BridgeVoC: Neural Vocoder with Schrödinger Bridge

Tong Lei[1,3], **Zhiyu Zhang**[4], **Rilin Chen**[3], **Meng Yu**[3], **Jing Lu**[1], **Chengshi Zheng**[2], **Dong Yu**[3] and **Andong Li**[2,*]

[1]Key Laboratory of Modern Acoustics, Nanjing University
[2] Key Laboratory of Noise and Vibration Research, Institute of Acoustics Chinese Academy of Sciences
[3]Tencent AI Lab
[4]National Mobile Communications Research Laboratory, Southeast University
tonglei@smail.nju.edu.cn, {liandong, cszheng}@mail.ioa.ac.cn, zhiyuzhang@seu.edu.cn,
rilinchen@tencent.com, lujing@nju.edu.cn, {raymondmyu, dyu}@global.tencent.com

Abstract

While previous diffusion-based neural vocoders typically follow a noise-to-data generation pipeline, the linear-degradation prior of the mel-spectrogram is often neglected, resulting in limited generation quality. By revisiting the vocoding task and excavating its connection with the signal restoration task, this paper proposes a 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. Quantitative experiments on LJSpeech and LibriTTS show that BridgeVoC achieves faster inference and surpasses existing diffusion-based vocoder baselines, while also matching or exceeding non-diffusion state-of-the-art methods across 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 [Wang *et al.*, 2017; Ren *et al.*, 2019; Tan *et al.*, 2024], text-to-audio [Huang *et al.*, 2023; Majumder *et al.*, 2024], singing voice synthesis [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; Liu *et al.*, 2022b].

In recent years, significant improvements in vocoding quality have been achieved because of the application of deep neural networks (DNNs). Auto-regressive (AR) methods such as WaveNet [Dieleman *et al.*, 2016; Oord *et al.*, 2018], SampleRNN [Mehri *et al.*, 2022], and LPCNet [Valin and Skoglund, 2019] often face challenges with slow generation speeds due to their sequential nature. Flow-based vocoder methods, such as WaveGlow [Prenger and Valle, 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. Additionally, non-autoregressive (NAR) methods like HiFi-GAN [Kong *et al.*, 2020] have emerged, offering parallel processing and enhanced efficiency.

Most recently, time-frequency (T-F) domain-based neural vocoders have gained prominence. In these methods, the network estimates the spectral magnitude and phase in the Short-Time Fourier Transform (STFT) domain, and the inverse STFT (iSTFT) operation is then utilized to generate waveforms. These T-F methods have demonstrated competitive performance and faster inference speeds compared to time-domain methods [Lee *et al.*, 2023; Hubert, 2024; Du *et al.*, 2024]. Typically, these non-diffusion methods gen-

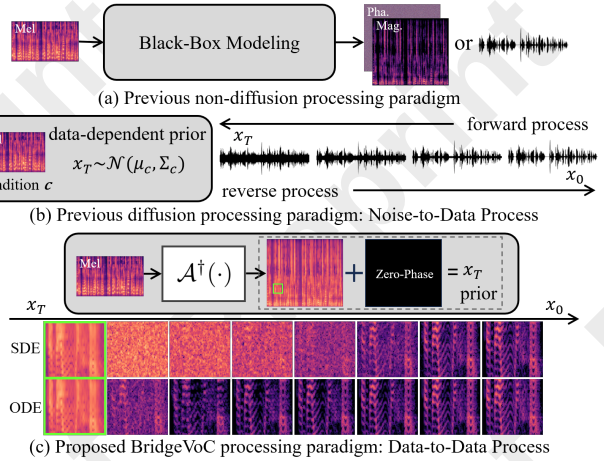

Figure 1: Illustrations of the various neural vocoder paradigms.

*Andong Li is the corresponding author.

erate waveforms by taking acoustic features, such as the mel-spectrogram, as input. They employ various generators to estimate the spectral magnitude and phase or directly produce the waveform, as illustrated in Figure 1(a).

Diffusion-based methods typically have slower inference speeds and lower objective metrics, but offer greater flexibility, diversity, and more natural-sounding audio than non-diffusion vocoders. For example, WaveGrad refines white Gaussian noise into high-fidelity audio via a gradient-based sampler conditioned on the mel-spectrogram, balancing inference speed and quality [Chen *et al.*, 2021]. DiffWave, a non-autoregressive diffusion model, efficiently generates high-fidelity audio through a Markov chain by optimizing a variational bound, requiring less computation and a smaller model size than WaveGrad, and excelling at unconditional generation [Kong *et al.*, 2021]. PriorGrad replaces DiffWave's standard Gaussian prior with a data-driven adaptive prior, enabling faster convergence and improved perceptual quality [Lee *et al.*, 2022]. Compared to PriorGrad and DiffWave, FreGrad achieves much faster training and inference, and a smaller model size, by operating in a simplified feature space and using frequency-aware components [Nguyen *et al.*, 2024]. As shown in Figure 1(b), diffusion vocoders start from random Gaussian noise and iteratively denoise it, conditioned on mel-spectrograms or other features, following a **noise-to-data** pipeline.

In this work, we revisit neural vocoding task and introduce the Schrödinger Bridge to establish a **data-to-data** process between target and corrupted spectrograms in the T-F domain from a restoration perspective rather than simple generation, as shown in Figure 1(c). Mel-spectrograms, derived from a linear-to-mel transform, can be projected back to the linear-scale domain using its pseudo-inverse [Lv *et al.*, 2024] based on the range-null decomposition (RND) theory, which provides strong structural information of the target. Our vocoding goal is to reconstruct ground-truth spectrograms from mel-spectrograms, addressing both the spectral compression and phase information problems. According to our rank analysis, the mel-domain conversion and reversion process tends to decrease the spectral rank, necessitating that the neural vocoding task increases the spectral rank to restore clean speech. In contrast, the speech denoising task exhibits an opposite trend. Therefore, this work offers a novel perspective to bridge the connection between waveform generation and the commonly used restoration techniques in speech enhancement[Lei *et al.*, 2025b]. Additionally, the multi-period discriminator [Kong *et al.*, 2020] and multi-resolution spectrogram discriminator [Won *et al.*, 2021] are employed to further improve the generation quality.

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 the 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 SE 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 $\{X, S, N\} \in \mathbb{C}^{T\times F}$ are the mixture, target, and noise signals; t and f index time and frequency.

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

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

where $\mathcal{A} \in \mathbb{R}^{F\times F_{mel}}$ is the linear mel filter, with $F_{mel} \ll F$ for compression. This transform discards phase and linearly compresses 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 $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$ is the range-space component and $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}$ is the null-space component. 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 vocoding 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 vocoding task. Recall that in the classical compressive sensing (CS) field [Zhang and 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. Next, we delve into the analysis from the perspective of the matrix rank.

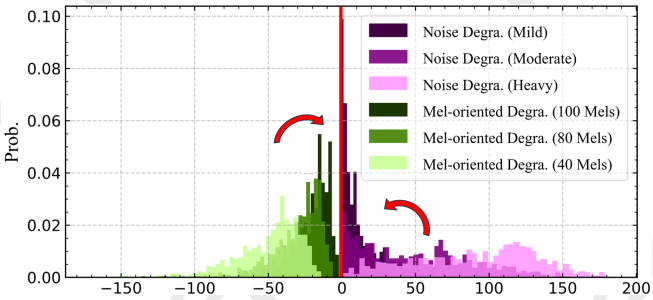

Figure 2: Relative rank difference with respect to the target spectrum for denoising and vocoding tasks. The ranks are calculated from the test set of the VoiceBank-DEMAND dataset. An absolute threshold η of 0.5 is set for rank calculation.

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:

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

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 the zero matrix. Mapping $S^{\dagger}$ to S is a restoration task similar to speech denoising, but while denoising (additive degradation) may increase spectral rank, vocoding (compression) reduces it. We illustrate these spectral rank changes below, defining $\mathcal{R}(\cdot) : \mathbb{R}^{T\times F} \to \mathbb{Z}$ as the matrix rank operation. By basic 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 calculate the rank difference between the degraded and target spectrum, defined 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 vocoding 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 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 operations of SB and the model’s training objectives. Finally, we describe the loss functions used in training.

3.1 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:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t, \quad \boldsymbol{x}_0 \sim p_0 = p_{\text{data}}, \tag{10}$$

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:

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{f}(\boldsymbol{x}_t, t) - g^2(t)\nabla \log p_t(\boldsymbol{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\boldsymbol{w}}_t, \quad \boldsymbol{x}_T \sim p_T \approx p_{\text{prior}}, \tag{11}$$

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] :

$$\mathbb{E}_{p_0(\boldsymbol{x}_0)p_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0),t}\left[\|s_\theta(\boldsymbol{x}_t, t) - \nabla \log p_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2\right], \tag{12}$$

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 $\boldsymbol{f}(\boldsymbol{x}_t, t) = f(t)\boldsymbol{x}_t$.

3.2 Schrödinger Bridge

The SB problem [Schrödinger, 1932; Bortoli *et al.*, 2021] originates from the optimization of path measures with constrained boundaries. For vocoding task, we define the target

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

Table 1: Demonstration of the noise schedules in BridgeVoC.

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:

$$\min_{p\in\mathcal{P}_{[0,T]}} D_{\mathrm{KL}}(p \parallel p_{\mathrm{ref}}), \quad s.t.\, p_0 = p_{\mathrm{S}},\ p_T = p_{\mathrm{S}^{\dagger}}, \tag{13}$$

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. (10), the SB problem is equivalent to a couple of forward-backward SDEs [Wang *et al.*, 2021; Chen *et al.*, 2022]:

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{f}(\boldsymbol{x}_t, t) + g^2(t)\nabla \log \Psi_t(\boldsymbol{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t, \quad \boldsymbol{x}_0 \sim p_{\mathrm{S}}, \tag{14}$$

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{f}(\boldsymbol{x}_t, t) - g^2(t)\nabla \log \widehat{\Psi}_t(\boldsymbol{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\boldsymbol{w}}_t, \quad \boldsymbol{x}_T \sim p_{\mathrm{S}^{\dagger}}, \tag{15}$$

where $\boldsymbol{f}$, g and $\boldsymbol{w}_t$ are from the forward SDE in Eq. (10). 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].

3.3 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 $p_{\mathrm{S}} = \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}_0, \epsilon_0^2\mathbf{I})$ 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 $\boldsymbol{x}_1$:

$$\widehat{\Psi}_t = \mathcal{N}_{\mathbb{C}}(\alpha_t\boldsymbol{x}_0, \alpha_t^2\sigma_t^2\boldsymbol{I}), \Psi_t = \mathcal{N}_{\mathbb{C}}(\bar{\alpha}_t\boldsymbol{x}_1, \alpha_t^2\bar{\sigma}_t^2\boldsymbol{I}), \tag{16}$$

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 $\boldsymbol{f}$ 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:

$$p_t = \Psi_t\widehat{\Psi}_t = \mathcal{N}\left(\frac{\alpha_t\bar{\sigma}_t^2\boldsymbol{x}_0 + \bar{\alpha}_t\sigma_t^2\boldsymbol{x}_1}{\sigma_1^2}, \frac{\alpha_t^2\bar{\sigma}_t^2\sigma_t^2}{\sigma_1^2}\boldsymbol{I}\right). \tag{17}$$

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 $\Delta\beta = \beta_1 - \beta_0$.

3.4 Loss Function

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, 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 can effectively improve the generation quality.

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

$$\mathcal{L}_d = \frac{1}{M}\sum_{m=1}^{M}\max\left(0, 1 - D_m\left(\mathbf{s}\right)\right) + \max\left(0, 1 + D_m\left(\tilde{\mathbf{s}}\right)\right), \tag{19}$$

$$\mathcal{L}_g = \frac{1}{M}\sum_{m=1}^{M}\max\left(0, 1 - D_m\left(\tilde{\mathbf{s}}\right)\right), \tag{20}$$

where $\tilde{\mathbf{s}} = \mathrm{iSTFT}(\tilde{\mathbf{S}}) \in \mathbb{R}^L$ denotes the reconstructed waveforms, $\mathrm{iSTFT}\left(\cdot\right)$ refers to the iSTFT operation, and M is the number of sub-discriminators. Discriminators includes multi-period discriminator [Kong *et al.*, 2020] and multi-resolution spectrogram discriminator [Won *et al.*, 2021; Lei *et al.*, 2025a]. Besides, the feature matching loss is also utilized:

$$\mathcal{L}_{fm} = \frac{1}{LM}\sum_{l,m}|\mathbf{f}_l^m\left(\tilde{\mathbf{s}}\right) - \mathbf{f}_l^m\left(\mathbf{s}\right)|, \tag{21}$$

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

$$\mathcal{L}_B = \mathcal{L}_{mse} + \lambda_{mel}\mathcal{L}_{mel} + \lambda_g\mathcal{L}_g + \lambda_{fm}\mathcal{L}_{fm}, \tag{22}$$

where λ_{mel}, λ_g, and λ_{fm} are the are the weight hyperparameters of corresponding loss. Detailed settings can be found in [Lei *et al.*, 2025c].

4 Experiments

4.1 Datasets

Two benchmarks are used in this study: LJSpeech [Keith and Linda, 2017] and LibriTTS [Heiga *et al.*, 2019]. LJSpeech contains 13,100 clean speech clips from a single female speaker at 22.05 kHz, partitioned into 12,500/100/500 clips for training, validation, and testing, following the VITS

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 2: 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 3: Ablation study of the signal reconstruction methods and net sizes on the LJSpeech benchmark.

repository. LibriTTS, sampled at 24 kHz, includes diverse recording conditions; we use the {*train-clean-100*, *train-clean-300*, *train-other-500*} subsets for training, *dev-clean+dev-other* for objective evaluation, and *test-clean+test-other* for subjective evaluation, as in [Lee *et al.*, 2023].

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. More ablation studies are conducted and can be found in the supplementary material.

For the weight hyperparameters in Eq. (22), 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. (22).

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.

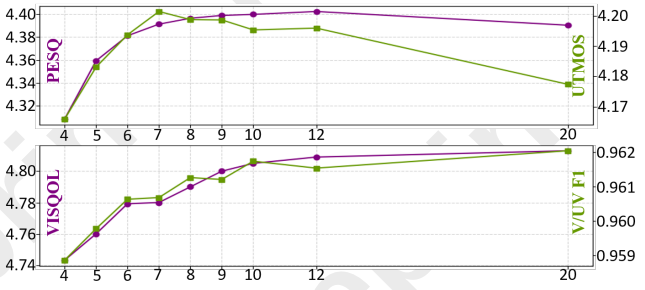

Figure 3: Metrics with different numbers of sampling steps during the reverse process on the test set of the LJSpeech dataset.

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 and 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	4.065	0.9782	**0.9716**	**4.863**	**4.296**	**0.0838**	20.69	34.43
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	0.9603	4.743	4.015	0.1118	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	19.46	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	4.824	4.262	0.1136	**15.04**	**32.72**

Table 4: 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 underlined.

Models	PESQ↑	Periodicity↓ RMSE	V/UV↑ F1	Pitch↓ RMSE	VISQOL↑
WaveGlow-256†	3.138	0.1485	0.9378	-	-
HiFiGAN-V1	3.056	0.1671	0.9212	52.53	4.721
iSTFTNet-V1	2.880	0.1672	0.9177	53.07	4.655
UnivNet-c32†	3.277	0.1305	0.9347	41.51	4.753
Avocodo	3.217	0.1611	0.9134	51.60	4.762
BigVGAN-base(1M steps)†	3.519	0.1287	0.9459	-	-
BigVGAN(1M steps)†	4.027	0.1018	0.9598	-	-
BigVGAN-base(5M steps)†	3.841	0.1073	0.9540	32.54	4.907
BigVGAN(5M steps)†	4.269	**0.0790**	**0.9670**	24.28	**4.963**
APNet	2.897	0.1586	0.9265	39.66	4.666
APNet2	2.834	0.1529	0.9227	46.37	4.582
Vocos†	3.615	0.1146	0.9484	35.58	4.879
PriorGrad	4.043	0.1277	0.9435	28.34	4.381
FreGrad	3.793	0.1443	0.9309	39.88	4.337
BridgeVoC-base(ours)	4.419	0.1021	0.9584	17.84	4.908
BridgeVoC(ours)	**4.459**	0.0980	0.9609	**14.89**	4.914

Table 5: 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↑	V/UV↑ F1	Pitch↓ RMSE	VISQOL↑	MUSHRA
Ground Truth	-	-	-	-	89.61±0.62
HiFiGAN-V1	3.090	0.9428	33.29	4.723	72.47±1.07
Vocos	3.684	0.9649	23.46	4.866	75.77±1.24
BigVGAN-base(5M steps)†	3.859	0.9649	28.85	4.893	80.23±0.99
BigVGAN(5M steps)†	4.282	**0.9722**	20.32	**4.958**	82.78±0.81
PriorGrad	3.911	0.9323	19.56	4.278	77.53±1.10
FreGrad	3.653	0.9268	27.93	4.201	78.06±1.11
BridgeVoC-base(ours)	4.323	0.9463	19.31	4.855	82.15±0.93
BridgeVoC(ours)	**4.334**	0.9473	**18.31**	4.863	*83.34±1.02

Table 6: 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).

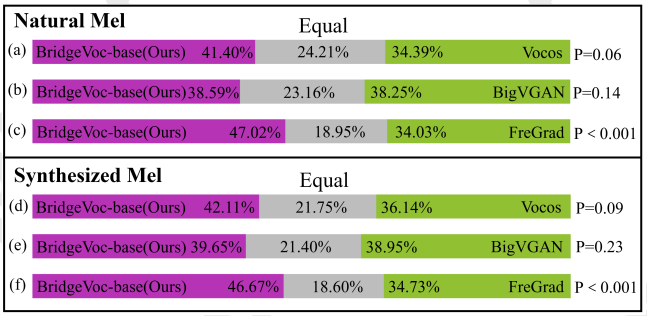

Figure 4: Average preference scores (in %) of ABX tests between BridgeVoC-base and two other baselines. (a)-(c) Mel-spectrograms are obtained from natural speech clips in the LibriTTS test set. (d)-(f) Mel-spectrograms are synthesized from F5-TTS [Chen *et al.*, 2024], where the transcripts are from the LibriTTS test set.

Ablation studies

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

Table 2 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 3 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. In addition, increasing the size of the network also improves the final output scores.

For the case of "gmax" / "+mmel" / "map" / "16.2M", Figure 3 shows the results of the number of reverse sampling

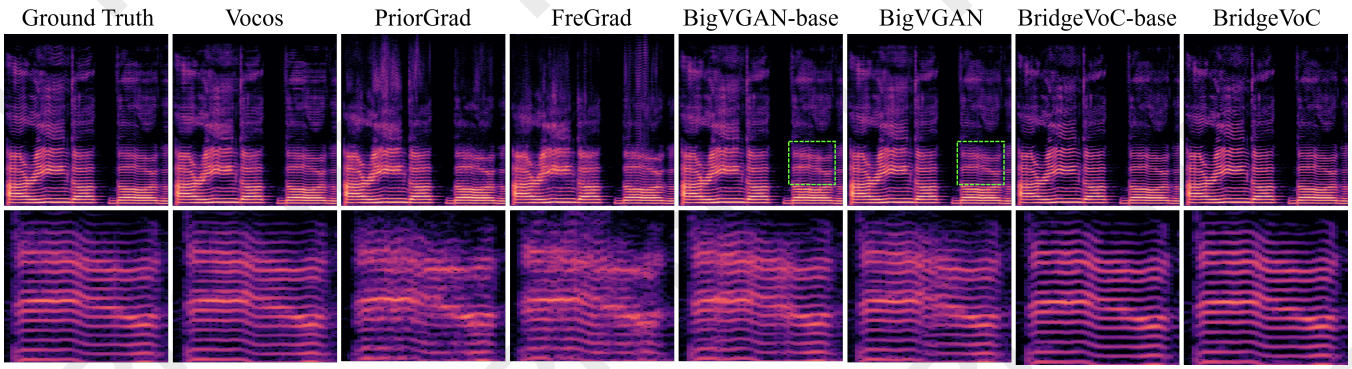

Figure 5: Spectral visualization of different vocoder methods. The audio clip is a singing voice from the MUSHDB18 test set.

steps ablations. We observe that increasing the number of steps improves some metrics, while others peak at a specific step count, consistent with findings in other diffusion-based studies [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, suggesting that BridgeVoC can further lower computational cost and speed up inference.

Comparisons with SoTA methods

Tables 4 and 5 present objective comparisons on the LJSpeech and LibriTTS datasets, revealing 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 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 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 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, validating the effectiveness of the proposed approach.

Table 6 presents the results on the out-of-domain test set. Compared to Table 5, 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.

The preference scores are shown in Figure 4. 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 higher subjective scores compared to FreGrad.

Figure 5 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.

5 Conclusions

In this paper, we introduce a novel time-frequency (T-F) domain-based diffusion neural vocoder that effectively bridges the gap between the data-to-data Schrödinger Bridge framework and range-null decomposition 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. By incorporating generative adversarial networks and optimizing various hyperparameters, our method achieves promising results in both objective and subjective evaluations. Extensive experiments on the LJSpeech and LibriTTS benchmarks demonstrate the efficacy and superiority of the proposed approach.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 12274221) and the Yangtze River Delta Science and Technology Innovation Community Joint Research Project (Grant No. 2024CSJGG1103).

References

[Ai and Ling, 2023] 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.

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

[Bortoli *et al.*, 2021] 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, pages 17695–17709. Curran Associates, Inc., 2021.

[Bunne *et al.*, 2023] 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*, pages 5802–5833. PMLR, 25–27 Apr 2023.

[Chen *et al.*, 2021] 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.

[Chen *et al.*, 2022] T. Chen, G. Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022.

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

[Chen *et al.*, 2024] 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.

[Choi *et al.*, 2021] 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.

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

[Du *et al.*, 2024] 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*, pages 66–80, 2024.

[Heiga *et al.*, 2019] 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.

[Hines *et al.*, 2015] A. Hines, J. Skoglund, and A. Kokaram. ViSQOL: an objective speech quality model. *EURASIP J. Audio Speech Music Process.*, pages 1–18, 2015.

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

[Huang *et al.*, 2023] 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*, pages 13916–13932. PMLR, 23–29 Jul 2023.

[Hubert, 2024] S. Hubert. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *Proc. ICLR*, 2024.

[Hwang *et al.*, 2025] 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.

[Kawahara *et al.*, 1999] 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.

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

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

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

[Kong *et al.*, 2020] 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, editors, *Proc. NeurIPS*, volume 33, pages 17022–17033. Curran Associates, Inc., 2020.

[Kong *et al.*, 2021] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *Proc. ICLR*, 2021.

[Kraft and Zölzer, 2014] 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.

[Laurent *et al.*, 2017] D. Laurent, S. Jascha, and B. Samy. Density estimation using real NVP. In *Proc. ICLR*, 2017.

[Lee *et al.*, 2022] 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.

[Lee *et al.*, 2023] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon. BigVGAN: A Universal Neural Vocoder with Large-Scale Training. In *Proc. ICLR*, 2023.

[Lei *et al.*, 2025a] Tong Lei, Qinwen Hu, Zhongshu Hou, and Jing Lu. Enhancing real-world far-field speech with supervised adversarial training. *Applied Acoustics*, 229:110407, 2025.

[Lei *et al.*, 2025b] Tong Lei, Qinwen Hu, Ziyao Lin, Andong Li, Rilin Chen, Meng Yu, Dong Yu, and Jing Lu. Fnsesbgan: Far-field speech enhancement with schrodinger bridge and generative adversarial networks. *arXiv preprint arXiv:2503.12936*, 2025.

[Lei *et al.*, 2025c] Tong Lei, Andong Li, Rilin Chen, Dong Yu, Meng Yu, Jing Lu, and Chengshi Zheng. Bridgevoc: Insights into using schrödinger bridge for neural vocoders. In *ICLR 2025 DeLTa Workshop*, 2025.

[Liu *et al.*, 2022a] H. Liu, W. Choi, X. Liu, Q. Kong, et al. Neural Vocoder is All You Need for Speech Superresolution. In *Proc. Interspeech*, pages 4227–4231, 2022.

[Liu *et al.*, 2022b] H. Liu, X. Liu, Q. Kong, Q. Tian, et al. VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration. In *Proc. Interspeech*, pages 4232–4236, 2022.

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

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

[Majumder *et al.*, 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, page 564–572, 2024.

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

[Meinard, 2015] M. Meinard. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

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

[Nguyen *et al.*, 2024] 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*, pages 10736–10740, 2024.

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

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

[Qian *et al.*, 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*, pages 5210–5219. PMLR, 2019.

[Rafii *et al.*, 2017] Z. Rafii, A. Liutkus, F. Stöter, S. Mimilakis, and R. Bittner. The MUSDB18 corpus for music separation. 2017.

[Rec, 2005] 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.

[Ren *et al.*, 2019] 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.

[Saeki *et al.*, 2022] 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.

[Schrödinger, 1932] 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, pages 269–310, 1932.

[Song *et al.*, 2021] 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.

[Taal *et al.*, 2011] 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.

[Tan *et al.*, 2024] 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.

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

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

[Wang *et al.*, 2021] 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*, pages 10794–10804. PMLR, 18–24 Jul 2021.

[Wang *et al.*, 2023] 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.

[Won *et al.*, 2021] J. Won, C. Daniel, and Y. Jaesam. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. Interspeech*, 2021.

[Yamagishi, 2012] J. Yamagishi. English multi-speaker corpus for cstr voice cloning toolkit, 2012.

[Zhang and Ghanem, 2018] J. Zhang and B. Ghanem. Istanet: Interpretable optimization-inspired deep network for image compressive sensing. In *Proc. CVPR*, pages 1828–1837, 2018.