitch generates the refined F_0 with the target voice style s. Note that we normalize F_0 only with the statistic of a single sentence for the fair zero-shot voice conversion scenario.

3.2.2. DiffVoice

We present DiffVoice, a conditional diffusion model for high-quality speech synthesis from content, target F_0, and target voice style. We also utilize a data-driven prior for the diffusion models to guide the inception. According to the source-filter theory [28], we first disentangle the speech components into a pitch and content representation. For a data-driven prior of DiffVoice, the source-filter encoder which consists of the source encoder E_{src} and filter encoder E_{ftr} reconstructs the intermediate Mel-spectrogram Z_m from the disentangled speech representation as $Z_m = Z_{src} + Z_{ftr}$, where $Z_{src} = E_{src}(F_0, s)$, $Z_{ftr} = E_{ftr}(content, s)$, and s denotes style embedding. Mel-spectrogram Z_m is regularized as follows:

$$\mathcal{L}_{rec} = \|X_{mel} - Z_m\|_1, \quad (7)$$

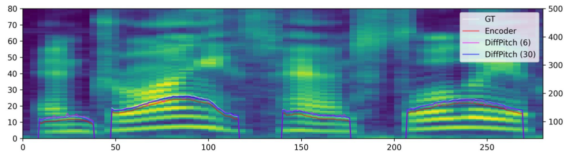

Figure 2: F_0 *reconstruction results on* F_0 *encoder and DiffPitch*

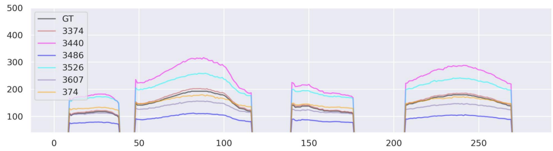

Figure 3: F_0 *generation results of target speech*

where X_{mel} is the Mel-spectrogram of the ground-truth speech. Subsequently, DiffVoice can utilize the source-filter encoder output Z_m as a prior, and use speaker representation s as condition to maximize speaker adaptation capacity. The following equation describes the forward process of DiffVoice:

$$\mathrm{d}X_{m,t} = \frac{1}{2}\beta_t(Z_m - X_{m,t})\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\mathbf{w}_t. \tag{8}$$

The reverse process of DiffVoice is defined by:

$$\mathrm{d}\hat{X}_{m,t} = \left(\frac{1}{2}(Z_m - \hat{X}_{m,t}) - s_{\theta_m}(\hat{X}_{m,t}, Z_m, t)\right)\beta_t \mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\bar{\mathbf{w}}_t. \tag{9}$$

In the forward process, a sample of noisy Mel-spectrogram $X_{m,t}$ is taken in the same manner as equation (4). Finally, the objective of training the Mel-spectrogram noise estimation network $s_{\theta m}$ is to optimize the score matching loss:

$$\mathcal{L}_m = \mathbb{E}_{X_0, X_t}\left[\lambda_t \left\| (s_{\theta_m}(X_{m,t}, Z_m, s, t)) - \nabla \log p_{t|0}(X_{m,t}|X_{m,0}) \right\|_2^2\right]. \tag{10}$$

During inference, the source-filter encoder takes a content representation from the source speech, target voice style s, and the converted F_0 from DiffPitch with the target voice style. The converted Mel-spectrogram Z_m from the source-filter encoder is used as a data-driven prior, and DiffVoice generates the converted speech conditioned with the target voice style.

3.3. Denoising models with masked prior

Although the data-driven prior can significantly improve the conversion performance, DiffVoice may rely on the reconstructed Mel-spectrogram in the source-filter encoder. To improve generalization performance of DiffVoice, we introduce a masked prior to the denoising diffusion models. Before fed to the DiffVoice, the prior Z_m is masked, and the diffusion network jointly learns the reconstruction and denoising process. Consequently, the model can reconstruct the masked area from the surrounding context. Specifically, we apply frequency masking by interpreting continuous pitch information from a contextual point of view.

4. Experiment and result

4.1. Experimental setup

4.1.1. Datasets and preprocessing

We train the model with a large-scale publicly available multi-speaker dataset, LibriTTS [29]. We utilize the *train-clean-360* and *train-clean-100* subsets of LibriTTS, which contain 110 hours of speech from 1,151 speakers. We additionally use *dev-clean-other* subsets of LibriTTS for validation. Then, we use VCTK dataset [30] to evaluate the zero-shot VC performance. We randomly select sentences from the paired speech of VCTK dataset. We downsample the audio to 16 kHz, and transform the audio into a log-scale Mel-spectrogram with 80 bins using short-time Fourier transform (STFT) and Mel-filters. We use a hop size of 320 and a window size of 1,280 to map the time-resolution of the self-supervised speech representation.

4.1.2. Training

We train the model using LibriTTS for 2M steps with a batch size of 64 on two NVIDIA A100 GPUs (five days), and use AdamW optimizer with the setting of [31], and implemented the learning rate schedule with a decay of $0.999^{1/8}$ at an initial learning rate of 5×10^{-5}. We segment the audio clip into 35,840 frame during training. For fine-tuning, we set the initial learning rate to 2×10^{-5}. A non-causal dilated WaveNet with 128 dimensions is used for all encoders, and DiffPitch uses the DiffWave with 64 dimensions and additional conditional layers for the pitch and style representations. DiffVoice consists of a 2D-UNet structure with the initial channel of 64 and three blocks, and the dimension of blocks are [64, 128, 256]. Following [5], we use the noise schedule parameters of β_0 and β_1 with 0.05 and 20 respectively. The masking ratio is set to 30% for the masked prior. For vocoder, we train the HiFi-GAN [32] with the same training dataset, and we only replace the discriminators with multi-scale STFT discriminator of EnCodec [33].

4.2. Analysis on F_0 prediction

Most previous VC systems utilize normalized or quantized F_0 for speaker-irrelevant pitch modeling. However, we estimate a raw F_0 with a target voice style for better speaker adaptation. We compared three F_0 prediction methods: F_0 transformation with a statistic of F_0, simple F_0 prediction with WaveNet, a diffusion-based F_0 prediction with DiffPitch. Figure 2 depicts that DiffPitch with 30 iteration steps has a similar F_0 contour with the ground-truth F_0. Figure 3 also show the diversity of pitch contours with different target voice styles. Hence, we utilize the converted F_0 by DiffPitch during VC with DiffVoice.

4.3. Zero-shot VC

We conduct various subjective and objective evaluation on the zero-shot VC scenario with three models: (1) autoencoder based VC model, AutoVC [1], (2) GAN based VC model, VoiceMixer, (3) unit-based end-to-end speech model, Speech Resynthesis (SR) [3], and (4) diffusion-based VC model, DiffVC[1]. Following [31], we conduct the naturalness and similarity mean opinion score (nMOS and sMOS, respectively). Table 1 depicts that our model has a better nMOS and sMOS than the others. Specifically, our model achieves significantly improved content consistency[2] and the speaker adaptation performance[3]. In addition, we conducted cross-lingual VC to demon-

[1]To train with the same settings as our model, we trained the speaker encoder of DiffVC with train-clean-100 and 360 with 1,151 speakers. Also, DiffVC* denotes the VC results using the official checkpoint. The official code implementation uses a speaker encoder trained with a large multi-speaker dataset (voxceleb1, voxceleb2, and LibriTTS-other) containing 8,371 speakers to extract and use the speaker embedding.

[2]We use an automatic speech recognition model, Whisper-large [34], and calculate the character error rate (CER) and word error rate (WER) on the 400 converted speeches with the text normalizer

[3]For 400×20 = 8,000 paired speeches, we measure the equal error rate (EER) of automatic speaker verification model [35]. We use Resemblyzer to calculate the speaker encoder cosine similarity (SECS).

Table 1: *Zero-shot VC results on unseen speakers from VCTK dataset*

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	Params.
GT	-	3.68±0.09	3.59±0.03	0.21	2.17	-	-	-
GT (Mel + Vocoder)	-	3.70±0.09	3.42±0.04	0.21	2.17	-	0.989	13M
AutoVC [1]	-	3.56±0.09	2.63±0.07	5.14	10.55	37.32	0.715	30M
VoiceMixer [2]	-	3.59±0.09	2.98±0.06	1.08	3.31	20.75	0.797	52M
SR [3]	-	3.51±0.10	2.83±0.06	5.14	10.55	37.32	0.715	15M
DiffVC*[5]	6 / 30	3.39±0.09 / 3.48±0.09	2.81±0.06 / 2.88±0.06	6.86 / 7.51	13.77 / 14.42	9.25 / 10.05	0.826 / 0.842	127M
DiffVC [5]	6 / 30	3.63±0.09 / 3.63±0.09	2.98±0.06 / 2.94±0.06	5.82 / 6.92	11.76 / 13.19	25.30 / 24.01	0.786 / 0.785	123M
Diff-HierVC (Ours)	6 / 30	**3.70±0.09 / 3.74±0.09**	3.03±0.06 / 3.02±0.06	**0.83 / 1.19**	**3.11 / 3.58**	3.29 / 3.66	0.861 / 0.860	18M
Diff-HierVC-Finetune (Ours)	6 / 30	3.65±0.09 / 3.66±0.09	**3.04±0.05 / 3.07±0.05**	0.97 / 1.34	3.15 / 3.75	**1.50 / 1.26**	**0.894 / 0.894**	18M

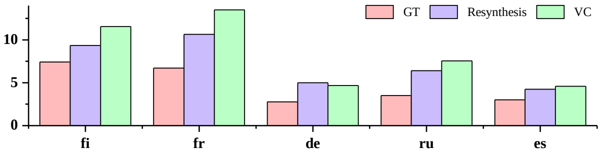

Figure 4: *CER results for zero-shot cross-lingual VC on CSS10*

strate zero-shot conversion performance in unseen languages. Figure 4 shows the robust generalization performance of our model in both resynthesis and VC scenarios, even for unseen languages. Furthermore, we fine-tune the model using only one sample per speaker. Fine-tuning with small steps (1,000 steps) can improve the performance of speaker adaptation. However, the model fine-tuned with more steps shows a lower robustness of content consistency by decreasing the CER and WER.

4.4. Ablation study

4.4.1. Pitch modeling

We compared three pitch modeling methods: DiffPitch, F_0 transformation with denormalization (Denorm.) [14], and a simple F_0 prediction with the F_0 Encoder. All methods utilize the same normalized F_0 of the source speech to convert the F_0 with target voice style. Although Denorm. could transform the normalized F_0 with the mean and variance of target speech, inaccurate F_0 extracted from target speech decreases the voice style transfer performance regarding CER and WER with a mispronunciation and inaccurate intonation as indicated in Table 2. In addition, using only the F_0 encoder decreases the voice style transfer performance with a higher EER than the DiffPitch even with the same WaveNet structure.

4.4.2. Data-driven prior and masked prior

As indicated in Table 2, the data-driven prior significantly improves the voice style transfer performance. However, we found that the diffusion models may rely on the performance of source-filter encoder and the diffusion models slightly reflect the conditional information to generate the converted speech. In addition, the inaccurate ground-truth F_0 extracted by YAPPT is sometimes fed to the models during training and inference. Employing masked prior improves the performance with better generalization on the diffusion models taking advantage of data-driven prior. In addition, an experiment was carried out to determine the suitable masking ratio, and as a result of Table 3, a masking ratio of 30% performed best.

4.4.3. Source-filter encoder

It is well known that disentangling the speech plays a important role in controlling speech representation. In this work, we adopt the source-filter (SF) encoder to disentangle the speech components and regulate the starting point of diffusion models. To evaluate the effectiveness of source-filter encoder, we replace the source-filter encoder with a single encoder. Using the single encoder decreases the performance of voice conversion of the entire model, which could not appropriately disentangle the speech representation and it results in the converted speech for a prior of DiffVoice having a lower speaker similarity with the target speech as indicated in Table 2.

Table 2: *Results of ablation study on zero-shot VC tasks with unseen speakers from VCTK dataset. For all methods, the number of sampling iterations is 6.*

Method	nMOS	sMOS	CER	EER	SECS
Diff-HierVC	**3.86±0.06**	**3.02±0.09**	0.83	**3.29**	**0.861**
Denorm. + DiffVoice	3.81±0.06	3.00±0.10	2.67	5.25	0.850
F_0 Encoder + DiffVoice	3.83±0.06	3.00±0.09	0.89	4.09	0.857
w.o Masked Prior	3.83±0.06	2.91±0.10	0.82	4.52	0.852
w.o Data-driven Prior	3.81±0.06	2.90±0.10	0.56	12.77	0.823
w.o SF Encoder	3.83±0.06	3.01±0.10	0.68	6.75	0.847
DiffPitch + SF Encoder	3.77±0.06	2.95±0.10	**0.30**	5.26	0.854

Table 3: *Results of ablation study on different masking ratio*

Metric	0%	10%	30%	50%	70%	90%
CER (↓)	0.82	0.70	0.83	0.86	0.89	0.96
EER (↓)	4.52	4.55	3.29	3.75	3.74	3.75

5. Conclusion

In this paper, we presented Diff-HierVC, a diffusion-based hierarchical VC system for high-fidelity converted pitch and Mel-spectrogram generation. DiffPitch improves performance in terms of speaker similarity and phonetic intelligibility. Then, DiffVoice restores high-quality speech through a denoising process. Subsequently, for better generalization of the diffusion model, we proposed a masked prior that can be robustly converted by considering the context and diffusion conditions. Consequently, our model outperformed the state-of-the-art in all metrics even with 6.8× fewer parameters, and we demonstrated the feasibility of building the zero-shot cross-lingual VC system, which can *Break Down Barriers* on various low-resource speech and language technologies. However, although our methods can significantly improve speaker adaptation quality, there are cases where the noise of input data is also considered as style. Hence, there is room for improvement towards high-quality and noise-free audio. In future work, we will decouple the noise and speech style with noise augmentation to generate high-fidelity audio even in a noisy environment.

6. Acknowledgements

This work was partly 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 ESTsoft Corp., Seoul, Korea.

7. 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.-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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] S. Zhao, H. Wang, T. H. Nguyen, and B. Ma, "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)*. IEEE, 2021, pp. 5969–5973.

[8] 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.

[9] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion," *Proc. Interspeech 2020*, pp. 3416–3420, 2020.

[10] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[11] J. Huang, W. Xu, Y. Li, J. Liu, D. Ma, and W. Xiang, "FlowCPCVC: A Contrastive Predictive Coding Supervised Flow Framework for Any-to-Any Voice Conversion," *Proc. Interspeech 2022*, pp. 2558–2562, 2022.

[12] R. Yuan, Y. Wu, J. Li, and J. Kim, "DeID-VC: Speaker De-identification via Zero-shot Pseudo Voice Conversion," in *Proc. Interspeech 2022*, 2022, pp. 2593–2597.

[13] A. Łańcucki, "FastPitch: Parallel Text-to-speech with Pitch Prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[14] 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)*. IEEE, 2020, pp. 6284–6288.

[15] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised Speech Decomposition via Triple Information Bottleneck," in *International Conference on Machine Learning*, 2020, pp. 7836–7846.

[16] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-Based Real Image Editing with Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.

[17] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, "MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 219–10 228.

[18] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," in *International Conference on Learning Representations*, 2021.

[19] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating Gradients for Waveform Generation," in *International Conference on Learning Representations*, 2021.

[20] S. Welker, J. Richter, and T. Gerkmann, "Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain," in *Proc. Interspeech 2022*, 2022, pp. 2928–2932.

[21] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," in *International Conference on Machine Learning*, 2021, pp. 8599–8608.

[22] R. Huang, M. W. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2022.

[23] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," in *International Conference on Learning Representations*.

[24] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[25] 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.

[26] K. Kasi and S. A. Zahorian, "Yet Another Algorithm for Pitch Tracking," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–361.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, p. 125.

[28] G. Fant, *Acoustic Theory of Speech Production*. Walter de Gruyter, 1970, no. 2.

[29] 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.

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

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.