Learning Neural Vocoder from Range-Null Space Decomposition

Andong Li[1,2], **Tong Lei**[3,4], **Zhihang Sun**[3], **Rilin Chen**[3], **Erwei Yin**[5,6], **Xiaodong Li**[1,2]
Chengshi Zheng[1,2*]
[1]Institute of Acoustics, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Tencent AI Lab
[4]Nanjing University
[5] Defense Innovation Institute, Academy of Military Sciences (AMS)
[6] Tianjin Artificial Intelligence Innovation Center (TAIIC)
cszheng@mail.ioa.ac.cn

Abstract

Despite the rapid development of neural vocoders in recent years, they usually suffer from some intrinsic challenges like opaque modeling, and parameter-performance trade-off. In this study, we propose an innovative time-frequency (T-F) domain-based neural vocoder to resolve the above-mentioned challenges. To be specific, we bridge the connection between the classical signal range-null decomposition (RND) theory and vocoder task, and the reconstruction of target spectrogram can be decomposed into the superimposition between the range-space and null-space, where the former is enabled by a linear domain shift from the original mel-scale domain to the target linear-scale domain, and the latter is instantiated via a learnable network for further spectral detail generation. Accordingly, we propose a novel dual-path framework, where the spectrum is hierarchically encoded/decoded, and the cross- and narrow-band modules are elaborately devised for efficient sub-band and sequential modeling. Comprehensive experiments are conducted on the LJSpeech and LibriTTS benchmarks. Quantitative and qualitative results show that while enjoying lightweight network parameters, the proposed approach yields state-of-the-art performance among existing advanced methods. Our code and the pretrained model weights are available at https://github.com/Andong-Li-speech/RNDVoC.

1 Introduction

Sound vocoder aims to reconstruct the audible time-domain waveforms using electronic and computational techniques, which is widely employed in text-to-speech (TTS) [Wang *et al.*, 2017; Huang *et al.*, 2022a; Li *et al.*, 2025], text-to-audio (TTA) [Liu *et al.*, 2023], and speech enhancement [Zhou *et al.*, 2024; Li *et al.*, 2022]. Compared with traditional digital signal processing (DSP)-based vocoders like STRAIGHT [Kawahara, 2006] and WORLD [Morise *et al.*, 2016], neural vocoders enjoy superior advantage in generation naturalness and quality, thereby garnering wide attention in recent years.

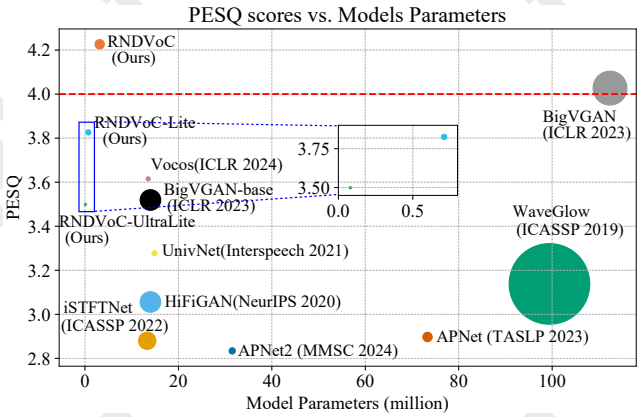

Figure 1: A case comparison in terms of PESQ score on the LibriTTS benchmark. A larger bubble denotes higher computational complexity. Note that both BigVGAN and our methods are trained for 1M steps herein.

In the primitive stage, neural vocoders typically generate waveforms sample by sample, and representative works include WaveNet [Van Den Oord *et al.*, 2016] and WaveRNN [Kalchbrenner *et al.*, 2018]. Despite the improvement, they often suffer from considerably slow inference efficiency due to their autoregressive processing characteristic. To alleviate the deficiency, some other schemes have subsequently been adopted, including knowledge distillation-based [Oord *et al.*, 2018], normalization flow-based [Ping *et al.*, 2020], and glottis-based [Juvela *et al.*, 2019]. Recently, generative adversarial network (GAN)-based neural vocoders have gained increasing attention due to their non-autogressive structure and high generation quality. Pioneering works include MelGAN [Kumar *et al.*, 2019] and HiFiGAN [Kong *et al.*, 2020], where the mel-spectrograms are gradually upsampled and recovered to waveforms via alternate residual blocks and upsampling layers. Except for the time domain based

*Chengshi Zheng is the corresponding author.

methods, time-frequency (T-F) domain based neural vocoders have been proposed in more recent years, which enjoy faster inference speed and promising generation quality [Siuzdak, 2024].

Despite the success of existing neural vocoders, some inherent challenges still exist and can impede the further progress. First, most of the end-to-end neural vocoders still suffer from the typical parameter-performance trade-off. For example, although achieving impressive performance, the number of parameters of BigVGAN is as high as 112M [Lee *et al.*, 2023], which is awkward and also inconvenient for practical applications. Second, while T-F domain based neural vocoders exhibit the advantage in inference efficiency, their performance often lags behind mainstream time-domain-based ones. For the first point, we argue that existing methods usually model the vocoder process as a black-box, neglecting the utilization the linear degradation prior of mel-spectrogram[1]. Therefore, excessive parameters are often required for overall target reconstruction. For the second point, short-time Fourier transform (STFT) operation explicitly decouples the frequency information among different sub-bands, and acoustic features usually lie in specific frequency regions. However, only full-band modeling is adopted in existing T-F domain based neural vocoders, neglecting the hierarchical characteristic of the T-F spectrum.

To this remedy, we propose a novel T-F domain based neural vocoder called **RNDVoC** to tackle the above-mentioned challenges. Specifically, we revisit the formulation of mel-spectrogram and establish the connection with range-null decomposition (RND) theory. By doing so, the reconstruction of target spectrum is thus decomposed into range-space modeling (RSM) and null-space modeling (NSM), where the former aims to project the original mel-spectrogram into the corresponding counterpart defined in the target linear-scale domain, and the latter is responsible for spectral details generation. In this way, we provide a more transparent and also elegant perspective for framework design. Besides, we devise an efficient dual-path structure, where the spectrum is hierarchically encoded and decoded, and the cross-band and narrow-band modules are alternate adopted for sub-band and sequential modeling. Experimental results show that with only only **2.8%** parameters and **8.17%** computational complexity, our method yields comparable and even better performance over BigVGAN-112M version in both objective and subjective scores, as well as nearly 10× speed-up on a CPU. Meanwhile, when the network parameters is further reduced as low as **0.08M**, the performance is still comparable to existing baselines. Figure 1 showcases an example of the PESQ scores using many existing representative methods on the LibriTTS benchmark. Our contributions can be summarized as four-fold:

- We provide a novel range-null space decompostion perspective to tackle the neural vocoder task. To our best knowledge, this is the first time to leverage the RND theory for audio generation.
- We devise a novel dual-path framework to encode/decode the spectrum hierarchically, and cross-band and narrow-band modules are employed for efficient sub-band and sequential modeling.
- We conduct extensive experiments to reveal the superiority of our method over existing baselines.
- We provide an ultra-light version with only around 80K trainable parameters. To our best knowledge, this is the smallest end-to-end neural vocoder up to now.

[1]Strictly speaking, mel-spectrogram is linearly compressed in the spectral magnitude, and the phase information is dropped.

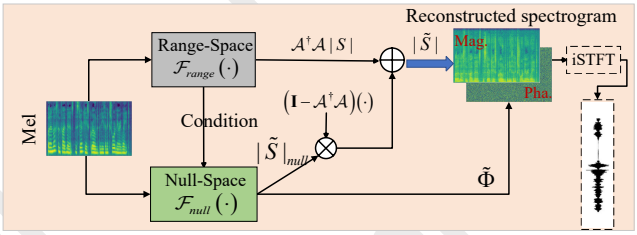

Figure 2: Illustrations of the proposed RNDVoC.

2 Related Works

2.1 DSP-based Methods

In conventional DSP-based vocoders, the speech is usually generated with statistical parameter estimation. In STRAIGHT [Kawahara, 2006], the excitation and resonant parameters are separately estimated for real-time speech manipulation and synthesis. In [Morise *et al.*, 2016], the F0, spectral envelop, and aperiodic parameters are first determined using analysis algorithms. A synthesis method is subsequently devised for the time-domain waveform generation. Although being simplicity, the synthesized speech quality is often unsatisfactory and may include some buzzing artifacts.

2.2 Neural Vocoder Methods

Autoregressive methods: In WaveNet [Van Den Oord *et al.*, 2016], consecutive WaveNet modules with increasing dilation rates are adopted for sample-level generation. WaveRNN [Kalchbrenner *et al.*, 2018], samples are autoregressively generated with RNN layers. In LPCNet [Valin and Skoglund, 2019], the linear prediction coefficients (LPC) are estimated to predict the next sample and a lightweight RNN is utilized for residual calculation. Despite the improvements over traditional DSP-based vocoders, the autoregressive nature can suffer from rather slow inference efficiency.

Flow-based methods: Normalization flow is regarded as a classical generation paradigm. Typical flow-based neural vocoders include WaveGlow [Prenger *et al.*, 2019], FlowWaveNet [Kim *et al.*, 2019], and RealNVP [Dinh *et al.*, 2016], where the bijective mapping is established between a normalization probability distribution and target data distribution through stacked invertible modules.

GAN-based methods: In MelGAN [Kumar *et al.*, 2019], the mel-spectrogram is gradually upsampled through consecutive upsampling layers and residual blocks. In HiFiGAN, to facilitate the periodic pattern generation, the multi-periodic and multi-scale discriminators are employed for improved speech

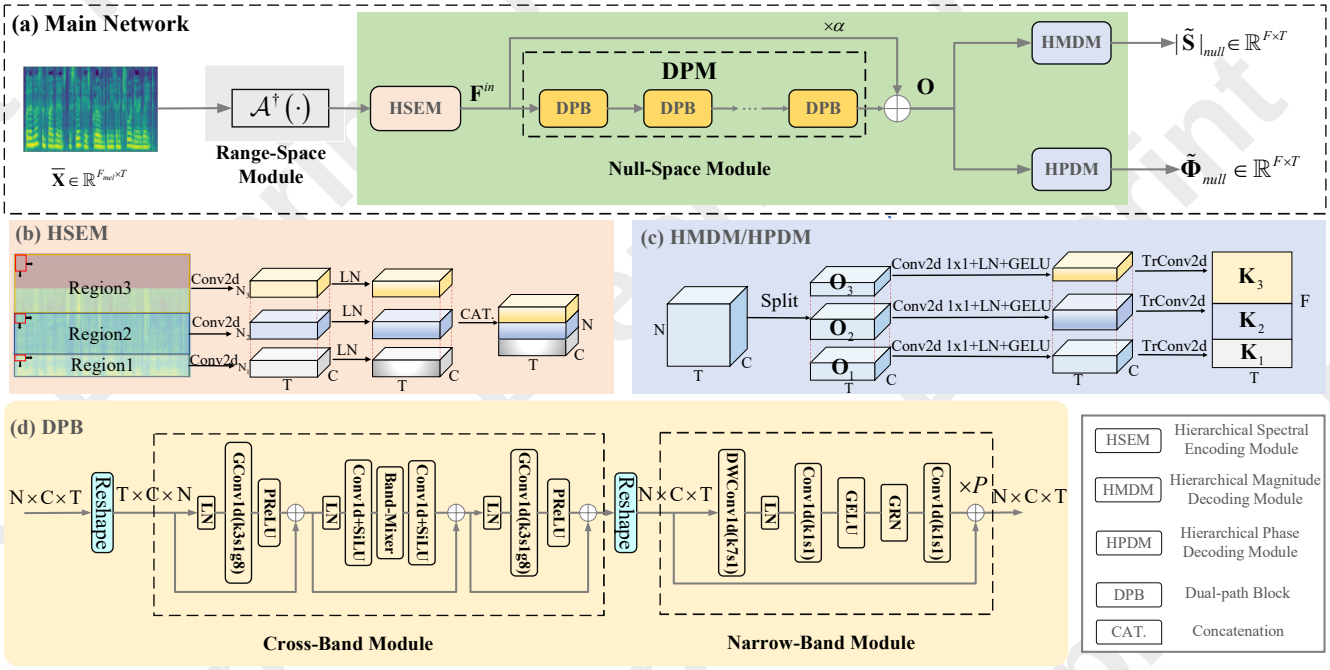

Figure 3: Network structure of the proposed RNDVoC. (a) Overall network structure of the RNDVoC. (b) Detailed structure of the hierarchical spectral encoding module (HSEM). (c) Detailed structure of the hierarchical magnitude decoding module (HMDM) and hierarchical phase decoding module (HPDM). (d) Detailed structure of the dual-path block (DPB). Diffferent modules are indicated with different colors.

generation. BigVGAN [Lee *et al.*, 2023] incorporates the periodic activation function and anti-aliased representation into the generator to boost the periodic components generation, and the network size is further scaled up to 112M for universal audio vocoding. Except for the above-mentioned time-domain methods, several works explore the feasibility in the T-F domain. Vocos [Siuzdak, 2024] stacks multiple ConvNext v2 blocks [Woo *et al.*, 2023] for magnitide and phase estimation. In [Ai and Ling, 2023], the magnitude and phase components are separately modeled with ResNet blocks.

Diffusion-based methods: Owing to the powerful generation capability of duffusion methods, some works also adopt diffusion for neural vocoder. Representative works include DiffWave [Kong *et al.*, 2021], WaveGrad [Chen *et al.*, 2021], and PriorGrad [Lee *et al.*, 2022], where the waveforms gradually degrade to Gaussian noise in the forward path and the target signals can be iteratively generated in the reverse process. Despite good performance, the implemention efficiency can be slow due to numerous iteration steps, and fast sampling strategies [Huang *et al.*, 2022b; Lu *et al.*, 2022] are required to reduce the inference cost.

3 Method

In this section, we first introduce the range-null decomposition (RND) theory and formulate the problem of the speech vocoder task, and then we demonstrate their interconnection. Finally, the proposed learning framework is elucidated.

3.1 Range-Null Space Decomposition

For a classical signal compression physical model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^D$, $\{\mathbf{y}, \mathbf{n}\} \in \mathbb{R}^d$ denote the target, observed, and noise signals, respectively, and $\mathbf{A} \in \mathbb{R}^{d \times D}$ denotes the compression matrix, with $d \ll D$. In the noise-free scenario, Eq.(1) can be simplified into $\mathbf{y} = \mathbf{A}\mathbf{x}$. If the pseudo-inverse of $\mathbf{A}$ is defined as $\mathbf{A}^\dagger \in \mathbb{R}^{D \times d}$, which satisfies $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} \equiv \mathbf{A}$, then the signal $\mathbf{x}$ can be decomposed into two orthogonal subspaces: one residing in the range-space of $\mathbf{A}$, and the other in the null-space:

$$\mathbf{x} \equiv \mathbf{A}^\dagger\mathbf{A}\mathbf{x} + \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x}, \tag{2}$$

where $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$ defines the range-space component and $\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x}$ corresponds to the remaining null-space component. For practical solution of $\mathbf{x}$, *i.e.*, $\tilde{\mathbf{x}}$, two consistency conditions should be satisfied: (1) Degradation consistency: $\mathbf{A}\tilde{\mathbf{x}} \equiv \mathbf{y}$, where the original information embedded in the compressed signal space remains unaltered after reconstruction. (2) Data consistency: $\tilde{\mathbf{x}} \sim p(\mathbf{x})$, where the estimation $\tilde{\mathbf{x}}$ should share the same data distribution to $\mathbf{x}$. Based on the above constraints, the solution can be expressed as:

$$\tilde{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y} + \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\hat{\mathbf{x}}, \tag{3}$$

where $\mathbf{A}^\dagger\mathbf{y}$ and $\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\hat{\mathbf{x}}$ correspond to the solutions in the range-space and null-space, respectively. Due to the orthogonality characteristic, $\tilde{\mathbf{x}}$ naturally satisfies the first condition, *i.e.*, $\mathbf{A}\tilde{\mathbf{x}} \equiv \mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{x} + \mathbf{A}\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\hat{\mathbf{x}} \equiv \mathbf{A}\mathbf{x} + \mathbf{0} \equiv \mathbf{y}$.

Therefore, we only need to consider the estimation of $\hat{\mathbf{x}}$ to meet the second condition, which can often be implemented with learnable networks.

3.2 Revisit Neural Vocoder Tasks

In this paper, we mainly consider the scenario when the mel-spectrogram serves as the acoustic feature due to its simplicity and commonality in neural vocoders. The degradation process of a log-scale mel-spectrogram can be given by:

$$\mathbf{X}^{mel} = \log\left(\mathcal{A}|\mathbf{S}|\right), \tag{4}$$

where $\mathbf{X}^{mel} \in \mathbb{R}^{F_m \times T}$ and $|\mathbf{S}| \in \mathbb{R}^{F \times T}$ denote the mel and target spectrograms, respectively, $\{F_m, F\}$ are the frequency size in the mel- and linear-scale, T is the frame size. $\mathcal{A} \in \mathbb{R}^{F_m \times F}$ refers to the mel-filter, which is instantiated by a linear compression matrix.

It is evident that Eq. (4) involves two degradation operations: ① Discarding the phase information, ② Magnitude compression with a linear operation. Intuitively, two steps are necessarily involved in the inverse process: (1) Estimating the phase spectrum, (2) Recovering the magnitude spectrum from the compressed observation.

In the preliminary literature, a neural network typically serves as the black-box to establish the mapping relation between target waveform/T-F spectrum and mel feature, which can be expressed as:

$$\tilde{\mathbf{s}} = \mathcal{F}_T\left(\mathbf{X}^{mel}, \Theta_1\right), \left\{|\tilde{\mathbf{S}}|, \tilde{\mathbf{\Phi}}\right\} = \mathcal{F}_{TF}\left(\mathbf{X}^{mel}, \Theta_2\right), \tag{5}$$

where $\mathcal{F}_T\left(\cdot, \Theta_1\right)$ and $\mathcal{F}_{TF}\left(\cdot, \Theta_2\right)$ refer to the mapping function of time-domain and T-F domain based neural vocoders, respectively. $\tilde{\mathbf{s}}$ is the estimated time-domain waveform, and $\left\{|\tilde{\mathbf{S}}|, \tilde{\mathbf{\Phi}}\right\}$ denote the estimated spectral magnitude and phase, respectively. Note that if the log-operation is absorbed into the left of Eq. (4), *i.e.*, $\overline{\mathbf{X}}^{mel} = \exp\left(\mathbf{X}^{mel}\right) = \mathcal{A}|\mathbf{X}|$, it then exhibits a similar format to the noise-free case of Eq. (1). In other words, the original magnitude spectrum recovery can be formulated into a classical compressive sensing (CS) problem [Baraniuk *et al.*, 2010], in which we attempt to reconstruct the target signal from a linearly-compressed representation. Following the explicit decomposition in Eq. (3), the abstract generation process of the target spectrum in Eq. (5) can be rewritten as:

$$|\tilde{\mathbf{S}}|_{range} = \mathcal{F}_{range}\left(\overline{\mathbf{X}}^{mel}\right) = \mathcal{A}^{\dagger}\overline{\mathbf{X}}^{mel}, \tag{6}$$

$$\left\{|\tilde{\mathbf{S}}|_{null}, \tilde{\mathbf{\Phi}}\right\} = \mathcal{F}_{null}\left(|\tilde{\mathbf{S}}|_{range}\right), \tag{7}$$

$$\begin{aligned} |\tilde{\mathbf{S}}| &= |\tilde{\mathbf{S}}|_{range} + \left(\mathbf{I} - \mathcal{A}^{\dagger}\mathcal{A}\right)|\tilde{\mathbf{S}}|_{null} \\ &= \mathcal{A}^{\dagger}\mathcal{A}|\mathbf{S}| + \left(\mathbf{I} - \mathcal{A}^{\dagger}\mathcal{A}\right)|\tilde{\mathbf{S}}|_{null}, \end{aligned} \tag{8}$$

$$\tilde{\mathbf{S}} = |\tilde{\mathbf{S}}|e^{j\tilde{\mathbf{\Phi}}}, \tag{9}$$

where $\mathcal{F}_{range}\left(\cdot\right)$ and $\mathcal{F}_{null}\left(\cdot\right)$ are the operations in the range- and null-spaces, respectively, $\mathcal{A}^{\dagger} \in \mathbb{R}^{F \times F_m}$ is the pseudo-inverse of $\mathcal{A}$. As shown in Figure 2, in $\mathcal{F}_{range}$, the original acoustic feature in the mel-domain is transformed into the target domain via the pseudo-inverse operation. In $\mathcal{F}_{null}$, the remaining magnitude details are estimated to supplement the overall spectral reconstruction. Compared with preliminary methods, the reconstruction format in Eq. (8) enjoys several advantages: First, we fully utilize the linear degradation of mel-spectrum as the prior to establish two orthogonal subspaces, which enhances the overall framework interpretability. Besides, due to the degradation consistency, the original acoustic feature embedded in the mel-spectrogram can be well preserved, mitigating the acoustic distortion in the reconstructed signals. One should note that as only the spectral magnitude is involved in degradation formulation, we enforce $\mathcal{F}_{null}$ to also estimate the phase component, as shown in Eq. (7). Besides, we notice that most recently, a similar pseudo-inverse strategy is adopted in [Lv *et al.*, 2024]. The difference is that here we endow the pseudo-inverse operation with a more intuitive physical explanation while it is only adopted as an empirical trick in [Lv *et al.*, 2024].

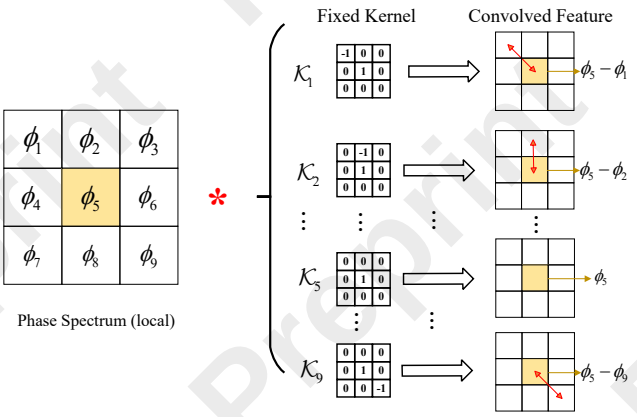

Figure 4: Illustration of the proposed omnidirectional phase loss.

3.3 Architecture Design of RNDVoC

As previously illustrated, we explicitly decouple two orthogonal sub-spaces and devise the network modules for estimation. Figure 3(a) presents detailed framework of the RNDVoC. Given the input mel-spectrogram, it is processed by the range-space module (RSM), where the pseudo-inverse operation is adopted to shift the mel into the linear-scale domain. After that, the null-space module (NSM) is utilized, and three modules are mainly involved: hierarchical spectral encoding module (HSEM), dual-path module (DPM), and two decoder branches, *i.e.*, hierarchical magnitude/phase decoding module (HMDM/HPDM), which will be introduced in a sequel.

Hierarchical Spectral Encoder and Decoder

Detailed network structure of the HSEM is shown in Figure 3(b). Given the input $|\tilde{\mathbf{S}}|_{range} \in \mathbb{R}^{F \times T}$, it is first split into I regions, and the i-th spectral region is notated as $|\tilde{\mathbf{S}}|_{range,i} \in \mathbb{R}^{F_i \times T}$. As the major acoustic features lie in the low/mid frequency regions (*e.g.* fundamental frequency F0), we follow the "from-fine-to-coarse" principle for region division, that is, we gradually increase F_i with the increase

Models	Domain	#Param. (M)	#MACs (Giga/5s)	Inference Speed		M-STFT↓	PESQ↑	MCD↓	Periodicity↓ RMSE	V/UV↑ F1	Pitch↓ RMSE	VISQOL↑
				CPU	GPU							
HiFiGAN-V1	T	13.94	152.90	0.1669(5.99×)	0.0069(145×)	1.1699	3.574	3.6711	0.1344	0.9474	33.6874	4.7706
iSTFTNet-V1	T	13.26	107.76	0.0949(10.54×)	0.0038(266×)	1.1883	3.535	3.6252	0.1356	0.9466	35.1055	4.7557
Avocodo	T	13.94	152.95	0.1611(6.21×)	0.0063(158×)	1.1653	3.604	3.6570	0.1386	0.9462	32.9887	4.7714
BigVGAN-base	T	13.94	152.90	0.3856(2.59×)	0.0368(27×)	0.9784	3.603	2.3314	0.1198	0.9562	30.2774	4.8217
BigVGAN	T	112.18	417.20	0.6674(1.50×)	0.0507(20×)	**0.9001**	**4.107**	**1.8769**	**0.0838**	**0.9716**	**20.6922**	**4.8699**
APNet	T-F	72.19	31.11	0.03(33.33×)	0.0022(458×)	1.2659	3.390	3.2847	0.1508	0.9454	23.0571	4.6947
APNet2	T-F	31.38	13.53	0.0187(53.48×)	0.0016(643×)	0.9815	3.492	2.8288	0.1126	0.9592	25.3629	4.7515
FreeV	T-F	18.19	<u>7.84</u>	<u>0.0139(71.94×)</u>	<u>0.0011(951×)</u>	1.0076	3.593	2.7502	0.1118	0.9603	25.9922	4.7427
Vocos	T-F	13.46	**5.80**	**0.0066(151.52×)**	**0.0007(1400×)**	1.0123	3.522	2.6699	0.1213	0.9559	29.1304	4.7741
RNDVoC(Ours)	T-F	**3.14**	34.10	0.0669(14.96×)	0.0043(233×)	<u>0.9103</u>	<u>3.987</u>	<u>2.0471</u>	<u>0.0854</u>	<u>0.9714</u>	<u>21.3183</u>	<u>4.8373</u>

Table 1: Objective comparisons among different baselines on the LJSpeech benchmark. T and T-F refer to time-domain and T-F domain-based methods, respectively. "$a\times$" 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. The best and second-best performances are respectively highlighted in **bold** and <u>underlined</u>, respectively.

in region index. For each spectral region, a separate Conv2d with stride is then utilized for spectral compression, followed by a layer-normalization (LN) layer [Lei Ba *et al.*, 2016]. The process can be formulated as:

$$\mathbf{F}_i^{in} = \text{LN}\left(\text{Conv2D}\left(|\tilde{\mathbf{S}}|_{range,i}\right)\right) \in \mathbb{R}^{N_i \times C \times T}, \quad (10)$$

where $\mathbf{F}_i^{in}$ is the compressed spectral feature of the i-th region, and $\{N_i, C\}$ denote the sub-band and channel size, respectively. Here C is set to 256. Similarly, for the decoding process, two branches are utilized for magnitude and phase estimation, respectively. Taking the magnitude branch as an example, detailed structure is shown in Figure 3(c). Given the input feature $\mathbf{O} \in \mathbb{R}^{N \times C \times T}$, it is first split into I regions, *i.e.*, $\{\mathbf{O}_1, \cdots, \mathbf{O}_I\}$. For each region, we first adopt a separate point-wise Conv2d layer, followed by LN and GELU activation layer. After that, a TrConv2d is utilized to recover the spectral target. The formulation can be expressed as:

$$\mathbf{K}_i = \text{TrConv2d}\left(\text{GELU}\left(\text{LN}\left(\text{Conv2d1x1}\left(\mathbf{O}_i\right)\right)\right)\right). \quad (11)$$

For the magnitude branch, the $\exp(\cdot)$ is applied to guarantee the non-negativity. For the phase branch, $\mathbf{K}_i$ is split into real and imaginary components, and Atan2 $(\cdot)$ is adopted for phase calculation. Note that in [Yu *et al.*, 2023], a similar band-split strategy is adopted, where the spectrum is split into N non-uniform sub-bands, and each sub-band is separately encoded and decoded. The difference is that here the number of regions I is usually far less than N[2], and thus fewer loop operations are required, leading to faster inference speed. Besides, as the network weights within one spectral region are shared, fewer trainable parameters are needed.

Dual-Path Block

To facilitate the spectral information modeling, $B = 6$ dual-path blocks (DPBs) are stacked, each of which is shown in Figure 3(d). Specifically, it consists of a cross-band module and a narrow-band module, namely corresponding to the time and sub-band modeling.

Given the input feature $\mathbf{F}^{(b)} \in \mathbb{R}^{N \times C \times T}$, it is first reshaped and sent to the cross-band module (CBM). Motivated by [Quan and Li, 2024], we adopt a light-weight design for cross-band modeling, and it involves three steps. First, a LN, followed by a group convolution with kernel being 3 along the sub-band axis and PReLU [He *et al.*, 2015], is adopted to model the correlation between neighboring sub-bands:

$$\mathbf{F}^{(b)'} = \mathbf{F}^{(b)} + \text{PReLU}\left(\text{GConv1d}\left(\text{LN}\left(\mathbf{F}^{(b)}\right)\right)\right). \quad (12)$$

After that, a Conv1d layer, followed by a SiLU activation function [Ramachandran *et al.*, 2017], is used to squeeze the channel size to $C^{'}$, which is four times smaller than C. Then a band-mixer is adopted for global sub-band shuffling, which is instantiated via a linear matrix operation. After that, another Conv1D+SiLU is adopted to recover the channel size. The process can be expressed as:

$$\mathbf{F}^{(b)''} = \mathbf{F}^{(b)'} + \text{SiLU}(\text{Conv1d}(\text{BandMixer}(\text{SiLU}(\text{Conv1d}(\text{LN}(\mathbf{F}^{(b)'})))))). \quad (13)$$

Finally, to fully utilize the correlation between adjacent sub-bands, another group convolution with kernel being 3 is adopted with residual connection:

$$\mathbf{F}^{(b+1)} = \mathbf{F}^{(b)''} + \text{PReLU}\left(\text{GConv1d}\left(\text{LN}\left(\mathbf{F}^{(b)''}\right)\right)\right). \quad (14)$$

For the narrow-band module (NBM), considering the success of ConvNext v2 blocks in neural vocoder task [Siuzdak, 2024], we stack $P = 2$ ConvNext v2 blocks to gradually capture long-term relations among adjacent frames. Different from the full-band modeling in previous works, here all sub-bands share the network parameters. Besides, to decrease the computation cost, the number of channels in both hidden and input layers remains unchanged, *i.e.*, C.

3.4 Loss Function

Following the settings in [Ai and Ling, 2023; Du *et al.*, 2023], we adopt both the reconstruction loss and adversarial loss for vocoder training. The reconstruction loss consists of log-amplitude loss $\mathcal{L}_a$, phase loss $\mathcal{L}_p$, real-imaginary loss $\mathcal{L}_{ri}$, mel-loss $\mathcal{L}_{mel}$, and consistency loss $\mathcal{L}_c$:

$$\mathcal{L}_{rec} = \lambda_a\mathcal{L}_a + \lambda_p\mathcal{L}_p + \lambda_{ri}\mathcal{L}_{ri} + \lambda_{mel}\mathcal{L}_{mel} + \lambda_c\mathcal{L}_c, \quad (15)$$

[2]In our practical settings, $I = 3$, and $N = 24$.

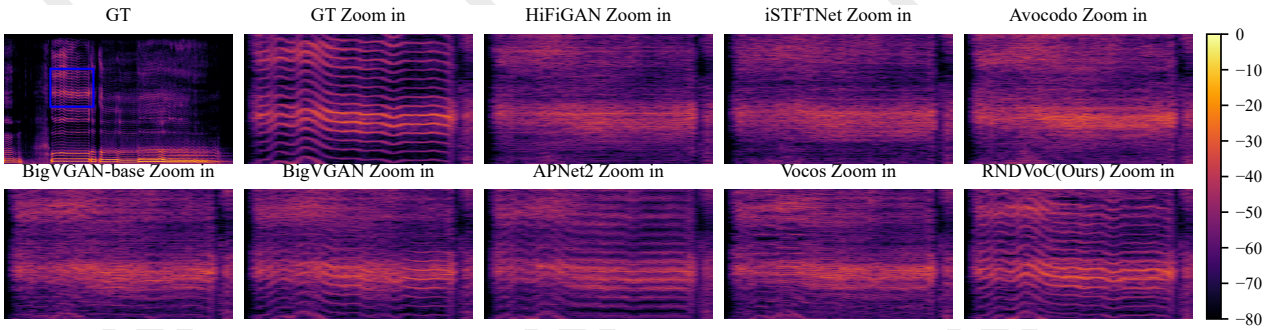

Figure 5: Spectral visualization of different vocoder methods. The audio clip is a singing voice from the MUSDB18 test set.

where $\{\lambda_a, \lambda_p, \lambda_{ri}, \lambda_{mel}, \lambda_c\}$ are the corresponding hyperparameters. In [Ai and Ling, 2023; Du *et al.*, 2023], an anti-wrapping loss is devised, and instantaneous phase (IP), group delay (GD) and instantaneous frequency (IF) are extracted via a specially-designed sparse matrix[3]. However, we find such calculation formula is usually time-consuming as most of the elements in the sparse matrix are zero and have no impact on the multiplied result. Besides, only two directions are considered for phase differential operation, resulting in limited phase relation capture. To this end, we propose a novel omni-directional phase loss, as shown in Figure 4. Specifically, we elaborately design nine 3×3 kernels with fixed parameters $\mathcal{K} = Cat\left(\mathcal{K}_1, \cdots, \mathcal{K}_9\right) \in \mathbb{R}^{9\times3\times3}$ to traverse the differential relations with adjacent eight T-F bins, and the fifth kernel is to return the IP. Therefore, with a simple convolution operation, the phase differential can be efficiently implemented as:

$$\Delta\mathbf{\Phi} = \mathbf{\Phi} * \mathcal{K}, \Delta\tilde{\mathbf{\Phi}} = \tilde{\mathbf{\Phi}} * \mathcal{K}, \tag{16}$$

where $\left\{\Delta\mathbf{\Phi}, \Delta\tilde{\mathbf{\Phi}}\right\} \in \mathbb{R}^{9\times F\times T}$, and "$*$" denotes the convolution operation . Then we can calculate the phase loss with the similar anti-wrapping loss to [Du *et al.*, 2023].

For the adversarial loss, the multi-period discriminator (MPD) [Kong *et al.*, 2020] and multi-resolution spectrogram discriminator (MRSD) [Jang *et al.*, 2021] are utilized. The hinge GAN is adopted, and the adversarial loss for the discriminator can be expressed as:

$$\mathcal{L}_D = \frac{1}{M}\sum_{m=1}^{M} \max\left(0, 1 - D_m\left(\mathbf{s}\right)\right) + \max\left(0, 1 + D_m\left(\tilde{\mathbf{s}}\right)\right), \tag{17}$$

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

$$\mathcal{L}_g = \frac{1}{M}\sum_{m=1}^{M} \max\left(0, 1 - D_m\left(\tilde{\mathbf{s}}\right)\right). \tag{18}$$

Besides, the feature matching loss is also utilized, given by:

$$\mathcal{L}_{fm} = \frac{1}{LM}\sum_{l,m} \left|\mathbf{f}_l^m\left(\tilde{\mathbf{s}}\right) - \mathbf{f}_l^m\left(\mathbf{s}\right)\right|, \tag{19}$$

[3]https://github.com/YangAi520/APCodec/blob/main/models.py

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	52.5285	4.7209
iSTFTNet-V1	2.880	0.1672	0.9177	53.0724	4.6548
UnivNet-c32†	3.277	0.1305	0.9347	41.5110	4.7530
Avocodo	3.217	0.1611	0.9134	51.5998	4.7620
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	32.5413	4.9067
BigVGAN(5M steps)†	**4.269**	0.0790	0.9670	24.2814	**4.9632**
APNet	2.897	0.1586	0.9265	39.6629	4.6659
APNet2	2.834	0.1529	0.9227	46.3732	4.5817
Vocos†	3.615	0.1146	0.9484	35.5844	4.8785
RNDVoC(Ours)	4.226	**0.0742**	**0.9698**	**23.9658**	4.9154

Table 2: 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.

where $\mathbf{f}_l^m\left(\cdot\right)$ denotes the l-th layer feature for m-th sub-discriminator. The final loss for generator is presented as:

$$\mathcal{L}_G = \mathcal{L}_{rec} + \lambda_g\mathcal{L}_g + \lambda_{fm}\mathcal{L}_{fm}, \tag{20}$$

where $\{\lambda_g, \lambda_{fm}\}$ are the corresponding hyperparameters. Detailed settings can be found in the supplementary material.

4 Experiments

4.1 Datasets

Two benchmarks are employed in this study, namely LJSpeech [Keith and Linda, 2017] and LibriTTS [Zen *et al.*, 2019]. The LJSpeech dataset includes 13,100 clean speech clips by a single female, and the sampling rate is 22.05 kHz. Following the division in the open-sourced VITS repository[4], $\{12500, 100, 500\}$ clips are used for training, valiation, and testing, respectively. The LibriTTS dataset covers diverse recording environments with the sampling rate of 24 kHz. Following the division in [Lee *et al.*, 2023], {*train-clean-100*, *train-clean-300*, *train-other-500*} are for model training. The subsets *dev-clean* + *dev-other* are for objective comparisons, and *test-clean* + *test-other* are for subjective evaluations.

[4]https://github.com/jaywalnut310/vits/tree/main/filelists

Models	GT	HiFiGAN-V1	Avocodo	BigVGAN-base	BigVGAN	APNet2	Vocos	RNDVoC
MUSHRA	89.45±0.45	69.99±1.15	68.16±1.26	74.27±1.12	79.33±0.92	54.95±1.11	73.18±1.15	**80.74±0.99

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

Table 3: MUSHRA scores among different methods on the LibriTTS benchmark. The confidence level is 95%, and we performed a t-test comparing RNDVoC with BigVGAN.

Ids	Setting	PESQ↑	MCD↓	Periodicity↓ RMSE	V/UV↑ F1	Pitch↓ RMSE
1	Baseline	3.987	2.0471	0.0854	0.9714	21.3183
2	Remove omnidirectional phase loss	3.892	2.2137	0.0889	0.9697	21.9124
3	Remove RND mode	3.655	2.4976	0.9611	0.1114	26.0120
4	Set matrice $\{\mathcal{A}, \mathcal{A}^\dagger\}$ as learnable	3.645	2.5989	0.9591	0.1164	25.6292

Table 4: Ablation studies conducted on the LJSpeech dataset.

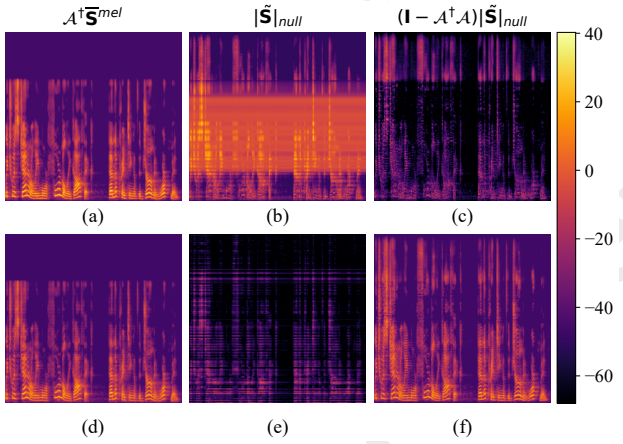

Figure 6: Spectral visualization of the range-space and nullspace with respect to whether $\{\mathcal{A}^\dagger, \mathcal{A}\}$ are fixed. (a)-(c) Thematrix parameters are fixed. (d)-(f) The matrix parametersare learnable.

4.2 Configurations

For the LJSpeech dataset, the number of mel-spectrogram F_m is set to 80, and the upper-bound frequency f_{max} is 8 kHz. For the LibriTTS dataset, 100 mel-bands are used with $f_{\max} = 12$ kHz. For both benchmarks, 1024-point FFT is set, with 1024 Hann window and 256 hop size.

A batch size of 16, a segment size of 16384, and an initial learning rate of 2e-4 are used for training. The AdamW optimizer [Loshchilov and Hutter, 2017] is employed, with $\{\beta_1 = 0.8, \beta_2 = 0.99\}$. The generator and discriminator are updated for 1 million steps, respectively.

4.3 Results and Analysis

In this study, various representative time-domain and T-F domain-based baselines are chosen for comparisons, including HiFiGAN [Kong *et al.*, 2020][5], iSTFTNet [Kaneko *et al.*, 2022][6], Avocodo [Bak *et al.*, 2023][7], WaveGlow [Prenger *et al.*, 2019][8], UnivNet [Jang *et al.*, 2021][9], BigVGAN [Lee *et al.*, 2023][10], APNet [Ai and Ling, 2023], APNet2 [Du *et al.*, 2023][11], FreeV [Lv *et al.*, 2024][12], and Vocos [Siuzdak, 2024][13].

Models	#Parm. (M)	#MACs (Giga/5s)	Datasets	PESQ↑	VISQOL↑
HiFiGAN-V2	0.92	9.6	LJSpeech	2.848	4.5419
		10.46	LibriTTS	2.308	4.3695
RNDVoC-Lite	0.71	9.54	LJSpeech	3.769	4.7817
		10.39	LibriTTS	3.834	4.8736
RNDVoC-UltraLite	0.08	1.66	LJSpeech	3.264	4.7006
		1.81	LibriTTS	3.499	4.8044

Table 5: Objective comparisons among lightweight neural vocoders.

Five metrics are involved in the objective evaluations: (1) Multi-resolution STFT (M-STFT) [Yamamoto *et al.*, 2020] is used to evaluate the spectral distance across multiple resolutions. (2) Wide-band version of Perceptual evaluation of speech quality (PESQ) [Rec, 2005] is chosen to assess the objective speech quality. (3) Mel-cepstral distortion (MCD) [Kubichek, 1993] measures the difference between mel-spectrograms through dynamic time wrapping (DTW). (4) Periodicity RMSE, V/UV F1 score, and pitch RMSE [Morrison *et al.*, 2022], which are regarded as major artifacts for non-autoregressive neural vocoders. (5) 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.

Except for objective metrics, we also conduct the MUSHRA testing based on the BeaqleJS platform [Kraft and Zölzer, 2014] for subjective evaluations. Thirty-five participants majoring in audio signal processing are engaged in the testing. For the former, each person rates the speech processed by different algorithms on a scale ranging from 0 to 100 in terms of the overall similarity compared to the reference clips.

Comparisons with SOTA Methods

Tables 1 and 2 present the objective comparisons on the LJSpeech and LibriTTS datasets, respectively. Several valuable observations can be made. First, T-F domain-based

[5]https://github.com/jik876/hifi-gan
[6]https://github.com/rishikksh20/iSTFTNet-pytorch
[7]https://github.com/ncsoft/avocodo
[8]https://github.com/NVIDIA/waveglow
[9]https://github.com/maum-ai/univnet
[10]https://github.com/NVIDIA/BigVGAN
[11]https://github.com/redmist328/APNet2
[12]https://github.com/BakerBunker/FreeV
[13]https://github.com/gemelo-ai/vocos

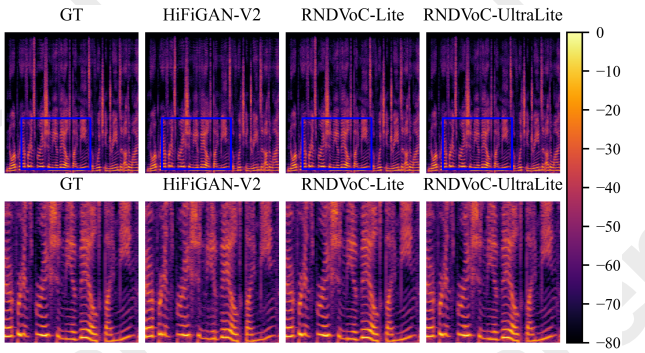

Figure 7: Spectral visualization generated by three light-weight neural vocoders, namely HiFiGAN-V2, RNDVoC-Lite, and RNDVoC-UltraLite. The audio clip is a speech voice from the LibriTTS test set.

methods exhibit overall faster inference speed over time-domain-based methods. This is mainly due to the use of STFT and its inverse transform, *i.e.*, iSTFT, and no upsampling opeartion is required. Second, T-F domain based methods possess overall notably less computation complexity, *e.g.*, 5.8 GMACs of Vocos vs. 152.9 GMACs of HiFiGAN. This has made T-F domain based neural vocoders attractive more recently. Third, compared with BigVGAN, the speech quality of existing T-F domain based neural vocoders are still notably inferior. Thanks to the fined-grained modeling along the time and sub-band axes, the proposed RNDVoC enjoys both light-weight network structure and promising performance. To be specific, with less than 3% trainable parameters and 10% computation cost, our approach yields comparable performance over BigVGAN in the LJSpeech dataset, and better performance in the LibriTTS benchmark. It is noteworthy that even compared with BigVGAN trained for 5M steps, our method still behaves competitive. It fully validates the effectiveness of the proposed approach.

Table 3 gives the MUSHRA results on the test set of LibriTTS dataset. One can observe that our RNDVoC is superior to BigVGAN with statistical difference ($p < 0.05$), further validating the advantage of our method in subjective quality. We notice that the subjective score of APNet2 is notably inferior to other methods. According to the feedback from the listeners, audible husky buzzing artifacts can arise in the generated utterances from APNet2, which leads to biased low score.

Figure 5 shows the spectral visualizations among different models, where the clip is a vocal voice from the out-of-distribution MUSDB18 [Rafii *et al.*, 2017] test set. Evidently, compared with other baselines, our approach can better recover the harmonic details.

Ablation Studies

Table 4 presents the ablation studies based on the LJSpeech dataset. First, we replace the proposed omnidirectional phase loss with that in [Du *et al.*, 2023], as shown from id1 to id2. One can observe clear performance degradation in objective metrics, indicating the effectiveness of the omnidirectional phase loss. Based on id2, we further remove the RND mode in target reconstruction, *i.e.*, the explicit superimposition between $|\tilde{\mathbf{S}}|_{range}$ and $\left(\mathbf{I} - \mathcal{A}^{\dagger}\mathcal{A}\right)|\tilde{\mathbf{S}}|_{null}$ is changed into the direct target mapping. From id2 to id3, notable performance degradation can be observed, indicating the significance of the proposed RND modeling. Then based on id2, we set the matrice $\{\mathcal{A}^{\dagger}, \mathcal{A}\}$ as learnable, trying to break the orthogonality between the two sub-spaces. From id2 to id4, we also observe the performance degradation, indicating the significance of orthogonality between two sub-spaces. In Figure 6, we visualize the estimated components form the range-space and null-space in terms of whether $\{\mathcal{A}^{\dagger}, \mathcal{A}\}$ are fixed. One can observe that if the parameters are fixed, the null-space, *i.e.*, $\left(\mathbf{I} - \mathcal{A}^{\dagger}\mathcal{A}\right)|\tilde{\mathbf{S}}|_{null}$ can capture the sparse and detailed harmonic structure, as shown in Figure 6(c). In contrast, in the learnable scenario, the estimation of the null-space is no longer sparse, indicating that the orthogonality property may not hold.

Toward Light-weight Design

To meet the requirements of edge-devices in practical scenarios, we investigate the potential of our method in light-weight design. Concretely, the number of DPB B is reduced into 4. When the channel size C is squeezed to 128 and 32, we obtain the light and ultra-lite version, namely abbreviated as **RNDVoC-Lite** and **RNDVoC-UltraLite**. Table 5 presents the performance of different light-weight models. Evidently, thanks to the efficacy of RND in information preservation, where the learning network only needs to concentrate on the null-space part, even only with 0.08M, our method still notably outperforms HiFiGAN-V2 in PESQ and VISQOL, which validates the effectiveness of our method.

In Figure 7, we present the spectral representations reconstructed by three lightweight vocoders. Notably, while HiFiGAN-V2 exhibits significant harmonic blurring (see the blue-boxed zoomed region), both RNDVoC-Lite and RNDVoC-UltraLite retain clear harmonic details. This contrast further highlights the superiority of our approach in lightweight design scenarios.

5 Conclusion

In this paper, we propose an innovative T-F domain-based neural vocoders. We bridge the connection with classical range-null decomposition theory, where the range-space aims to convert the original acoustic feature in the mel-scale domain into the target linear-scale domain, and null-space serves to reconstruct the remaining spectral details. Based on that, a novel dual-path network structure is devised, where the cross-band and narrow-band modules are devised to efficiently model the relations in the sub-band and time axes. Extensive experiments are conducted on the LJSpeech and LibriTTS benchmarks to validate the efficacy of the proposed method.

References

[Ai and Ling, 2023] Yang Ai and Zhen-Hua 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.

[Bak *et al.*, 2023] Taejun Bak, Junmo Lee, Hanbin Bae, Jinhyeok Yang, Jae-Sung Bae, and Young-Sun Joo. Avocodo: Generative adversarial network for artifact-free vocoder. In *Proc. AAAI*, volume 37, pages 12562–12570, 2023.

[Baraniuk *et al.*, 2010] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56(4):1982–2001, 2010.

[Chen *et al.*, 2021] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating Gradients for Waveform Generation. In *Proc. ICLR*, 2021.

[Dinh *et al.*, 2016] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[Du *et al.*, 2023] Hui-Peng Du, Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. APNet2: High-Quality and High-Efficiency Neural Vocoder with Direct Prediction of Amplitude and Phase Spectra. In *Proc. NCMSC*, pages 66–80. Springer, 2023.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. In *Proc. ICCV*, pages 1026–1034, 2015.

[Hines *et al.*, 2015] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. ViSQOL: an objective speech quality model. *EURASIP J. Audio Speech Music Process.*, 2015:1–18, 2015.

[Huang *et al.*, 2022a] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. In *Proc. IJCAI*, pages 4157–4163, 2022.

[Huang *et al.*, 2022b] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proc. ACMMM*, pages 2595–2605, 2022.

[Jang *et al.*, 2021] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. UnivNet: A neural Vocoder with Multi-resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. Interspeech*, pages 2207–2211, 2021.

[Juvela *et al.*, 2019] Lauri Juvela, Bajibabu Bollepalli, Vassilis Tsiaras, and Paavo Alku. 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, 2019.

[Kalchbrenner *et al.*, 2018] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *Proc. ICML*, pages 2410–2419. PMLR, 2018.

[Kaneko *et al.*, 2022] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform. In *Proc. ICASSP*, pages 6207–6211. IEEE, 2022.

[Kawahara, 2006] Hideki Kawahara. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

[Keith and Linda, 2017] I. Keith and J. Linda. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. Accessed: 2025-01-12.

[Kim *et al.*, 2019] Sungwon Kim, Sang-gil Lee, Jongyoon Song, Jaehyeon Kim, and Sungroh Yoon. FloWaveNet: A Generative Flow for Raw Audio. In *Proc. ICML*, pages 3370–3378. PMLR, 2019.

[Kong *et al.*, 2020] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, pages 17022–17033, 2020.

[Kong *et al.*, 2021] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *Proc. ICLR*, 2021.

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

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

[Kumar *et al.*, 2019] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Proc. NeurIPS*, 32, 2019.

[Lee *et al.*, 2022] Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *Proc. ICLR*, 2022.

[Lee *et al.*, 2023] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A Universal Neural Vocoder with Large-Scale Training. In *Proc. ICLR*, 2023.

[Lei Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.

[Li *et al.*, 2022] Andong Li, Shan You, Guochen Yu, Chengshi Zheng, and Xiaodong Li. Taylor, Can You Hear Me Now? A Taylor-Unfolding Framework for Monaural

Speech Enhancement. In *Proc. IJCAI*, pages 4193–4200, 2022.

[Li *et al.*, 2025] Andong Li, Zhihang Sun, Fengyuan Hao, Xiaodong Li, and Chengshi Zheng. Neural vocoders as speech enhancers. *arXiv preprint arXiv:2501.13465*, 2025.

[Liu *et al.*, 2023] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *Proc. ICML*, pages 21450–21474. PMLR, 2023.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2022] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Proc. NeurIPS*, 35:5775–5787, 2022.

[Lv *et al.*, 2024] Yuanjun Lv, Hai Li, Ying Yan, Junhui Liu, Danming Xie, and Lei Xie. FreeV: Free Lunch For Vocoders Through Pseudo Inversed Mel Filter. In *Proc. Interspeech*, pages 3869–3873, 2024.

[Morise *et al.*, 2016] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans Inf Syst*, 99(7):1877–1884, 2016.

[Morrison *et al.*, 2022] Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. Chunked Autoregressive GAN for Conditional Waveform Synthesis. In *Proc. ICLR*, 2022.

[Oord *et al.*, 2018] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *Proc. ICML*, pages 3918–3926. PMLR, 2018.

[Ping *et al.*, 2020] Wei Ping, Kainan Peng, Kexin Zhao, and Zhao Song. Waveflow: A compact flow-based model for raw audio. In *Proc. ICML*, pages 7706–7716. PMLR, 2020.

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

[Quan and Li, 2024] Changsheng Quan and Xiaofei Li. SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 32:1310–1323, 2024.

[Rafii *et al.*, 2017] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation. 2017.

[Ramachandran *et al.*, 2017] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 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.

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

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

[Van Den Oord *et al.*, 2016] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016.

[Wang *et al.*, 2017] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech*, pages 4006–4010, 2017.

[Woo *et al.*, 2023] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proc. CVPR*, pages 16133–16142, 2023.

[Yamamoto *et al.*, 2020] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. ICASSP*, pages 6199–6203. IEEE, 2020.

[Yu *et al.*, 2023] Jianwei Yu, Yi Luo, Hangting Chen, Rongzhi Gu, and Chao Weng. High Fidelity Speech Enhancement with Band-split RNN. In *Proc. Interspeech*, pages 2483–2487, 2023.

[Zen *et al.*, 2019] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech*, pages 1526–1530, 2019.

[Zhou *et al.*, 2024] Rui Zhou, Xian Li, Ying Fang, and Xiaofei Li. Mel-FullSubNet: Mel-Spectrogram Enhancement for Improving Both Speech Quality and ASR. *arXiv preprint arXiv:2402.13511*, 2024.