DDPMVC: Non-parallel any-to-many voice conversion using diffusion encoder

Ryuichi Hatakeyama*,Kohei Okuda† and Toru Nakashika‡
* The University of Electro-Communications, Tokyo
E-mail: r.hatakeyama@uec.ac.jp
† The University of Electro-Communications, Tokyo
E-mail: okuda-k@dym.jp
‡ The University of Electro-Communications, Tokyo
E-mail: nakashika@uec.ac.jp

***Abstract*—In this paper, we propose DDPMVC, a voice conversion (VC) model for non-parallel data that utilizes the diffusion model. The distinctive feature of diffusion models lies in their high expressive power for high-dimensional data and their ability to learn more stably compared to traditional generative models. Although several VC methods using diffusion models have been proposed, only VoiceGrad meets the non-parallel and any-to-many condition, which makes it easy to handle in training and inference. DDPMVC is regarded as an extension of VoiceGrad that incorporates a rule-based diffusion process into the encoder. By using the encoder to convert speech into latent variables that have less speaker information, it is expected to improve the accuracy of the non-parallel VC. Experimental results demonstrated that the performance of DDPMVC surpassed that of VoiceGrad in terms of the mel cepstral distortion and speaker similarity.**

I. Introduction

Voice conversion (VC) is a technology that enables the transformation of a speaker identity into that of another without altering the speech content. It is categorized based on the type of data utilized: parallel data, which involves aligning the same spoken content across different speakers, and non-parallel data, which involves the arbitrary alignment of spoken content. Conversion targets include those from one specific speaker to another (`One-to-One`) and from specific multiple speakers to other specific multiple speakers (`Many-to-Many`). However, converting any speaker to specific multiple speakers (`Any-to-Many`) is deemed the ideal scenario.

Currently, research on VC predominantly utilizes deep learning models, with variational autoencoders (VAE) [1] and generative adversarial networks (GAN) [2] being notable examples. However, recent advancements have introduced diffusion models for voice conversion, demonstrating superior performance over traditional deep learning approaches such as VAE and GAN [3]–[5]. The diffusion-based VC [4] utilizes the continuous-time denoising diffusion probabilistic model (DDPM) [6], [7], distinguished by its use of an average voice extractor in the encoder. Similarly, DiffSVC [3] utilizes DDPM for VC, specifically targeting the conversion of singing voices. However, the diffusion-based VC requires parallel data for Many-to-Many VC models, and DiffSVC, while aimed at singing voices, does not support multiple speakers in Any-to-One conversions. In contrast, VoiceGrad [5] utilizes a score-based generative modeling (SBM) [8] within diffusion models for VC, achieving Any-to-Many conversions using non-parallel data. Non-parallel VC methods generally assume that the speech is generated by speaker-independent (phonetic) information accompanied with speaker information, and consist of an encoder that extracts phonetic information from speech and a decoder that generates speech from it given a target speaker's label [9]. VoiceGrad performs sampling without the use of an encoder. Consequently, it attempts the conversion to the target speaker without separating the speaker characteristics of the source speaker. This approach can result in suboptimal conversion accuracy, particularly when there are significant differences between the characteristics of the source and target speakers. Therefore, there is potential for improvement in the conversion accuracy of VoiceGrad under these conditions. In this study, we propose DDPMVC, a voice conversion model based on the diffusion model that incorporates a rule-based diffusion process into the encoder to extract speaker-independent information, building upon VoiceGrad's conversion process.Using a diffusion model for both the encoder and decoder in voice conversion simplifies the process and reduces training effort. We conducted a comparative analysis of the quality of conversion accuracy between DDPMVC and VoiceGrad.

II. Diffusion Model

A. Score-based generative modeling (SBM)

The score-based generative modeling (SBM) [8] utilizes scores to efficiently compute maximum likelihood estimates in high-dimensional data, addressing the challenges associated with these calculations. Given a likelihood $p(\boldsymbol{x})$ for an input variable $\boldsymbol{x}$, the gradient of the log likelihood $\log p(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, referred to as the score $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$, is denoted by the function $s(\boldsymbol{x})$. The common training approach for SBM is denoising score matching (DSM) [10], which focuses on learning scores related to the conditional probability during perturbation, rather than directly learning the scores. The

denoising score matching equation for SBM is as follows:

$$L_{\mathrm{SBM}}(\theta)=\sum_{l=1}^{L} w_l \mathbb{E}_{\boldsymbol{x},\tilde{\boldsymbol{x}}}\left[\left\|\frac{\boldsymbol{x}-\tilde{\boldsymbol{x}}}{\sigma_l^2}-\boldsymbol{s}_\theta(\tilde{\boldsymbol{x}},\sigma_l)\right\|^2\right], \quad (1)$$

where $\boldsymbol{x} \sim p(\boldsymbol{x})$, $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{x}, \sigma_l^2 \boldsymbol{I})$, and w_l represents the weight of the loss for each noise step. This approach utilizes perturbed distributions, disturbed by adding noise of various magnitudes to the data distribution, to calculate the score on these distributions. To align with Eq. (1), sampling is performed for each perturbed distribution using various noises through the Langevin Monte Carlo method. This is referred to as annealed Langevin dynamics [8].

B. Denoising diffusion probabilistic model (DDPM)

The denoising diffusion probabilistic model (DDPM) [7] consists of a diffusion process that gradually adds noise to the data $\boldsymbol{x}_0$ and a reverse diffusion process that traces back the diffusion process to remove the added noise. Viewing the noise-added data as latent variables, maximum likelihood estimation is performed as a latent variable model, and data sampling is conducted by sequentially sampling the latent variables through the reverse diffusion process. The diffusion process considers a Markov process that gradually adds noise to the data $\boldsymbol{x}_0$ and obtains a sequence of noise-added data $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$. Specifically, this is expressed by

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)=\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \quad (2)$$

where $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$ and $\{\beta_1, ..., \beta_T\}$ are noise parameters that control the variance magnitude. The reverse diffusion process is a generative process that starts from pure noise $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$ and gradually removes noise to generate data. Each step is normally distributed, and their means and variances are characterized by a model with parameters θ. This process is described by

$$p_\theta(\boldsymbol{x}_{0:T})=p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \quad (3)$$

where $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t))$. The parameters of this model are obtained by performing maximum likelihood estimation. The objective function is given by Eq. (4), where w_t represents the weight at each time step:

$$L(\theta)=\sum_{t=1}^{T} w_t \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)\|^2\right]. \quad (4)$$

With real data $\boldsymbol{x}_0$, $\boldsymbol{x}_{1:T}$ are considered latent variables. Since this involves maximum likelihood estimation with latent variables, it is performed by maximizing the variational lower bound (ELBO) [1] of the log-likelihood.

C. Continuous diffusion model

Song et al. (2021) [11] have extended discrete-time models (e.g., SBM and DDPM) to continuous time, defining them as stochastic differential equations (SDE). The continuous-time expression for SBM, eqrefeq:vesde, is called VE SDE, and the continuous-time expression for DDPM, eqrefeq:vpsde, is called VP SDE.

$$d\boldsymbol{x}=\sqrt{\frac{d[\sigma(t)^2]}{dt}}d\boldsymbol{w} \quad (5)$$

$$d\boldsymbol{x}=-\frac{1}{2}\beta(t)dt+\sqrt{\beta(t)}d\boldsymbol{w} \quad (6)$$

Here, $\boldsymbol{w}$ denotes the standard Wiener process or Brownian motion, and $d\boldsymbol{w}$ represents a normal distribution with a mean of 0 and a variance of τ in a small time τ. In continuous time, t is set to $t \in [0, 1]$, and the small time is treated as $\Delta t = 1/N$, taking the limit as $N \to \infty$. The difference between VE SDE and VP SDE lies in the method of adding noise. In VE SDE, only the noise increases while keeping the input the same, with the variance diverging. In VP SDE, the input gradually disappears, and the noise gradually increases to a certain magnitude, keeping the variance constant. According to [12], the SDEs for the reverse diffusion processes of SBM and DDPM are respectively given by Eqs. (7) and (8):

$$d\boldsymbol{x}=\left[-\frac{d[\sigma(t)^2]}{dt}\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})\right]dt+\sqrt{\frac{d[\sigma(t)^2]}{dt}}d\boldsymbol{w}, \quad (7)$$

$$d\boldsymbol{x}=\left[-\frac{1}{2}\beta(t)\boldsymbol{x}-\beta(t)\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})\right]dt+\sqrt{\beta(t)}d\boldsymbol{w}. \quad (8)$$

Learning utilizes the conditional probability of the diffusion process (diffusion kernel) $p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$. The diffusion kernel can also be expressed as a normal distribution [13]. The diffusion kernels for VE SDE and for VP SDE are respectively given by

$$p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))=\mathcal{N}(\boldsymbol{x}(t);\boldsymbol{x}(0),[\sigma(t)^2-\sigma(0)^2]\boldsymbol{I}), \quad (9)$$

$$p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))=\mathcal{N}(\boldsymbol{x}(t);\boldsymbol{x}(0)e^{-\frac{1}{2}\gamma(t)},\boldsymbol{I}-\boldsymbol{I}e^{-\gamma(t)}), \quad (10)$$

where $\gamma(t)=\int_0^t \beta(s)ds$. By estimating the score using this diffusion kernel, data can be sampled in the reverse diffusion process using the estimated score. The objective function is given by Eq. (11), where $\lambda(t)$ represents the weighting at each time step:

$$\begin{aligned}\theta^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathbb{E}_t \big\{ &\lambda(t)\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{\boldsymbol{x}(t)|\boldsymbol{x}(0)} \\ &\left[||\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}(t),t)-\nabla_{\boldsymbol{x}(t)}\log p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))||_2^2\right]\big\}.\end{aligned} \quad (11)$$

Sampling for the SDE is performed using the Euler-Maruyama method (EMM) [14] and predictor-corrector (PC) sampling [11]. The Euler-Maruyama method is a technique for obtaining samples by utilizing the score $\boldsymbol{s}_\theta(\boldsymbol{x}, t)$, estimated through score matching, in the first-order approximation of the reverse diffusion process (as in Eqs. (7) and (8)). Additionally, the PC sampling is a sampling technique composed of two parts: a predictor, which utilizes a numerical solver for the discretized reverse diffusion SDE, and a corrector, which

utilizes a score-based Markov Chain Monte Carlo (MCMC) method. The algorithm executes the predictor process N times, and for each execution, the corrector process is executed M times. The predictor involves the use of reverse diffusion samplers [11], which are the discretized equations of the reverse diffusion SDE. The corrector is executed to correct the results of the predictor. Techniques for the corrector include sampling by the Langevin Monte Carlo method and annealed Langevin dynamics.

III. Conventional Method: VoiceGrad

In this section, we briefly review VoiceGrad [5], which is related to our proposed method. VoiceGrad is a VC model based on SBM and DDPM. It based on SBM learns scores using Eq. (12) adapted for multiple speakers in DSM, and performs acoustic feature sampling through iterative updates of annealed Langevin dynamics (ALD) using the estimated scores. Additionally, It based on DDPM learns noise using Eq. (13) adapted for multiple speakers, and performs acoustic feature sampling through the reverse diffusion [5] using the estimated noise.

$$L_{DSM}(\theta) = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{k,\boldsymbol{x},\tilde{\boldsymbol{x}}}\left[\left\|\frac{\boldsymbol{x}-\tilde{\boldsymbol{x}}}{\sigma_l} - \sigma_l \boldsymbol{s}_\theta(\tilde{\boldsymbol{x}},\sigma_l,k)\right\|^2\right], \tag{12}$$

$$L_{DPM}(\theta) = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_l,l,k) - \boldsymbol{\epsilon}\right\|^2\right]. \tag{13}$$

where let $k \sim p(k), \boldsymbol{x} \sim p(\boldsymbol{x}|k), \tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{x}, \sigma^2\boldsymbol{I})$. The index k corresponds to the speaker, and the target speaker is conditioned by k. The acoustic feature used is the mel spectrogram. The key characteristic of this model is the absence of an encoder. It is trained on non-parallel data and supports Any-to-Many conversions. However, during sampling, since the input speech retains not only phonetic information but also speaker information, there is a possibility that the speaker characteristics of the source speech may remain after conversion.

IV. Proposed Method: DDPMVC

In this work, we propose a VC method that utilizes a rule-based diffusion process as an encoder (diffusion encoder) and reverse diffusion sampling as a decoder, inspired by the VC process of VoiceGrad. By employing a rule-based diffusion process in the encoder, it is expected that it will be possible to remove the speaker's information from the input speech, leaving only the phonetic information. The noise added during the diffusion process is intended to be mild enough to disrupt the fine structures that represent the speaker's information in the input speech without destroying the phonetic information. Traditional VC methods using diffusion models [3], [4] remove the speaker identity from the input speech using a different technique in the encoder, and then input the speech, which retains only phonetic information, into a decoder that uses a diffusion model to produce a voice converted to the target speaker's identity. On the other hand, DDPMVC integrates the encoder and decoder within a unified diffusion model framework, allowing for holistic optimization and thus, it is expected to improve conversion accuracy. Furthermore, by utilizing continuous-time diffusion models, it is believed that the discretization errors that occur in SBM and DDPM can be reduced, potentially leading to better conversion accuracy compared to VoiceGrad. Moreover, we anticipate that the use of continuous-time diffusion models will reduce the discretization errors encountered in SBM and DDPM.

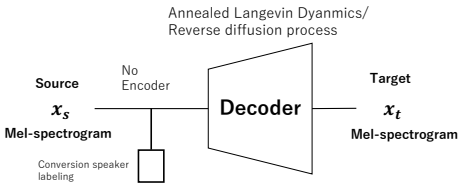

Fig. 1. Schematic diagram of VoiceGrad.

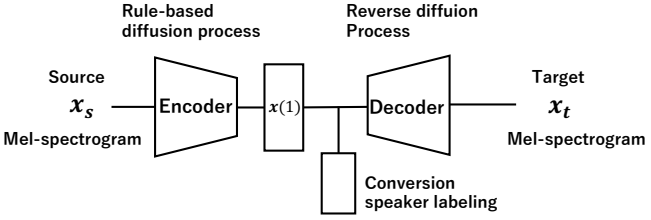

Fig. 2. Schematic diagram of DDPMVC.

The acoustic features in DDPMVC are 80-dimensional mel-spectrograms. The model is trained using these features, and during conversion, it outputs the acoustic features converted to the target speaker. Additionally, the mel-spectrogram $x_{d,m}$, where d is the d-th mel filter bank channel and m is the m-th frame index, is normalized using the mean ϕ_d and variance ξ_d as $x_{d,m} \leftarrow \frac{x_{d,m}-\phi_d}{\xi_d}$. The conversion process is shown in Fig. 2.

The mel-spectrograms $\boldsymbol{x}_s$ of the source speech is subjected to noise addition through a rule-based diffusion process (diffusion encoder) in the encoder. This involves adding noise to the extent that the content of the speech is somewhat preserved. The purpose of this is to minimize the possibility of incorrect exploration during sampling, thereby preventing alterations in the speech content as much as possible. The diffusion encoder utilizes the diffusion kernels given by Eqs. (9) and (10). The speaker-independent mel-spectrograms $\boldsymbol{x}(1)$, to which noise has been added, undergoes reverse diffusion sampling of DDPM, VE/VP SDE in the decoder, ultimately yielding the target speaker's mel-spectrograms $\boldsymbol{x}_t$. The objective function for training is defined by Eq. (13) for DDPM and by Eq. (11) for VE/VP SDE. Finally, the transformed acoustic features are utilized to generate speech using HiFi-GAN[15].

V. Experiments

A. Experiment setup

The dataset utilized for this study was the CMU Arctic voice corpus [16], which contains recordings of English-speaking individuals. Four speakers were selected for training, comprising

TABLE I
COMPARISON OF MEL CEPSTRAL DISTORTION (MCD) WITH 95% CONFIDENCE INTERVALS BETWEEN VOICEGRAD AND DDPMVC. "PREDICTOR" INDICATES THAT REVERSE DIFFUSION SAMPLERS ARE USED FOR SAMPLING.

Type	SBM-VG	DDPM-VG	DDPMVC				
			DDPM	Predictor		PC	
				VE SDE	VP SDE	VE SDE	VP SDE
M→M	7.42 ± 0.07	7.06 ± 0.08	6.50 ± 0.07	7.25 ± 0.08	6.50 ± 0.07	7.02 ± 0.06	$\mathbf{6.48 \pm 0.07}$
F→F	6.71 ± 0.09	6.19 ± 0.09	$\mathbf{5.86 \pm 0.08}$	6.40 ± 0.08	5.90 ± 0.09	6.64 ± 0.08	6.11 ± 0.09
M→F	7.62 ± 0.09	7.03 ± 0.11	6.59 ± 0.08	7.57 ± 0.10	6.60 ± 0.08	7.31 ± 0.07	$\mathbf{6.50 \pm 0.08}$
F→M	7.15 ± 0.07	6.76 ± 0.10	$\mathbf{6.24 \pm 0.07}$	6.99 ± 0.07	6.27 ± 0.08	6.90 ± 0.06	6.31 ± 0.07
All	7.28 ± 0.05	6.81 ± 0.06	$\mathbf{6.34 \pm 0.05}$	7.13 ± 0.06	6.36 ± 0.05	7.02 ± 0.04	6.37 ± 0.04

TABLE II
COMPARISON OF VOICE CONVERSION FOR UNKNOWN SPEAKERS.

Type	SBM-VG	DDPM-VG	DDPMVC				
			DDPM	Predictor		PC	
				VE SDE	VP SDE	VE SDE	VP SDE
M→M	7.14 ± 0.09	6.79 ± 0.10	$\mathbf{6.25 \pm 0.10}$	6.79 ± 0.10	6.33 ± 0.10	6.95 ± 0.08	6.45 ± 0.10
F→F	6.65 ± 0.08	6.00 ± 0.08	5.75 ± 0.07	6.24 ± 0.08	$\mathbf{5.74 \pm 0.08}$	6.53 ± 0.06	6.01 ± 0.08
M→F	7.40 ± 0.07	6.73 ± 0.08	$\mathbf{6.26 \pm 0.08}$	7.27 ± 0.08	6.30 ± 0.08	7.07 ± 0.07	6.38 ± 0.09
F→M	7.36 ± 0.08	6.91 ± 0.11	6.45 ± 0.09	7.15 ± 0.10	$\mathbf{6.44 \pm 0.10}$	7.05 ± 0.07	6.53 ± 0.09
All	7.14 ± 0.06	6.60 ± 0.06	$\mathbf{6.18 \pm 0.05}$	6.86 ± 0.07	6.20 ± 0.06	6.90 ± 0.05	6.34 ± 0.05

two male speakers (`aew`, `ksp`) and two female speakers (`clb`, `lnh`), with 128 utterances for training and 30 utterances for testing each. Additionally, to assess the accuracy of VC for arbitrary speakers, 30 utterances from one male speaker (`bdl`) and one female speaker (`slt`) were added for sampling. The audio data was downsampled to 16 kHz and converted into 32-dimensional mel cepstrums as acoustic features. The WORLD vocoder [17] was used for voice analysis and synthesis. The network architecture utilized for training was based on U-Net [18], consisting of approximately 44.72 million parameters.

B. Objective evaluation

To assess the effectiveness of DDPMVC's encoder and the continuous diffusion model, experiments were conducted in stages. First, we compared VoiceGrad using SBM (SBM-VoiceGrad) with VoiceGrad using DDPM (DDPM-VoiceGrad). The settings for each VoiceGrad model were based on [5], with SBM having 11 time steps ($L = 11$) and DDPM having 20 time steps ($L = 20$). Next, to verify the effectiveness of DDPMVC's diffusion encoder, we added it to DDPM-VoiceGrad. By directly comparing it with DDPM-VoiceGrad, we could assess the impact of the diffusion encoder. This additional model is also treated as DDPMVC by definition. It employed the diffusion process used in VP SDE's DDPMVC, with the noise schedule for VP SDE set to $\beta(t) \in [0.1, 2]$. The clarity of the speech after adding noise was evaluated using Short-Time Objective Intelligibility (STOI) [19]. It is an objective evaluation method that takes clean and noisy speech as input and outputs a score from 0.0 to 1.0, where a higher STOI indicates better intelligibility. The clarity of all the noisy test data showed a STOI of 0.52 ± 0.10, indicating that the noise added was weak enough to preserve some speech content. Finally, we conducted comparative experiments using DDPMVC. It was set to $N = 1000$ time steps, with the noise schedule for VE SDE set to $\sigma(t) \in [0.01, 0.85]$ and for VP SDE to $\beta(t) \in [0.1, 2]$. The sampling methods used were Reverse diffusion sampling and PC sampling. All these experiments used the trained speakers' speech as test data. For PC sampling, it used reverse diffusion samplers as the predictor and annealed Langevin dynamics as the corrector.

Additionally, experiments were conducted on unknown speakers who were not part of the training dataset. The specifications for SBM-VoiceGrad and DDPM-VoiceGrad remained as previously described. For DDPMVC, both VESDE and VPSDE were utilized, with Reverse diffusion sampling and PC sampling used as the sampling methods.

TABLE III
RESULTS OF SUBJECTIVE EVALUATION.

	MOS	DMOS
Ground Truth	4.39 ± 0.18	–
SBM-VG	1.95 ± 0.15	1.90 ± 0.13
DDPM-VG	$\mathbf{3.12 \pm 0.13}$	2.98 ± 0.18
DDPMVC	2.62 ± 0.15	$\mathbf{3.29 \pm 0.16}$

C. Subjective evaluation

In the subjective evaluation experiment, 13 participants conducted mean opinion score (MOS) tests to assess the naturalness of the converted speech and degradation mean opinion score (DMOS) tests to assess their similarity to the target speakers. The methods compared included Ground Truth, VoiceGrad using SBM and DDPM, and DDPMVC utilizing reverse diffusion samplers of VP SDE. For the evaluation of naturalness, 50 test questions were presented, comprising generated speech of the trained speakers from the three methods compared, along with the ground truth. The naturalness of the sound quality was rated on a 5-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). For the evaluation of speaker similarity, participants were given 50 test questions to assess the similarity between the trained speakers, converted using the three methods, and the target speakers. The similarity between the target and the converted speech was rated on a 5-point scale (1: Unlikely, 2: Not very likely, 3: Fairly likely, 4: Likely, 5: Definitely). Participants were asked to evaluate the results ignoring noise.

D. Results

Mel cepstral distortion (MCD) was used as an objective evaluation metric. The patterns of VC conducted were male-to-male (M→M), female-to-female (F→F), male-to-female (M→F), and female-to-male (F→M), forming four pairs. All results were calculated with a 95% confidence interval. First, when comparing SBM-VoiceGrad (SBM-VG) and DDPM-VoiceGrad (DDPM-VG), it can be seen from Table I that DDPM-VoiceGrad outperforms SBM-VoiceGrad in all conversion patterns. This indicates that DDPM achieves higher accuracy in voice conversion during the VoiceGrad transformation process. Next, the effectiveness of the diffusion encoder is examined. According to Table I, comparing DDPM-VoiceGrad with the version that includes a diffusion encoder (DDPMVC, DDPM) shows that the latter achieves higher accuracy in all conversion patterns. This demonstrates the effectiveness of the diffusion encoder utilized in DDPMVC. It is thought that the diffusion encoder helps to separate speaker characteristics to some extent, and by performing reverse diffusion from a state distanced from the original speaker, it becomes easier to approach the target speaker. Next, the results of DDPMVC with continuous diffusion model are reviewed. When compared to SBM-VoiceGrad, it is evident that DDPMVC outperforms SBM-VoiceGrad in all sampling methods. Compared to DDPM-VoiceGrad, DDPMVC utilizing VESDE showed lower accuracy, but DDPMVC utilizing VPSDE demonstrated higher accuracy. When compared to Enc+DDPM-VG, the DDPMVC with VPSDE utilizing PC sampling outperformed in conversions from male to male and male to female. However, it showed lower accuracy in female-to-female, female-to-male conversions, and overall evaluation, indicating that the effectiveness of the continuous diffusion model is limited.

The results for unknown speakers are presented in Table II. DDPM-based DDPMVC showed the best overall performance. Furthermore, considering DDPMVC as a whole, it is clear that it surpasses the performance of SBM-VoiceGrad. Compared to DDPM-VoiceGrad, DDPMVC with VPSDE and DDPMVC with DDPM showed higher accuracy in each sampling method. Although the overall MCD values are lower than those for known speakers, this is because the pre-conversion voices were already similar to the target voices, resulting in lower initial MCD values.

The results of the subjective evaluation are shown in Table III. In terms of naturalness evaluation (MOS), DDPMVC showed better naturalness compared to SBM-VoiceGrad but was inferior to DDPM-VoiceGrad. However, in the similarity evaluation (DMOS), which is more important than naturalness in VC, DDPMVC outperformed VoiceGrad, indicating that DDPMVC's voices are closer to the target voices. These results are presumably due to the fact that the converted speech by VoiceGrad remained closely similar to the input speech with minimal conversion, whereas the converted speech by DDPMVC underwent a degree of conversion. As a result, more artifacts occurred with DDPMVC during conversion, which leads to less naturalness in converted speech by DDPMVC than VoiceGrad. It is worth noting that some techniques used in state-of-the-art VC methods such as phoneme posteriorgram (PPG) [20], [21], self-supervised learning (SSL) [22]–[24] could improve the overall MOS; our aim is to compare the accuracy of mel-spectrogram conversion by DDPMVC with a baseline, VoiceGrad.

VI. Conclusions

In this paper, we proposed DDPMVC, an Any-to-Many voice conversion model for non-parallel data that utilizes a rule-based diffusion process as the encoder and a continuous-time diffusion model's reverse diffusion process as a decoder. Compared to the conventional method VoiceGrad, DDPMVC, especially with VPSDE, showed superior results in MCD evaluation. However, in terms of MOS evaluation, it resulted in an inferior performance compared to VoiceGrad. In the evaluation of DMOS, DDPMVC achieved closer scores to the target speech, suggesting that while the conversion became closer to the target speaker, this might have resulted in a reduction in naturalness. Going forward, we aim to improve the model such that it can enhance the naturalness of the converted speech while maintaining the similarity to the target speaker. Furthermore, by utilizing methods such as PPG to make the intermediate representations even more speaker-independent, it is expected that naturalness can be increased, leading to further improvements in accuracy.

Acknowledgment

This work was supported by JSPS KAKENHI under Grant 24H00715.

References

[1] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2022. arXiv: 1312.6114 [stat.ML].

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014.

[3] S. Liu, Y. Cao, D. Su, and H. Meng, "DiffSVC: A diffusion probabilistic model for singing voice conversion," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 741–748. DOI: 10.1109/ASRU51503.2021.9688219.

[4] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, *Diffusion-based voice conversion with fast maximum likelihood sampling scheme*, 2022. arXiv: 2109.13821 [cs.SD].

[5] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and S. Seki, *Voicegrad: Non-parallel any-to-many voice conversion with annealed langevin dynamics*, 2024. arXiv: 2010.02977 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2010.02977.

[6] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 2256–2265.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[8] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," *Interspeech*, pp. 5289–5293, 2017.

[10] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011. DOI: 10.1162/NECO_a_00142.

[11] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2021. arXiv: 2011.13456 [cs.LG].

[12] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982, ISSN: 0304-4149.

[13] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations* (Institute of Mathematical Statistics Textbooks). Cambridge University Press, 2019. DOI: 10.1017/9781108186735.

[14] X. Mao, "The truncated Euler–Maruyama method for stochastic differential equations," *Journal of Computational and Applied Mathematics*, vol. 290, pp. 370–384, 2015, ISSN: 0377-0427. DOI: https://doi.org/10.1016/j.cam.2015.06.002.

[15] J. Kong, J. Kim, and J. Bae, *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*, 2020. arXiv: 2010.05646 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2010.05646.

[16] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[17] M. Masanori, Y. Fumiya, and O. Kenji, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016. DOI: 10.1587/transinf.2015EDP7457.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217. DOI: 10.1109/ICASSP.2010.5495701.

[20] S.-w. Park, D.-y. Kim, and M.-c. Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data," *arXiv*, 2020. eprint: 2005.03295 (eess.AS).

[21] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6. DOI: 10.1109/ICME.2016.7552917.

[22] W.-C. Huang, Y.-C. Wu, T. Hayashi, and T. Toda, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," *arXiv*, 2020. eprint: 2010.12231 (eess.AS).

[23] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "FragmentVC: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," *arXiv*, 2021. eprint: 2010.14150 (eess.AS).

[24] J.-h. Lin, Y. Y. Lin, C.-M. Chien, and H.-y. Lee, "S2VC: A framework for any-to-any voice conversion with self-supervised pretrained representations," *arXiv*, 2021. eprint: 2104.02901 (eess.AS).