RETHINKING THE JOINT ESTIMATION OF MAGNITUDE AND PHASE FOR TIME-FREQUENCY DOMAIN NEURAL VOCODERS

Lingling Dai**, Andong Li**, Tong Lei†, Meng Yu‡, Xiaodong Li**, Chengshi Zheng**

⋆ Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
†Tencent AI Lab, Shen Zhen, China
‡Tencent AI Lab, Bellevue, WA, USA
*University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

Time-frequency (T-F) domain-based neural vocoders have shown promising results in synthesizing high-fidelity audio. Nevertheless, it remains unclear on the mechanism of effectively predicting magnitude and phase targets jointly. In this paper, we start from two representative T-F domain vocoders, namely Vocos and APNet2, which belong to the single-stream and dual-stream modes for magnitude and phase estimation, respectively. When evaluating their performance on a large-scale dataset, we accidentally observe severe performance collapse of APNet2. To stabilize its performance, in this paper, we introduce three simple yet effective strategies, each targeting the topological space, the source space, and the output space, respectively. Specifically, we modify the architectural topology for better information exchange in the topological space, introduce prior knowledge to facilitate the generation process in the source space, and optimize the backpropagation process for parameter updates with an improved output format in the output space. Experimental results demonstrate that our proposed method effectively facilitates the joint estimation of magnitude and phase in APNet2, thus bridging the performance disparities between the single-stream and dual-stream vocoders.

Index Terms— Neural Vocoder, Joint Phase and Magnitude Estimation, Architectural Topology

1. INTRODUCTION

Neural vocoders aim to reconstruct audible waveforms from intermediate acoustic features or hidden representations using deep neural networks (DNNs) [1]. Driven by the escalating demands of audio-related applications, they have garnered substantial attention and achieved remarkable progress in recent years [2, 3, 4, 5, 6, 7]. Existing methods can be roughly categorized into time-domain and time-frequency (T-F) domain approaches, where the latter have gained increasing prominence due to the appealing advantages in perceptual quality and inference efficiency [8, 9, 10].

Unlike consecutive upsampling operations, T-F domain-based neural vocoders typically estimate the real and imaginary (RI) parts or magnitude-phase (MP) pairs, followed by the inverse short-time Fourier transform (iSTFT) for target waveform generation. Therefore, a core challenge lies in *how to effectively predict magnitude and phase targets jointly?* When inspecting related audio processing tasks *e.g.*, speech enhancement (SE) and bandwidth extension (BWE), we notice that two typical architectural topologies are usually adopted: *single-stream* and *dual-stream*. For the former, magnitude and phase share most deep modeling units, with separate output heads for dual-target prediction [11, 12, 13, 14, 15, 16, 17]. For
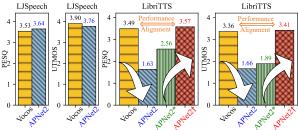

Fig. 1. The performance illustration of Vocos and APNet2 on the LJSpeech dataset and LibriTTS dataset, where APNet2* is aligned with Vocos in implementation details and APNet2† denotes our improved version.

the latter, by contrast, dual-stream architectures employ independent modeling streams with minimal or no inter-stream interactions for MP prediction [18, 19, 20]. Nonetheless, it still remains an open question to determine an optimal architecture modality.

When switching back to the vocoder task, we compare two representative models: Vocos [1] and APNet2 [18]. The reasons are two-fold. First, they belong to the single-stream and dual-stream, respectively. Besides, both of them utilize the ConvNext block [21] and its variant, *i.e.*, ConvNext v2 [22], as the basic modeling unit. We evaluate their performance on two benchmarks: LJSpeech [23] and LibriTTS [24], and PESQ and UTMOS scores are shown in Fig. 1. Surprisingly, while both models exhibit comparable reconstruction quality on the single-speaker LJSpeech dataset, APNet2 suffers from significant performance collapse on LibriTTS, a larger benchmark with diverse acoustic recordings. To investigate the root, we first adjust APNet2 to align with Vocos in terms of basic modeling unit and training configurations, yielding APNet2*, however, a large performance gap still exists[1]. Therefore, we conjecture the "dual-stream design" to be the primary reason to cause the performance collapse in the vocoder task. Further, for more diverse generation scenarios, due to the inherent wrapping property of phase, in the existing dual-stream architecture modality, the phase branch may lack **adequate** feature guidance from the magnitude branch[2], which seems especially significant for waveform reconstruction in the T-F domain.

To this end, we propose to resolve the performance collapse observed in existing "dual-stream" based neural vocoders with three simple yet effective strategies, each targeting the **topological space**,

---

[1] Similar performance gap is observed in other large-scale datasets, *e.g.*, Libriheavy dataset [25].

[2] Although some feature interactions strategies can be introduced for guidance, we observe marginal improvements, as presented in Table 2.
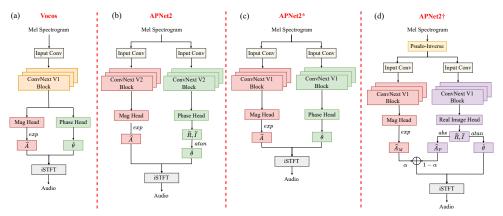
Fig. 2. The model structure illustrations of Vocos, APNet2, APNet2* and APNet2†.

source space, and **output space**. Specifically, for the topological space, we propose a straightforward and intuitive strategy by modifying the architectural topology with more information interaction and weight sharing between the magnitude and phase streams. For the source space, when the compressed Mel spectrum is considered as the input, it may be unsuitable to serve as the information source, as it is misaligned with the magnitude spectrum at the T-F bin level. Motivated by [10, 12], we project the Mel spectrum into the range space of the target spectrum by applying the pseudo-inverse Mel filter, providing a more aligned and informative input for joint MP prediction. For the output space, instead of direct phase prediction or its trigonometric representations, we enforce the phase branch to generate partial magnitude components, enabling cross-stream magnitude optimization from the magnitude branch. This mechanism can establish implicit feature dependency between the two streams. Comprehensive results demonstrate that our proposed strategies effectively mitigate the performance collapse in APNet2. Our contributions are summarized as follows:

❏ (1) We conduct quantitative experiments to investigate the performance disparities across different architectural topologies, taking Vocos and APNet2 as a typical example.

❏ (2) We reformulate the generation and optimization process in magnitude and phase modeling from three distinct spaces.

❏ (3) We conduct extensive experiments to validate the effectiveness of the proposed strategies, which effectively improve the model performance on joint magnitude and phase estimation.
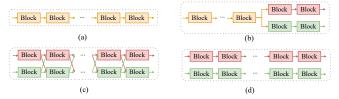

Fig. 3. Overall diagram of different structural designs. (a) Shared structure between magnitude and phase estimations. (b) Partially shared structure between magnitude and phase estimations. (c) Two-stream structure with shuffled connections between magnitude and phase estimations. (d) Two-stream structure with independent connections between magnitude and phase estimations.

2. METHODOLOGY

2.1. Rethinking Frequency Vocoders: A Topology Perspective

For T-F domain-based neural vocoders, direct or implicit estimation of magnitude and phase is crucial for high-quality audio generation.

In this paper, we choose two open-sourced vocoders, namely Vocos[3] and APNet2[4], which possess relatively similar network modeling units, loss functions, training pipelines, but different in architectural topology, as a typical example to investigate the effective way of joint estimation of MP pairs, and the overall network diagrams are shown in Fig. 2(a)-(b), respectively.

We evaluate the generation quality of the two vocoders on two common benchmarks: the LJSpeech [23] and LibriTTS [24] datasets. The former is a single-speaker dataset whose duration is around 24 h, and the latter is a more acoustic-diverse dataset, containing around 585 h duration recorded from 2,456 speakers. As illustrated in Fig. 1, Vocos and APNet2 achieve comparable PESQ and UTMOS scores on the LJSpeech dataset. However, APNet2 undergoes severe performance collapse on the more diverse and larger LibriTTS benchmark[5], as shown in id1-id2 of Table 1. To isolate the key factor that led to the phenomenon, we first conduct quantitative experiments by gradually substituting the network unit, output type, and loss functions in APNet2 with those in Vocos. As presented in id3-id5 of Table 1, after gradually excluding all the possible factors above, we do not observe substantial performance improvement of APNet2, which points to the remaining factor: **architectural topology**, or more specifically, the information interaction mechanism between magnitude and phase targets.

2.2. Optimizing in the Topological Space

When inspecting the generation process of MP pairs in Vocos and APNet2, one can observe they belong to Fig. 3(a) and (d), respectively. Concretely, in Vocos, magnitude and phase share the main modeling units, and are only independent at the output heads, whereas in APNet2, independent modeling streams are utilized for magnitude and phase, respectively, and no interactions are involved. Recently, in addressing the phase estimation challenges due to its intrinsic wrapping characteristics, feature guidance from the magnitude branch has been proven helpful [11, 26]. Therefore, we attempt to optimize the architecture of APNet2 from the topological space. To be specific, in Fig. 3(b)-(c), we present two topology variants, which enable more information interaction or weight sharing between the magnitude and phase streams. In Fig. 3(b), the magnitude and phase branches share partial feature modeling while leaving separate modeling capabilities for both targets. Moreover, in Fig. 3(c),

---

Table 1. The experimental results of transforming APNet2 to Vocos on the LibriTTS dataset. "Direct" denotes directly estimate the phase spectrum, and "Atan" denotes estimating the phase via the Atan function. "V-L" and "A-L" denotes adopting the loss setups by Vocos and APNet2, respectively.

Ids	Model	Unit	Output	Loss	Topology	PESQ↑	UTMOS↑	VISQOL↑	MCD↓	M-STFT↓	V/UV F1↑	Periodicity RMSE↓	Pitch RMSE↓	Param. (M)
1	Vocos	ConvNext v1	Direct	V-L	Shared	**3.487**	**3.356**	**4.861**	3.260	0.921	0.945	**0.124**	32.846	13.53
2		ConvNext v2	Atan	A-L		1.626	1.663	4.397	6.173	1.674	0.752	0.314	254.807	31.57
3	APNet2	ConvNext v1	Atan	A-L	Separate	2.194	2.203	4.558	5.215	1.371	0.893	0.194	93.412	31.53
4		ConvNext v1	Direct	A-L		1.529	1.334	4.294	7.183	2.162	0.697	0.365	236.895	26.54
5		ConvNext v1	Direct	V-L		2.556	1.886	4.755	4.232	1.123	0.905	0.193	66.027	26.54

Table 2. The experimental results of different architectural topologies on the LibriTTS dataset. * denotes that the model is only different from Vocos in architectural topology. "R" indicates the number of weight-sharing layers.

Ids	Model	Topology	PESQ↑	UTMOS↑	VISQOL↑	MCD↓	M-STFT↓	V/UV F1↑	Periodicity RMSE↓	Pitch RMSE↓	Param. (M)
1	Vocos-D	Shared	**3.607**	**3.461**	**4.883**	**3.145**	**0.897**	**0.952**	**0.114**	34.694	29.73
2	APNet2*	Separate	2.556	1.886	4.755	4.232	1.123	0.905	0.193	66.027	26.54
3	APNet2*-SF	Shuffle	3.400	3.344	4.861	3.297	0.903	0.946	0.126	35.622	26.54
4	APNet2*-PS	Partially Shared (R=2)	3.382	3.182	4.863	3.373	0.912	0.943	0.130	35.921	28.99
5	APNet2*-PS	Partially Shared (R=4)	3.419	3.136	4.862	3.359	0.914	0.944	0.128	**32.628**	30.72
6	APNet2*-PS	Partially Shared (R=6)	3.461	3.367	4.873	3.265	0.894	0.948	0.120	31.980	31.07

we introduce shuffled operations between the magnitude and phase branches at various layer levels, with information exchange from previous layers while maintaining independent processing.

2.3. Optimizing in the Source Space

Apart from substituting a more effective architectural topology for joint magnitude and phase prediction, we also consider solutions with minimal modifications on model structures for further improvements. Instead of directly posing feature sharing or information interaction mechanism from the topological space, introducing a magnitude-related term that aligns with the target magnitude in the source space seems to alleviate the need for information exchange from the magnitude stream. Additionally, adopting prior knowledge as input features also benefits model convergence and performance improvement [27, 28, 13, 29, 30].

Motivated by [10, 12], we use the pseudo-inverse mel-spectrogram as an input feature, which serves as an effective representation of range space. Let $\mathbf{S} \in \mathbb{R}^{F_m \times T}$ denote the mel-spectrogram, where F_m is the number of mel bins and T is the number of frames. The pseudo-inverse mel-spectrogram $\mathbf{S}_{\text{prior}} \in \mathbb{R}^{F \times T}$, with F being the number of STFT frequency bins, is computed as:

$$\mathbf{S}_{\text{prior}} = \mathbf{W}_{\text{mel}}^{\dagger}\mathbf{S}, \qquad (1)$$

where $\mathbf{W}_{\text{mel}}^{\dagger} \in \mathbb{R}^{F \times F_m}$ is the pseudo-inverse of the mel filter bank matrix. This operation reconstructs a coarse linear spectrogram from the mel-spectrogram, providing richer frequency information as input to the vocoder. By taking $\mathbf{S}_{\text{prior}}$ as the input, the model benefits from a more informative initialization for both magnitude and phase prediction, as well as finer guidance for the phase stream.

2.4. Optimizing in the Output Space

In T-F domain-based neural vocoders, the magnitude is usually relatively easier for prediction due to its clear structural characteristic, while phase representation remains an active and open area of research. For instance, a mainstream of models predicts phase with the Atan operation on network outputs [31]. More recently, Vocos [1] proposed to estimate the phase directly and apply the cosine and sine calculations for RI estimation. However, they still face the following challenges:

1) **Wrapping issue**: The phase spectrogram is inherently wrapped within the range of $(-\pi, \pi]$, leading to discontinuities that can complicate the learning process for neural networks.

2) **Multiple-solution problem**: $\text{Atan}(k\tilde{I}, k\tilde{R})$ or $\theta + 2h\pi$ with different values of $k \in \mathbb{R}$ or $h \in \mathbb{Z}$ target to one final result but with different network outputs, making the optimization more difficult.

3) **Uncertain coherence with magnitude**: The generation of phase is less dependent on the generated magnitude for architectural topologies with non-shared parts, which may lead to inconsistent results when reconstructing the waveform.

As illustrated in Fig. 2(d), we propose a simple yet effective modification on APNet2 by enforcing the phase components to estimate partial magnitude components, which facilitates the joint parameter update in the backpropagation process. Specifically, we utilize the estimated real component $\tilde{R}$ and imaginary component $\tilde{I}$ to add a magnitude-related item $\hat{A}_p$ for the final magnitude prediction:

$$\hat{A} = \alpha\hat{A}_M + (1 - \alpha)\,\hat{A}_p, \qquad (2)$$

$$\hat{A}_p = \sqrt{\tilde{R}^2 + \tilde{I}^2}, \qquad (3)$$

where α is a trainable weighting hyper-parameter. During backpropagation, taking the MSE magnitude-loss function as an example, the gradient flow of the network parameter Φ is computed as:

$$\frac{\partial\left(\hat{A} - A\right)^2}{\partial\Phi} = 2\left(\hat{A} - A\right)\left[\frac{\partial\left(\alpha\hat{A}_M\right)}{\partial\Phi_M} + \frac{\partial\left((1 - \alpha)\,\hat{A}_P\right)}{\partial\Phi_P}\right], \qquad (4)$$

where the magnitude-related rather than phase-related quantities are introduced to drive the overall parameter update, thereby alleviating the optimization challenges brought by the incorrect phase estimation. Furthermore, merely the magnitude loss term allows the whole network parameters to be updated, which indirectly enhances coherence between phase and magnitude generation. Besides, the value of RI components is also constrained by magnitude under such optimization, thereby relieving the problem of multiple solutions.

3. EXPERIMENTAL SETUP

The experiments are based on LibriTTS benchmark. Following the data split principle in BigVGAN [7], we utilize all the training subsets {*train-clean-100, train-clean-360, train-other-500*} for

Table 3. The experimental results of making modifications from the source space and output space on the LibriTTS dataset. † denotes that the model further adopts the prior source and the MI-RI output type. "Prior" denotes use pseudo-inverse mel-spectrogram as input.

Ids	Model	Topology	Source	Output	PESQ↑	UTMOS↑	VISQOL↑	MCD↓	M-STFT↓	V/UV F1↑	Periodicity RMSE↓	Pitch RMSE↓	Param. (M)
1			Raw	Direct	2.556	1.886	4.755	4.232	1.123	0.905	0.193	66.027	26.54
2			Prior	Direct	3.197	2.987	4.846	3.500	0.944	0.936	0.138	42.139	29.50
3	APNet2*	Separate	Raw	MI-RI	3.436	3.202	4.869	3.367	0.929	0.948	0.123	30.394	26.80
4				MB	3.436	3.202	4.869	3.368	0.929	0.948	0.123	30.348	
5	Vocos†-D	Shared	Prior	MI-RI	3.576	3.369	4.880	3.247	0.897	0.948	0.121	35.721	32.35
6	APNet2†	Separate	Prior	MI-RI	3.569	**3.410**	**4.896**	**3.199**	**0.883**	0.948	0.117	29.580	29.76
7	APNet2†-SF	Shuffle	Prior	MI-RI	3.510	3.325	4.873	3.282	0.900	0.949	0.120	34.863	29.76
8	APNet2†-PS	Partially Shared (R=2)	Prior	MI-RI	**3.610**	**3.410**	4.886	3.219	**0.883**	0.950	**0.114**	30.602	30.95
9	APNet2†-PS	Partially Shared (R=4)	Prior	MI-RI	3.573	3.396	4.887	3.239	0.884	**0.951**	0.115	**27.686**	32.90
10	APNet2†-PS	Partially Shared (R=6)	Prior	MI-RI	3.511	3.314	4.872	3.289	0.900	0.948	0.120	31.799	33.46

training, and {*test-clean*, *test-other*} subsets for evaluation. We employ official implementations of Vocos and APNet2. Note that the performance may be different from the official checkpoints as we remove the amplitude augmentation during training and align the learning rate to 5e-4 for fair comparisons. For objective evaluations, we employ eight objective metrics: perceptual evaluation of speech quality (PESQ) [32], UTMOS [33], Virtual Speech Quality Objective Listener (VISQOL) [34], mel-cepstral distortion (MCD) [35], multi-resolution STFT loss (M-STFT) [36], Periodicity Root Mean Square Error (RMSE), V/UV F1 score, and pitch RMSE [37].

4. RESULTS AND ANALYSIS

4.1. Evaluation on Different Architectural Topologies

To verify the effectiveness of the feature-sharing mechanism introduced between the magnitude and phase streams in the topological space, we conduct comprehensive experiments on models of different architectural topologies with close parameter sizes. We denote the partially shared structure as APNet2*-PS and the two-stream structure with shuffled connections as APNet2*-SF, and a large version of Vocos as Vocos-D. As presented in Table 2, one can observe that APNet2*-PS and APNet2*-SF significantly improve the performance of APNet2, revealing that the choice of architectural topologies can significantly impact the performance of frequency-domain neural vocoders, as it influences the way magnitude and phase information are processed and integrated during audio generation.

4.2. Evaluation of Source Space and Output Space

In Table 3, we first provide ablation studies on the effectiveness of modifications in the source space and the output space, respectively. By introducing the pseudo-inverse mel-spectrogram in the source space, the performance of APNet2* improves significantly, with PESQ increasing from 2.556 to 3.197, indicating the effectiveness of applying magnitude initialization. Additionally, after replacing the MI-RI output format, the performance of APNet2* also improves. As we separate the magnitude item from the magnitude branch (MB), one can observe only a very slight difference from the final results, revealing that the magnitude component $\hat{A}_P$ from phase branch attributes merely no influence in the forward process of audio generation and our proposed MI-RI contributes positively to the parameter update in the backpropagation process when the magnitude and phase streams involves insufficient information interaction. We further apply our proposed method in source space and output space to all the above-mentioned architectural topologies. As presented in Table 3, our proposed method effectively improves the performance of models with less effective information interaction between magnitude and phase streams, and also minimizes their performance disparities with Vocos-D.
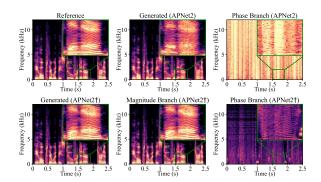

Fig. 4. Spectrogram visualization of the reference, generated audio signal by APNet2† and APNet2, and the intermediate results of APNet2† and APNet2, respectively.

4.3. Visualization of the Intermediate Results

In Fig. 4, we present the spectrogram visualization of the reference, the generated audio signal from APNet2† and APNet2, and the intermediate results of APNet2† and APNet2, where the magnitude components are derived separately from the magnitude branch and the phase branch. It can be observed that the magnitude component $\hat{A}_P$ from the phase branch in APNet2† exhibits clear harmonic structures that align with the spectrogram of the generated audio, whereas in APNet2, the magnitude component derived from the phase branch does not exhibit clear structural details. This finding further demonstrates the effective improvement of our proposed method in enhancing the coherence between magnitude and phase. Additionally, the magnitude component $\hat{A}_M$ from the magnitude branch illustrates no significant difference compared with the final generated spectrogram, further verifying that our proposed MI-RI does not affect the forward process.

5. CONCLUSIONS

In this paper, we revisit the joint magnitude and phase estimation mechanism on two typical neural vocoders: Vocos and APNet2. To bridge their performance gaps, we propose simple yet effective modifications in three distinct spaces, including the topological space, the source space, and the output space. Specifically, we propose to enhance the information interaction in the topological space, introduce prior knowledge to the source space, and optimize the backpropagation process with an improved output format in the output space. Experimental results demonstrate that our proposed method effectively aligns the performance of APNet2 with Vocos, both jointly and exclusively.

6. REFERENCES

[1] Hubert Siuzdak, "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," in *Proc. ICLR*, 2024, pp. 1–15.

[2] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al., "Seed-tts: A family of high-quality versatile speech generation models," in *arXiv:2406.02430*, 2024.

[3] Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D Plumbley, "AudioSR: Versatile audio super-resolution at scale," in *Proc. ICASSP*, 2024, pp. 1076–1080.

[4] Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang, "VoiceFixer: Toward general speech restoration with neural vocoder," in *arXiv:2109.13731*, 2021.

[5] Andreev, Pavel and Alanov, Aibek and Ivanov, Oleg and Vetrov, Dmitry, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *Proc. ICASSP*, 2023, pp. 1–5.

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

[7] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon, "Bigvgan: A universal neural vocoder with large-scale training," in *Proc. ICLR*, 2022.

[8] 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*, 2022, pp. 6207–6211.

[9] Dinh Son Dang, Tung Lam Nguyen, Bao Thang Ta, Tien Thanh Nguyen, Thi Ngoc Anh Nguyen, Dang Linh Le, Nhat Minh Le, and Van Hai Do, "LightVoc: An Upsampling-Free GAN Vocoder Based On Conformer And Inverse Short-time Fourier Transform," in *Proc. Interspeech*, 2023, pp. 3043–3047.

[10] 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*, 2024, pp. 3869–3873.

[11] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng, "PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network," in *Proc. AAAI*, 2020, vol. 34, pp. 9458–9465.

[12] Andong Li, Tong Lei, Zhihang Sun, Rilin Chen, Erwei Yin, Xiaodong Li, and Chengshi Zheng, "Learning Neural Vocoder from Range-Null Space Decomposition," in *Proc. IJCAI*, 2025.

[13] Guochen Yu, Xiguang Zheng, Nan Li, Runqiang Han, Chengshi Zheng, Chen Zhang, Chao Zhou, Qi Huang, and Bing Yu, "BAE-Net: a Low Complexity and High Fidelity Bandwidth-Adaptive Neural Network for Speech Super-Resolution," in *Proc. ICASSP*, 2024, pp. 571–575.

[14] Ke Tan and DeLiang Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[15] Chenggang Zhang, Jinjiang Liu, and Xueliang Zhang, "A Complex Spectral Mapping with Inplace Convolution Recurrent Neural Networks For Acoustic Echo Cancellation," in *Proc. ICASSP*, 2022, pp. 751–755.

[16] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. ICASSP*, 2023, pp. 1–5.

[17] Houjian Guo, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro, "Quickvc: A lightweight vits-based any-to-many voice conversion model using istft for faster conversion," in *Proc. ASRU*, 2023, pp. 1–7.

[18] 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*, 2023, pp. 66–80.

[19] Ye-Xin Lu, Yang Ai, Hui-Peng Du, and Zhen-Hua Ling, "Towards High-Quality and Efficient Speech Bandwidth Extension with Parallel Amplitude and Phase Prediction," in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

[20] 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.*, vol. 31, pp. 2145–2157, 2023.

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proc. CVPR*, 2022, pp. 11976–11986.

[22] 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*, 2023, pp. 16133–16142.

[23] Keith Ito and Linda Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[24] 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*, 2019, pp. 1526–1530.

[25] Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey, "Libriheavy: A 50,000 hours asr corpus with punctuation casing and context," in *Proc. ICASSP*. IEEE, 2024, pp. 10991–10995.

[26] Chengyu Zheng, Xiulian Peng, Yuan Zhang, Sriram Srinivasan, and Yan Lu, "Interactive speech and noise modeling for speech enhancement," in *Proc. AAAI*, 2021, vol. 35, pp. 14549–14557.

[27] Jean-Marc Valin and Jan Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.

[28] Kun Song, Yongmao Zhang, Yi Lei, Jian Cong, Hanzhao Li, Lei Xie, Gang He, and Jinfeng Bai, "DSPGAN: a GAN-based universal vocoder for high-fidelity TTS by time-frequency domain supervision from DSP," in *Proc. ICASSP*, 2023, pp. 1–5.

[29] Peng Liu, Dongyang Dai, and Zhiyong Wu, "RFWave: Multi-band Rectified Flow for Audio Waveform Reconstruction," in *arXiv:2403.05010*, 2024.

[30] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. ICASSP*, 2021, pp. 6633–6637.

[31] Donald S. Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2016.

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

[33] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[34] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte, "Visqol: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, pp. 1–18, 2015.

[35] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993, vol. 1, pp. 125–128.

[36] 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*, 2020, pp. 6199–6203.

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