REVISITING VOCOS: THAT PHASINESS BUSINESS IN TIME-FREQUENCY NEURAL VOCODING

Ünal Ege Gaznepoğlu [1*], *Frank Zalkow* [2], *Mohammad Joshaghani* [2]
Emanuël A.P. Habets [1,2], *Nils Peters* [3], *Christian Dittmar* [2]

[1]International Audio Laboratories Erlangen, Germany
[2]Fraunhofer Institute for Integrated Circuits IIS, Germany
[3]Dept. of Electronic & Electrical Engineering, Trinity College Dublin, Ireland

ABSTRACT

Recently, neural vocoders utilizing time-frequency representations have been approaching the state-of-the-art quality of time-domain neural vocoders. Vocos is a notable example due to its computational efficiency, but its audio quality lags behind the time-domain vocoders and the reasons remain debated. Thus, in this study, we revisit Vocos from a phase reconstruction perspective. First, we quantify the gap between time-domain and time-frequency domain vocoders using bandlimited mel spectrograms as inputs. Later, via an ablation study, we verify the Vocos architecture is effective for magnitude modeling, but less so for phase. We then adapt the Vocos backbone to predict phase differences, a precursor for phase reconstruction, and identify 1D convolutional layers are hindering their accurate prediction. Our findings indicate that future research needs to focus on inductive biases that allow the architecture to better model the time-frequency structure of speech signals, without sacrificing the support for arbitrary input representations.

Index Terms— neural vocoders, phase recovery

1. INTRODUCTION

Neural vocoders synthesize time-domain audio signals from a lossy representations, typically mel spectrograms. State-of-the-art time-domain neural vocoders can reconstruct high-quality audio signals but are computationally expensive due to the cascaded temporal upsampling [1]. Approaches including multi-band synthesis can reduce this cost at the expense of some audio quality [2, 3]. Vocos [4] avoids neural upsampling by predicting complex-valued short-time Fourier transform (STFT) coefficients for synthesis via inverse STFT. However, its audio quality is slightly below the state-of-the-art time-domain vocoders [5], resulting in artifacts resembling *phasiness*, the loss of clarity associated also with classical methods [6].

Subsequent studies have examined Vocos, reaching differing interpretations and proposing divergent remedies. The original author of Vocos hypothesizes[1] that since phase recovery is an autoregressive task, the convolutional neural network (CNN)-based architecture is not suitable for it [7]. Notably, the magnitude and phase predictions of Vocos were found to deviate substantially from the ground truth [5, 8]. The inverse STFT partially compensates these errors through overlap–add redundancy, similar to the classical Griffin–Lim algorithm [9]. Few works report performance gains by incorporating loss terms between predicted and ground-truth magnitude and phase spectrograms [10–12] and using a pseudo-inverse for recovering the linear-scale magnitude spectrograms [12]. In WaveNeXt, the authors argue the inverse STFT is unnecessary because ConvNeXt blocks could directly predict the time-domain signal [5]. In WaveHax, the architecture is in the spotlight: ConvNeXt blocks with 2D convolutions refine a harmonic prior to exploit local time-frequency structure [8]. To summarize, the exact reasons for Vocos' phase modeling issues and the best way to address them remain unclear. Furthermore, the aforementioned studies have not compared Vocos to time-domain vocoders on an equal footing, preventing a clear assessment of the quality gap and its causes. So, in this work, we investigate why phase modeling is a bottleneck for time-frequency domain vocoders. We do so by comparing Vocos to the state-of-the-art time-domain vocoders, and later by analyzing Vocos through a phase reconstruction lens. This work has the following contributions:

- We compare Vocos to BigVGAN and find that the audio quality gap is still present, even after controlling for confounding factors such as training loss and discriminators.
- Training Vocos variants with oracle knowledge of either the magnitude or phase spectrograms reveals that 1D convolutions are effective for magnitude modeling, but less so for phase modeling.
- Then, we investigate whether the Vocos architecture can predict phase differences, a precursor for phase reconstruction that does not require autoregression, and find that it is not effective for this task either. We later show that switching to 2D convolutions substantially improves performance on this particular subtask.

2. METHODOLOGY (VOCODING)

2.1. Input representation

We use 80-band log mel spectrograms (for the frequency range 0–8 kHz) as input, at a sampling rate of 22.05 kHz. Bandlimited spectrograms could accentuate the gap between time-domain and time-frequency domain vocoders, since time-frequency domain vocoders are known to struggle with generating harmonic structure [8].

2.2. Model architectures

Vocos. Fig. 1 outlines the Vocos architecture. It consists of a Conv1D layer, followed by a stack of ConvNeXt [13] blocks, and a 'Head' block that interprets the network outputs as STFT coefficients. Finally, the time-domain waveform is reconstructed using the inverse STFT with a Hann window. In practice, we found that

*Work performed during an internship at Fraunhofer IIS.
[1]https://openreview.net/forum?id=vY9nzQmQBw

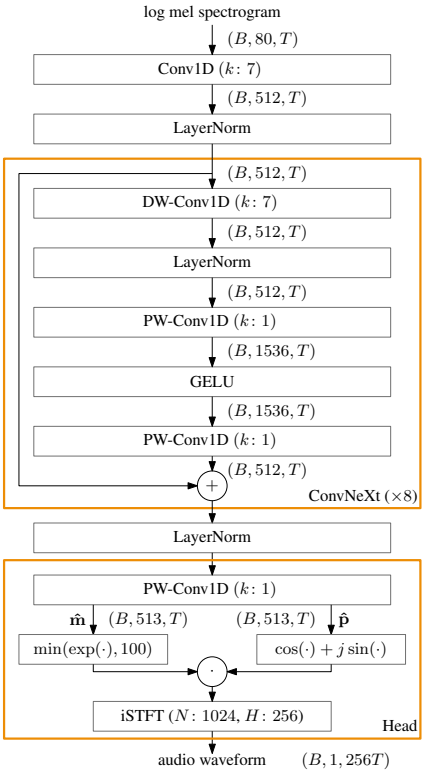

Fig. 1. The Vocos architecture, predicting log-magnitude and phase spectrograms ($\hat{\mathbf{m}}$, $\hat{\mathbf{p}}$). DW/PW denote depthwise/pointwise operations. Batch size and number of time frames are indicated by B, T.

the magnitude clamping prevents the network from predicting accurate magnitude spectrograms. Fig. 2 shows the histogram of the maximum STFT magnitude for all training samples, with -1 dBFS amplitude scaling. The default clamping threshold of 100 is set too low to accurately represent up to 41 % of the training samples. Based on the histogram, we adopt a threshold of 400, and advise future work to carefully consider the choice of this hyperparameter.

BigVGANv2. We retrained BigVGAN to serve as a benchmark time-domain vocoder [1]. An official checkpoint `bigvgan_v2_22khz_80band_fmax8k_256x` matches our input representation but was trained on a larger dataset, so we report results for both that checkpoint and our retrained model.

2.3. Discriminator

In our vocoder trainings, we utilize the multi-period discriminator [14] with periods of (2, 3, 5, 7, 11), EnCodec MS-STFT discriminator [15] for FFT sizes of (128, 256, 512, 1024, 2048), and the MS-SB-CQT discriminator [16].

3. METHODOLOGY (PHASE RECONSTRUCTION)

In Vocos, the Head is the only block that imposes phase-related inductive bias, by (elementwise) wrapping the predicted values to the $[-\pi, \pi)$ range. However, spectrograms are characterized by a much stronger inductive bias of consistency, in the form of a coupling between the time and frequency gradients of log-magnitude and phase spectrograms [17, 18]. So, time-frequency bins are influenced by neighboring elements in the time-frequency plane, and in the case of harmonic overtones, even by components spaced further apart along the frequency axis.

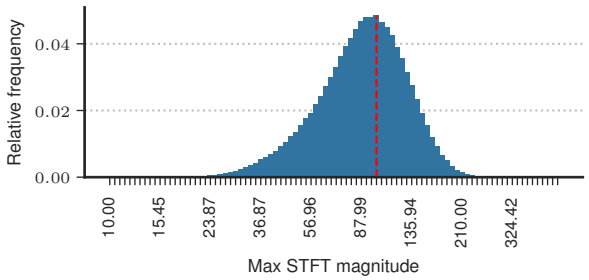

Fig. 2. Histogram of the maximum ground-truth STFT magnitude for all training samples (with -1 dBFS amplitude scaling). Dashed line indicates the Vocos' default clamping threshold of 100.

These relations are baked in DSP-based phase reconstruction methods, such as the Phase Gradient Heap Integration (PGHI) [18] as well as in some neural network-based approaches [19–22]. In essence, these methods first predict the phase differences (i.e., phase gradients) from a given log-magnitude spectrogram, and later integrate them across time and frequency to obtain a phase estimate. As mentioned above, Vocos does none of these operations explicitly, but instead offloads them to the ConvNeXt blocks. This warrants an investigation into whether the Vocos architecture can predict phase differences, and if not, which components are responsible for this failure. Hence, later on, we adapt the Vocos architecture to predict phase differences, and train it with a loss that directly compares the predicted and ground-truth phase differences.

3.1. Phase differences

Phase reconstruction methods often use the frequency phase differences (FPD) (also called group delay), time phase differences (TPD) (also called instantaneous frequency), and baseband phase differences (BPD), which are given by

$$\text{FPD}[m, n] = \mathcal{W}(\phi[m, n] - \phi[m-1, n]) \,, \tag{1}$$

$$\text{TPD}[m, n] = \mathcal{W}(\phi[m, n] - \phi[m, n-1]) \,, \tag{2}$$

$$\text{BPD}[m, n] = \mathcal{W}(\text{TPD}[m, n] - 2\pi \frac{mH}{N}) \,, \tag{3}$$

where $\mathcal{W}(\cdot)$ denotes the principal value wrapping operation, and ϕ denotes the phase spectrogram, for frequency bin m and time index n. The STFT hop size and FFT size are denoted by H and N, respectively. It is commonly assumed that BPD are easier to model with CNNs than TPD since cumulative contributions from the linear phase term (due to time shifts between frames) are removed [23].

3.2. Modeling and integrating phase differences

We adapt the Vocos backbone to predict BPD and FPD, by altering the first Conv1D layer to take a 513-channel log-magnitude spectrogram, and the 1026-channel output is split into BPD and FPD. For comparison, we also design a 2D convolutional layer variant with 6 ConvNeXt blocks and with number of channels as (32, 64) instead of (512, 1536). The two-stage phase reconstruction method proposed by Masuyama et al., whose architecture is shown in Fig. 3, serves as a benchmark. To train these models, we use the linear wrapping-aware loss [11, 24], given by

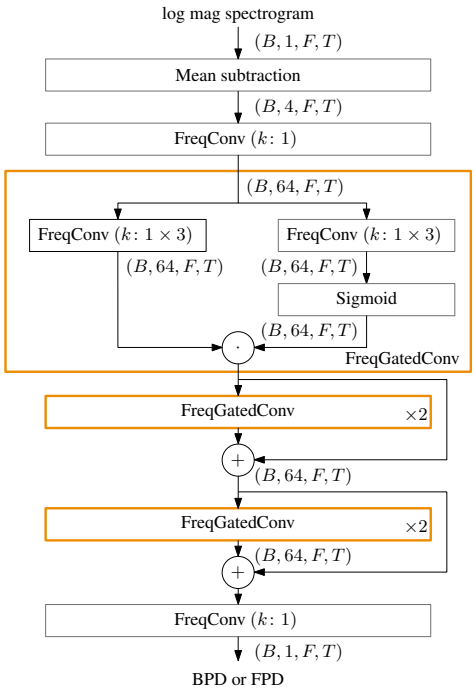

Fig. 3. Benchmark phase difference prediction architecture, using causal 2D convolutions over time and frequency dimensions [21].

$$\mathcal{L}_{\text{wa}}(\hat{x}, x) = \left\| \hat{x} - x - 2\pi \,\text{round}\left(\frac{\hat{x} - x}{2\pi}\right) \right\|_1 . \tag{4}$$

For evaluation, the predicted phase differences need to be converted into a phase spectrogram $\hat{\phi}$. We do so by using the least-squares phase reconstruction method of Masuyama et al. [21].

4. EXPERIMENTAL SETUP

4.1. Datasets

For our trainings, we used LibriTTS [25] subsets (train-clean-100, train-clean-360, train-other-500), similar to BigVGAN and Vocos [1, 4]. This dataset contains approximately 585 h of speech data from 2456 speakers. We employed the random amplitude augmentation as proposed for Vocos, namely, scaling the audio so that the maximum amplitude is within $[-6, -1)$ dBFS. For validation and testing, we used LibriTTS subsets dev-clean (40 speakers, 9 h) and test-clean (39 speakers, 8.5 h), respectively. In addition, we used 20 proprietary German utterances from two speakers for evaluation. While reporting the results, we use mnemonics D1 and D2 for test-clean and the out-of-distribution German datasets, respectively. We resampled all audio files to 22.05 kHz using Torchaudio's resampling function.

4.2. Losses and training strategy

We use the following losses in our vocoding experiments: 1) the multi-resolution mel spectrogram loss $\mathcal{L}_{\text{mel}}$, 2) the hinge adversarial loss $\mathcal{L}_{\text{adv}}$, and 3) the feature matching loss $\mathcal{L}_{\text{fm}}$ of DAC [26]. The discriminators were trained using the hinge critic loss of DAC, symbolized as $\mathcal{L}_{\text{critic}}$. For training phase difference prediction models, we use the wrapping-aware loss (4). The training settings for both lines of experiments are summarized in Table 1.

Table 1. Training strategy for neural vocoding and phase difference prediction experiments, largely based on [26] and [21].

Parameter	Vocoding	Phase Diff.
Batch size (B)	16	32
Utterance length [samples]	16384	16384
Generator Loss	$15\mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{adv}} + 2\mathcal{L}_{\text{fm}}$	$\mathcal{L}_{\text{wa}}$
Discriminator Loss	$\mathcal{L}_{\text{critic}}$	N/A
Optimizer(s)	AdamW	RAdam
Opt. params	lr: 2e$-$4, β: (0.9, 0.999)	lr: 6e$-$4
LR scheduler(s)	cosine annealing	cosine annealing
Training duration [steps]	2M	400k

5. RESULTS

5.1. Comparison of vocoders

Objective evaluation. We report five objective evaluation metrics: the multi-resolution log-mel spectrogram loss $\mathcal{L}_{\text{mel}}$, SCOREQ using the ground-truth audio as reference [27], voiced-unvoiced F1 score, periodicity RMSE [7], and log-spectral convergence (LSC) [21]

$$\text{LSC}(\hat{\phi}, \mathbf{A}) = 20\log_{10}\left(\frac{\|\mathbf{A} - |\text{STFT}(\text{iSTFT}(\mathbf{A}e^{j\hat{\Phi}}))|\|_{\text{Fro}}}{\|\mathbf{A}\|_{\text{Fro}}}\right), \tag{5}$$

where $\mathbf{A}$ and $\hat{\phi}$ denote the ground-truth magnitude spectrogram and the predicted phase spectrogram (obtained by applying STFT to the predicted waveform), respectively, and $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm. The results are reported in Table 2.

Even though retraining on LibriTTS degrades the objective metrics slightly, BigVGANv2 outperforms Vocos. Compared to the official Vocos recipe, using the outlined state-of-the-art vocoder training recipe yields mixed results for Vocos: improving $\mathcal{L}_{\text{mel}}$, deteriorating SCOREQ, and having negligible effects on pitch metrics and LSC.

Subjective evaluation. Furthermore, we conducted a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test to assess the subjective quality, using webMUSHRA [28]. Sixteen participants rated 14 out of the 20 aforementioned German audio samples (dataset D2). For the lower anchor, we used pseudo-inverted mel spectrograms and the phase estimates by the PGHI algorithm [18], as also done in a previous work [29]. The test stimuli are available at our accompanying website[2]

Results, shown in Fig. 4, corroborate that BigVGANv2 is significantly better than Vocos at generating waveforms from band-limited mel spectrograms. Changes to the magnitude clamping and adopting a state-of-the-art training recipe yield a significant improvement for Vocos, but not enough to reach the performance of BigVGANv2. Interestingly, this outcome is only reflected by one of the objective metrics, namely $\mathcal{L}_{\text{mel}}$. The limited bandwidth of the underlying Wav2Vec2.0 representation might be rendering SCOREQ less sensitive to the differences between the two models.

5.2. Learning isolated STFT components with Vocos

Results from the previous section indicate that the gap between Vocos and BigVGANv2 is still present, even after controlling for confounding factors. To investigate the reasons for this gap, we trained two Vocos variants: 1) Vocos-Mag, which predicts only the log magnitude spectrogram $\hat{\mathbf{m}}$, and $\hat{\mathbf{p}}$ is set to the ground-truth phase; and 2) Vocos-Phase predicts the phase spectrogram $\hat{\mathbf{p}}$, and $\hat{\mathbf{m}}$ is set to

[2] https://audiolabs-erlangen.de/resources/NLUI/2026-IWAENC-vocoder

Table 2. Objective evaluation results, comparing Vocos to time-domain vocoders. BigVGANv2 (official) is the official checkpoint, and Vocos (official†) is obtained by retraining Vocos using its official training recipe, but with the input representation explained in Section 2.1.

Model	$\mathcal{L}_{\text{mel}}$ (↓)		SCOREQ (↓)		V/UV F1 (↑)		Periodicity RMSE (↓)		LSC [dB] (↓)	
	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
BigVGANv2 (official)	0.767	0.722	.092	.130	.967	.958	.0665	.0635	−18.94	−22.01
Vocos (official†)	1.389	1.463	.161	.213	.938	.917	.1261	.1380	−12.68	−13.54
BigVGANv2	0.846	0.815	.100	.149	.964	.955	.0713	.0666	−18.06	−20.32
Vocos	1.308	1.309	.188	.258	.940	.919	.1262	.1374	−12.39	−13.43
Vocos-Mag	0.495	0.552	.073	.147	.987	.987	.0273	.0269	−29.18	−31.04
Vocos-Phase	1.560	1.634	.495	.614	.921	.901	.1755	.1777	−10.39	−11.20

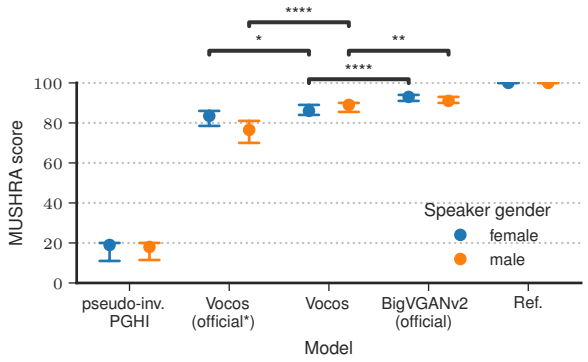

Fig. 4. Subjective listening test results (MUSHRA). The significance markers ('*': $p < 0.05$, '**': $p < 0.01$, and '****': $p < 0.0001$) are based on two-sided Mann-Whitney-Wilcoxon tests with Holm-Bonferroni correction.

the ground-truth log magnitudes. The objective metrics for these experiments are also reported in Table 2, showing that Vocos-Mag outperforms any other experiment in this paper, and Vocos-Phase performing the worst. We draw a number of conclusions from these results:

- Vocos architecture can easily learn to predict a magnitude spectrogram that is compatible with the ground-truth phase. One dimensional convolutions seem to be sufficient for this task, calling into question modifications such as using a pseudo-inverse for recovering the linear scale magnitude spectrograms [12]
- In contrast, predicting a phase spectrogram that is compatible with the ground-truth magnitudes is significantly more challenging, to the point that joint prediction of magnitude and phase spectrograms yields better results than predicting the phase spectrogram alone. While it is known that Vocos struggles with phase modeling, our findings show that the freedom to predict an 'inconsistent' magnitude spectrogram (instead of being forced to predict the ground truth) is instrumental for Vocos' plausible waveform synthesis. This implies that approaches such as regularizing the predicted STFT coefficients to be closer to the ground-truth ones [11, 12, 30] may be counterproductive.

5.3. Phase difference prediction with the Vocos backbone

To isolate the source of Vocos' phase errors, we investigate if the Vocos architecture can predict phase differences from ground-truth log-magnitude spectrograms. This is important because according to the underlying signal model, phase differences are a precursor for phase reconstruction, and estimating them from the log-magnitude spectrogram does not require autoregression. So, if the Vocos architecture is not effective for this task, it would indicate that the issue is not just with autoregression, but also with the inductive biases of the architecture itself.

Table 3. Impact of architectural choices on phase difference prediction performance.

Model	# params.	# FLOPs/s	$\mathcal{L}_{\text{wa}}$ (↓)	LSC [dB] (↓)
Vocos (Conv1D)	15.0M	3.0B	.646	−20.24
Vocos (Conv2D)	37.1K	3.5B	.112	−29.56
Benchmark model [21]	247K	21.8B	.238	−26.12

The results are reported in Table 3. The metrics indicate that the Vocos backbone with Conv1D layers is not effective for predicting phase differences. On the other hand, switching to Conv2D layers substantially improves the performance, even outperforming the benchmark method of Masuyama et al. [21]. This suggests that the inductive biases of the Vocos architecture, in particular the use of 1D convolutions, bottleneck its ability to model the time-frequency structure of speech signals, which is crucial for phase reconstruction.

However, replacing 1D convolutions with 2D ones poses some challenges. The frequency bins become a spatial dimension, narrow receptive fields of the 2D convolutions would not allow the model to learn harmonic structures spanning a wider frequency range. Therefore, future research should focus on inductive biases that allow the architecture to better model the time-frequency structure of speech signals, without requiring ad-hoc techniques such as harmonic priors [8] that rely on pitch information and restrict applicability to input representations with local time-frequency structure.

6. CONCLUSION

In this paper, we investigated the limitations of the Vocos architecture. We showed that using a bandlimited mel spectrogram as input is an informative benchmark for comparing vocoders, accentuating the gap between time-domain and time-frequency domain vocoders. We found that even after controlling for confounding factors, Vocos still lags behind BigVGAN, a state-of-the-art time-domain vocoder. Then, to investigate the reasons for this gap, we trained Vocos variants with oracle knowledge of either the magnitude or phase spectrograms. Our results showed that the inconsistencies in Vocos-predicted magnitude spectrograms are "not a bug, but a feature", as the freedom to predict an inconsistent magnitude spectrogram is instrumental for compensating errors in the phase. Finally, we showed that fixing the Vocos architecture might not require autoregression, but rather better inductive biases for modeling the time-frequency structure of speech signals.

7. REFERENCES

[1] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *Proc. Intl. Conf. on Learning Representations (ICLR)*, 2023.

[2] A. Mustafa, J. Büthe, S. Korse, K. Gupta, G. Fuchs, and N. Pia, "A streamwise GAN vocoder for wideband speech coding at very low bit rate," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[3] M. Kawamura, Y. Shirahata, R. Yamamoto, and K. Tachibana, "Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time Fourier transform," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[4] H. Siuzdak, "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," in *Proc. Intl. Conf. on Learning Representations (ICLR)*, 2024.

[5] T. Okamoto, H. Yamashita, Y. Ohtani, T. Toda, and H. Kawai, "WaveNeXt: ConvNeXt-based fast neural vocoder without ISTFT layer," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.

[6] J. Laroche and M. Dolson, "Phase-vocoder: About this phasiness business," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1997.

[7] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive GAN for conditional waveform synthesis," in *Proc. Intl. Conf. on Learning Representations (ICLR)*, 2022.

[8] R. Yoneyama, A. Miyashita, R. Yamamoto, and T. Toda, "Wavehax: Aliasing-free neural waveform synthesis based on 2d convolution and harmonic prior for reliable complex spectrogram estimation," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 33, 2025.

[9] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Audio, Speech and Sig. Proc.*, vol. 32, no. 2, 1984.

[10] Y. Lv, H. Li, Y. Yan, J. Liu, D. Xie, and L. Xie, "FreeV: Free lunch for vocoders through pseudo inversed mel filter," in *Proc. Interspeech Conf.*, 2024.

[11] H.-P. Du, Y. Ai, R.-C. Zheng, Y.-X. Lu, and Z.-H. Ling, "Is GAN necessary for mel-spectrogram-based neural vocoder?" *IEEE Signal Process. Lett.*, vol. 32, 2025.

[12] A. Li *et al.*, "Learning neural vocoder from range-null space decomposition," in *Proc. International Joint Conf. on Artificial Intelligence*, 2025.

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[14] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," presented at the Proc. Neural Information Processing Systems (NeurIPS), 2020.

[15] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. on Machine Learning Research (TMLR)*, 2023.

[16] Y. Gu, X. Zhang, L. Xue, and Z. Wu, "Multi-scale subband constant-q transform discriminator for high-fidelity vocoder," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[17] F. Auger, É. Chassande-Mottin, and P. Flandrin, "On phase-magnitude relationships in the short-time Fourier transform," *IEEE Signal Processing Letters*, vol. 19, no. 5, 2012.

[18] Z. Průša, P. Balazs, and P. L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 5, 2017.

[19] A. Marafioti, N. Holighaus, N. Perraudin, and P. Majdak, "Adversarial generation of time-frequency features with application in audio synthesis," in *Proc. Intl. Conf. on Machine Learning (ICML)*, 2019.

[20] B. Di Giorgi, M. Levy, and R. Sharp, "Mel spectrogram inversion with stable pitch," in *Proc. Intl. Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[21] Y. Masuyama, K. Yatabe, K. Nagatomo, and Y. Oikawa, "Online phase reconstruction via DNN-based phase differences estimation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 31, 2023.

[22] A. Fernandez, J. Azcarreta, Ç. Bilen, and J. Monge Alvarez, "Efficient neural and numerical methods for high-quality online speech spectrogram inversion via gradient theorem," in *Proc. Interspeech Conf.*, 2025.

[23] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, 2014.

[24] Y. Ai and Z.-H. Ling, "Low-latency neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses for speech generation tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, 2024.

[25] H. Zen *et al.*, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech Conf.*, 2019.

[26] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proc. Neural Information Processing Systems (NeurIPS)*, 2023.

[27] A. Ragano, J. Skoglund, and A. Hines, "SCOREQ: Speech quality assessment with contrastive regression," in *Proc. Neural Information Processing Systems (NeurIPS)*, 2024.

[28] M. Schoeffler *et al.*, "webMUSHRA — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[29] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. ISCA Speech Synthesis Workshop (SSW)*, 2019.

[30] H.-P. Du, Y.-X. Lu, Y. Ai, and Z.-H. Ling, "APNet2: High-quality and high-efficiency neural vocoder with direct prediction of amplitude and phase spectra," in *Proc. National Conf. Man-Machine Speech Communication (NCMMSC)*, 2024.