BRIDGEVOC: INSIGHTS INTO USING SCHRÖDINGER BRIDGE FOR NEURAL VOCODERS

Tong Lei & Rilin Chen & Meng Yu & Dong Yu
Tencent AI Lab
Beijing, China; Shenzhen, China; Bellevue, WA, USA
{fayelei, rilinchen}@tencent.com
{raymondmyu, dyu}@global.tencent.com

Andong Li & Chengshi Zheng
Key Laboratory of Noise and Vibration Research
Institute of Acoustics Chinese Academy of Sciences
Beijing, China
{liandong, cszheng}@mail.ioa.ac.cn

Tong Lei & Jing Lu
Key Laboratory of Modern Acoustics
Nanjing University, Nanjing, China
{tonglei@smail, lujing@}nju.edu.cn

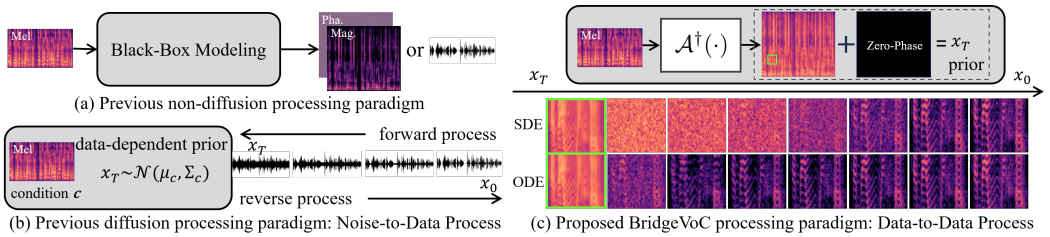
ABSTRACT

While previous diffusion-based neural vocoders typically follow a noise-to-data generation pipe-line, the linear-degradation prior of the mel-spectrogram is often neglected, resulting in limited generation quality. By revisiting the vocoder task and excavating its connection with the signal restoration task, this paper proposes a novel time-frequency (T-F) domain-based neural vocoder with the Schrödinger Bridge, called **BridgeVoC**, which is the first to follow the data-to-data generation paradigm. Specifically, the mel-spectrogram can be projected into the target linear-scale domain and regarded as a degraded spectral representation with a deficient rank distribution. Based on this, the Schrödinger Bridge is leveraged to establish a connection between the degraded and target data distributions. During the inference stage, starting from the degraded representation, the target spectrum can be gradually restored rather than generated from a Gaussian noise process. We conduct extensive experiments on the LJSpeech and LibriTTS benchmarks. Quantitative and qualitative results demonstrate that the proposed method enjoys faster inference speed and outperforms existing diffusion-based vocoder baselines, while also achieving competitive or better performance compared to other non-diffusion state-of-the-art methods across multiple evaluation metrics.

1 INTRODUCTION

Neural vocoders are essential for generating high-quality waveforms from acoustic features, playing a crucial role in speech and audio generation tasks such as text-to-speech (TTS) (Wang et al., 2017; Ren et al., 2020; 2019; Tan et al., 2024), text-to-audio (TTA) (Huang et al., 2023; Majumder et al., 2024), singing voice synthesis (SVS) (Liu et al., 2022c; Hwang et al., 2025), voice conversion (Qian et al., 2019; Choi et al., 2021), audio editing (Wang et al., 2023), and speech enhancement (SE) (Liu et al., 2022a;b). The core challenge lies in their ability to faithfully reconstruct high-fidelity waveforms while maintaining computational efficiency - a dual objective that continues to drive research innovation in this field.

The evolution of vocoding techniques has been significantly accelerated by deep neural networks (DNNs). Early auto-regressive (AR) approaches like WaveNet (Dieleman et al., 2016; Oord et al., 2018), SampleRNN (Mehri et al., 2022), and LPCNet (Valin & Skoglund, 2019) achieved remarkable quality but suffered from inherent latency due to sequential generation. Flow-based vocoder methods, such as WaveGlow (Prenger et al., 2019), FlowWaveNet (Kim et al., 2019), and RealNVP (Laurent et al., 2017), address these issues by enabling faster generation speeds and improved performance through bijective mappings between a normalized probability distribution and the target data distribution using stacked invertible modules. Non-autoregressive (NAR) methods like HiFiGAN (Kong et al., 2020) have emerged, offering parallel processing and enhanced efficiency.



(a) Previous non-diffusion processing paradigm

(b) Previous diffusion processing paradigm: Noise-to-Data Process

(c) Proposed BridgeVoC processing paradigm: Data-to-Data Process

Figure 1: Illustrations of the various neural vocoder paradigms.

A paradigm shift emerged with time-frequency (T-F) domain approaches that leverage spectral processing in the STFT domain. Methods such as BigVGAN (Lee et al., 2023), Vocos (Hubert, 2024), and APNet2 (Du et al., 2024) demonstrated superior inference speeds by directly estimating spectral components (magnitude and phase) followed by iSTFT reconstruction, effectively decoupling temporal resolution challenges from neural network processing. As illustrated in Figure 1(a), these frameworks typically employ hybrid architectures combining learned spectral transformations with deterministic signal processing components.

Recent advances in generative modeling have introduced diffusion-based vocoders that trade computational efficiency for exceptional audio naturalness. WaveGrad is a conditional waveform generation model that refines white Gaussian noise into high-fidelity audio using a gradient-based sampler conditioned on the mel-spectrogram, effectively balancing the inference speed and sample quality (Chen et al., 2021). DiffWave is a non-autoregressive diffusion probabilistic model that efficiently converts white Gaussian noise into high-fidelity audio through a Markov chain by optimizing a variational bound on data likelihood. It provides a significantly smaller model size and computational resource requirement compared to WaveGrad while excelling in unconditional generation tasks (Kong et al., 2021). Unlike DiffWave, which uses a standard Gaussian prior, PriorGrad employs an adaptive prior based on data statistics, resulting in faster convergence and improved perceptual quality (Lee et al., 2022). Compared with PriorGrad and DiffWave, FreGrad enjoys significantly faster training and inference speeds, and a smaller model size, by operating on a simplified feature space and incorporating frequency-aware components (Nguyen et al., 2024). The processing paradigm of diffusion-based vocoders is fundamentally illustrated in Figure 1(b). Using the mel-spectrogram or other acoustic features as the condition, these vocoders usually start from a random Gaussian distribution and gradually approximate the target distribution by iterative denoising process, which essentially follows a **noise-to-data** pipeline.

In this work, we revisit the neural vocoding task and introduce the Schrödinger Bridge (SB) framework to establish a **data-to-data** process between the target spectrogram in the T-F domain and a corrupted spectrogram. This approach is formulated from a general restoration perspective, rather than the conventional generative paradigm, as illustrated in Figure 1(c). Specifically, mel-spectrograms—derived through a linear-to-mel transform—are projected back to the linear-scale domain using their pseudo-inverse (Lv et al., 2024), based on the range-null decomposition (RND) theory. This projection provides robust structural information about the target spectrogram, enabling a principled approach to vocoding. The core objective of our vocoding framework is to reconstruct ground-truth spectrograms from mel-spectrograms, addressing two critical challenges: spectral compression and phase reconstruction. Through rank analysis, we observe that the mel-domain conversion and reversion process tends to reduce the spectral rank, necessitating that the neural vocoding task increase the spectral rank to restore clean speech. This stands in contrast to speech denoising tasks, which exhibit the opposite trend. This insight establishes a novel connection between waveform generation and speech restoration techniques, offering a unified perspective that bridges these traditionally distinct domains. To further enhance generation quality, we incorporate advanced discriminative components, including the multi-period discriminator (MPD) (Kong et al., 2020) and the multi-resolution spectrogram discriminator (MRSD) (Won et al., 2021). These discriminators operate at multiple temporal and spectral resolutions, ensuring fine-grained perceptual quality and improved fidelity in the synthesized waveforms. The contributions of this paper are summarized as follows:

- BridgeVoC is the first T-F domain-based vocoder with the Schrödinger Bridge (SB) framework, exploring a data-to-data process rather than the conventional noise-to-data process in the previous literature.

- BridgeVoC introduces a novel perspective on bridging waveform generation and restoration, a connection not investigated in the preliminary literature.
- By integrating the SB framework with multi-mel losses and a generative adversarial network (GAN), BridgeVoC achieves performance comparable to the state-of-the-art model BigVGAN, addressing the limitations of diffusion models in achieving excellent objective metrics.

2 MOTIVATION

In this section, we start with the fundamental signal models to elucidate how we transition from the conditional mel-to-waveform paradigm to the spectrum-to-spectrum restoration paradigm. Firstly, through the RND theory, a novel insight is provided to convert the mel-spectrogram back to degraded counterpart in the linear-scale spectrogram. Subsequently, rank analysis reveal contrasting rank trends between vocoding and denoising tasks. This observation inspired us to apply restoration methods commonly used in speech enhancement to the vocoding task.

2.1 SIGNAL MODELS

The signal model of the speech denoising task in the T-F domain is represented as:

$$X_{t,f} = S_{t,f} + N_{t,f}, \tag{1}$$

where $\{X, S, N\} \in \mathbb{C}^{T\times F}$ denote the mixture, target, and noise signals, respectively. The subscripts $t \in \{1, ..., T\}$ and $f \in \{1, ..., F\}$ represent the time and frequency indices, respectively.

For the vocoder task, mel-spectrograms $Y^{mel} \in \mathbb{R}^{T\times F_{mel}}$ are obtained through the signal model

$$Y^{mel} = |S|\mathcal{A}, \tag{2}$$

where $\mathcal{A} \in \mathbb{R}^{F\times F_{mel}}$ denotes the linear mel filter. F_{mel} is the mel size and typically satisfies $F_{mel} \ll F$ for a compressed representation. The transform indicates that 1) the phase part is discarded, and 2) a linear compression is applied in the frequency dimension.

2.2 RANGE-NULL SPACE DECOMPOSITION

For a classical signal compression physical model in the noise-free scenario, the target $\mathbf{x} \in \mathbb{R}^D$ and the observed signals $\mathbf{y} \in \mathbb{R}^d$ can be simplified into $\mathbf{y} = \mathbf{A}\mathbf{x}$. If the pseudo-inverse of $\mathbf{A} \in \mathbb{R}^{d\times D}$ is defined as $\mathbf{A}^\dagger \in \mathbb{R}^{D\times d}$, which satisfies $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} \equiv \mathbf{A}$ and $d \ll D$, then the signal $\mathbf{x}$ can be decomposed into two orthogonal sub-spaces:

$$\mathbf{x} \equiv \mathbf{A}^\dagger\mathbf{A}\mathbf{x} + \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x}, \tag{3}$$

where $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$ defines the range-space component and $\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x}$ corresponds to the remaining null-space component. By comparing Eq. (2) and Eq. (3), we notice the mel-spectrogram can be converted into the range space, *i.e.*, the first term on the right-hand side of the equal sign in Eq. (3), by left-multiplying the pseudo-inverse of $\mathcal{A}$, *i.e.*, $\mathcal{A}^\dagger$. Since the null-space component is unknown in practice, the vocoder task can be formulated into the target estimation problem given the range-space component as the prior input, which is actually a classical signal recovery problem. Thanks to the powerful capability of the generative approach, we can effectively recover the remaining null-space component. Therefore, the RND theory provides us a different perspective to rethink the vocoder task. Recall that in the classical compressive sensing (CS) field (Zhang & Ghanem, 2018), a similar target is shared, where the target signal can be recovered from a linearly-compressed representation with the help of the structural sparseness prior. In the following part, we delve into the analysis from the perspective of the matrix rank.

2.3 RANK ANALYSIS

Following the RND, we use the pseudo-inverse to map mel-spectrograms back to the original linear-scale domain, despite imperfections due to information loss, non-unique inverse mapping, approximation limitations, and lack of phase information(Meinard, 2015). This process is formulated as

$$\hat{Y} = Y^{mel}\mathcal{A}^\dagger = |S|\mathcal{A}\mathcal{A}^\dagger, \tag{4}$$

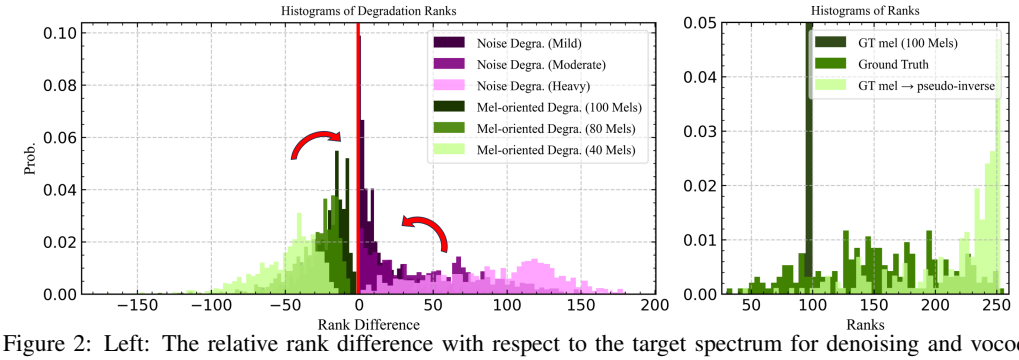


Figure 2: Left: The relative rank difference with respect to the target spectrum for denoising and vocoding tasks. The ranks are calculated from the test set of the VoiceBank-DEMAND dataset. An absolute threshold η of 0.5 is set for rank calculation; Right: The ranks of mel- and pseudo-inverse spectrograms on the dev-clean and dev-other subsets of the LibriTTS dataset, with the upper limit of the y-axis truncated for clarity.

where $\mathcal{A}^\dagger \in \mathbb{R}^{F_m \times F}$ is the pseudo-inverse transform matrix satisfying $\mathcal{A}\mathcal{A}^\dagger\mathcal{A} \equiv \mathcal{A}$. The linear-scale representation $\hat{Y} \in \mathbb{R}^{T \times F}$ matches the feature dimensions of the target signals S. By appending a zero-phase component to $\hat{Y}$, we can obtain its complex form $S^\dagger \in \mathbb{C}^{T \times F}$:

$$S^\dagger = \hat{Y} + i \cdot \mathbf{0}, \tag{5}$$

where $\mathbf{0} \in \mathbb{R}^{T \times F}$ is a zero matrix. Mapping $S^\dagger$ to S is a restoration problem akin to speech denoising, differing in that denoising involves additive degradation and can increase the spectral rank, while vocoding involves the signal compression and thus decreases the spectral rank. We demonstrate these spectral rank changes with proofs, defining $\mathcal{R}(\cdot) : \mathbb{R}^{T \times F} \to \mathbb{Z}$ as the matrix rank operation. Using fundamental matrix rank properties, we have

$$\mathcal{R}(|X|) \approx \mathcal{R}(|S| + |N|) \leq \mathcal{R}(|S|) + \mathcal{R}(|N|), \tag{6}$$

$$\mathcal{R}(\hat{Y}) = \mathcal{R}(|S|\mathcal{A}\mathcal{A}^\dagger) \leq \min\{\mathcal{R}(|S|), \mathcal{R}(\mathcal{A}\mathcal{A}^\dagger)\}. \tag{7}$$

In Eqs. (6)-(7), the phase component is omitted, as the rank is associated with eigenvalues, which are more closely related to signal energy. Eq. (6) provides an upper bound on the rank of the mixture spectrum X. This implies that after adding noise N, the upper bound of the matrix rank tends to increase, and the stronger the noise, the higher the upper bound. For Eq. (7), it is deduced that with the decrease in the number of mel bands, *i.e.*, $\mathcal{R}(\mathcal{A}\mathcal{A}^\dagger)$ decreases, the rank $\mathcal{R}(\hat{Y})$ tends to decrease. These two disparities in the rank distribution between noise-induced and mel-oriented degradations are visualized in Figure 2, where we define the rank difference between the degraded and target spectrum as

$$\Delta\mathcal{R}^{denoising} = \mathcal{R}(|X|) - \mathcal{R}(|S|), \tag{8}$$

$$\Delta\mathcal{R}^{vocoding} = \mathcal{R}(\hat{Y}) - \mathcal{R}(|S|). \tag{9}$$

The noise degradation employs three levels: “mild”, “moderate”, and “heavy” with decreasing signal-to-noise ratios (SNRs). For vocoding, we use three mel-band configurations (40, 80, and 100) to represent varying spectral compression. An STFT operation results in 257-dimensional features. Higher noise level has higher spectral rank and hinders sparsity, while higher mel-band compression leads to a negative rank difference. Therefore, from the perspective of the matrix rank, the vocoder and speech enhancement can share a similar goal, *i.e.*, decrease the rank difference between the degraded and target spectra, further motivating us to address the vocoder task with the restoration paradigm.

3 BridgeVoC

In this section, we introduce BridgeVoC, an SB-based T-F domain vocoder. We begin with a brief overview of the most commonly used diffusion models, specifically score-based generative models (SGMs), including the forward and reverse stochastic differential equations (SDE) and the score matching objective of the score network. Then we define the paired data for the restoration task based on the signal model described in Section 2.3. Next, we detail the specific operations of SB and the model’s training objectives and provide a comprehensive description of the loss functions employed during training. However, the description of **SGMs** and the **loss functions** is provided in the **Appendix** due to space limitations and conventional usage.

3.1 SCHRÖDINGER BRIDGE

The SB problem (Schrödinger, 1932; Bortoli et al., 2021) originates from the optimization of path measures with constrained boundaries. For vocoder task, we define the target distribution p_{S} to be equal to the data distribution p_{data}, and we consider the distribution of $S^{\dagger}$, denoted as $p_{\mathrm{S}^{\dagger}}$, to be the prior distribution. Considering p_0, p_T the marginal distributions of p at boundaries, SB is defined as minimization of the Kullback-Leibler (KL) divergence:

$$\min_{p \in \mathcal{P}_{[0,T]}} D_{\mathrm{KL}}(p \parallel p_{\mathrm{ref}}), \quad s.t.\ p_0 = p_{\mathrm{S}},\ p_T = p_{\mathrm{S}^{\dagger}}, \tag{10}$$

where $\mathcal{P}_{[0,T]}$ is the space of path measures on a finite time index $[0, T]$ with p_{ref} the reference path measure. When p_{ref} is defined by the same form of forward SDE as SGMs in Eq. (15), the SB problem is equivalent to a couple of forward-backward SDEs (Wang et al., 2021; Chen et al., 2022):

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{f}(\boldsymbol{x}_t, t) + g^2(t)\nabla \log \Psi_t(\boldsymbol{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t, \quad \boldsymbol{x}_0 \sim p_{\mathrm{S}}, \tag{11}$$

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{f}(\boldsymbol{x}_t, t) - g^2(t)\nabla \log \widehat{\Psi}_t(\boldsymbol{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\boldsymbol{w}}_t, \quad \boldsymbol{x}_T \sim p_{\mathrm{S}^{\dagger}}, \tag{12}$$

where $\boldsymbol{f}$, g and $\boldsymbol{w}_t$ are from the forward SDE in Eq. (15). With Ψ_t and $\widehat{\Psi}_t$ the optimal forward and reverse drifts, the marginal distribution of the SB state $\boldsymbol{x}_t$ can be expressed as $p_t = \widehat{\Psi}_t \Psi_t$. Typically, SB is not fully tractable; closed-form solutions exist only when the families of p_{ref} are strictly limited (Bunne et al., 2023; Chen et al., 2023).

3.2 SCHRÖDINGER BRIDGE BETWEEN PAIRED DATA

We assume the maximum time $T = 1$ for convenience. Exploring the tractable SB between Gaussian-smoothed paired data with linear drift in SDE, we consider Gaussian boundary conditions $p_{\mathrm{S}} = \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}, \epsilon_0^2 \mathbf{I})$ and $p_{\mathrm{S}^{\dagger}} = \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}_1, e^{2\int_0^1 f(\tau)\mathrm{d}\tau}\epsilon_0^2 \boldsymbol{I})$. As $\epsilon_0 \to 0$, $\widehat{\Psi}_t$ and Ψ_t converge to the tractable solution between the target data $\boldsymbol{x}_0$ and the corrupted data $\boldsymbol{x}_1$:

$$\widehat{\Psi}_t = \mathcal{N}_{\mathbb{C}}(\alpha_t \boldsymbol{x}_0, \alpha_t^2 \sigma_t^2 \boldsymbol{I}), \Psi_t = \mathcal{N}_{\mathbb{C}}(\bar{\alpha}_t \boldsymbol{x}_1, \alpha_t^2 \bar{\sigma}_t^2 \boldsymbol{I}), \tag{13}$$

where $\alpha_t = e^{\int_0^t f(\tau)\mathrm{d}\tau}$, $\bar{\alpha}_t = e^{-\int_t^1 f(\tau)\mathrm{d}\tau}$, $\sigma_t^2 = \int_0^t \frac{g^2(\tau)}{\alpha_\tau^2}\mathrm{d}\tau$ and $\bar{\sigma}_t^2 = \int_t^1 \frac{g^2(\tau)}{\alpha_\tau^2}\mathrm{d}\tau$ are determined by f and g in the reference SDE, which are analogous to the noise schedule in SGMs (Kingma et al., 2021). The marginal distribution of the SB also has a tractable form:

$$p_t = \Psi_t \widehat{\Psi}_t = \mathcal{N}\left(\frac{\alpha_t \bar{\sigma}_t^2 \boldsymbol{x}_0 + \bar{\alpha}_t \sigma_t^2 \boldsymbol{x}_1}{\sigma_1^2}, \frac{\alpha_t^2 \bar{\sigma}_t^2 \sigma_t^2}{\sigma_1^2}\boldsymbol{I}\right). \tag{14}$$

Several noise schedules (Chen et al., 2023; Ante et al., 2024), such as variance-preserving (VP), variance-exploding (VE) and gmax, are listed in Table 1 with $\Delta\beta = \beta_1 - \beta_0$.

Following the approach in (Ante et al., 2024), we let the neural model B_θ directly predict the target data, using both reconstruction and adversarial losses as the training criteria similar to the SE tasks, where S denotes the target signal and $\tilde{S} = B_\theta(\boldsymbol{x}_t, \boldsymbol{x}_T, t)$ represents the current estimate produced by the neural network. We empirically observe that the introduction of adversarial loss effectively improves generation quality. The specific loss functions are detailed in the **Appendix**.

Sch.	$f(t)$	$g^2(t)$	α_t	σ_t^2
gmax	0	$\beta_0 + t\Delta\beta$	1	$\frac{1}{2}\Delta\beta t^2 + \beta_0 t$
Scaled VP	$-\frac{1}{2}(\beta_0 + t\Delta\beta)$	$c(\beta_0 + t\Delta\beta)$	$e^{-\frac{1}{2}\int_0^t(\beta_0+\tau\Delta\beta)\mathrm{d}\tau}$	$c(e^{\int_0^t(\beta_0+\tau\Delta\beta)\mathrm{d}\tau} - 1)$
VE	0	ck^{2t}	1	$\frac{c\left(k^{2t}-1\right)}{2\log(k)}$

Table 1: Demonstration of the noise schedules in BridgeVoC.

4 EXPERIMENTS

4.1 DATASETS

Two benchmarks are utilized in this study: LJSpeech (Keith & Linda, 2017) and LibriTTS (Heiga et al., 2019). The LJSpeech dataset comprises 13,100 clean speech clips from a single female

speaker, with a sampling rate of 22.05 kHz. Consistent with the partitioning in the publicly available VITS repository, the dataset is divided into $\{12500, 100, 500\}$ clips for training, validation, and testing, respectively. The LibriTTS dataset encompasses a variety of recording environments with a sampling rate of 24 kHz. Following the partitioning scheme in (Lee et al., 2023), the subsets {*train-clean-100*, *train-clean-300*, *train-other-500*} are used for model training. The subsets *dev-clean* + *dev-other* are employed for objective evaluation, while *test-clean* + *test-other* are used for subjective evaluations. To evaluate the generalization capability of neural vocoders, the VCTK dataset (Yamagishi, 2012) is utilized for out-of-distribution evaluations, where around 200 clips are randomly selected from the dataset for evaluations.

4.2 CONFIGURATIONS

Since the bridge between the target data S and the corrupted data $S^\dagger$ can be viewed as a restoration task, it is intuitive to choose the noise-conditional score network (NCSN++) (Song et al., 2021) as the backbone neural model. Our ablation study experimented with three sizes of NCSN++, with trainable parameter counts of 16.2M, 36.5M, and 64.9M, respectively. The number of the sampling in the reverse process is empirically set to 10.

In terms of noise schedulers, $\beta_0 = 0.01$ and $\beta_1 = 20$ are set for both gmax and scaled VP types. For VE type, we use $k = 2.6$ and $c = 0.40$, and for scaled VP type, we use $c = 0.30$. The processing time for the proposed SB is set to $T = 1$ with $t_{\text{min}} = 10^{-4}$. The reverse SDE and the probability flow Ordinary Differential Equation (ODE) (Chen et al., 2022) samplers are chosen in the inference stage. Ablation studies are conducted and can be found in the **Appendix**.

For the weight hyperparameters of the losses in Eq. (24), λ_{mel}, λ_g and λ_{fm} are 0.1, 10.0 and 10.0, respectively. "+GAN" refers to the inclusion of the loss terms $\mathcal{L}_g$ and $\mathcal{L}_{fm}$ in Eq. (24).

We train all models for 1 million steps, except for BigVGAN, which is trained for 5 million steps. The training configurations for the T-F domain SE models are aligned with those of APNet2 and BigVGAN. For feature extraction, we employ a 1024-point FFT, a Hann window of length 1024, and a hop size of 256. For the LJSpeech dataset, we utilize 80 mel-bands with the upper-bound frequency f_{max} set to 8 kHz, meaning the model is required to conduct a super-resolution task to generate the spectral component over 8 kHz. For LibriTTS, the mel-bands and upper-bound frequency are set to 100 and 12 kHz, respectively.

4.3 RESULTS AND ANALYSIS

For vocoding performance comparisons, we select popular vocoding models as baselines, including time-domain methods (BigVGAN (Lee et al., 2023), HiFiGAN (Kong et al., 2020)), T-F domain methods (Vocos (Hubert, 2024), FreeV (Lv et al., 2024), APNet2 (Du et al., 2024)), and diffusion-based methods (DiffWave (Kong et al., 2021), PriorGrad (Lee et al., 2022), and FreGrad (Nguyen et al., 2024)). To compare the model efficiency, we calculate the number of model parameters (#Params) and real-time factor (RTF) which is measured on a single Tesla V100 GPU.

Eight metrics are involved in the objective evaluations: (1) Wide-band version of Perceptual evaluation of speech quality (PESQ) (Rec, 2005) serves to assess the objective speech quality. (2) Extended Short-Time Objective Intelligibility (ESTOI) (Taal et al., 2011) measures the intelligibility of speech. (3) Periodicity RMSE, V/UV F1 score, F0, and pitch RMSE (Morrison et al., 2022; Kawahara et al., 1999) are regarded as major artifacts for non-autoregressive neural vocoders. (4) Virtual Speech Quality Objective Listener (VISQOL) (Hines et al., 2015) predicts the Mean Opinion Score-Listening Quality Objective (MOS-LQO) score by evaluating the spectro-temporal similarity. (5) UTMOS (Saeki et al., 2022) is used to obtain subjective scores related to the perceived quality of speech, providing an objective approximation of human judgment.

For subjective evaluations, we employ the MUSHRA and ABX testing methodologies based on the BeaqleJS platform (Kraft & Zölzer, 2014). A total of 19 participants, all specializing in audio signal processing, are involved in the testing. In the MUSHRA test, each participant is required to rate the speech processed by various algorithms on a scale from 0 to 100, based on the overall similarity to a reference. In the ABX test, participants are asked to select the clip they prefer in terms of overall speech quality, or choose "equal" if no preference can be given.

Models	Domain	#Param. (M)	#MACs (Giga/5s)	Inference Speed	PESQ↑	ESTOI↑	V/UV↑ F1	VISQOL↑	UTMOS↑	Periodicity↓ RMSE	Pitch↓ RMSE	F0↓ RMSE
HiFiGAN-V1	T	14.0	152.90	0.0092	3.574	0.8892	0.9474	4.771	4.219	0.1344	33.69	36.23
BigVGAN-base	T	14.0	152.90	0.0395	3.603	0.9569	0.9562	4.822	4.210	0.1198	30.28	39.21
BigVGAN	T	112.4	417.20	0.0584	<u>4.065</u>	<u>0.9782</u>	**0.9716**	**4.863**	**4.296**	**0.0838**	20.69	<u>34.43</u>
APNet2	T-F	31.5	13.53	0.0027	3.476	0.9412	0.9592	4.752	3.985	0.1126	25.36	41.76
Vocos	T-F	13.5	5.80	0.0009	3.522	0.9455	0.9559	4.774	3.970	0.1213	29.13	36.56
FreeV	T-F	18.3	7.84	0.0015	3.593	0.9474	<u>0.9603</u>	4.743	4.015	<u>0.1118</u>	25.99	39.09
DiffWave	T	6.91	231.07×200	0.8738	3.652	0.9321	0.9375	4.325	3.871	0.1585	27.42	37.84
FreGrad	T	2.62	34.42×50	0.3959	3.774	0.9475	0.9432	4.450	3.933	0.1413	24.17	36.72
PriorGrad	T	2.62	71.43×50	0.8874	3.961	0.9579	0.9506	4.509	4.004	0.1283	<u>19.46</u>	36.07
BridgeVoC-base(ours)	T-F	16.2	113.79×10	0.1747	4.418	0.9883	0.9576	4.817	4.237	0.1160	15.24	32.94
BridgeVoC(ours)	T-F	64.8	450.45×10	0.5409	**4.440**	**0.9896**	0.9598	<u>4.824</u>	<u>4.262</u>	0.1136	**15.04**	**32.72**

Table 2: Results of objective evaluations on the dev-clean and dev-other subset of LJSpeech dataset. "#Param." denotes the number of trainable parameters. Metrics with ↓ indicate that lower values are better. The inference speed on a GPU is evaluated based on a single Tesla V100. The computational complexity of the diffusion methods needs to be multiplied × by the number of reverse sampling steps. The best and second-best performances are namely highlighted in **bold** and <u>underlined</u>.

Models	PESQ↑	Pitch↓ RMSE	VISQOL↑
WaveGlow-256†	3.138	-	-
HiFiGAN-V1	3.056	52.53	4.721
iSTFTNet-V1	2.880	53.07	4.655
UnivNet-c32†	3.277	41.51	4.753
Avocodo	3.217	51.60	4.762
BigVGAN-base(1M steps)†	3.519	-	-
BigVGAN(1M steps)†	4.027	-	-
BigVGAN-base(5M steps)†	3.841	32.54	4.907
BigVGAN(5M steps)†	4.269	<u>24.28</u>	**4.963**
APNet	2.897	39.66	4.666
APNet2	2.834	46.37	4.582
Vocos†	3.615	35.58	4.879
PriorGrad	4.043	28.34	4.381
FreGrad	3.793	39.88	4.337
BridgeVoC-base(ours)	4.419	17.84	4.908
BridgeVoC(ours)	**4.459**	**14.89**	<u>4.914</u>

Table 3: Objective comparisons among baselines on the LibriTTS benchmark. "-" denotes the results are not reported, and † denotes the results are calculated using the open-sourced model checkpoints.

Models	PESQ↑	Pitch↓ RMSE	VISQOL↑	MUSHRA
Ground Truth	-	-	-	89.61±0.62
HiFiGAN-V1	3.090	33.29	4.723	72.47±1.07
Vocos	3.684	23.46	<u>4.866</u>	75.77±1.24
BigVGAN-base†	3.859	28.85	4.893	80.23±0.99
BigVGAN†	<u>4.282</u>	20.32	**4.958**	82.78±0.81
PriorGrad	3.911	<u>19.56</u>	4.278	77.53±1.10
FreGrad	3.653	27.93	4.201	78.06±1.11
BridgeVoC-base	4.323	19.31	4.855	82.15±0.93
BridgeVoC	**4.334**	**18.31**	4.863	*__83.34__±__1.02__

Table 4: Metric comparisons on VCTK. All models are pretrained on the LibriTTS dataset. For the MUSHRA test, with a confidence level of 95%, we performed a t-test comparing BridgeVoC with BigVGAN, yielding a p-value of less than 0.05 (*p<0.05).

4.3.1 COMPARISONS WITH SOTA METHODS

Tables 2 and 3 present objective comparisons on the LJSpeech and LibriTTS datasets, revealing several key observations. First, the T-F domain-based methods exhibit faster inference speeds compared to the time-domain methods, primarily due to the use of STFT and its inverse transform, iSTFT, which eliminate the need for upsampling operations. Second, the T-F domain-based methods generally have significantly lower computational complexity, e.g., 5.8 GMACs for Vocos versus 152.9 GMACs for HiFiGAN, making them increasingly attractive. Third, despite these advantages, the speech quality of these existing T-F domain-based neural vocoders remains inferior to that of BigVGAN. Fourth, previous diffusion-based methods start from noise in the time domain and use the mel-spectrogram as a diffusion condition, failing to fully leverage the prior information of the mel-spectrogram. The proposed BridgeVoc, however, benefits from the prior structural information provided by the pseudo-inverse operation and the combination of the T-F domain-based Schrödinger bridge and auxiliary losses. This allows BridgeVoc to achieve both relatively fast inference speeds and promising performance. Notably, even when compared to BigVGAN trained for 5 million steps on the LibriTTS benchmark, our method remains competitive, fully validating the effectiveness of the proposed approach.

Table 4 presents the results on the out-of-domain test set. Compared to Table 3, the relative advantage of BridgeVoc over BridgeVoc-base in objective metrics slightly decreases. This is because the amount of data in LibriTTS is probably insufficient for a large NCSN++ network. The MUSHRA results on the test set of the VCTK dataset reveal that our BridgeVoc is statistically superior to BigVGAN ($p < 0.05$), further demonstrating the advantage of our method in achieving subjective quality close to the ground truth signal.

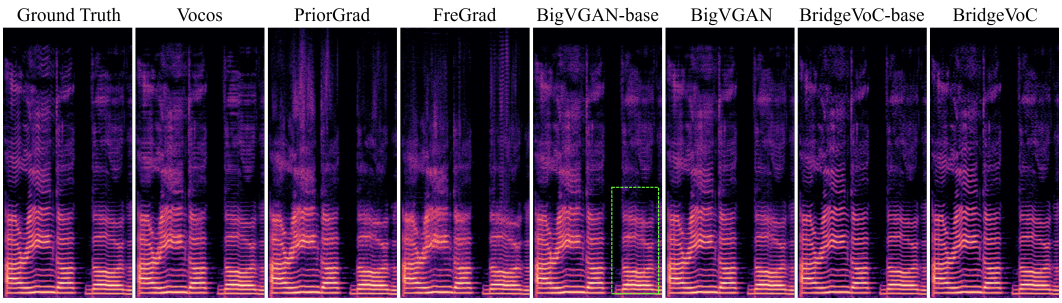


Figure 3: Spectral visualization of different vocoder methods. The audio clip is a singing voice from the MUSHDB18 test set.

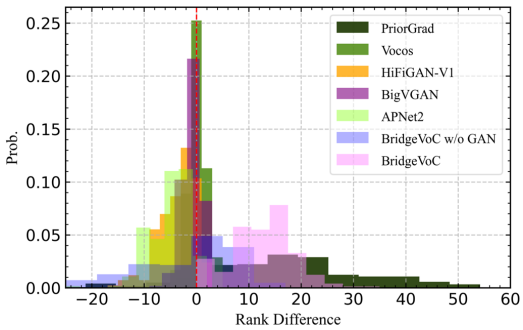


Figure 4: The rank differences between different models and the Ground Truth on the dev-clean and dev-other subsets of the LibriTTS dataset.

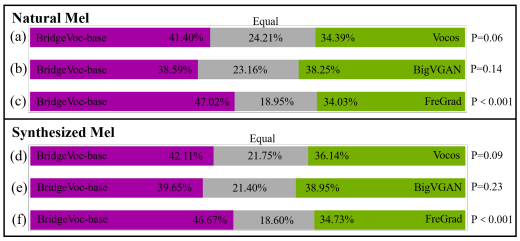


Figure 5: Average preference scores (in %) of ABX tests between BridgeVoC-base and two other baselines. (a)-(c) Mel-spectrograms are obtained from natural speech clips in the LibriTTS test set. (d)-(f) Mel-spectrograms are synthesized from F5-TTS (Chen et al., 2024), where the transcripts are from the LibriTTS test set.

Figure 3 presents spectral visualizations of different models for a vocal clip from the out-of-distribution MUSDB18 (Rafii et al., 2017) test set. Our approach more effectively recovers harmonic details and avoids artificial harmonic fluctuations compared to other baselines, particularly BigVGAN-base. Subjective experiments revealed that some listeners reported "strange pitch shifts" relative to the ground truth in the MUSHRA experiments, with most instances traced back to BigVGAN-base. While BigVGAN also shows some "artificial generation" artifacts, their extent is significantly reduced. The rank differences between various models and the Ground Truth are shown in Figure 4. Diffusion models, PriorGrad and BridgeVoC, mostly have positive rank differences, indicating they generate more rank information than the Ground Truth. In contrast, APNet2 and HiFiGAN mostly have negative rank differences. Vocos and BigVGAN results are close to zero, indicating their inferred ranks are similar to the Ground Truth. Although using GAN in BridgeVoC eliminates high-frequency artifacts (Figure 8), it also increases meaningless rank.

The preference scores are shown in Figure 5. For both nature and synthesized mel cases, the preference performance of the BridgeVoC-base is significantly better over FreGrad ($p < 0.001$), and is not significantly different from BigVGAN and Vocos ($p > 0.05$). Note that we choose PriorGrad as the baseline diffusion model because the Mean Opinion Score (MOS) experiments in (Nguyen et al., 2024) indicate that PriorGrad achieves slightly higher subjective scores compared to FreGrad.

5 Conclusions

In this work, we present a novel time-frequency (T-F) domain-based diffusion neural vocoder that seamlessly integrates the data-to-data Schrödinger Bridge framework with range-null decomposition (RND) theory. Our approach involves converting the original acoustic features from the mel-scale domain to the target linear-scale domain using the range-space component, while the null-space component reconstructs the remaining spectral details through a diffusion generation process. To enhance synthesis quality, we incorporate generative adversarial networks (GANs) and conduct systematic optimization of hyperparameters. Comprehensive experiments on the LJSpeech and LibriTTS benchmarks demonstrate the effectiveness of our method, achieving state-of-the-art performance in both objective metrics and subjective evaluations.

REFERENCES

Y. Ai and Z. Ling. Apnet: An all-frame-level neural vocoder incorporating direct prediction of amplitude and phase spectra. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 31:2145–2157, 2023. doi: 10.1109/TASLP.2023.3277276.

J. Ante, K. Roman, B. Jagadeesh, and G. Boris. Schrödinger bridge for generative speech enhancement. In *Proc. Interspeech*, pp. 1175–1179, 2024.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Proc. NeurIPS*, volume 34, pp. 17695–17709. Curran Associates, Inc., 2021.

C. Bunne, Y. Hsieh, M. Cuturi, and A. Krause. The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *Proc. AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5802–5833. PMLR, 25–27 Apr 2023.

N. Chen, Y. Zhang Zha, H. Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating Gradients for Waveform Generation. In *Proc. ICLR*, 2021. URL https://openreview.net/forum?id=NsMLjcFaO8O.

T. Chen, G. Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nioAdKCEdXB.

Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

Z. Chen, G. He, K. Zheng, X. Tan, and J. Zhu. Schrodinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*, 2023.

H.S. Choi, J. Lee, W. Kim, J. Lee, et al. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Proc. NeurIPS*, 34:16251–16265, 2021.

S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016.

H. Du, Y. Lu, Y. Ai, and Z. Ling. APNet2: High-Quality and High-Efficiency Neural Vocoder with Direct Prediction of Amplitude and Phase Spectra. In *Proc. MMSC*, pp. 66–80, 2024.

Y. Guo, C. Du, Z. Ma, X. Chen, and K. Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11121–11125. IEEE, 2024.

Z. Heiga, C. Rob, J.-W. Ron, D. Viet, et al. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech*, 2019. URL https://arxiv.org/abs/1904.02882.

A. Hines, J. Skoglund, A. Kokaram, and N. Harte. ViSQOL: an objective speech quality model. *EURASIP J. Audio Speech Music Process.*, pp. 1–18, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Proc. NeurIPS*, 33:6840–6851, 2020.

R. Huang, J. Huang, D. Yang, Y. Ren, et al. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. In *Proc. ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932. PMLR, 23–29 Jul 2023.

S. Hubert. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *Proc. ICLR*, 2024. URL https://openreview.net/forum?id=vY9nzQmQBw.

J. Hwang, S. Lee, and S. Lee. Hiddensinger: High-quality singing voice synthesis via neural audio codec and latent diffusion models. *Neural Netw.*, 181:106762, 2025.

H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.*, 27(3-4):187–207, 1999.

I. Keith and J. Linda. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

S. Kim, S. Lee, J. Song, J. Kim, and S. Yoon. FloWaveNet : A generative flow for raw audio. In *Proc. ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3370–3378. PMLR, 09–15 Jun 2019.

D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. *Proc. NeurIPS*, 34: 21696–21707, 2021.

J. Kong, J. Kim, and J. Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Proc. NeurIPS*, volume 33, pp. 17022–17033. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf.

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *Proc. ICLR*, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

S. Kraft and U. Zölzer. BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality. In *Linux Audio Conference, Karlsruhe, DE*, 2014.

D. Laurent, S. Jascha, and B. Samy. Density estimation using real NVP. In *Proc. ICLR*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

S. Lee, H. Kim, C. Shin, X. Tan, et al. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *Proc. ICLR*, 2022. URL https://openreview.net/forum?id=_BNiN4IjC5.

S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon. BigVGAN: A Universal Neural Vocoder with Large-Scale Training. In *Proc. ICLR*, 2023. URL https://openreview.net/forum?id=iTtGCMDEzS_.

H. Liu, W. Choi, X. Liu, Q. Kong, et al. Neural Vocoder is All You Need for Speech Superresolution. In *Proc. Interspeech*, pp. 4227–4231, 2022a. doi: 10.21437/Interspeech.2022-11017.

H. Liu, X. Liu, Q. Kong, Q. Tian, et al. VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration. In *Proc. Interspeech*, pp. 4232–4236, 2022b. doi: 10.21437/Interspeech.2022-11026.

J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proc. AAAI*, volume 36, pp. 11020–11028, 2022c.

Y. Lv, H. Li, Y. Yang, J. Liu, et al. FreeV: Free Lunch For Vocoders Through Pseudo Inversed Mel Filter. In *Proc. Interspeech*, pp. 3869–3873, 2024.

N. Majumder, C. Hung, D. Ghosal, W. Hsu, et al. Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization. In *Proc. ACMMM*, MM '24, pp. 564–572, 2024. doi: 10.1145/3664647.3681688.

S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, et al. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. In *Proc. ICLR*, 2022.

S. Mehta, R. Tu, J. Beskow, É. Székely, and G. Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11341–11345. IEEE, 2024.

M. Meinard. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015. ISBN 978-3-319-21944-8.

M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, et al. Chunked Autoregressive GAN for Conditional Waveform Synthesis. In *Proc. ICLR*, 2022.

Tan Dat Nguyen, Ji-Hoon Kim, Youngjoon Jang, Jaehun Kim, and Joon Son Chung. Fregrad: Lightweight and Fast Frequency-Aware Diffusion Vocoder. In *Proc. ICASSP*, pp. 10736–10740, 2024. doi: 10.1109/ICASSP48485.2024.10447251.

A. Oord, Y. Li, I. Babuschkin, K. Simonyan, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *Proc. ICML*, pp. 3918–3926. PMLR, 2018.

R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proc. ICASSP*, pp. 3617–3621. IEEE, 2019.

K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *Proc. ICML*, pp. 5210–5219. PMLR, 2019.

Z. Rafii, A. Liutkus, F. Stöter, S. Mimilakis, and R. Bittner. The MUSDB18 corpus for music separation, 2017. URL https://doi.org/10.5281/zenodo.1117372.

ITUT Rec. P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, CH–Geneva*, 41:48–60, 2005.

Y. Ren, Y. Ruan, X. Tan, T. Qin, et al. Fastspeech: Fast, robust and controllable text to speech. In *Proc. NeurIPS*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f63f65b503e22cb970527f23c9ad7db1-Paper.pdf.

Y. Ren, C. Hu, X. Tan, T. Qin, et al. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari. UTMOS: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

E. Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pp. 269–310, 1932.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proc. ICLR*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

C.-H. Taal, R. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.*, 19(7):2125–2136, 2011.

X. Tan, J. Chen, H. Liu, J. Cong, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4234–4245, 2024. doi: 10.1109/TPAMI.2024.3356232.

J. Valin and J. Skoglund. LPCNet: Improving neural speech synthesis through linear prediction. In *Proc. ICASSP*, pp. 5891–5895. IEEE, 2019.

G. Wang, Y. Jiao, Q. Xu, Y. Wang, and C. Yang. Deep Generative Learning via Schrödinger Bridge. In *Proc. ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10794–10804. PMLR, 18–24 Jul 2021.

Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, et al. Tacotron: Towards end-to-end speech synthesis. *Proc. Interspeech*, pp. 4006, 2017.

Y. Wang, Z. Ju, X. Tan, L. He, et al. Audit: Audio editing by following instructions with latent diffusion models. *Proc. NeurIPS*, 36:71340–71357, 2023.

J. Won, C. Daniel, Y. Jaesam, K. Bongwan, and K. Juntae. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. Interspeech*, 2021. URL https://api.semanticscholar.org/CorpusID:235435945.

J. Yamagishi. English multi-speaker corpus for cstr voice cloning toolkit, 2012. URL http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html/.

J. Zhang and B. Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1828–1837, 2018.

A Appendix

A.1 Score-Based Generative Models

Given a data distribution $p_{\text{data}}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d$, SGMs (Song et al., 2021) are built on a continuous-time diffusion process defined by a forward SDE:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t, \quad \boldsymbol{x}_0 \sim p_0 = p_{\text{data}}, \tag{15}$$

where $t \in [0, T]$ is a finite time index, $\boldsymbol{x}_t \in \mathbb{R}^d$ is the state of the process, $\boldsymbol{f}$ is a vector-valued drift term, g is a scalar-valued diffusion term, and $\boldsymbol{w}_t \in \mathbb{R}^d$ is a standard Wiener process. To ensure that the boundary distribution is a Gaussian prior distribution $p_{\text{prior}} = \mathcal{N}(\mathbf{0}, \sigma_T^2 \boldsymbol{I})$, we construct the drift term $\boldsymbol{f}$ and the diffusion term g accordingly. This construction guarantees that the forward SDE has a corresponding reverse SDE:

$$\begin{aligned} \mathrm{d}\boldsymbol{x}_t = [\boldsymbol{f}(\boldsymbol{x}_t, t) - g^2(t)\nabla \log p_t(\boldsymbol{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\boldsymbol{w}}_t, \\ \boldsymbol{x}_T \sim p_T \approx p_{\text{prior}}, \end{aligned} \tag{16}$$

where $\bar{\boldsymbol{w}}_t$ is the reverse-time Wiener process, and $\nabla \log p_t(\boldsymbol{x}_t)$ is the *score function* of the marginal distribution p_t. To enable inference generated data samples at $t = 0$, we can replace the score function with a score network $s_\theta(\boldsymbol{x}_t, t)$ and solve it reversely from p_{prior} at $t = T$. A score network is usually learned by the denoising score matching objective (Song et al., 2021) :

$$\mathbb{E}_{p_0(\boldsymbol{x}_0)p_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0),t}\left[\|s_\theta(\boldsymbol{x}_t, t) - \nabla \log p_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2\right], \tag{17}$$

where $t \sim \mathcal{U}(0, T)$ and $p_{t|0}$ is the conditional transition distribution from $\boldsymbol{x}_0$ to $\boldsymbol{x}_t$, determined by the pre-defined forward SDE and analytical for a linear drift $\boldsymbol{f}(\boldsymbol{x}_t, t) = f(t)\boldsymbol{x}_t$.

A.2 Loss Function

Given that we employ the pseudo-inverse to map mel-spectrograms back to the original uncompressed linear-scale spectrogram, the extraction of amplitude information in the mel domain can assist the model in better reconstructing the original linear-scale information. Therefore, the reconstruction losses include both the mean-square error (MSE) loss $\mathcal{L}_{mse}$ and the mel loss $\mathcal{L}_{mel}$ following the settings in (Ai & Ling, 2023; Du et al., 2024). The former is defined as the MSE between $\tilde{S}$ and S in the STFT domain:

$$\mathcal{L}_{mse} = \frac{1}{FT}\sum_{f,t}\left\|\tilde{S}_{f,t} - S_{f,t}\right\|_2^2. \tag{18}$$

The mel loss measures the mean absolute error (MAE) between the mel-spectrograms of the estimated $\tilde{Y}^{mel}$ and target waveforms Y^{mel}:

$$\mathcal{L}_{mel} = \frac{1}{F_{mel}T}\sum_{f,t}\left\|\tilde{Y}_{f,t}^{mel} - Y_{f,t}^{mel}\right\|_1. \tag{19}$$

For the multi-mel loss, we compute the sum of mel losses across seven different configurations:

$$\mathcal{L}_{mel} = \mathcal{L}_{mel_0} + \mathcal{L}_{mel_1} + ... + \mathcal{L}_{mel_6}. \tag{20}$$

These configurations vary in the Fast Fourier Transform size (n_{fft}) and the number of mel frequency bins (n_{mels}), which are set to (32, 64, 128, 256, 512, 1024, 2048) and (5, 10, 20, 40, 80, 160, 210), respectively. For all configurations, the upper-bound frequency (f_{max}) is fixed at half the sampling

rate, while the window size and hop size are set to n_{fft} and $n_{\text{fft}}/4$, respectively. The single mel loss defined in Eq. (19) corresponds to the specific configuration where $n_{\text{fft}} = 1024$ and $n_{\text{mels}} = 160$.

The adversarial losses includes the hinge GANs of discriminators D_m and generator B_θ, denoted as $\mathcal{L}_d$ and $\mathcal{L}_g$, respectively:

$$\mathcal{L}_d = \frac{1}{M}\sum_{m=1}^{M} \max\left(0, 1 - D_m\left(\mathbf{s}\right)\right) + \max\left(0, 1 + D_m\left(\tilde{\mathbf{s}}\right)\right), \tag{21}$$

$$\mathcal{L}_g = \frac{1}{M}\sum_{m=1}^{M} \max\left(0, 1 - D_m\left(\tilde{\mathbf{s}}\right)\right), \tag{22}$$

where $\tilde{\mathrm{s}} = \text{iSTFT}(\tilde{\mathbf{S}}) \in \mathbb{R}^L$ denotes the reconstructed waveforms, $\text{iSTFT}\left(\cdot\right)$ refers to the iSTFT operation, and M is the number of sub-discriminators. Discriminators includes multi-period discriminator (MPD) (Kong et al., 2020) and multi-resolution spectrogram discriminator (MRSD) (Won et al., 2021).

The discriminator settings for the adversarial loss include two components: Multi-Period Discriminator (MPD) and Multi-Resolution Spectral Discriminator (MRSD). The MPD captures variations in audio periodic patterns using five sub-discriminators. Each sub-discriminator reshapes the 1D raw audio waveform into a 2D format based on predefined period values, which are set to $\{2, 3, 5, 7, 11\}$. The MRSD consists of three sub-discriminators, where the magnitude spectrum serves as the input. This input is then fed into a stack of Conv2d layers to compute the discriminative scores. The configurations for $\{$window size, hop size, $n_{\text{fft}}\}$ in the three sub-discriminators are (512, 128, 512), (1024, 256, 1024), and (2048, 512, 2048), respectively.

Besides, the feature matching loss is also utilized:

$$\mathcal{L}_{fm} = \frac{1}{LM}\sum_{l,m} |\mathbf{f}_l^m\left(\tilde{\mathbf{s}}\right) - \mathbf{f}_l^m\left(\mathbf{s}\right)|, \tag{23}$$

where $\mathbf{f}_l^m(\cdot)$ denotes the l-th layer feature for the m-th sub-discriminator. Finally, the loss for the neural model is

$$\mathcal{L}_B = \mathcal{L}_{mse} + \lambda_{mel}\mathcal{L}_{mel} + \lambda_g\mathcal{L}_g + \lambda_{fm}\mathcal{L}_{fm}, \tag{24}$$

where λ_{mel}, λ_g, and λ_{fm} are the are the weight hyperparameters of corresponding loss.

A.3 Ablation studies

To determine the optimal configuration of diffusion hyperparameters and network settings for BridgeVoC, we conducted ablation experiments on the LJSpeech benchmark.

Table 5 presents the test performance with various combinations of losses and noise schedules when the network parameter count is 16.2M. From the experimental results, it is evident that the introduction of auxiliary losses, single mel loss "+mel" and multi-mel loss "+mmel", can significantly enhance the model's performance. Furthermore, adding GAN on top of "+mmel" further improves the WB-PESQ score by 0.016. Correspondingly, other metrics also show certain improvements. When comparing Scaled VP and VE under the "+mmel+GAN" condition, gmax emerges as the optimal choice for the majority of indicators. Additionally, when the sampler is switched from the reverse SDE to the probability flow ODE, there is a slight degradation in performance.

Table 6 lists the results for the methods of reconstructing the signal from the network output and varying the network size under the settings of "gmax", "+mmel+GAN", and "SDE". "map" and "crm" denote that the network output is the complex spectrum mapping and the complex mask, respectively. "decouple" indicates that the network outputs the amplitude and phase of the signal separately, which are then coupled to form the output signal. The results indicate that the "crm"configuration is optimal for our task, rather than the previously default "map" form used in the NCSN++ network. Additionally, increasing the network size also improves the final output scores.

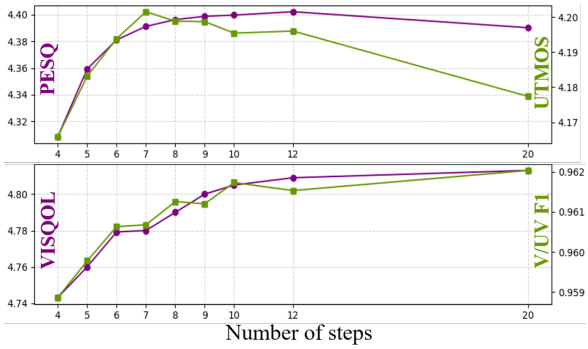


Figure 6: Metrics with different numbers of sampling steps during the reverse process on the test set of the LJSpeech dataset.

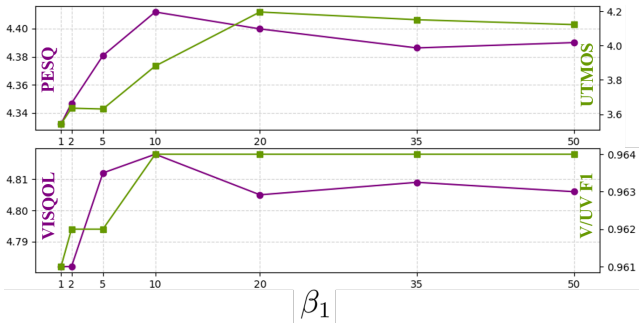


Figure 7: Metrics with different values of β_1.

Shedules	Losses	Sampler	PESQ	VISQOL	UTMOS
gmax	mse	SDE	4.005	4.182	3.966
Scaled VP	mse	SDE	4.207	4.389	3.804
VE	mse	SDE	4.195	4.421	3.640
gmax	+mel	SDE	4.314	4.681	4.062
gmax	+mmel	SDE	4.400	4.805	4.195
gmax	+mmel	ODE	4.311	4.778	4.203
gmax	+mmel+GAN	SDE	**4.416**	4.798	**4.217**
Scaled VP	+mmel+GAN	SDE	4.379	4.796	3.987
VE	+mmel+GAN	SDE	4.370	**4.816**	3.796

Table 5: Ablation study of loss function and noise schedules on the LJSpeech benchmark.

Recon.	#Param.(M)	PESQ	VISQOL	UTMOS
map	16.2	4.416	4.798	4.217
crm	16.2	4.418	4.817	4.237
decouple	16.2	4.369	4.764	3.765
crm	36.5	4.431	4.807	4.258
crm	64.9	4.440	4.824	4.262

Table 6: Ablation study of the signal reconstruction methods and net sizes on the LJSpeech benchmark.

For the case of "gmax" / "+mmel" / "map" / "16.2M", Figure 6 shows the results of the number of reverse sampling steps ablations. We found that for certain metrics, increasing the number of reverse sampling steps yields better results. However, for other metrics, a specific number of steps achieves the highest score. Similar observations have been reported in other diffusion-based works (Ho et al., 2020). This phenomenon maybe due to the trade-off between the granularity of the sampling process and the accumulation of numerical errors. As the number of sampling steps increases, the model can more accurately capture the underlying data distribution, leading to improved performance for some metrics. However, beyond a certain point, the benefits of additional steps may be outweighed by the increased potential for error accumulation, resulting in a decline in performance for other metrics. This finding also implies that 10 steps are adequate for BridgeVoC, while reducing the number of steps to 7 does not lead to a substantial performance decline. Therefore, BridgeVoC has the potential to further reduce computational complexity and increase inference speed.

For the hyperparameter β_1 in the noise schedule, prior studies Bortoli et al. (2021); Chen et al. (2023); Ante et al. (2024) have empirically explored values ranging from 1 to 50. To determine the optimal value of β_1 in this work, we conducted an ablation experiment under the configuration "gmax" / "+mmel" / "map" / "16.2M". As shown in Figure 7, the results reveal that there is no significant difference in the three metrics—PESQ, VISQOL, and V/UV F1—within the range of 10

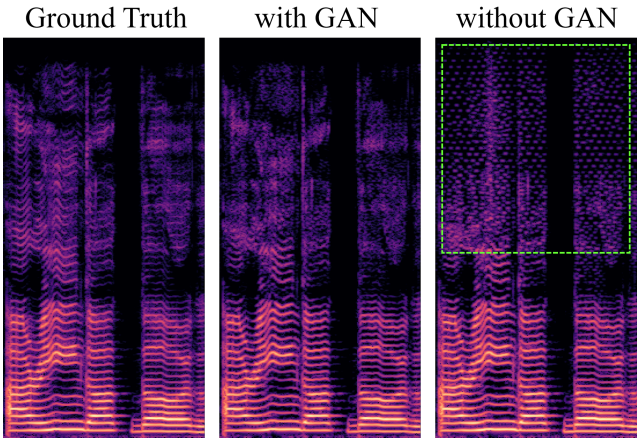


Figure 8: Spectral visualization of BrdigeVoc-base (with GAN) and BrdigeVoc-base without GAN. The audio clip is a singing voice from the MUSHDB18 test set.

Sampling Methods	Losses	PESQ↑	ESTOI↑	V/UV↑ F1	VISQOL↑	UTMOS↑	Periodicity↓ RMSE	Pitch↓ RMSE	F0↓ RMSE
Score Matching	mse	4.211	0.9859	0.9528	4.767	3.436	**0.0733**	14.32	37.60
Rectified Flow Matching	mse	4.317	**0.9890**	**0.9711**	4.711	3.595	0.0903	15.84	32.99
Optimal-transport Flow Matching	mse	4.309	0.9884	0.9708	4.709	3.586	0.0911	15.46	32.96
Schrödinger Bridge	mse	4.005	0.9791	0.9534	4.182	3.966	0.1438	**13.45**	33.20
Schrödinger Bridge	mse+mmel	**4.400**	0.9835	0.9644	**4.805**	**4.195**	0.1047	14.19	**31.75**

Table 7: Metric comparisons of several Probabilistic Sampling Methods on LJSpeech. The best and second-best performances are namely highlighted in **bold** and underlined.

to 35. However, when β_1 increases from 10 to 20, a noticeable improvement in UTMOS is observed. Consequently, we chose $\beta_1 = 20$ as the optimal value for the final experimental setup.

A.4 Comparisons between BridgeVoC w/ and w/o GAN

Analysis of the inference results from BridgeVoc without GAN reveals the presence of a regular dotted spectral distribution in the high-frequency region, as shown in Figure 8. Compared with the output from BridgeVoc with GAN, i.e., BridgeVoc-base, it becomes evident that the dotted structure mainly appears in frequency bands where clear prior band structures are not obtainable from the mel inversion. This indicates that it is challenging to accurately recover high-frequency components even with the incorporation of GAN.

In the ablation study of the loss function, detailed in the main text, objective metrics such as PESQ and VISQOL show that the dotted structures do not significantly affect the intrusive evaluation scores. However, they do have a noticeable impact on UTMOS, a non-intrusive metric that better reflects perceptual quality. Therefore, it can be inferred that the regular dotted structure we observe is the representation of the posterior energy probability in regions where prior information is unclear. It reflects the tendency of the model to distribute energy in a sparse and structured manner to approximate the high-frequency content.

A.5 Comparative Analysis of Sampling Strategies in Probabilistic Modeling

We also compared several probabilistic sampling methods, such as score matching with an Ornstein-Uhlenbeck SDE with a variance-exploding (OUVE) and a one-step corrector Song et al. (2021), rectified flow matching Guo et al. (2024), and optimal-transport flow matching Mehta et al. (2024), within the same training framework. Both flow matching methods employed an Euler sampler for the ODE. Unlike the score network, which takes $\nabla \log p_t$ as its training objective, or the flow matching network that predicts the velocity field to transform the initial distribution into the target distribution, Schrödinger Bridge methods can directly predict the target data through model parameterization. This characteristic enables the incorporation of auxiliary loss functions.

The experimental results presented in Table 7 show that when the loss function is solely mean squared error (MSE), each of the four sampling strategies exhibits a slight advantage in one or two specific objective metrics. However, performance improves across all metrics except for Pitch RMSE when the Schrödinger Bridge method incorporates an auxiliary multi-mel loss. Notably, the UTMOS score increases from 3.966 to 4.196, indicating that the multi-mel loss helps capture features that are both measurable and perceptually relevant to human listeners.