



Vocoder-Projected Feature Discriminator

Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Yuto Kondo

NTT Corporation, Japan

takuhiro.kaneko@ntt.com

Abstract

In text-to-speech (TTS) and voice conversion (VC), acoustic features, such as mel spectrograms, are typically used as synthesis or conversion targets owing to their compactness and ease of learning. However, because the ultimate goal is to generate high-quality waveforms, employing a vocoder to convert these features into waveforms and applying adversarial training in the time domain is reasonable. Nevertheless, upsampling the waveform introduces significant time and memory overheads. To address this issue, we propose a *vocoder-projected feature discriminator (VPFD)*, which uses vocoder features for adversarial training. Experiments on diffusion-based VC distillation demonstrated that a pretrained and frozen vocoder feature extractor with a single upsampling step is necessary and sufficient to achieve a VC performance comparable to that of waveform discriminators while reducing the training time and memory consumption by 9.6 and 11.4 times, respectively.¹

Index Terms: voice conversion, efficient training, generative adversarial networks, diffusion model, knowledge distillation

1. Introduction

Text-to-speech (TTS) and voice conversion (VC) are techniques designed to generate speech from text and speech inputs, respectively. In TTS and VC, a widely adopted approach is the two-stage framework, where the first model generates acoustic features (e.g., mel spectrograms) from input data (e.g., text or acoustic features) and the second model, called the vocoder, synthesizes the waveform from the generated acoustic features. Compared to an end-to-end approach, the two-stage approach offers advantages such as more compact learning and greater portability of individual modules, which have led to intensive research.

Realistic acoustic features must be generated to synthesize high-quality speech. Generative adversarial network (GAN) [1]-based adversarial training, where an acoustic feature generator is trained adversarially with a discriminator, has been widely adopted to achieve this objective (e.g., [2–12]). Specifically, considering that the ultimate goal is to synthesize high-quality waveforms, it is reasonable to focus on improving the quality of the generated acoustic features in the time domain. To achieve this objective, a previous study [12] proposed the vocoder waveform discriminator (VWD), which converts acoustic features into a waveform using a vocoder (e.g., [13–20]) and subsequently distinguishes between the real and synthesized waveforms using a waveform discriminator (e.g., [13, 15, 17, 21]), as shown in Figure 1(a). The effectiveness of VWD

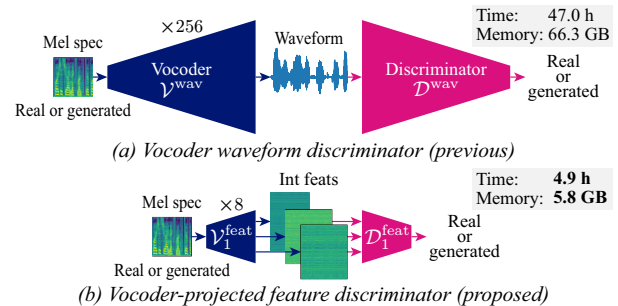


Figure 1: Comparison between (a) vocoder waveform discriminator (VWD, previous) and (b) vocoder-projected feature discriminator (VPFD, proposed). In VPFD, the discriminator is constructed using intermediate features (int feats) from a pre-trained vocoder, thereby bypassing the need to learn an effective representation for waveform synthesis while achieving faster training and reduced memory consumption.

has been demonstrated, for example, in diffusion-based VC distillation [12], where it enhances speech quality and speaker similarity while stabilizing GAN training—achievements that were challenging using typical mel-spectrogram discriminators. This can be attributed to the fact that mel-spectrogram discriminators must not only learn to distinguish between real and generated data but also capture a representation that reflects the reality of the waveform. VWD alleviates the latter requirement by employing a pretrained vocoder, thereby facilitating the aforementioned benefits. However, the limitation of this approach is that upsampling from acoustic features to the waveform, e.g., by 256 times, introduces time and memory overheads. Consequently, training is infeasible under conditions with limited time and computational resources.

To overcome this limitation, we propose a novel discriminator called the *vocoder-projected feature discriminator (VPFD)*, which distinguishes between real and generated data using the intermediate features of the vocoder rather than the waveform obtained from the output of the vocoder, as shown in Figure 1(b). The core idea is to reduce the computation time and memory consumption by utilizing intermediate features derived from fewer upsampling steps (e.g., 8 times) instead of the waveform obtained through full upsampling (e.g., 256 times).

The two key questions in this approach are as follows: *Q1. To what extent can upsampling be reduced?* *Q2. How should the vocoder feature extractor be handled during training?* In particular, regarding the latter, inspired by Projected GAN [22–24], we explored the importance of pretraining and freezing parameters. In our experiments, we addressed these questions by applying VPFD to diffusion-based VC distillation [12]. The results show that a pretrained and frozen vocoder

¹Audio samples are available at <https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/vpfd/>.

with a single upsampling is necessary and sufficient to achieve VC performance comparable to that obtained with VWD. By reducing upsampling, *VPFD* reduces the training time and memory consumption by factors of 9.6 and 11.4, respectively. Improving the quality of acoustic features is essential for various tasks. Therefore, the findings of this study are expected to have broader applications.

The remainder of this paper is organized as follows. Section 2 outlines the problem setting, describes the baseline VWD, and introduces the proposed *VPFD* and its application to diffusion-based VC distillation. Section 3 presents the experimental results. Finally, Section 4 concludes the paper and discusses potential future research.

2. Method

2.1. Problem setting

As discussed in Section 1, this study focused on the acoustic feature generator \mathcal{G} (first-stage model), which generates acoustic features \mathbf{x}^g from the input data \mathbf{z} (e.g., text or acoustic features) within the two-stage framework of TTS and VC. The goal is to enhance the quality of $\mathbf{x}^g = \mathcal{G}(\mathbf{z})$, ensuring that it is comparable to the real acoustic features \mathbf{x}^r in the training set. To achieve this, we employed GAN [1]-based adversarial training. In Sections 2.2 and 2.3, we discuss the components of the previous VWD [12] and proposed *VPFD*, both of which are designed to achieve this objective.

2.2. Preliminary: Vocoder waveform discriminator

In the above-mentioned problem, the ultimate goal is to generate high-quality waveforms; therefore, it is reasonable to improve the realism of acoustic features in the time domain. To achieve this objective, VWD [12] converts \mathbf{x}^r and \mathbf{x}^g into waveforms using a vocoder \mathcal{V}^{wav} (e.g., [13–20]) and distinguishes them using a waveform discriminator \mathcal{D}^{wav} (e.g., [13, 15, 17, 21]). The loss function (specifically, the least squares GAN form [25]) is expressed as

$$\mathcal{L}_D^{\text{VWD}} = \mathbb{E}_{\mathbf{x}^r} (\mathcal{D}^{\text{wav}}(\mathcal{V}^{\text{wav}}(\mathbf{x}^r)) - 1)^2 + \mathbb{E}_{\mathbf{z}} (\mathcal{D}^{\text{wav}}(\mathcal{V}^{\text{wav}}(\mathbf{x}^g)))^2, \quad (1)$$

$$\mathcal{L}_G^{\text{VWD}} = \mathbb{E}_{\mathbf{z}} (\mathcal{D}^{\text{wav}}(\mathcal{V}^{\text{wav}}(\mathbf{x}^g)) - 1)^2, \quad (2)$$

where \mathcal{D}^{wav} attempts to distinguish \mathbf{x}^r and \mathbf{x}^g in the time domain by minimizing $\mathcal{L}_D^{\text{VWD}}$, whereas \mathcal{G} attempts to generate \mathbf{x}^g that can deceive \mathcal{D}^{wav} in the time domain by minimizing $\mathcal{L}_G^{\text{VWD}}$. To stabilize the GAN training, feature matching (FM) loss [13, 26, 27] is also commonly used as follows:

$$\mathcal{L}_{\text{FM}}^{\text{VWD}} = \mathbb{E}_{\mathbf{x}^r, \mathbf{z}} \sum_{i=1}^{M^{\text{wav}}} \frac{1}{N_i^{\text{wav}}} \|\mathcal{D}_i^{\text{wav}}(\mathcal{V}^{\text{wav}}(\mathbf{x}^r)) - \mathcal{D}_i^{\text{wav}}(\mathcal{V}^{\text{wav}}(\mathbf{x}^g))\|_1, \quad (3)$$

where M^{wav} is the number of layers in \mathcal{D}^{wav} , and $\mathcal{D}_i^{\text{wav}}$ and N_i^{wav} denote the features and number of features in the i -th layer of \mathcal{D}^{wav} , respectively. This loss encourages \mathbf{x}^g to be closer to \mathbf{x}^r in the feature space of \mathcal{D}^{wav} .

2.3. Proposal: Vocoder-projected feature discriminator

As discussed in Section 1, VWD improves the quality of the synthesized waveform and stabilizes GAN training. However, upsampling from acoustic features to the waveform (e.g., 256 times) results in increased training time and memory consumption. To overcome this limitation while retaining the advantages

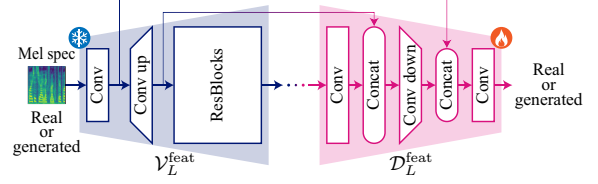


Figure 2: Network architecture of vocoder-projected feature discriminator (*VPFD*). *VPFD* is composed of $\mathcal{V}_L^{\text{feat}}$ and $\mathcal{D}_L^{\text{feat}}$. *Conv*, *ResBlocks*, and *Concat* indicate a convolution layer, residual blocks [28], and a channel concatenation operation, respectively.

of VWD, that is, bypassing the need to learn an effective representation for waveform synthesis, we propose *VPFD*, which distinguishes between \mathbf{x}^r and \mathbf{x}^g using the intermediate features of a pretrained vocoder \mathcal{V}^{wav} . Let the intermediate feature extractor up to the L -th upsampling layer of the vocoder be denoted by $\mathcal{V}_L^{\text{feat}}$, and let the discriminator that performs discrimination using these features be denoted by $\mathcal{D}_L^{\text{feat}}$. The more the upsampling steps are reduced, the more the training time and memory consumption can be reduced. The necessary number of upsampling steps, i.e., the setting of L , is experimentally evaluated in Section 3.2. The training objective and network architecture of *VPFD* are detailed below.

Training objective. In *VPFD*, the adversarial loss (Equations 1 and 2) and FM loss (Equation 3) are rewritten with the replacement of the modules as follows:

$$\mathcal{L}_D^{\text{VPFD}_L} = \mathbb{E}_{\mathbf{x}^r} (\mathcal{D}_L^{\text{feat}}(\mathcal{V}_L^{\text{feat}}(\mathbf{x}^r)) - 1)^2 + \mathbb{E}_{\mathbf{z}} (\mathcal{D}_L^{\text{feat}}(\mathcal{V}_L^{\text{feat}}(\mathbf{x}^g)))^2, \quad (4)$$

$$\mathcal{L}_G^{\text{VPFD}_L} = \mathbb{E}_{\mathbf{z}} (\mathcal{D}_L^{\text{feat}}(\mathcal{V}_L^{\text{feat}}(\mathbf{x}^g)) - 1)^2, \quad (5)$$

$$\mathcal{L}_{\text{FM}}^{\text{VPFD}_L} = \mathbb{E}_{\mathbf{x}^r, \mathbf{z}} \sum_{i=1}^{M_L^{\text{feat}}} \frac{1}{N_{L,i}^{\text{feat}}} \|\mathcal{D}_{L,i}^{\text{feat}}(\mathcal{V}_L^{\text{feat}}(\mathbf{x}^r)) - \mathcal{D}_{L,i}^{\text{feat}}(\mathcal{V}_L^{\text{feat}}(\mathbf{x}^g))\|_1, \quad (6)$$

where M_L^{feat} is the number of layers in $\mathcal{D}_L^{\text{feat}}$, and $\mathcal{D}_{L,i}^{\text{feat}}$ and $N_{L,i}^{\text{feat}}$ denote the features and number of features in the i -th layer of $\mathcal{D}_L^{\text{feat}}$, respectively.

Network architecture. Based on the finding for Projected GAN [22–24] that multiscale feature extraction is effective, we employ a network architecture based on U-Net [29], as shown in Figure 2.² In $\mathcal{D}_L^{\text{feat}}$, downsampling is performed at the same rate as upsampling in $\mathcal{V}_L^{\text{feat}}$. During downsampling, the kernel size is set to twice the downsampling rate, while in other cases, it is set to 21. The number of output channels is set to equal the number of layers in $\mathcal{V}_L^{\text{feat}}$ at the same scale. A leaky rectified linear unit [30] is used as the activation function, and weight normalization [31] is applied to all the convolution layers. In the default setting, $\mathcal{V}_L^{\text{feat}}$ is pretrained and its parameters are fixed during training. In contrast, the parameters of $\mathcal{D}_L^{\text{feat}}$ are optimized during training. The effectiveness of this strategy is experimentally validated in Section 3.2.

2.4. Application to distillation of diffusion-based VC

In the experiments (Section 3), we validated *VPFD* by applying it to the adversarial distillation of diffusion-based VC [12].

²More specifically, we employ an “inverted” U-Net. While the original U-Net [29] utilizes a structure that performs downsampling followed by upsampling, it should be noted that in this case, we use a structure that performs upsampling followed by downsampling.

3. Experiments

3.1. Experimental setup

In this distillation, VoiceGrad [32], a denoising diffusion probabilistic model (DDPM) [33]-based nonparallel VC, is distilled into a one-step diffusion-based VC (FastVoiceGrad; FVG) by leveraging both GANs [1] and diffusion models [34]. We evaluated the performance when VWD, which was originally used in FVG, was replaced with *VPFD*. In this section, we first provide a brief overview of VoiceGrad, followed by a description of FVG configured using *VPFD*.

VoiceGrad. VoiceGrad is a nonparallel VC model that generates \mathbf{x}^g from \mathbf{x}^r , conditioned on a speaker embedding \mathbf{s} [35] and content embedding \mathbf{p} [36], using diffusion and reverse diffusion processes. During training, reconstruction is performed using \mathbf{x}^r , \mathbf{s} , and \mathbf{p} extracted from the speech of the same speaker, whereas during inference, conversion is performed using \mathbf{x}^r and \mathbf{p} extracted from the source speaker’s speech and \mathbf{s} extracted from the target speaker’s speech. The diffusion and reverse diffusion processes are described below.

Diffusion process. During this process, \mathbf{x}^r ($= \mathbf{x}_0$) is gradually transformed into noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ over T steps (where $T = 1000$ in practice). Owing to the reproductivity of the normal distribution and a reparameterization trick [37], the t -step diffused data \mathbf{x}_t ($t \in \{1, \dots, T\}$) are expressed as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad (7)$$

where $\bar{\alpha} = \prod_{i=1}^t \alpha_i$, $1 - \alpha_i$ represents the noise variance at the i -th step, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse diffusion process. During this process, \mathbf{x}_t is gradually denoised towards \mathbf{x}_0 . This denoising process is given by

$$\boldsymbol{\mu}_\theta(\mathbf{x}, t, \mathbf{s}, \mathbf{p}) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{s}, \mathbf{p}) \right), \quad (8)$$

where $\boldsymbol{\epsilon}_\theta$ is a denoising function parameterized by θ .

FVG with VPFD. In FVG, a pretrained VoiceGrad model is used as an initial condition, and its denoising process (i.e., Equation 8) is distilled such that its output remains feasible even when this process is performed only once. This is achieved by distilling the model using adversarial loss [1, 25] and score distillation loss [38]. For clarity, we denote the parameters of the student model (i.e., FVG) and teacher model (i.e., VoiceGrad) by ϕ and θ , respectively.

Adversarial loss. The adversarial loss (including the FM loss) configured with *VPFD* is defined by Equations 4–6, where \mathbf{x}^g is generated by $\boldsymbol{\mu}_\phi$ (calculated as in Equation 8) and $\mathbf{z} = \mathbf{x}^r$. We denote this \mathbf{x}^g as \mathbf{x}_ϕ^g .

Score distillation loss. The score distillation loss is expressed as

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{t, \mathbf{x}^r} \sqrt{\bar{\alpha}_t} \|\mathbf{x}_\phi^g - \mathbf{x}_\theta^g\|, \quad (9)$$

where $\mathbf{x}_\theta^g = \boldsymbol{\mu}_\theta(\text{sg}(\mathbf{x}_{\phi,t}^g), t, \mathbf{s}, \mathbf{p})$, sg represents the stop gradient operation, $\mathbf{x}_{\phi,t}^g$ is the t -step diffused \mathbf{x}_ϕ^g , and $t \in \{1, \dots, T\}$. This loss encourages \mathbf{x}_ϕ^g to match the output obtained by diffusing and then reversely diffusing it using the teacher model $\boldsymbol{\mu}_\theta$.

Total objective. The total objective function is expressed as

$$\mathcal{L}_{\mathcal{D}} = \mathcal{L}_{\mathcal{D}}^{\text{VPFD}_L}, \quad (10)$$

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\mathcal{G}}^{\text{VPFD}_L} + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}}^{\text{VPFD}_L} + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}}, \quad (11)$$

where λ_{FM} and λ_{distill} are weighting hyperparameters, set to 2 and 45, respectively, in the experiments, following the FVG study [12]. $\mathcal{D}_L^{\text{feat}}$ is optimized by minimizing $\mathcal{L}_{\mathcal{D}}$, whereas \mathcal{G} is optimized by minimizing $\mathcal{L}_{\mathcal{G}}$.

Data. The experimental setup related to the data followed that of the FVG study [12]. We evaluated *VPFD* for one-shot any-to-any VC. The main experiments (Sections 3.2 and 3.3) were conducted using the VCTK dataset [39], which includes utterances from 110 English speakers. Generalizability, independent of the dataset, was evaluated using the LibriTTS dataset [40], which contains utterances from approximately 1,100 English speakers (Section 3.4). For the unseen-to-unseen VC scenario, 10 speakers and 10 sentences were excluded for evaluation, while the remaining data were used for training. The audio clips were downsampled to 22.05 kHz, from which 80-dimensional log-mel spectrograms were extracted with an FFT size of 1024, hop size of 256, and window size of 1024. These log-mel spectrograms were used as conversion targets.

Implementation. For a fair comparison, we implemented our model (*FVG+VPFD_L*) based on FVG [12], with the only modification being the replacement of the discriminator from VWD with *VPFD_L*. Both $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\mu}_\phi$ were implemented using the U-Net [29] architecture, consisting of 12 convolution layers with 512 hidden channels, two downsampling/upsampling processes, gated linear units [41], and weight normalization [31]. \mathbf{s} was extracted using a speaker encoder [35], while \mathbf{p} was obtained using a bottleneck feature extractor [36]. For \mathcal{V} , we utilized a pretrained HiFi-GAN V1 generator [15].³ The baseline VWD used \mathcal{D}^{wav} , which combines the multiperiod discriminator (MPD) [15] and multiresolution discriminator (MRD) [17]. In contrast, the proposed *VPFD* used $\mathcal{D}^{\text{feat}}$, implemented as shown in Figure 2. The models were trained using the Adam optimizer [42] with a batch size of 32, learning rate of 0.0002, β_1 of 0.5, and β_2 of 0.9 for 100 epochs on VCTK and 50 epochs on LibriTTS. The waveforms were synthesized using \mathcal{V} .

Evaluation metrics. To efficiently investigate various models, we primarily employed objective metrics and conducted a subjective evaluation for the most critical comparison (Table 4). The following six objective metrics were used: (1) *UTMOS* \uparrow [43]: the predicted mean opinion score (MOS) is tuned to assess the quality of synthesized speech. (2) *DNSMOS* \uparrow [44]: the predicted MOS is optimized to evaluate the quality of noise-suppressed speech. (3) Character error rate (*CER* \downarrow) by Whisper-large-v3 [45]: measures speech intelligibility. (4) Speaker encoder speaker similarity (*SECS* \uparrow) by Resemblyzer⁴: quantifies speaker similarity. (5) *Time* \downarrow : training time (in hours) measured on a single A100 GPU. (6) *Memory* \downarrow : maximum memory consumption (in gigabytes) during training. For each metric, \uparrow indicates that a larger value is better, while \downarrow indicates that a smaller value is better.

3.2. Ablation study

Q1. To what extent can upsampling be reduced? Initially, we investigated the required number of upsampling steps L . The results are presented in Table 1. We found that *FVG+VPFD₀* significantly degrades *DNSMOS* and *SECS* because $\mathcal{V}_0^{\text{feat}}$ consists of only a single convolution layer, which is inadequate for extracting effective features. In contrast, *FVG+VPFD_L* with $L \geq 1$ achieved *UTMOS*, *DNSMOS*, *CER*, and *SECS* comparable to those of *FVG*. Notably, *FVG+VPFD₁* achieved this while reducing the training time and memory consumption by

³<https://github.com/jik876/hifi-gan>

⁴<https://github.com/resemble-ai/Resemblyzer>

factors of 9.6 and 11.4, respectively. To investigate why one upsampling is sufficient, we visualized the vocoder features in Figure 3. As shown in (c), periodic structures, which are crucial for waveform representation but absent in the mel-spectrogram (a) and features without upsampling (b), emerge with a single upsampling, suggesting its effectiveness. Based on these findings, we set $L = 1$ for subsequent experiments.

Table 1: Comparison of performance with varying numbers of upsampling steps. As the color transitions from red to white to blue, the scores worsen.

Model	UTMOS \uparrow	DNSMOS \uparrow	CER \downarrow	SECS \uparrow	Time \downarrow	Memory \downarrow
FVG	3.96	3.77	1.3	0.847	47.0	66.3
FVG+VPFD ₀	3.98	3.66	1.3	0.843	1.5	2.8
FVG+VPFD ₁	3.99	3.79	1.2	0.851	4.9	5.8
FVG+VPFD ₂	3.99	3.79	1.2	0.850	14.1	17.9
FVG+VPFD ₃	3.98	3.78	1.2	0.849	23.0	31.1
FVG+VPFD ₄	3.98	3.78	1.3	0.849	31.8	44.3

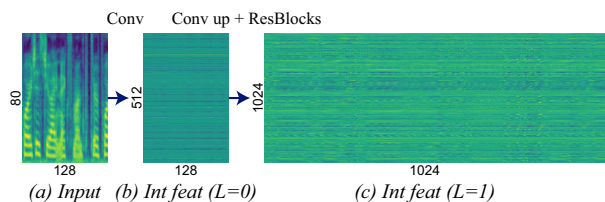


Figure 3: Comparison of input and intermediate features. **Best viewed when zoomed in.** The aspect is adjusted for clarity.

Q2. How should the vocoder feature extractor $\mathcal{V}_1^{\text{feat}}$ be handled during training? Next, we investigated the handling of $\mathcal{V}_1^{\text{feat}}$ during the training. In particular, we examined the effects of pretraining and freezing parameters. The results in Table 2 demonstrate that both strategies are crucial for all metrics. Based on these findings, we employed pretrained and frozen $\mathcal{V}_1^{\text{feat}}$ in subsequent experiments.

Table 2: Analysis of the importance of pretraining and freezing the vocoder feature extractor.

Pretrained	Frozen	UTMOS \uparrow	DNSMOS \uparrow	CER \downarrow	SECS \uparrow	Time \downarrow	Memory \downarrow
		3.62	3.54	1.3	0.814	4.9	5.8
✓		3.74	3.64	1.3	0.837	4.9	5.8
	✓	3.97	3.68	1.3	0.843	4.9	5.8
✓	✓	3.99	3.79	1.2	0.851	4.9	5.8

3.3. Comparative study

Comparison with other training acceleration techniques. This study accelerated training by reducing the upsampling steps in the discriminator. To evaluate this approach, we compared it with several alternatives: (1) *FVG_{early}*, which reduces the number of training epochs from 100 to 10, thereby matching the training time of FVG+VPFD₁. (2) *FVG w/o MRD*, which ablates MRD in VWD. (3) *FVG w/o MPD*, which ablates MPD in VWD. (4) *FVG+MelD_{small}*, which replaces VWD with a mel-spectrogram discriminator (MelD) with an architecture similar to that of MRD. (5) *FVG+MelD_{large}*, which increases the number of channels in *FVG+MelD_{small}* to match the training time of FVG+VPFD₁. The results in Table 3 indicate the following: (1) The performance decreases as the number of training epochs decreases. (2) and (3) Ablating the discriminator in VWD degrades the performance and does not significantly reduce the training time. (4) and (5) *FVG+MelD* fails to achieve a high DNSMOS, regardless of the model size. These results indicate that *FVG+VPFD₁* is most effective for reducing training time while maintaining VC performance.

Table 3: Comparison with other training acceleration techniques. As the color transitions from red to white to blue, the scores worsen.

Model	UTMOS \uparrow	DNSMOS \uparrow	CER \downarrow	SECS \uparrow	Time \downarrow	Memory \downarrow
FVG	3.96	3.77	1.3	0.847	47.0	66.3
(1) FVG _{early}	3.82	3.72	1.3	0.843	4.7	66.3
(2) FVG w/o MRD	3.95	3.75	1.2	0.845	32.1	43.6
(3) FVG w/o MPD	3.87	3.70	1.3	0.847	40.3	56.7
(4) FVG+MelD _{small}	3.98	3.66	1.3	0.845	1.9	3.1
(5) FVG+MelD _{large}	3.99	3.68	1.3	0.845	4.9	6.1
FVG+VPFD ₁	3.99	3.79	1.2	0.851	4.9	5.8

Subjective evaluation. For a comprehensive evaluation, we conducted MOS tests for 90 different speaker/sentence pairs to evaluate speech quality (*qMOS* on a five-point scale: 1 = bad, 2 = poor, 3 = fair, 4 = good, and 5 = excellent) and speaker similarity (*sMOS* on a four-point scale: 1 = different (sure), 2 = different (not sure), 3 = same (not sure), and 4 = same (sure)). In both tests, we compared *FVG+VPFD₁* with *FVG* to assess the effect of replacing VWD with *VPFD*. Ground-truth speech and speech converted using DiffVC-30 [46], a widely used baseline, were included as anchor samples. For each test, more than 1,000 responses were collected from 11 participants. As shown in Table 4, *FVG+VPFD₁* achieved performance comparable to *FVG* in both objective and subjective evaluations.

Table 4: Subjective evaluations with 95% confidence intervals. Objective scores are also included as a reference.

Model	qMOS \uparrow	sMOS \uparrow	UTMOS \uparrow	DNSMOS \uparrow	CER \downarrow	SECS \uparrow
Ground truth	4.43 \pm 0.08	3.57 \pm 0.08	4.14	3.75	0.1	0.871
DiffVC-30	3.60 \pm 0.10	2.37 \pm 0.12	3.76	3.75	5.4	0.802
FVG	3.61 \pm 0.09	2.64 \pm 0.12	3.96	3.77	1.3	0.847
FVG+VPFD ₁	3.63 \pm 0.10	2.69 \pm 0.12	3.99	3.79	1.2	0.851

3.4. Generalizability analysis

To investigate dataset dependency, we also conducted experiments using the LibriTTS dataset [40]. The results presented in Table 5 demonstrate a similar tendency as in the VCTK dataset; that is, *FVG+VPFD₁* achieves VC performance comparable to that of *FVG* while reducing both the training time and memory consumption by factors of 9.6 and 11.4, respectively.

Table 5: Results on LibriTTS dataset. \dagger Ground-truth converted speech does not necessarily exist in LibriTTS. Therefore, source speech was used to calculate the scores.

Model	UTMOS \uparrow	DNSMOS \uparrow	CER \downarrow	SECS \uparrow	Time \downarrow	Memory \downarrow
Ground truth \dagger	4.06	3.70	0.6	–	–	–
FVG	3.94	3.75	1.2	0.843	176.3	66.3
FVG+VPFD ₁	4.06	3.76	1.1	0.847	18.4	5.8

4. Conclusion

We proposed *VPFD*, which uses vocoder features for adversarial training. The results show that a pretrained and frozen vocoder feature extractor with a single upsampling step is necessary and sufficient to achieve comparable VC performance while significantly reducing training time and memory usage. Given the broad use of adversarial training with acoustic features in TTS and VC, we see potential for applying this approach to other tasks in future research.

5. References

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.
- [2] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *ICASSP*, 2017.
- [3] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *Interspeech*, 2017.
- [4] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 84–96, 2017.
- [5] 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," in *Interspeech*, 2017.
- [6] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *EUSIPCO*, 2018.
- [7] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *SLT*, 2018.
- [8] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion," in *Interspeech*, 2019.
- [9] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Non-parallel voice conversion with augmented classifier star generative adversarial networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2982–2995, 2020.
- [10] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "MaskCycleGAN-VC: Learning non-parallel voice conversion with filling in frames," in *ICASSP*, 2021.
- [11] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Interspeech*, 2021.
- [12] T. Kaneko, H. Kameoka, K. Tanaka, and Y. Kondo, "FastVoiceGrad: One-step diffusion-based voice conversion with adversarial conditional diffusion distillation," in *Interspeech*, 2024.
- [13] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS*, 2019.
- [14] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, 2020.
- [15] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.
- [16] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in *SLT*, 2021.
- [17] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Interspeech*, 2021.
- [18] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform," in *ICASSP*, 2022.
- [19] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *ICLR*, 2023.
- [20] T. Kaneko, H. Kameoka, K. Tanaka, and S. Seki, "iSTFTNet2: Faster and more lightweight iSTFT-based neural vocoder using 1D-2D CNN," in *Interspeech*, 2023.
- [21] —, "Wave-U-Net Discriminator: Fast and lightweight discriminator for generative adversarial network-based speech synthesis," in *ICASSP*, 2023.
- [22] A. Sauer, K. Chitta, J. Müller, and A. Geiger, "Projected GANs converge faster," in *NeurIPS*, 2021.
- [23] A. Sauer, K. Schwarz, and A. Geiger, "StyleGAN-XL: Scaling StyleGAN to large diverse datasets," in *SIGGRAPH*, 2022.
- [24] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis," in *ICML*, 2023.
- [25] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.
- [26] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *ICML*, 2016.
- [27] K. Oyamada, H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and H. Ando, "Generative adversarial network-based approach to signal reconstruction from magnitude spectrogram," in *EUSIPCO*, 2018.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.
- [30] A. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop*, 2013.
- [31] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *NIPS*, 2016.
- [32] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and S. Seki, "VoiceGrad: Non-parallel any-to-many voice conversion with annealed langevin dynamics," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 2213–2226, 2024.
- [33] J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," in *NeurIPS*, 2020.
- [34] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.
- [35] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.
- [36] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1717–1728, 2021.
- [37] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *ICLR*, 2014.
- [38] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *ECCV*, 2024.
- [39] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.
- [40] 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," in *Interspeech*, 2019.
- [41] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017.
- [42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [43] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022," in *Interspeech*, 2022.
- [44] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2021.
- [45] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.
- [46] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *ICLR*, 2022.