DegVoC: Revisiting Neural Vocoder from a Degradation Perspective

Andong Li[1,2], Tong Lei[3], Lingling Dai[1,2], Kai Li[4], Rilin Chen[3], Meng Yu[3], Xiaodong Li[1,2], Dong Yu[3], Chengshi Zheng[1,2]*

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Tencent AI Lab
[4]Tsinghua University, Beijing,China
cszheng@mail.ioa.ac.cn

Abstract

Existing neural vocoders have demonstrated promising performance by leveraging Mel-spectrum as an acoustic feature for conditional audio generation. Nonetheless, they remain constrained by an inherent "performance-cost" dilemma that significantly hinders the development of this field. This paper revisits this foundational task from a novel degradation perspective, where Mel-spectrum is regarded as a special signal degradation process from the target spectrum. Drawing inspiration from traditional sparse signal recovery problems, we propose DegVoC, a GAN-based neural vocoder with a two-step solution procedure. First, by exploiting degradation priors, we attempt to retrieve the initial spectral structure from Mel-domain representations as an initial solution via a simple linear transformation. Based on that, we introduce a deep prior solver that accounts for the heterogeneous distribution of sub-bands in the time-frequency domain. A convolution-style attention module with a large kernel size is specially devised for efficient inter-frame and inter-band contextual modeling. With 3.89 M parameters and substantially reduced inference complexity, DegVoC achieves state-of-the-art performance across objective and subjective evaluations, outperforming existing GAN-, DDPM- and flow-matching-based baselines.

Introduction

Thanks to the proliferation of large language models (LLMs) and generative diffusion models, burgeoning progress has been made in the field of audio generation (Shen et al. 2023; Kondratyuk et al. 2024). The audio vocoder is regarded as one of the fundamental techniques, where target time-domain waveforms are reconstructed from acoustic features via electronic and computational methods (Dudley 1939; Kawahara, Masuda-Katsuse, and De Cheveigne 1999). Compared with traditional digital signal processing (DSP)-based vocoders like STRAIGHT (Kawahara 2006) and WORLD (Morise, Yokomori, and Ozawa 2016), neural vocoders have demonstrated remarkable performance in both generation quality and naturalness, thus attracting extensive research attention in recent years (Dudley 1939; Siuzdak 2024; Li et al. 2025c,a).

*The corresponding author.

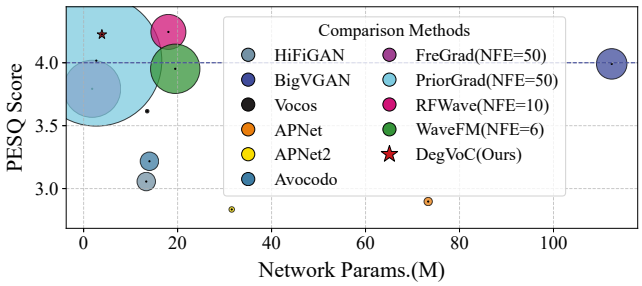


Figure 1: A case example on the LibriTTS benchmark. A larger bubble/star indicates higher inference cost (formatted in MACs). NFE denotes the number of function evaluations.

Early neural vocoders typically generate waveforms in an autoregressive manner (Van Den Oord et al. 2016). Despite significant progress, they exhibit rather slow inference speed, which can heavily hinder their adoption in practical applications (Kalchbrenner et al. 2018). Subsequent advancements have been made by incorporating normalization flow (Luong and Tran 2021) and glottis (Juvela et al. 2019) techniques. Nonetheless, they can still suffer from inefficient inference and limited quality improvements. Later, generative adversarial network (GAN)-based neural vocoders have become mainstream due to their balance between quality and efficiency, falling into two categories: time-domain methods (Yang et al. 2021; Kong, Kim, and Bae 2020; Lee et al. 2023) and time-frequency (T-F) domain methods (Siuzdak 2024; Ai and Ling 2023). More recently, generative diffusion-based vocoders have begun to gain more attention due to the powerful capability of diffusion models in computer vision and multi-modality generation fields. By establishing a bijective Markov chain between the Gaussian and audio signal distributions, target audio can be gradually estimated via an iterative denoising function in the reverse process (Kong et al. 2021; Lee et al. 2022). Further performance improvement has been achieved by incorporating flow matching (FM) and its variants (Liu, Dai, and Wu 2025; Luo, Miao, and Duan 2025).

Although existing neural vocoders have achieved exceptional performance, they still face an inherent *Performance-Cost Dilemma*, which stems from two core factors. First, the vocoder task is usually formulated as a **conditional gener-**

ation task, where Mel-spectrum serves as the acoustic clue for target generation from a random Gaussian distribution. While intuitively challenging, a large number of network parameters are often required for accurate acoustic modeling (Lee et al. 2023). Second, although diffusion models can potentially achieve a higher upper-bound, multiple reverse sampling steps inevitably lead to an inefficient inference process and high computational costs, limiting their deployment in edge scenarios (Liu, Dai, and Wu 2025). Therefore, a natural question arises: *How can we elegantly break this embarrassing dilemma?*

Fortunately, we have uncovered the "key" by delving into the degradation prior of the Mel representation and the hierarchical distribution prior of the target T-F spectrum along the frequency dimension. Specifically, we re-examine Mel modeling from the perspective of signal degradation and formulate the generation process as an inverse optimization problem, *i.e.*, **a classical signal retrieval task** rather than target generation from scratch. Furthermore, considering the heterogeneous distribution among different frequency regions, we propose an uneven sub-band division and merging strategy, and a large-kernel convolutional attention module (LKCAM) is devised to model inter-frame and inter-band dependencies, leveraging large receptive fields for long-term temporal-spectral context. Based on that, a novel GAN-based neural vocoder named **DegVoC** is proposed. It not only offers high reconstruction quality and low computational complexity but also enables fast inference speed. Extensive experiments show that with only **3.89 M** parameters, it outperforms BigVGAN-112 M in both objective and subjective performance. Moreover, with only **7.6%** of the computational complexity and approximately **8.3**$\times$ speed-up, it even surpasses RFWave (Liu, Dai, and Wu 2025), a recent state-of-the-art (SoTA) method based on FM. To our best knowledge, this is the first time the degradation perspective of the Mel spectrum has been revealed, and a GAN-based neural vocoder has achieved SoTA performance under such a light-weight configuration. Figure 1 showcases the PESQ comparison of different methods on the LibriTTS benchmark. Our contributions can be summarized as follows:

❑ We revisit the vocoder task and reformulate the generation process as a classical signal restoration task.

❑ We propose a novel light-weight model that enables both high-quality waveform generation and fast inference.

❑ We conduct extensive experiments, and both objective and subjective experiments validate the superiority of the proposed method over existing GAN and diffusion baselines.

Related Works

Autoregressive Models: The WaveNet (Van Den Oord et al. 2016) was proposed to model the generation probability of the current sample based on previous samples. Dilated convolutions were stacked to gradually expand the receptive field. In (Valin and Skoglund 2019), the LPCNet was proposed, using linear prediction coefficients (LPCs) as auxiliary input to alleviate the network modeling burden. Despite the improvements over traditional DPS-based vocoders, they still suffer from inefficient generation speed due to the inherent sample-level modeling mechanism.

GAN-based Models: Existing GAN-based vocoders can be simply classified into two types. For time-domain-based, the canonical MelGAN (Kumar et al. 2019) was proposed by incorporating residual blocks as the generator and multi-scale discriminators. Subsequently, the HiFiGAN was proposed by designing multi-period discriminators (MPDs) to enhance generation quality. In (Lee et al. 2023), a periodic-based activation function was designed and the network is scaled up to 112 M to further strengthen waveform generation quality. Due to the decoupling characteristic of different frequency components in the Fourier representation, T-F domain-based vocoders have received increasing attention recently. In (Kaneko et al. 2022), the iSTFTNet works by early-stopping and transforming to the waveform through the inverse short-time Fourier transform (iSTFT). In (Siuzdak 2024), stacked ConvNext blocks are adopted to directly model the target magnitude and phase spectra. Further improvements were achieved by incorporating the dual-path branches and modified ConvNext blocks (Du et al. 2023).

Diffusion-based Models: Recently, diffusion models have been applied to audio generation. In (Kong et al. 2021), denoising diffusion probability models (DDPMs) were adopted, where target waveforms are gradually recovered through multiple iterations in the reverse stage. In (Lee et al. 2022), a data-dependent adaptive prior was explored to structure the prior distribution, yielding remarkable performance with fewer number of function evaluations (NFEs). Most recently, FM has attracted substantial attention due to its simple formulation and inference paradigm. In (Lee, Choi, and Lee 2025), natural periodic features of waveform signals were introduced, and a period-aware flow matching estimator was used to capture multi-periodic features. Meanwhile, Liu *et al.* modeled different sub-band distributions in the T-F domain using rectified flow with 10 sampling steps (Liu, Dai, and Wu 2025).

Revisiting Audio Vocoder Task

Formally, the log-scale Mel-spectrum $\mathbf{X}^{mel}$ can be defined with the following physical model:

$$\mathbf{X}^{mel} = \log\left(\mathcal{A}\left|\mathbf{S}\right|\right), \tag{1}$$

where $|\mathbf{S}| \in \mathbb{R}^{F\times F}$ denotes the magnitude spectrum of the target waveform in the T-F domain, and $\mathcal{A} \in \mathbb{R}^{F_m\times T}$ is the Mel filter matrix, instantiated as a linear compression matrix. $\{F, F_m, T\}$ represent the frequency, mel-frequency, and frame sizes, respectively. Absorbing the logarithm operation into the left-hand side, we obtain:

$$\mathbf{Y} = \exp\left(\mathbf{X}^{mel}\right) = \mathcal{A}\left|\mathbf{S}\right|. \tag{2}$$

Compared with the target complex spectrum $\mathbf{S} \in \mathbb{C}^{F\times T}$, the Mel-spectrum $\mathbf{Y}$ undergoes two key degradations: ① Loss of phase information; ② Linear magnitude compression. Consequently, the process from Mel to target can be regarded as the corresponding inverse process, *i.e.*, phase retrieval and magnitude recovery, which are considered as classical signal restoration problems (Peer, Welker, and Gerkmann 2023; Baraniuk et al. 2010; Dai et al. 2025).

To bridge $\mathbf{Y}$ and $\mathbf{S}$, we generalize the Mel-spectrum to the complex domain by treating magnitude as a special case

in the complex domain, *i.e.*, $\underline{\mathbf{Y}} = \mathbf{Y}\exp(j\mathbf{0})$, where $\mathbf{0}$ is an all-zero phase spectrum. Besides, we introduce a residual term $\mathbf{E}$, which reflects the phase gap between the target spectrum and that of all-zero phase, *i.e.*, $\mathbf{E} = |\mathbf{S}|\exp(j\mathbf{0}) - \mathbf{S}$, then we rewrite Eq.(2) as:

$$\underline{\mathbf{Y}} = \mathcal{A}(\mathbf{S} + |\mathbf{S}|\exp(j\mathbf{0}) - \mathbf{S}) = \mathcal{A}\mathbf{S} + \mathbf{E}^{'}, \quad (3)$$

where $\mathbf{E}^{'} = \mathcal{A}\mathbf{E} \in \mathbb{C}^{F_m \times T}$ denotes the compressed residual in the Mel domain. Figure 2 visualizes the Mel spectrum and the compressed residual term $\mathbf{E}^{'}$. One can observe that $\mathbf{E}^{'}$ shares a similar structural pattern to the Mel spectrum, which can be explained as it mainly reflects the phase difference and is weighted by the target spectral magnitude. Notably, Eq.(3) can be regarded as a classical optimization problem for signal recovery, where we aim to restore $\mathbf{S}$ from the degraded observation $\underline{\mathbf{Y}}$. Using maximum a posteriori (MAP), the optimization formulation for $\mathbf{S}$ is:

$$\log P(\mathbf{S}|\underline{\mathbf{Y}}) \propto \log P(\underline{\mathbf{Y}}|\mathbf{S}) + \log P(\mathbf{S}), \quad (4)$$

where $\log P(\underline{\mathbf{Y}}|\mathbf{S})$ is the log-likelihood and $\log P(\mathbf{S})$ is the log-prior of the target spectrum. Assuming the residual term $\mathbf{E}^{'}$ follows a zero-mean time-varying complex Gaussian (TVCG) distribution (Dong et al. 2023; Li et al. 2024), we rewrite Eq.(4) as:

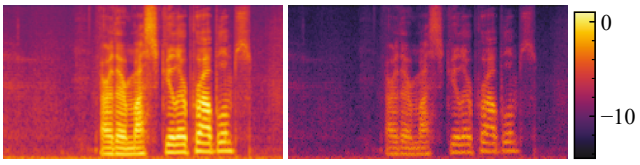
$$\mathbf{S}^{\star} = \arg\min_{\mathbf{S}} \frac{1}{\sigma_t^2} \|\underline{\mathbf{Y}} - \mathcal{A}\mathbf{S}\|_2^2 + \alpha\mathcal{G}(\mathbf{S}), \quad (5)$$

where σ_t is the time-varying variance and $\mathcal{G}(\cdot)$ is a regularization function for $\mathbf{S}$. While such an optimization problem is common in traditional sparse signal estimation field such as compressive sensing (CS) (Baraniuk et al. 2010), they usually adopt a recursive solution following such a two-stage paradigm:

- *Initialization step*: Recovering basic structure characteristic from the degraded observation via a simple linear operation, *e.g.*, matrix transpose or pseudo-inverse.
- *Alternating update step*: Further restoring the signal details depending on the predefined prior assumption of $\mathcal{G}(\cdot)$, *e.g.*, L_p sparseness, which corresponds to the traditional ISTA algorithm (Beck and Teboulle 2009). And some optimization methods like proximal gradient descent (PGD) (Zhang et al. 2022) or ADMM (Afonso, Bioucas-Dias, and Figueiredo 2011) are adopted for iterative estimation.

Recently, the employment of DNNs has notably facilitated the development of signal reconstruction by modeling the prior term in a data-driven manner, that is, instead of the manual prior assumption, a neural network serves as the deep regularizer to learn the complicated prior of the target signals (Zhang and Ghanem 2018). Inspired by that, we reformulate the solution of $\mathbf{S}$ into two sub-procedures:

- **Initialization solver**: By virtue of the linear degradation prior, we can transform the Mel-spectrum into linear T-F domain as the initialization solution.
- **Deep prior solver**: By feasibly excavating the target prior by a specially devised neural network, we further recover the remaining spectral details.



(a) Mel spectrum (b) Compressed residual $\mathbf{E}^{'}$

Figure 2: Spectral visualizations of the Mel spectrum and the compressed residual term $\mathbf{E}^{'}$. The logarithm operation is adopted for better visualization.

Notably, unlike conventional conditional generation paradigm, we frame the vocoder task as optimization from degraded observations, following traditional signal recovery pipelines. One may notice that a similar pseudo-inverse operation was mentioned in (Lv et al. 2024). The main difference lies in that here we invest it from a feasible optimization perspective while it is adopted as an empirical trick in (Lv et al. 2024). More optimization analysis are provided in **Appendix**.

Proposed Framework

Based on the above analysis, we propose a novel neural vocoder framework named **DegVoC**, which inherits the two-step optimization pipeline, as shown in Figure 3(a). In the first step, we attempt to recover the basic spectral structure in the linear T-F scale. And a deep prior solver is devised to effectively restore the remaining spectral details. We introduce the two procedures in the following two parts.

Initialization Solver

Leveraging the linear degradation prior, we can simply retrieve the basic spectral structure, and three choices are provided for magnitude initialization:
(1) Matrix Transpose: Motivated by the common scheme in the CS field, we view $\mathcal{A}$ as a type of sampling matrix (You et al. 2021) and use the transpose of $\mathcal{A}$ for initialization, *i.e.*, $\mathcal{A}^{\mathsf{T}}\mathbf{Y}$, where $(\cdot)^{\mathsf{T}}$ denotes the transpose operation.
(2) Matrix Pesudo-inverse: Since F_m usually meets $F_m \ll F$, it is nearly impossible to restore the spectrum perfectly, we use the pseudo-inverse as alternative, *i.e.*, $\mathcal{A}^{\dagger}\mathbf{Y}$, where $\mathcal{A}^{\dagger} \in \mathbb{R}^{F \times F_m}$ is the pseudo-inverse of $\mathcal{A}$. In Appendix, we show that such a operation is the range-space representation of the target spectral magnitude.
(3) Learnable: In the previous two strategies, fixed matrix is adopted for recovery. To adapt to the model optimization, based on (1) and (2), we set the matrix weights to be learnable and it can thus be optimized together with the deep prior solver jointly.

For phase initialization, we simply set it to all-zero, and observe it is adequate to achieve good performance in our internal experiments.

Deep Prior Solver

Previous literature usually adopt full-band modules (*e.g.*, ResNet (He et al. 2016) or ConvNext (Liu et al. 2022)

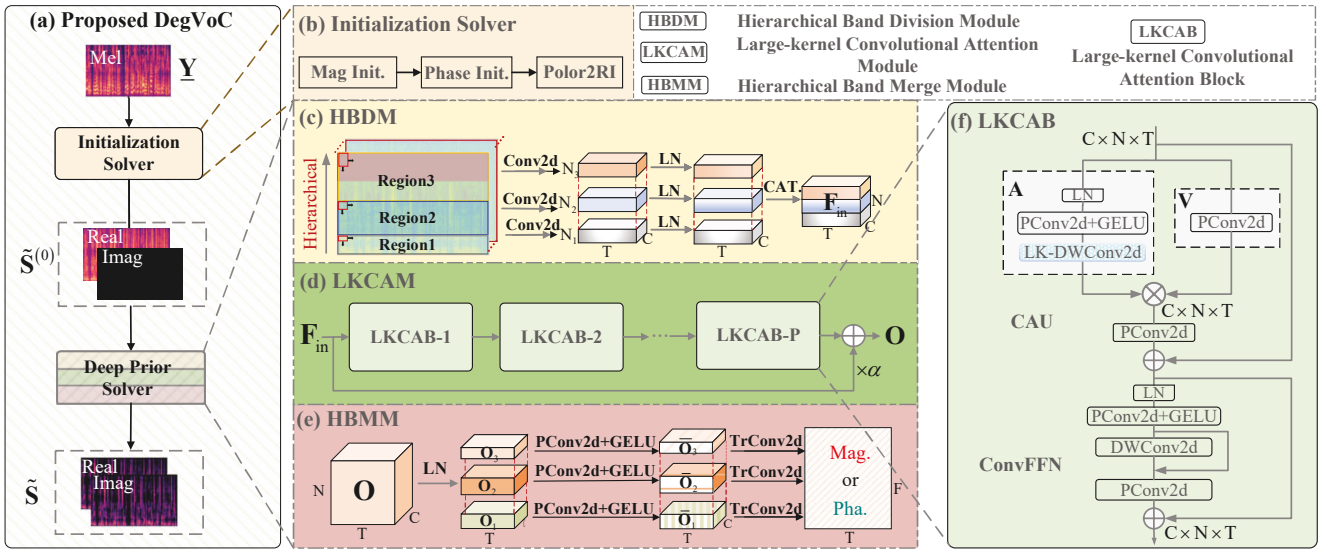


Figure 3: (a) Overall diagram of the proposed DegVoC, which follows a two-step optimization tactic. First, the initialization solver is utilized to recover the overall spectral structure of the spectrum from the Mel domain representation. In the deep prior solver, we leverage the heterogeneous distribution prior of different sub-bands and devise a novel network, where the sub-bands are hierarchically encoded and interacted via large convolution kernels to facilitate the effective restoration of target spectrum. (b) Internal structure of the initialization solver. (c) Internal structure of the HBDM. (d) Internal structure of the LKCAM. (e) Internal structure of the HBMM. (f) Internal structure of the LKCAB. Different modules are indicated with different colors.

blocks) for target spectrum estimation, which neglects the hierarchical prior of the spectrum in the T-F domain. For example, harmonic components usually concentrate on low- and mid-frequency regions and foundation frequency (F_0) usually $\leq$ 500 Hz. Therefore, we contend that it seems more reasonable to encode and model the sub-band distributions, and similar concept has also been proposed more recently (Liu, Dai, and Wu 2025).

As shown in Figure 3(a), the deep prior solver consists of three parts: hierarchical band division module (HBDM), large kernel convolutional attention module (LKCAM), and hierarchical band merge module (HBMM), which are illustrated below. Note that in contrast to previous recursive-based literature (Masuyama et al. 2019; Liu et al. 2024), only one iteration step is employed, and reasons are two-fold. First, recursive estimation can incur inefficiency in the inference stage. Second, we argue that by elaborate network design, amendable performance can be obtained by one-pass.

HBDM/HBMM Detailed structure of the HBDM is shown in Figure 3(c). Given the output of the initialization solver, *i.e.*, $\tilde{\mathbf{S}}^{(0)} \in \mathbb{C}^{F\times T}$, we first convert it into the real-valued version by concatenating along the channel axis:

$$\underline{\tilde{\mathbf{S}}}^{(0)} = \text{Cat}\left(\mathcal{R}\left(\tilde{\mathbf{S}}^{(0)}\right), \mathcal{I}\left(\tilde{\mathbf{S}}^{(0)}\right)\right) \in \mathbb{R}^{2\times F\times T}, \quad (6)$$

where $\text{Cat}(\cdot)$ denotes the concatenation operation, and $\{\mathcal{R}, \mathcal{I}\}$ are the real and imaginary operations, respectively. Considering that harmonic components lie in the low- and mid-frequency regions, we adopt an uneven strategy for sub-band division. Concretely, totally K regions are set[1]. For the k-th region, a separate Conv2d is applied to compress the frequency size, followed by layer normalization:

$$\mathbf{F}_{in,k} = \text{LN}\left(\text{Conv2d}\left(\underline{\tilde{\mathbf{S}}}_k^{(0)}\right)\right) \in \mathbb{R}^{C\times N_k\times T}, \quad (7)$$

where $\{C, N_k\}$ denote the number of channels and the compressed sub-bands, respectively. Here C is set to 256. After that, we concatenate all the compressed representations as:

$$\mathbf{F}_{in} = \text{Cat}\left(\mathbf{F}_{in,1}, \cdots, \mathbf{F}_{in,K}\right) \in \mathbb{R}^{C\times N\times T}, \quad (8)$$

where $N = \sum_k N_k$. For spectral decoding, an opposite process is adopted, as shown in Figure 3(e). For the input $\mathbf{O} \in \mathbb{R}^{C\times N\times T}$, it is first split into K regions, and each feature region $\mathbf{O}_k$ will first pass a pointwise Conv2d (PConv2d), LN, and Gaussian error linear unit (GELU) (Hendrycks and Gimpel 2016). After that, the TrConv2d is utilized for target estimation. Note that for spectral magnitude estimation, the exponential function is adopted to ensure non-negativity, and the Atan2 $(\cdot)$ function is utilized for phase estimation.

LKCAM In the T-F domain, both inter-frame and inter-band relations are significant for spectrum estimation. On one hand, audio signals exhibit strong sequential property, and contextual or long-term information can profit the modeling of current frame (Tan, Chen, and Wang 2018). On the other hand, recent success on bandwidth extension (BWE) have revealed the modeling dependency among different frequency bands, and the generation of high frequency components can be effectively guided by the low-/mid-frequency

[1]In this paper, we empirically set $K = 3$, and $N = 24$.

Models	Type	NFE	#Param. (M)	#MACs (Giga/5s)	Inference CPU	Inference GPU	M-STFT↓	PESQ↑	MCD↓	V/UV↑ F1	Period.↓ RMSE	VISQOL↑	WER↓
HiFiGAN-V1	GAN	1	14.01	166.41	5.88×	157×	1.1039	3.056	4.284	0.921	0.167	4.721	7.89
iSTFTNet-V1		1	13.33	117.30	10.66×	299×	1.1565	2.880	4.480	0.918	0.167	4.655	7.54
Avocodo		1	14.01	166.47	5.29×	148×	1.1130	3.217	4.314	0.913	0.161	4.762	7.25
BigVGAN-base		1	14.01	166.41	2.33×	23×	0.8841	3.521	3.034	0.941	0.131	4.869	6.77
BigVGAN		1	112.39	454.08	1.54×	17×	0.8105	3.991	2.547	0.955	0.104	4.934	**6.29**
APNet		1	73.33	33.92	31.97×	499×	1.2725	2.897	4.150	0.927	0.159	4.666	7.54
APNet2		1	31.52	14.79	58.82×	749×	1.1429	2.834	4.153	0.923	0.153	4.582	7.44
Vocos†		1	13.53	**6.35**	**142.76**×	**1538**×	0.8544	3.615	3.105	0.948	0.115	4.879	6.61
FreGrad	DDPM	50	**1.82**	1589	2.32×	35×	1.3973	3.793	6.025	0.929	0.143	4.699	8.26
PriorGrad		50	2.70	8364	0.56×	8×	1.3926	4.017	5.854	0.938	0.130	4.737	7.52
RFWave†	FM	10	18.04	598	1.92×	26×	0.9552	**4.251**	2.319	0.961	0.103	4.775	6.50
WaveFM†		6	19.50	1188	0.66×	13×	0.8421	3.954	2.582	0.955	0.105	**4.943**	6.48
DegVoC(Ours)	GAN	1	3.89	45.62	18.20×	217×	**0.8028**	4.225	**2.209**	**0.966**	**0.088**	4.928	**6.29**

Table 1: Objective comparisons among different baselines on the LibriTTS benchmark. "$a\times$" denotes the speed-up ratio over real-time. The inference speed on a CPU is evaluated based on a CPU Intel(R) Core(TM) i7-14700F. For GPU, it is based on NVIDIA GeForce RTX 4060 Ti. $(\cdot)^\dagger$ denotes using the pre-trained checkpoint from the original paper for inference. ↓ means the lower the better, and ↑ means the higher the better. The best and second-best performances are respectively highlighted in **bold** and underlined, respectively. The learnable option is adopted in the initialization solver.

Models	M-STFT	PESQ	MCD	V/UV F1	VISQOL
HiFiGAN-V1	1.2102	2.907	2.733	0.876	4.570
Avocodo	1.1971	3.065	2.703	0.866	4.641
BigVGAN-base	1.0253	3.492	1.871	0.940	4.739
BigVGAN	0.8385	3.892	1.400	**0.946**	4.898
APNet2	1.2021	2.450	2.768	0.862	4.456
Vocos†	0.8975	3.522	1.949	0.937	4.848
RFWave†	0.8792	**4.154**	1.143	0.924	1.750
WaveFM†	0.8503	3.856	1.553	0.944	**4.926**
DegVoC(Ours)	**0.7688**	4.147	**1.114**	0.941	4.900

Table 2: Objective comparisons among other baselines on the out-of-distribution samples from the EARS corpus. WER is not reported due to the lack of text transcripts.

counterparts (Yun, Kim, and Lee 2025). As a remedy, we propose a novel convolution-style attention module with large kernel sizes, as shown in Figure 3(d).

Specifically, it consists of P blocks, where P is set to 8 to span a large spectral receptive field. The detailed structure of each block is presented in Figure 3(f), which consists of two parts, namely convolutional attention unit (CAU) and ConvFFN. For the first part, inspired by (Hou et al. 2024), we replace the traditional self-attention (SA) with a convolutional modulation layer. Given the input of $\mathbf{H}^{(p)} \in \mathbb{R}^{C\times N\times T}$, instead of calculating the similarity score matrix $\mathbf{A}$, we adopt a convolution operation to modulate the feature of the value $\mathbf{V}$, given by:

$$\mathbf{Z}^{(p)} = \mathbf{A} \otimes \mathbf{V}, \tag{9}$$

$$\mathbf{A} = \text{LKDWConv2d}\left(\text{GELU}\left(\text{PConv2d}\left(\text{LN}\left(\mathbf{H}^{(p)}\right)\right)\right)\right), \tag{10}$$

$$\mathbf{V} = \text{PConv2d}\left(\mathbf{H}^{(p)}\right), \tag{11}$$

where $\text{LKDWConv2d}(\cdot)$ denotes the depthwise Conv2d with large kernel (I_f, I_t) along the frequency and frame axes, respectively. Here we empirically set $\{I_f, I_t\} = \{9, 11\}$ to enable large receptive field. $\otimes$ is the elementwise multiplication. In Sec. , we further ablate the kernel size, as well as the comparisons with SA and other typical basic structures. For ConvFFN, inspired by (Zhou et al. 2023), we insert a DWConv2d with residual connection to empower detail encoding.

Loss Function

Following the settings in (Li et al. 2025b), both reconstruction and adversarial losses are adopted. For the former, we include amplitude loss $\mathcal{L}_a$, real and imaginary loss $\mathcal{L}_{ri}$, phase loss $\mathcal{L}_p$, Mel loss $\mathcal{L}_m$, and consistency loss $\mathcal{L}_c$:

$$\mathcal{L}_{rec} = \lambda_a\mathcal{L}_a + \lambda_{ri}\mathcal{L}_{ri} + \lambda_p\mathcal{L}_p + \lambda_m\mathcal{L}_m + \lambda_c\mathcal{L}_c, \tag{12}$$

where $\{\lambda_a, \lambda_{ri}, \lambda_p, \lambda_m, \lambda_c\} = \{45, 45, 100, 45, 45\}$ are the weighting coefficients.

For the adversarial loss, MPDs and multi-resolution STFT discriminators (MRDs) (Jang et al. 2021) are adopted, and the hinge loss is utilized:

$$\mathcal{L}_D = \frac{1}{M}\sum_{m=1}^{M}\max\left(0, 1 + D_m\left(\tilde{s}\right)\right), \tag{13}$$

where D_m is the m-th sub-discriminator. For the generator, the adversarial loss is given by:

$$\mathcal{L}_g = \frac{1}{M}\sum_{1}^{M}\max\left(0, 1 - D_m\left(\tilde{s}\right)\right). \tag{14}$$

Besides, we also incorporate the feature-matching loss $\mathcal{L}_{fm}$, and the final loss for generator can be rewritten as:

$$\mathcal{L}_G = \mathcal{L}_{rec} + \lambda_g\mathcal{L}_g + \lambda_{fm}\mathcal{L}_{fm}, \tag{15}$$

where $\{\lambda_g, \lambda_{fm}\} = \{1, 1\}$.

Models	GT	HiFiGAN-V1	Vocos	BigVGAN	WaveFM	RFWave	DegVoC(Ours)
AISHELL3	4.04±0.05	3.45±0.07	3.61±0.07	3.69±0.07	3.26±0.06	3.66±0.06	**3.78±0.03
MUSDB18(Vocals)	4.00±0.06	3.07±0.06	3.14±0.04	3.22±0.06	3.16±0.07	3.70±0.06	*3.77±0.06

Note: $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: MUSHRA scores among different methods on the out-of-distribution AISHELL3 and MUSDB18 benchmarks. The confidence level is 95%, and we performed a t-test comparing DegVoC and RFWave.

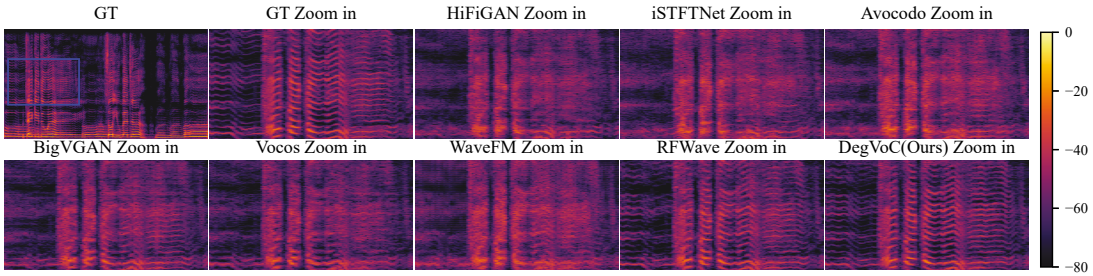


Figure 4: Spectral visualization by different methods. The audio clip is a vocal sound from the MUSDB18 test set.

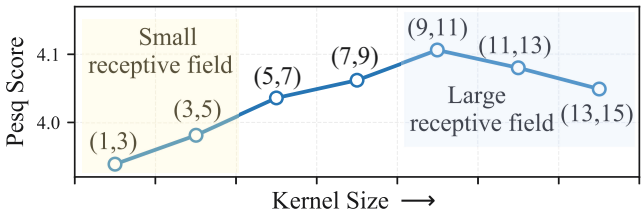


Figure 5: PESQ performance under different $\{I_f, I_t\}$ configurations. The number of LKCAB P is set to 6.

Experimental Setups

Datasets

The LibriTTS (Zen et al. 2019) benchmark is utilized for training, which covers diverse acoustic environments as well as more clips, rendering it challenging for vocoder task. The target sampling rate is 24 kHz. Following the setup in (Lee et al. 2023), {*train-clean-100*, *train-clean-360*, *train-other-500*} are utilized for training, around 208 clips from *dev-clean-other* are for objective metric evaluations.

Training Configurations

In terms of the Mel feature extraction, F_m = 100, and the upper-bound frequency f_{max} is set to 12 kHz. The window size is 1024, with 75% overlap between adjacent frames. And 1024-point FFT is adopted. For network training, the batch size is 16, and the segment size is chunked to 16,384. The learning rate is initialized at 2e-4, following a cosine scheduler for update in the epoch level. The AdamW scheme (Loshchilov and Hutter 2017) is adopted for optimization with $\{\beta_1, \beta_2\} = \{0.8, 0.99\}$. We train the model for 2 M steps, where the generator and discriminators are updated for 1 M steps, respectively.

Evaluation Metrics

Seven objective metrics are employed for evaluations: (1) Multi-resolution STFT (M-STFT) (Yamamoto, Song, and Kim 2020) evaluates the spectral distance across multiple T-F resolutions. (2) Wide-band MOS-LQO score based on the ITU-P.862.2 standard using the *python-pesq* toolkit[2], which evaluates the speech quality. (3) Mel-cepstral distance (MCD) (Kubichek 1993) measures the Mel-spectrum difference through dynamic time wrapping. (4) V/UV F1 score and periodicity RMSE (Morrison et al. 2022) evaluate the major artifacts for non-autoregressive neural vocoders. (5) Word error rate (WER), computed using OpenAI's Whisper-Base (English)[3]. (6) Virtual Speech Quality Objective Listener (VISQOL) (Hines et al. 2015) gives the MOS-LQO score by evaluating the spectro-temporal similarity.

For subjective evaluations, we conduct the MUSHRA testing based on the BeaqleJS platform (Kraft and Zölzer 2014). Thirty participants majoring in audio engineering are involved, and twenty-two valid results are collected. To alleviate the rating randomness, we scale the score range into $[1, 5]$ with the interval of 0.5. Each person should rate each audio clip in terms of overall similarity compared to the corresponding reference.

Results and Analysis

Comparisons with SoTA methods

We compare our method with 12 advanced baselines in Table 1: HiFiGAN-V1 (Kong, Kim, and Bae 2020), iSTFTNet-V1 (Kaneko et al. 2022), Avocodo (Bak et al. 2023), BigVGAN base and large versions (Lee et al. 2023), APNet (Ai and Ling 2023), APNet2 (Du et al. 2023), Vocos (Siuzdak 2024), FreGrad (Nguyen et al. 2024), PriorGrad (Lee et al. 2022), RFWave (Liu, Dai, and Wu 2025), and WaveFM (Luo, Miao, and Duan 2025).

From the quantitative results in Table 1, several important observations can be made. First, BigVGAN yields the best performance among GAN-based baselines by introducing the periodic activation function and model scaling-up,

[2]https://github.com/ludlows/PESQ/tree/master
[3]https://github.com/openai/whisper

IDs	Init.	Learnable	M-STFT	PESQ	MCD	VISQOL
1	T	✗	1.0187	3.957	2.542	4.797
2	P	✗	0.8320	4.142	2.139	**4.931**
3	T	✓	0.8284	4.110	2.373	4.919
4	P	✓	**0.8028**	**4.225**	**2.209**	4.928

Table 4: Ablation studies *w.r.t.* initialization scheme. {T, P} denotes matrix transpose and pseudo-inverse operations.

which effectively facilitates the harmonic component generation within the waveform. However, the large computational cost also heavily impedes the inference efficiency. In contrast, by resorting to spectral magnitude and phase estimation, existing T-F domain-based methods, *e.g.*, Vocos, enjoys prominent advantage in processing speed, despite at the cost of performance suboptimality over BigVGAN. Second, thanks to the utilization of DDPM and FM, notable improvements are achieved. For example, RFWave obtains the highest PESQ score and secondary performance in MCD, V/UV F1 and Periodicity RMSE, which are closely concerned with the acoustic distortion. Third, with the GAN paradigm, the proposed method yields overall the best objective performance over baselines, as well as light-weight designs in network parameters and inference cost. This reveals the remarkable trade-off between performance and processing efficiency. Moreover, our method yields the lowest WER, indicating the well preservation in semantic information.

To reveal the superiority of our method in out-of-distribution scenarios, the EARS corpus (Richter et al. 2024), a benchmark with diverse emotions, is adopted. We randomly select 200 samples as the test set and all models are trained from LibriTTS corpus. The metric results are shown in Table 2. Again, our method enjoys overall superior performance over other baselines, which reveals the effectiveness of the proposed framework, where the "initialization and target prior learning" pipeline is introduced and large kernels are introduced for efficient inter-frame and inter-band learning. Figure 4 visualizes the reconstructed waveforms in the T-F domain by different vocoders. One can observe that while most baselines can incur blurring in the regions with rich harmonic components, our method demonstrates the best performance in recovering harmonic details.

In Table 3, we report the subjective scores on two out-of-distribution benchmarks, namely AISHELL3 (Shi et al. 2021) and MUSDB18 (Rafii et al. 2017). The former is a mandarin speech corpus and the latter is a music dataset. For each corpus, 15 samples are randomly selected for evaluations[4]. One can observe that for both audio types, the proposed method yields the highest scores, and outperforms RFWave with statistical difference ($p < 0.1$), further validating the advantage of out method in subjective quality. It is noteworthy that the performance gap of RFWave and our method over other baselines seems larger in the music scenario, which is attributed to the richer harmonic components for song vocals over natural speech signals (see Figure 4 for clear illustrations). Therefore, further investigations are still required for existing neural vocoders in the music reconstruction scenarios.

[4]For MUSDB18, we choose the vocal type for evaluations.

Sets	M-STFT	PESQ	MCD	V/UV F1	VISQOL
Set-A	0.9474	3.704	3.032	0.943	4.866
Set-B	0.8634	4.012	2.561	0.958	4.904
Ours	**0.8028**	**4.225**	**2.209**	**0.966**	**4.928**

Table 5: Ablation studies *w.r.t.* basic network structure.

Ablation Studies

In this section, we investigate the choice in initialization scheme mentioned in Sec. , the kernel size and basic structure within LKCAM, and the results are reported on the *dev-clean-other* of the LibriTTS corpus. **(1) Initialization scheme**: Table 4 shows the results of different initialization methods mentioned in Sec. . For fixed cases, when the pseudo-inverse matrix is adopted, improved performance can be observed. When the matrix weights are set to be learnable, while notable improvements are given for the transpose scheme, close scores are obtained for the pseudo-inverse version. Therefore, we choose the pseudo-inverse type with learnable attribute by default. As shown in Appendix, the pseudo-inverse operation recovers the range-space of the target spectral magnitude, which aligns with classical signal recovery theory (Baraniuk et al. 2010). **(2) Kernel size**: In Figure 5, we present the PESQ scores under different $\{I_f, I_t\}$ configurations within LKCAB. One can observe that a suitable receptive field is crucial for spectrum reconstruction. While a small receptive field is detrimental to the contextual frame and sub-band modeling, a too large receptive field can neglect the local details modeling. Collectively, $\{9, 11\}$ seems to be an optimal choice. **(3) Basic structure**: In Table 5, we compare the proposed LKCAB with another two options. For Set-A, the proposed CAU is replaced by the self-attention operation. For Set-B, we replace LKCAB with the ConvNext v2 block (Woo et al. 2023) and similar computational complexity is kept to ensure fair comparisons. One can observe that our method yields the best performance on objective metrics, indicating both the effectiveness and efficiency of the proposed convolutional-attention calculation via large kernel size. More experimental results and analysis are provided in **Appendix**.

Conclusion

In this paper, we propose a novel T-F domain-based neural vocoder for audio generation. Specifically, we revisit the vocoder task from a degradation perspective, and a two-step optimization procedure is devised. First, in the initialization stage, we roughly recover the basic spectral structure from the Mel-domain representation. After that, leveraging the heterogeneous distributions of sub-bands, a large-kernel convolutional attention module is devised for efficient modeling *w.r.t.* inter-frame and inter-bands. Extensive experiments are conducted, and both quantitative and qualitative results validate the superiority of the proposed method over both existing GAN- and diffusion-based baselines.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62501588.

References

Afonso, M. V.; Bioucas-Dias, J. M.; and Figueiredo, M. A. 2011. An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE Trans. Image Process.*, 20(3): 681–695.

Ai, Y.; and Ling, Z.-H. 2023. 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.

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

Baraniuk, R. G.; Cevher, V.; Duarte, M. F.; and Hegde, C. 2010. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56(4): 1982–2001.

Beck, A.; and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1): 183–202.

Dai, L.; Li, A.; Han, Z.; Zheng, C.; and Li, X. 2025. BAPEN: Towards Versatile Audio Phase Retrieval. In *Proc. ACMMM*, 8293–8302.

Dong, W.; Wu, J.; Li, L.; Shi, G.; and Li, X. 2023. Bayesian Deep Learning for Image Reconstruction: From structured sparsity to uncertainty estimation. *IEEE Signal Process. Mag.*, 40(1): 73–84.

Du, H.-P.; Lu, Y.-X.; Ai, Y.; and Ling, Z.-H. 2023. Apnet2: High-quality and high-efficiency neural vocoder with direct prediction of amplitude and phase spectra. In *Proc. NCMMSC*, 66–80. Springer.

Dudley, H. 1939. Remaking speech. *J. Acoust. Soc. Am.*, 11(2): 169–177.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. CVPR*, 770–778.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hines, A.; Skoglund, J.; Kokaram, A. C.; and Harte, N. 2015. ViSQOL: an objective speech quality model. *EURASIP J. Audio Speech Music Process.*, 2015(1): 13.

Hou, Q.; Lu, C.-Z.; Cheng, M.-M.; and Feng, J. 2024. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12): 8274–8283.

Jang, W.; Lim, D.; Yoon, J.; Kim, B.; and Kim, J. 2021. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. Interspeech*, 2207–2211.

Juvela, L.; Bollepalli, B.; Tsiaras, V.; and Alku, P. 2019. Glotnet—a raw waveform model for the glottal excitation in statistical parametric speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 27(6): 1019–1030.

Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.; Dieleman, S.; and Kavukcuoglu, K. 2018. Efficient neural audio synthesis. In *Proc. ICML*, 2410–2419. PMLR.

Kaneko, T.; Tanaka, K.; Kameoka, H.; and Seki, S. 2022. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform. In *Proc. ICASSP*, 6207–6211. IEEE.

Kawahara, H. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *acoust. sci. technol.*, 27(6): 349–353.

Kawahara, H.; Masuda-Katsuse, I.; and De Cheveigne, A. 1999. 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.

Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Schindler, G.; Hornung, R.; Birodkar, V.; Yan, J.; Chiu, M.-C.; et al. 2024. VideoPoet: a large language model for zero-shot video generation. In *Proc. ICML*, 25105–25124.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, 17022–17033.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *Proc. ICLR*, 1–12.

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

Kubichek, R. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, 125–128. IEEE.

Kumar, K.; Kumar, R.; De Boissiere, T.; Gestin, L.; Teoh, W. Z.; Sotelo, J.; De Brebisson, A.; Bengio, Y.; and Courville, A. C. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Proc. NeurIPS*, 1–12.

Lee, S.-g.; Kim, H.; Shin, C.; Tan, X.; Liu, C.; Meng, Q.; Qin, T.; Chen, W.; Yoon, S.; and Liu, T.-Y. 2022. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *Proc. ICLR*, 1–12.

Lee, S.-g.; Ping, W.; Ginsburg, B.; Catanzaro, B.; and Yoon, S. 2023. BigVGAN: A Universal Neural Vocoder with Large-Scale Training. In *Proc. ICLR*, 1–13.

Lee, S.-H.; Choi, H.-Y.; and Lee, S.-W. 2025. PeriodWave: Multi-Period Flow Matching for High-Fidelity Waveform Generation. In *Proc. ICLR*, 1–15.

Li, A.; Chen, R.; Gu, Y.; Weng, C.; and Su, D. 2024. Opine: Leveraging a Optimization-Inspired Deep Unfolding Method for Multi-Channel Speech Enhancement. In *Proc. ICASSP*, 11376–11380. IEEE.

Li, A.; Lei, T.; Chen, R.; Li, K.; Yu, M.; Li, X.; Yu, D.; and Zheng, C. 2025a. BridgeVoC: Revitalizing Neural Vocoder from a Restoration Perspective. *arXiv preprint arXiv:2511.07116*.

Li, A.; Lei, T.; Sun, Z.; Chen, R.; Yin, E.; Li, X.; and Zheng, C. 2025b. Learning Neural Vocoder from Range-Null Space Decomposition. In *Proc. IJCAI*, 8131–8140.

Li, A.; Sun, Z.; Hao, F.; Li, X.; and Zheng, C. 2025c. Neural Vocoders as Speech Enhancers. *arXiv preprint arXiv:2501.13465*.

Liu, H.; Baoueb, T.; Fontaine, M.; Le Roux, J.; and Richard, G. 2024. Gla-grad: A griffin-lim extended waveform generation diffusion model. In *Proc. ICASSP*, 11611–11615. IEEE.

Liu, P.; Dai, D.; and Wu, Z. 2025. RFWave: Multi-band Rectified Flow for Audio Waveform Reconstruction. In *Proc. ICLR*, 1–16.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proc. CVPR*, 11976–11986.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, T.; Miao, X.; and Duan, W. 2025. WaveFM: A High-Fidelity and Efficient Vocoder Based on Flow Matching. In *Proc. NAACL*, 2187–2198.

Luong, M.; and Tran, V. A. 2021. Flowvocoder: A small footprint neural vocoder based normalizing flow for speech synthesis. *arXiv preprint arXiv:2109.13675*.

Lv, Y.; Li, H.; Yan, Y.; Liu, J.; Xie, D.; and Xie, L. 2024. FreeV: Free Lunch For Vocoders Through Pseudo Inversed Mel Filter. In *Proc. Interspeech*, 3869–3873.

Masuyama, Y.; Yatabe, K.; Koizumi, Y.; Oikawa, Y.; and Harada, N. 2019. Deep griffin–lim iteration. In *Proc. ICASSP*, 61–65. IEEE.

Morise, M.; Yokomori, F.; and Ozawa, K. 2016. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans Inf Syst*, 99(7): 1877–1884.

Morrison, M.; Kumar, R.; Kumar, K.; Seetharaman, P.; Courville, A.; and Bengio, Y. 2022. Chunked Autoregressive GAN for Conditional Waveform Synthesis. In *Proc. ICLR*, 1–13.

Nguyen, T. D.; Kim, J.-H.; Jang, Y.; Kim, J.; and Chung, J. S. 2024. Fregrad: Lightweight and fast frequency-aware diffusion vocoder. In *Proc. ICASSP*, 10736–10740. IEEE.

Peer, T.; Welker, S.; and Gerkmann, T. 2023. DiffPhase: Generative diffusion-based STFT phase retrieval. In *Proc. ICASSP*, 1–5. IEEE.

Rafii, Z.; Liutkus, A.; Stöter, F.-R.; Mimilakis, S. I.; and Bittner, R. 2017. The MUSDB18 corpus for music separation.

Richter, J.; Wu, Y.-C.; Krenn, S.; Welker, S.; Lay, B.; Watanabe, S.; Richard, A.; and Gerkmann, T. 2024. EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation. In *Proc. Interspeech*, 4873–4877.

Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; and Bian, J. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Shi, Y.; Bu, H.; Xu, X.; Zhang, S.; and Li, M. 2021. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Proc. Interspeech*, 2756–2760.

Siuzdak, H. 2024. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *Proc. ICLR*, 1–13.

Tan, K.; Chen, J.; and Wang, D. 2018. Gated residual networks with dilated convolutions for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 27(1): 189–198.

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

Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K.; et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12: 1.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proc. CVPR*, 16133–16142.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. ICASSP*, 6199–6203. IEEE.

Yang, G.; Yang, S.; Liu, K.; Fang, P.; Chen, W.; and Xie, L. 2021. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *Proc. SLT*, 492–498. IEEE.

You, D.; Zhang, J.; Xie, J.; Chen, B.; and Ma, S. 2021. COAST: Controllable arbitrary-sampling network for compressive sensing. *IEEE Trans. Image Process.*, 30: 6066–6080.

Yun, J.-H.; Kim, S.-B.; and Lee, S.-W. 2025. Flowhigh: Towards efficient and high-quality audio super-resolution with single-step flow matching. In *Proc. ICASSP*, 1–5. IEEE.

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

Zhang, J.; and Ghanem, B. 2018. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proc. CVPR*, 1828–1837.

Zhang, J.; Zhang, Z.; Xie, J.; and Zhang, Y. 2022. High-throughput deep unfolding network for compressive sensing MRI. *IEEE J Sel Topics Signal Process.*, 16(4): 750–761.

Zhou, Y.; Li, Z.; Guo, C.-L.; Bai, S.; Cheng, M.-M.; and Hou, Q. 2023. Srformer: Permuted self-attention for single image super-resolution. In *Proc. ICCV*, 12780–12791.