DDDM-VC: Decoupled Denoising Diffusion Models with Disentangled Representation and Prior Mixup for Verified Robust Voice Conversion

Ha-Yeong Choi[1*], **Sang-Hoon Lee**[2,3*], **Seong-Whan Lee**[1†]
[1]Department of Artificial Intelligence, Korea University, Seoul, Korea
[2]Department of Artificial Intelligence, Ajou University, South Korea
[3]Department of Software and Computer Engineering, Ajou University, South Korea
[1]{hayeong, sw.lee}@korea.ac.kr, [2,3]{sanghoonlee}@ajou.ac.kr

Abstract

Diffusion-based generative models have recently exhibited powerful generative performance. However, as many attributes exist in the data distribution and owing to several limitations of sharing the model parameters across all levels of the generation process, it remains challenging to control specific styles for each attribute. To address the above problem, we introduce decoupled denoising diffusion models (DDDMs) with disentangled representations, which can enable effective style transfers for each attribute in generative models. In particular, we apply DDDMs for voice conversion (VC) tasks, tackling the intricate challenge of disentangling and individually transferring each speech attributes such as linguistic information, intonation, and timbre. First, we use a self-supervised representation to disentangle the speech representation. Subsequently, the DDDMs are applied to resynthesize the speech from the disentangled representations for style transfer with respect to each attribute. Moreover, we also propose the prior mixup for robust voice style transfer, which uses the converted representation of the mixed style as a prior distribution for the diffusion models. The experimental results reveal that our method outperforms publicly available VC models. Furthermore, we show that our method provides robust generative performance even when using a smaller model size. Audio samples are available at https://hayeong0.github.io/DDDM-VC-demo/.

1 Introduction

Denoising diffusion models (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Song et al. 2021) have achieved significant success in image generation tasks (Ramesh et al. 2022; Saharia et al. 2022b). Diffusion models have also attracted increasing interest in the audio domain in recent years, owing to their ability to synthesize high-quality speech (e.g., Mel-spectrogram and audio). Various applications employ diffusion models, such as text-to-speech (TTS) (Popov et al. 2021; Kim, Kim, and Yoon 2022a,b), neural vocoder (Kong et al. 2021; Chen et al. 2021; Huang et al. 2022), speech enhancement (Han and Lee 2022), and voice conversion (VC) (Liu et al. 2021; Popov et al. 2022).

Although diffusion models have achieved success in most speech applications owing to their powerful generative performance, there remains room for improvement in conventional methods. As data include many attributes, it is difficult to control specific styles for each attribute with a single denoiser that shares the model parameters across all levels of generation process. To reduce this burden in the image generation domain, eDiff-i (Balaji et al. 2022) subdivides the single denoiser into multiple specialized denoisers that originate from the single denoiser progressively according to specific iterative steps. However, a limitation still exists in controlling each attribute within entirely the same conditioning framework for every iteration, which results in a lack of controllability.

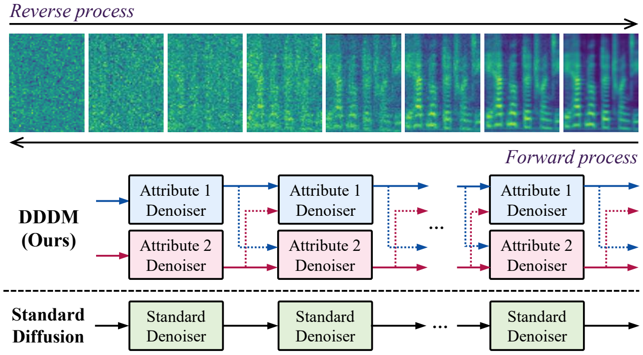


Figure 1: Speech synthesis in DDDM and standard diffusion model. Although a single denoiser with same parameter is used for all denoising steps in standard diffusion models, we subdivide the denoiser into multiple denoiser for each attribute by utilizing self-supervised representation. For each intermediate time step, each denoiser focuses on removing the single noise from its own attribute.

To address the above issues, we first present decoupled denoising diffusion models (DDDMs) with disentangled representations. As illustrated in Figure 1, we disentangle the denoiser into specific attribute-conditioned denoisers to improve the model controllability for each attribute. Subsequently, each denoiser focuses on the noise from its own attribute at the same noise level and removes the noise at each intermediate time step. To demonstrate the effectiveness of DDDMs, we focus on the VC tasks that still face challenges in disentangling and controlling each speech attribute (Choi et al. 2021). VC is a task for transferring or controlling

[*]These authors contributed equally.
[†]Corresponding author

the voice style while maintaining the linguistic information. As speech consists of various attributes such as linguistic information, intonation, and timbre, it remains challenging to transfer the voice style in zero/few-shot scenarios.

Based on the DDDMs, we present DDDM-VC which can effectively transfer and control the voice style for each attribute. We first utilize the self-supervised representation to disentangle the speech representation based on the source-filter theory (Fant 1970). Subsequently, we resynthesize the speech for each attribute from the disentangled representation using DDDMs. We also propose the prior mixup, a novel verified robust voice style transfer training scenario that uses the converted speech as a prior distribution for the diffusion model that is generated from the mixed speech representation, and restores the source speech. Thus, although DDDM-VC is trained by reconstructing the source speech, the prior mixup can reduce the train-inference mismatch problem for VC tasks. We demonstrate that DDDMs can effectively transfer the voice style even with lower model parameters compared to the state-of-the-art VC model (Popov et al. 2022). Furthermore, the experimental results reveal the effectiveness of speaker adaptation in the zero/one-shot scenarios. The main contributions of this study are as follows:

- We propose decoupled denoising diffusion models (DDDMs), which can effectively control the style for each attribute in generative models by decoupling attributes and adopting the disentangled denosiers.
- To demonstrate the effectiveness of DDDMs, We present DDDM-VC, which can disentangle and resynthesize speech for each attribute with self-supervised speech representation. Furthermore, we propose a prior mixup to improve voice style transfer performance.
- Our model provides better performance in both many-to-many and zero-shot voice style transfer compared with the state-of-the-art VC model. We can also successfully adapt to novel voice with a single sample.

2 Background

Denoising diffusion models have significantly improved various generative tasks such as image generation (Ramesh et al. 2022; Rombach et al. 2022), image inpainting (Saharia et al. 2022a; Lugmayr et al. 2022), and audio generation (Chen et al. 2021; Kong et al. 2021; Huang et al. 2022). These models typically consist of a forward process that gradually adds random noise, and a reverse process that progressively removes random noise and restores the original sample.

Unlike the original diffusion model that uses a discrete-time diffusion process by Markov chains (Ho, Jain, and Abbeel 2020), the score-based generative model uses a stochastic differential equation (SDE)-based continuous-time diffusion process (Song et al. 2021). The stochastic forward process is defined as follows:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)\mathrm{d}\mathbf{w} \ , \tag{1}$$

where $f(.,t)$ is the drift coefficient of the $\mathbf{x}(t)$, $g(\mathrm{t})$ is the diffusion coefficient, and $\mathbf{w}$ denote the Brownian motion. The reverse-time SDE can be expressed as:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\nabla_x \log p_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}} \ , \tag{2}$$

Method	EER (↓)	SECS (↑)
Source-Filter Encoder + GAN	10.25	0.831
Source-Filter Encoder + Diffusion	7.75	0.846

Table 1: Speaker adaptation results of GAN and diffusion

where $\bar{\mathbf{w}}$ is Brownian motion for the time flowing in backward, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ represents the score function. To estimate $\mathbf{s}_\theta(\mathbf{x}, t) \simeq \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, score-based diffusion model is trained with score matching objective:

$$\begin{aligned} \theta^* = \arg\min_\theta \mathbb{E}_t \Big\{ & \lambda(t) \mathbb{E}_{\mathbf{x}(\mathbf{0})} \mathbb{E}_{\mathbf{x}(t)|\mathbf{x}(\mathbf{0})} \\ & \left[\| s_\theta(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0)) \|_2^2 \right] \Big\} \ . \end{aligned} \tag{3}$$

Diffusion vs. GAN On the other hand, generative adversarial networks (GAN) have also shown a powerful generative performance in speech domain. (Lee et al. 2021b; Choi et al. 2021; Bak et al. 2023). Nevertheless, it is well known that there exists a trade-off between fidelity and diversity (Huang et al. 2023), producing high-quality samples but not covering the entire distribution (Dhariwal and Nichol 2021). In preliminary experiment, we compare each method (Choi et al. 2021; Popov et al. 2022) using the same encoder in Table 1. The results showed that the diffusion-based VC model has shown a better speaker adaptation performance than GAN-based VC models. In this regard, we chose diffusion-based VC model (Popov et al. 2022) as a baseline model.

3 Decoupled Denoising Diffusion Models

To effectively control the style for each attribute in generative models, we propose decoupled denoising diffusion models (DDDMs) with multiple disentangled denoisers. Although an ensemble of diffusion models was presented in (Balaji et al. 2022), only a single expert is used at the specific denoising step in this method. In contrast, we investigate the decomposition of diffusion models in a single denoising step. Specifically, more than one attribute denoiser is used at any given point. Unlike the general diffusion process, which employs a single denoiser, we subdivide the denoiser into N denoisers with disentangled representations. Following the use of data-driven priors in (Popov et al. 2022), we use a disentangled representation of an attribute Z_n as the prior for each attribute denoiser. Therefore, the forward process can be expressed:

$$dX_{n,t} = \frac{1}{2}\beta_t(Z_n - X_{n,t})dt + \sqrt{\beta_t}dW_t \ , \tag{4}$$

where $n \in [1, N]$, n denotes each attribute, N is the total number of attributes, β_t regulates the amount of stochastic noise and W_t is the forward Brownian motion. Reverse trajectories exist for the given forward SDE of each attribute (4). The reverse process of each disentangled denoiser can be defined as follows:

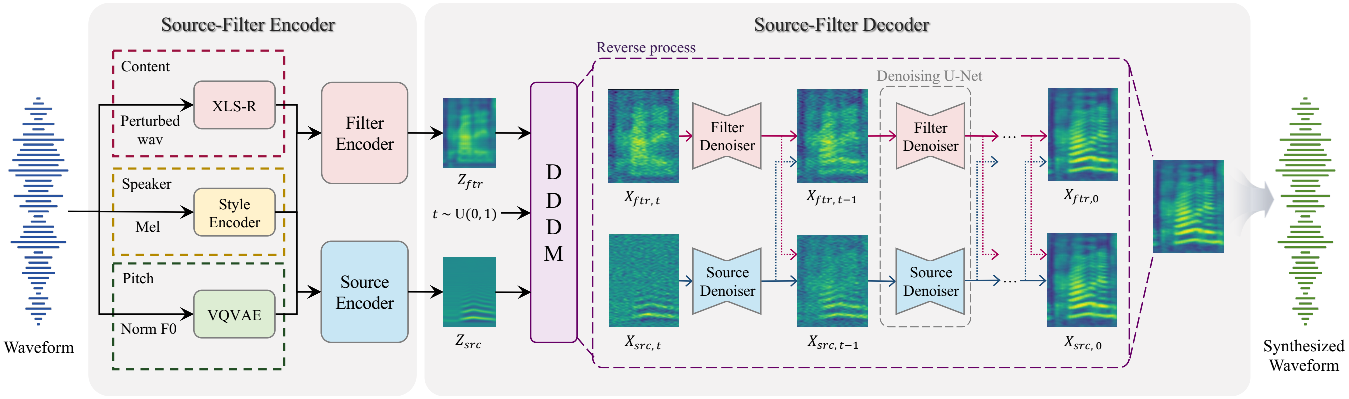


Figure 2: Overall framework of DDDM-VC.

$$d\hat{X}_{n,t} = \left(\frac{1}{2}(Z_n - \hat{X}_{n,t}) - \sum_{n=1}^{N} s_{\theta_n}(\hat{X}_{n,t}, Z_n, t)\right)\beta_t dt + \sqrt{\beta_t}\, d\bar{W}_t, \tag{5}$$

where $t \in [0,1]$, s_{θ_n} represents the score function of each attribute n parameterized by θ_n and $\bar{W}_t$ denotes the backward Brownian motion. The forward process (4) that generates a noisy sample $X_{n,t}$ with each prior attribute n is as follows:

$$p_{0t}(X_{n,t}|X_0) = \mathcal{N}\left(e^{-\frac{1}{2}\int_0^t \beta_s ds} X_0 + \left(1 - e^{-\frac{1}{2}\int_0^t \beta_s ds}\right) Z_n, \left(1 - e^{-\int_0^t \beta_s ds}\right) \mathrm{I}\right), \tag{6}$$

where I is the identity matrix. The distribution (6) is Gaussian, thus we have the following equation:

$$\nabla \log p_{0t}(X_{n,t}|X_0) = -\frac{X_{n,t} - X_0(e^{-\frac{1}{2}\int_0^t \beta_s ds}) - Z_n(1 - e^{-\frac{1}{2}\int_0^t \beta_s ds})}{1 - e^{-\int_0^t \beta_s ds}}. \tag{7}$$

The reverse process (5) is trained by optimizing the parameter θ_n using the following objective:

$$\theta_n^* = \underset{\theta_n}{\arg\min} \int_0^1 \lambda_t \mathbb{E}_{X_0, X_{n,t}} \| s_{\theta_n}(X_{n,t}, Z_n, t) - \nabla \log p_{0t}(X_{n,t}|X_0)\|_2^2 dt, \tag{8}$$

where $\theta = [\theta_1, \cdots, \theta_N]$ and $\lambda_t = 1 - e^{-\int_0^t \beta_s ds}$. Furthermore, we derive fast sampling using the ML-SDE solver (Popov et al. 2022), which maximizes the log-likelihood of forward diffusion with the reverse SDE solver. We extend DDDMs to DDDM-VC to control the voice style for each attribute in the following Section. In addition, we show that DDDMs can be applied to audio mixing by leveraging multiple denoisers to blend the sound and speech with the desired balance in Appendix H.

4 DDDM-VC

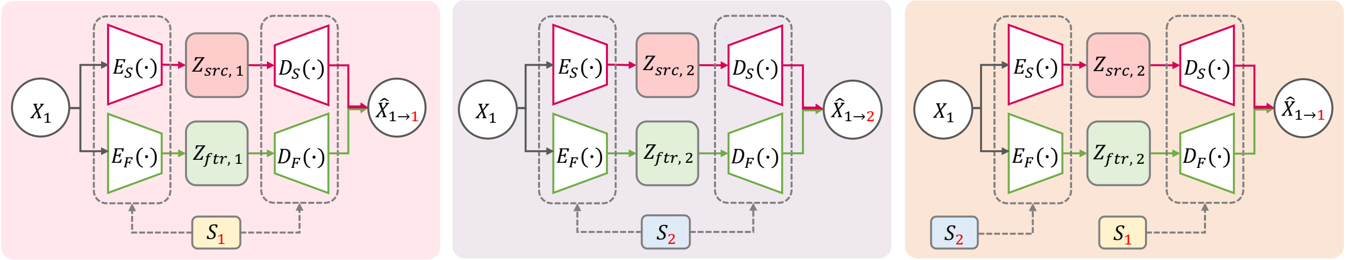
DDDM-VC consists of a source-filter encoder and source-filter decoder as illustrated in Figure 2. We first disentangle the speech using self-supervised speech representations as in subsection 4.1. Thereafter, we use these disentangled speech representations to control each attribute and to generate high-quality speech with the proposed disentangled denoiser as explained in subsection 4.2. Furthermore, we propose the prior mixup for a robust voice conversion scenario in subsection 4.3.

4.1 Speech Disentanglement

Content Representation To extract the content representation relating to the phonetic information, we utilize self-supervised speech representations. Unlike (Polyak et al. 2021) utilizing the discrete representation of audio from HuBERT or using language-dependent representation such as phonetic posteriorgram, we use a continuous representation of audio from XLS-R, which is Wav2Vec 2.0 trained with a large-scale cross-lingual speech dataset for robust zero-shot cross-lingual VC. Furthermore, before fed to the filter encoder, audio is perturbed to remove the content-independent information following (Choi et al. 2021). As (Lee et al. 2022b) demonstrated that the representation from the middle layer of XLS-R contains substantial linguistic information, we adopt this representation as the content representation.

Pitch Representation Following (Polyak et al. 2021), we extract the fundamental frequency (F0) from the audio using YAPPT algorithm (Kasi and Zahorian 2002) to encode the intonation such as the speaker-irrelevant speaking style. The F0 from each sample is normalized for each speaker for speaker-independent pitch information, and VQ-VAE is used to extract the vector-quantized pitch representation. For a fair comparison, we normalize the F0 for each sentence, not for a speaker, during inference.

Speaker Representation VC transfers the voice style, and our goal is to achieve robust zero-shot voice style transfer from novel speakers. To this end, we use style encoder (Min et al. 2021) that can extract the speaker representation from the Mel-spectrogram of the target speech. The extracted



(a) Resynthesis (b) Conversion (c) Prior Mixup

Figure 3: (a) Speech resynthesis from disentangled speech representations (training). (b) Voice conversion from converted speech representations (inference). (c) Prior mixup for better speaker adaptation quality. To reduce the train-inference mismatch problem, the decoder also learns to convert the randomly converted representations into input speech during training.

speaker representation is averaged per sentence for global speaker representation, and fed to all encoders and decoders for the speaker adaptation.

4.2 Speech Resynthesis

Source-filter Encoder In this work, we simply define the speech attributes according to the source-filter theory (Fant 1970). The filter encoder takes the content and speaker representations, whereas the source encoder takes the pitch and speaker representations. Previously, (Lee et al. 2022a) demonstrated that the data-driven prior in the diffusion process can simply guide the starting point of the reverse process. (Popov et al. 2022) adopted an average phoneme-level Mel encoder for voice conversion with a data-driven prior. However, this method requires a text transcript to extract the phoneme-level average Mel-spectrogram and pre-trained average Mel-encoder, and the smoothed Mel representation results in mispronunciation. To achieve a substantially more detailed prior, we use the entirely reconstructed source and filter Mel-spectrograms, Z_{src} and Z_{ftr} which are regularized by the target Mel-spectrogram X_{mel} as follows:

$$\mathcal{L}_{rec} = \|X_{mel} - (Z_{src} + Z_{ftr})\|_1, \tag{9}$$

where

$$Z_{src} = E_{src}(pitch, s), \;\; Z_{ftr} = E_{ftr}(content, s). \tag{10}$$

It is worth noting that the disentangled source and filter Mel-spectrograms from the disentangled representations are simply converted with different speaker representation s. Thus, we utilize the converted source and filter Mel-spectrogram as each prior in each denoiser for VC.

Source-filter Decoder We utilize disentangled denoisers for the source and filter representations based on our DDDMs. The source decoder takes a source representation Z_{src} as a prior and the filter decoder takes a filter representation Z_{ftr} as a prior. Subsequently, each denoiser is trained to generate a target Mel-spectrogram from each prior with the same noise, which is conditioned on a speaker representation. Each denoiser can focus on removing the single noise from its own attribute. The forward process is expressed as:

$$dX_{src,t} = \frac{1}{2}\beta_t(Z_{src} - X_{src,t})dt + \sqrt{\beta_t}dW_t, \tag{11}$$

$$dX_{ftr,t} = \frac{1}{2}\beta_t(Z_{ftr} - X_{ftr,t})dt + \sqrt{\beta_t}dW_t, \tag{12}$$

where $t \in [0,1]$, $X_{src,t}$ and $X_{ftr,t}$ are the generated noisy samples with each prior attribute (i.e., source-related and filter-related attribute respectively). For the given forward SDE of each attribute (11) and (12), there exist reverse trajectories. The reverse process is expressed as:

$$\begin{aligned} d\hat{X}_{src,t} = \bigg(&\frac{1}{2}(Z_{src} - \hat{X}_{src,t}) - \Big(s_{\theta_{src}}(\hat{X}_{src,t}, Z_{src}, s, t) \\ &+ s_{\theta_{ftr}}(\hat{X}_{ftr,t}, Z_{ftr}, s, t)\Big)\bigg)\beta_t dt + \sqrt{\beta_t}d\bar{W}_t, \end{aligned} \tag{13}$$

$$\begin{aligned} d\hat{X}_{ftr,t} = \bigg(&\frac{1}{2}(Z_{ftr} - \hat{X}_{ftr,t}) - \Big(s_{\theta_{ftr}}(\hat{X}_{ftr,t}, Z_{ftr}, s, t) \\ &+ s_{\theta_{src}}(\hat{X}_{src,t}, Z_{src}, s, t)\Big)\bigg)\beta_t dt + \sqrt{\beta_t}d\bar{W}_t, \end{aligned} \tag{14}$$

where $s_{\theta_{src}}$ and $s_{\theta_{ftr}}$ denote the score function parameterized by θ_{src} and θ_{ftr} respectively.

4.3 Prior Mixup

Although the speech can be disentangled into several attributes and resynthesized with high-quality using the self-supervised representation and diffusion processes, we still train the model by only reconstructing or using the input speech as the target speech in both the reconstruction and diffusion processes, which induces the train-inference mismatch problem. In non-parallel voice conversion scenario, the ground-truth of the converted speech does not exist; Thus, the model is trained only by reconstructing the source speech. However, as we convert the source speech with a different voice style for VC, we shift our focus from reconstruction to conversion even in the training scenario.

To achieve this, we propose a prior mixup in the diffusion process, which uses the randomly converted representation instead of the reconstructed representation as a prior distribution as illustrated in Figure 3-(c). Specifically, because the source-filter encoder can also be trained to reconstruct a source and filter of speech from the disentangled representation, the converted source and filter can be obtained with the randomly selected speaker style s_r as follows:

$$Z_{src,r} = E_{src}(pitch, s_r),\ Z_{ftr,r} = E_{ftr}(content, s_r). \tag{15}$$

Subsequently, the randomly converted source and filter, $Z_{src,r}$ and $Z_{ftr,r}$ are used as the prior for each denoiser as below:

$$dX_{src,t} = \frac{1}{2}\beta_t(Z_{src,r} - X_{src,t})dt + \sqrt{\beta_t}dW_t\ , \tag{16}$$

$$dX_{ftr,t} = \frac{1}{2}\beta_t(Z_{ftr,r} - X_{ftr,t})dt + \sqrt{\beta_t}dW_t\ . \tag{17}$$

The reverse process for the given forward SDE of each attribute (16) and (17) is expressed as:

$$\begin{aligned} d\hat{X}_{src,t} = \Big(&\frac{1}{2}(Z_{src,r} - \hat{X}_{src,t}) - s_{\theta_{src}}(\hat{X}_{src,t}, Z_{src,r}, s_o, t) \\ &- s_{\theta_{ftr}}(\hat{X}_{ftr,t}, Z_{ftr,r}, s_o, t)\Big)\beta_t dt + \sqrt{\beta_t}d\bar{W}_t, \end{aligned} \tag{18}$$

$$\begin{aligned} d\hat{X}_{ftr,t} = \Big(&\frac{1}{2}(Z_{ftr,r} - \hat{X}_{ftr,t}) - s_{\theta_{ftr}}(\hat{X}_{ftr,t}, Z_{ftr,r}, s_o, t) \\ &- s_{\theta_{src}}(\hat{X}_{src,t}, Z_{src,r}, s_o, t)\Big)\beta_t dt + \sqrt{\beta_t}d\bar{W}_t, \end{aligned} \tag{19}$$

where s_o is the original speaker style.

Hence, the prior mixup can alleviate the train-inference mismatch problem as the model is trained to convert the converted speech into the source speech even when reconstructing the source speech. Moreover, the voice style can be adapted in the source-filter decoder when the source-filter encoder may not execute VC effectively during inference. The entire model, including the style encoder, source-filter encoder, and decoder without pre-trained XLS-R and F0 VQ-VAE, is jointly trained in an end-to-end manner with Equation (8) for each attribute and Equation (9).

In order to verify that the training-inference mismatch can be resolved in the decoder, Table shows the conversion result using the reconstruction Mel (not converted Mel) as a prior that has not been converted from the encoder to the target. Without Prior Mixup, the data-driven prior restricts the function of the diffusion decoder as speech enhancement, which just enhances the audio quality. However, the diffusion decoder, which is trained with Prior Mixup, can also convert the voice style even with the wrong prior by conditioning the target voice style in the diffusion decoder.

5 Experiment and Result

5.1 Experimental Setup

Datasets We used the large-scale multi-speaker LibriTTS dataset (Zen et al. 2019) to train the model. The *train-clean-360* and *train-clean-100* of LibriTTS, which consist of 245 hours of audio samples for 1,151 speakers, were used for training. Thereafter, we evaluated VC performance on LibriTTS and VCTK dataset (Veaux et al. 2017) for many-to-many and zero-shot VC scenarios.

Preprocessing We resampled the audio from the sampling rate of 24,000 Hz to 16,000 Hz using the Kaiser-best algorithm of torchaudio Python package. We use the downsampled audio waveform as the input for XLS-R (0.3B) (Babu et al. 2022) to extract the self-supervised speech representation. For the target speech and the input of speaker encoder, we used log-scale Mel-spectrogram with 80 bins. To map the time frames between the self-supervised representation and Mel-spectrogram without any interpolation, Mel-spectrogram was transformed with hop size of 320, window size of 1280, and 1280-point Fourier transform.

Training For reproducibility, we attached the source code of DDDM-VC in the Supplementary materials. We trained DDDM-VC using the AdamW optimizer (Loshchilov and Hutter 2019) with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and weight decay $\lambda = 0.01$, and applied the learning rate schedule with a decay of $0.999^{1/8}$ at an initial learning rate of 5×10^{-5}. We train all models including ablation study with a batch size of 64 for 200 epochs. Architecture details are described in Appendix A. For prior mixup, we mixed the speaker representation using binary selection between the original and shuffled representations in the same batch. For zero-shot voice conversion, we did not fine-tune the model. For one-shot speaker adaptation, we fine-tuned the model with only one sentence of novel speakers for 500 steps with optimizer initialization and an initial learning rate of 2×10^{-5}. We used the pre-trained Vocoder to convert the Mel-spectrogram into waveform. For vocoder, we used HiFi-GAN V1 (Kong, Kim, and Bae 2020) as an generator, and we used multi-scale STFT-based discriminators (MS-STFTD) of EnCodec (Défossez et al. 2022) which use a complex-valued STFT with real and imaginary components.

5.2 Evaluation Metrics

Subjective Metrics We measured the mean opinion score (MOS) for the speech naturalness and speaker similarity in VC tasks. At least 20 listeners rated each sample from the source and converted speech on a scale of 1 to 5 for the speech naturalness MOS (nMOS). At least 20 listeners rated the target and converted speech on a scale of 1 to 4 for the speaker similarity MOS (sMOS).

Objective Metrics We calculated the character error rate (CER) and word error rate (WER) using Whisper (Radford et al. 2022) which is public available automatic speech recognition (ASR) model[1] with large-scale multi-lingual and multitask supervision for the content consistency measurement. We evaluated the equal error rate (EER) of automatic speaker verification (ASV) model (Kwon et al. 2021), which is trained with large-scale speech recognition dataset, VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) for the speaker similarity measurement. Furthermore, we determined the speaker encoder cosine similarity (SECS) for the additional similarity measurement. As VCTK provided a paired utterance per speaker, we also evaluated the Mel-cepstral

[1] https://github.com/openai/whisper. We used a large model of Whisper with 1,550M parameters, and used a presented text normalizer before calculating the CER and WER.

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	Params. (↓)	Real-time (↑)
GT	-	3.82±0.05	3.44±0.03	0.54	1.84	-	-	-	-
GT (Mel + Vocoder)	-	3.81±0.05	3.23±0.05	0.60	2.19	-	0.986	13M	-
AutoVC (Qian et al. 2019)	-	3.62±0.05	2.44±0.04	5.34	8.53	33.30	0.703	30M	×99.13
VoiceMixer (Lee et al. 2021a)	-	3.75±0.05	2.74±0.05	2.39	4.20	16.00	0.779	52M	×123.03
SR (Polyak et al. 2021)	-	3.62±0.05	2.55±0.04	6.63	11.72	33.30	0.693	15M	×177.22
DiffVC (Popov et al. 2022)	6	3.77±0.05	2.72±0.05	7.28	12.80	10.50	0.817	123M	× 20.06
DiffVC (Popov et al. 2022)	30	3.77±0.05	2.77±0.05	7.99	13.92	11.00	0.817	123M	×4.63
DDDM-VC-Small (Ours)	6	3.75±0.05	2.75±0.05	3.25	5.80	6.25	0.826	21M	×28.73
DDDM-VC-Small (Ours)	30	**3.79±0.05**	**2.81±0.05**	4.25	7.51	6.25	0.827	21M	×6.65
DDDM-VC-Base (Ours)	6	3.75±0.05	2.75±0.05	**1.75**	**4.09**	**4.00**	0.843	66M	×22.75
DDDM-VC-Base (Ours)	30	**3.79± 0.05**	2.80±0.05	2.60	5.32	4.24	**0.845**	66M	×5.09

Table 2: Many-to-many VC results on seen speakers from LibriTTS dataset

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	MCD$_{13}$ (↓)
GT	-	4.28±0.06	3.87±0.03	0.21	2.17	-	-	-
GT (Mel + Vocoder)	-	4.03±0.07	3.82±0.03	0.21	2.17	-	0.989	0.67
AutoVC (Qian et al. 2019)	-	2.49±0.09	1.88±0.08	5.14	10.55	37.32	0.715	5.01
VoiceMixer (Lee et al. 2021a)	-	3.43±0.08	2.63±0.08	1.08	3.31	20.75	0.797	4.49
SR (Polyak et al. 2021)	-	2.58±0.10	2.03±0.07	2.12	6.18	27.24	0.750	5.12
DiffVC (Popov et al. 2022)	6	3.48±0.07	2.62±0.08	5.82	11.76	25.30	0.786	4.82
DiffVC (Popov et al. 2022)	30	3.62±0.07	2.50±0.07	6.92	13.19	24.01	0.785	5.00
DDDM-VC-Small (Ours)	6	3.76±0.07	2.99±0.07	1.27	3.77	6.51	0.852	**4.39**
DDDM-VC-Small (Ours)	30	3.84±0.06	2.96±0.07	1.95	4.70	6.89	0.851	4.55
DDDM-VC-Base (Ours)	6	3.74±0.07	2.98±0.07	**1.00**	**3.49**	**6.25**	0.856	4.42
DDDM-VC-Base (Ours)	30	**3.88±0.06**	**3.05±0.07**	1.77	4.35	6.49	**0.858**	4.54
DDDM-VC-Fine-tuning (Ours)	6	3.74±0.07	**3.07±0.07**	1.26	3.80	**0.81**	**0.910**	**4.27**
DDDM-VC-Fine-tuning (Ours)	30	3.86±0.07	3.06±0.07	1.87	4.63	0.82	0.913	4.38

Table 3: Zero-shot VC results on unseen speakers from VCTK dataset. We additionally report the one-shot speaker adaptation result of DDDM-VC-Base model (DDDM-VC-Fine-tuing) which is fine-tuned with only single sample per speaker for 500 steps.

distortion (MCD). We produced all possible pairs from the converted and target speech (400×20 = 8,000), and calculated all the evaluation metrics.

5.3 Many-to-Many Voice Conversion

We performed the many-to-many VC task with seen speakers during the training, and compared our models with various VC models. As indicated in Table 2, DDDM-VC-Small also outperformed the other models in all subjective and objective metrics without ASR results. Although VoiceMixer had a lower CER and WER, it had a lower voice style transfer performance in terms of the EER and SECS. Furthermore, we compared the converted speech generated with 6 and 30 iterations to evaluate the performance with fast sampling. Although the objective results of the model with 6 iterations were better than those of the model with 30 iterations, the model with 30 iterations achieved better performance in both the nMOS and sMOS evaluations. Thus, the audio quality was perceptually improved and the generated samples had better diversity with the stochastic iterative processes.

5.4 Zero-shot Voice Conversion

We also report the results of the zero-shot VC tasks. As indicated in Table 3, our models significantly outperformed the baseline models in terms of speaker similarity. In particular, only the DDDM-VC models could adapt the voice style with novel speakers in terms of EER and SECS. We found that increasing iteration steps improved the diversity of converted speech in that CER, WER, and EER were increased, but the nMOS was consistently improved. We analyzed the effectiveness of each proposed component in the ablation study. In addition, we can control each attribute by transferring different styles to each attribute respectively as indicated in Appendix E.

5.5 One-shot Speaker Adaptation

For better speaker adaptation, we additionally fine-tuned our model on the VCTK dataset. We only used one sample per speaker, which is under ten seconds per speaker. As indicated in Table 3, the speaker similarity in terms of EER and SECS is consistently improved but the CER increased after the model overfitted the small training samples.

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	Params. (↓)
DDDM-VC-Small (Ours)	30	-	-	4.25	7.51	6.25	0.827	21M
DDDM-VC-Base (Ours)	30	3.76±0.05	3.08±0.05	2.60	5.32	4.24	0.845	66M
w/o Prior Mixup	30	3.79±0.05	3.03±0.05	3.28	5.66	7.99	0.821	66M
w/o Disentangled Denoiser	30	3.76±0.05	3.00±0.05	3.20	5.57	9.75	0.815	36M
w/o Normalized F0	30	3.78±0.05	3.00±0.05	3.27	5.88	10.25	0.811	33M
w/o Data-driven Prior	30	3.83±0.05	2.87±0.05	2.32	4.86	19.25	0.786	66M

Table 4: Results of ablation study on many-to-many VC tasks with seen speakers from LibriTTS.

Prior Mixup	Encoder Output (Prior)	EER (↓)	SECS (↑)
✓	Recon. Mel	48.34	0.677
✗	Recon. Mel	7.10	0.852

Table 5: Ablation study for Prior Mixup with wrong prior.

Params.	EER (↓)	SECS (↑)	CER (↓)	WER (↓)
36 M	8.78/7.78	0.847/0.852	0.55/0.84	2.92/3.26
170 M	9.00/7.32	0.843/0.851	0.58/0.70	2.84/3.05
340 M	10.25/8.50	0.840/0.844	0.66/0.88	3.04/3.29

Table 6: Objective evaluation of the impact of scaling up model parameters on the Prior Mixup (without/with). We train each model with LibriTTS-train-960 dataset and evaluate the zero-shot VC performance on VCTK dataset.

5.6 Ablation Study

Prior Mixup We trained the DDDM-VC model without the prior mixup to clarify the reduction in the train-inference mismatch. As indicated in Table 4, the prior mixup could improve the generalization performance with better speaker adaptation in that the EER of the model with the prior mixup decreased and the SECS increased. However, the naturalness was slightly decreased, which can occur in VC since it does not take into account the target rhythm on the fixed-length of input speech. The research on the rhythm conversion could address this issue and we leave it for the future work. To further verify that the trainging-inference mismatch can be resolved in the decoder with prior mixup, we compared the VC results using the reconstructed (not converted) Mel-spectrogram as a prior in Table 5. Without Prior Mixup, the data-driven prior restricts the function of the diffusion decoder as speech enhancement, which just enhances the audio quality. However, the diffusion decoder, which is trained with Prior Mixup, can also convert the voice style even with the wrong prior by conditioning the target voice style in the diffusion decoder. In other words, our proposed diffusion decoder performs more than mere enhancement; it facilitates robust style adaptation. We also analyze scaling up the VC system and how prior mixup could improve the generalization performance regardless of model size (without information bottleneck) in Table 6. Table 6 shows that the diffusion models without prior mixup have also troubles in scale-up in that the large-scale model could learn to estimate the random noise from the noised sample by ignoring the conditioning. Table 6 shows that scaling the diffusion-based model with prior mixup increases the performance of voice style transfer.

Disentangled Denoiser We observed that removing the disentangled denoiser (employing only a single denoiser) decreased the performance in all metrics. It indicates that the disentangled denoiser can improve the model performance by effectively adapting each representation to the target voice style, compared to a single denoiser.

Normalized F0 We determined that removing the normalized F0 conditioning decreases the VC performance. Without the pitch contour, the encoder may not disentangle the content information of the speech effectively, resulting in a degradation of the VC performance. As it is difficult to reconstruct the speech from the perturbed speech representation, the use of additional pitch information that can be extracted from the ground-truth speech may improve the stability of the model.

Data-driven Prior As noted in (Lee et al. 2022a), a data-driven prior can improve the performance of diffusion model. We minimize the L1 distance of Mel-spectrogram between the ground-truth Mel-spectrogram and output of the source-filter encoder as Equation (10) for the data-driven prior. Each output from the source and filter encoder was used for the prior of each diffusion model, which was disentangled by the source-filter theory. Although nMOS was reported slightly lower, the performance of speaker adaptation significantly increased with data-driven prior. In the VC tasks, using the converted Mel-spectrogram performs better than using the average Mel-spectrogram (Popov et al. 2022). Besides, we think that the enhanced prior through normalizing flow (Kim et al. 2020) may also improve the performance of models.

6 Conclusion

We have presented DDDMs for the robust control of various data components in diffusion models. We successfully demonstrated that DDDMs can improve the style transfer performance in VC tasks. DDDM-VC can convert the voice style even in zero-shot voice style transfer tasks by improving the speaker adaptation quality significantly. We have also proposed the prior mixup, which can improve the robustness of style control by learning to restore the data from converted representations for better generalization with reduced train-inference mismatch. Furthermore, we demonstrated that our model can robustly convert the voice with high-quality regardless of the model size. The small model also achieved better performance than state-of-the-art VC models.

7 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), No.2019-0-01371, Development of Brain-inspired AI with Human-like Intelligence, No.2021-0-02068, Artificial Intelligence Innovation Hub, No.2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation, and IITP-2024-No.RS-2023-00255968, the Artificial Intelligence Convergence Innovation Human Resources Development) and ESTsoft Corp., Seoul, Korea.

References

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

Bak, T.; Lee, J.; Bae, H.; Yang, J.; Bae, J.-S.; and Joo, Y.-S. 2023. Avocodo: Generative adversarial network for artifact-free vocoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12562–12570.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

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

Choi, H.-S.; Lee, J.; Kim, W.; Lee, J.; Heo, H.; and Lee, K. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34: 16251–16265.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*, 1086–1090.

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

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Fant, G. 1970. *Acoustic theory of speech production*. 2. Walter de Gruyter.

Han, S.; and Lee, J. 2022. NU-Wave 2: A General Neural Audio Upsampling Model for Various Sampling Rates. In *Interspeech*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, R.; Lam, M. W.; Wang, J.; Su, D.; Yu, D.; Ren, Y.; and Zhao, Z. 2022. 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*.

Huang, R.; Ren, Y.; Jiang, Z.; Cui, C.; Liu, J.; and Zhao, Z. 2023. FastDiff 2: Revisiting and Incorporating GANs and Diffusion Models in High-Fidelity Speech Synthesis. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6994–7009.

Kasi, K.; and Zahorian, S. A. 2002. Yet another algorithm for pitch tracking. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, I–361.

Kim, H.; Kim, S.; and Yoon, S. 2022a. Guided-TTS: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, 11119–11133.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33: 8067–8077.

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

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33: 17022–17033.

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

Kwon, Y.; Heo, H. S.; Lee, B.-J.; and Chung, J. S. 2021. 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)*.

Lee, S.-g.; Kim, H.; Shin, C.; Tan, X.; Liu, C.; Meng, Q.; Qin, T.; Chen, W.; Yoon, S.; and Liu, T.-Y. 2022a. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *International Conference on Learning Representations*.

Lee, S.-H.; Kim, J.-H.; Chung, H.; and Lee, S.-W. 2021a. VoiceMixer: Adversarial voice style mixup. *Advances in Neural Information Processing Systems*, 34: 294–308.

Lee, S.-H.; Kim, S.-B.; Lee, J.-H.; Song, E.; Hwang, M.-J.; and Lee, S.-W. 2022b. 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*.

Lee, S.-H.; Yoon, H.-W.; Noh, H.-R.; Kim, J.-H.; and Lee, S.-W. 2021b. 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*, volume 35, 13198–13206.

Liu, S.; Cao, Y.; Su, D.; and Meng, H. 2021. DiffSVC: A diffusion probabilistic model for singing voice conversion. In

2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 741–748.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Min, D.; Lee, D. B.; Yang, E.; and Hwang, S. J. 2021. Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, 7748–7759.

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

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, 8599–8608.

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

Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; and Hasegawa-Johnson, M. 2019. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, 5210–5219.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; et al. 2022b. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

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

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