GOMPSNR: Reflourish the Signal-to-Noise Ratio Metric for Audio Generation Tasks

Lingling Dai[1,2], **Andong Li**[1,2*], **Cheng Chi**[1,2], **Yifan Liang**[1,2], **Xiaodong Li**[1,2], **Chengshi Zheng**[1,2*]

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{dailingling, liandong, chicheng2021, liangyifan, lxd, cszheng}@mail.ioa.ac.cn

Abstract

In the field of audio generation, signal-to-noise ratio (SNR) has long served as an objective metric for evaluating audio quality. Nevertheless, recent studies have shown that SNR and its variants are not always highly correlated with human perception, prompting us to raise the questions: *Why does SNR fail in measuring audio quality?* And *how to improve its reliability as an objective metric?* In this paper, we identify the inadequate measurement of phase distance as a pivotal factor and propose to reformulate SNR with specially designed phase-distance terms, yielding an improved metric named GOMPSNR. We further extend the newly proposed formulation to derive two novel categories of loss function, corresponding to magnitude-guided phase refinement and joint magnitude-phase optimization, respectively. Besides, extensive experiments are conducted for an optimal combination of different loss functions. Experimental results on advanced neural vocoders demonstrate that our proposed GOMPSNR exhibits more reliable error measurement than SNR. Meanwhile, our proposed loss functions yield substantial improvements in model performance, and our well-chosen combination of different loss functions further optimizes the overall model capability.

Code — https://github.com/lingling-dai/GOMPSNR

Introduction

In the development of audio technology, objective metrics have played a pivotal role in evaluating audio quality by providing standardized and quantitative measurements. Over the years, numerous metrics have been proposed, either with a reference (intrusive) or without a reference (non-intrusive). Among these metrics, the signal-to-noise ratio (SNR) or signal-to-distortion ratio (SDR) has been widely applied in audio signal processing and audio generation tasks, including speech enhancement (SE) (Hao et al. 2021), bandwidth extension (BWE) (Hauret et al. 2023; Zhang et al. 2021), and blind speech separation (BSS) (Luo and Mesgarani 2019).

Based on the strict mathematical formulation, SNR provides a straightforward measure of the difference between the estimated signal and the reference signal by computing the ratio of the power of the desired signal to that of the background noise or distortion. During the successive development of SNR, variants such as segmental SNR (segSNR) (Richards 1965), frequency weighted segSNR ($\text{segSNR}_{\text{fw}}$) (Tribolet et al. 1978), and scale-invariant SNR (SI-SNR) (Roux et al. 2019) have been proposed to achieve more accurate and reliable measurements by taking into account the inherent characteristics of the audio signal. However, an increasing number of studies reveal that SNR and its variants may fail to provide consistent results with other perceptual metrics or subjective evaluations (Li et al. 2021; Kumar et al. 2020; Zhang et al. 2025; Erdogan et al. 2023; Wang et al. 2024), and they are gradually being marginalized as a standard metric. Meanwhile, other metrics with similar mathematical formulations, such as Mel-cepstral distortion (MCD) (Kubichek 1993) and Multi-resolution short-time Fourier transform (M-STFT) (Yamamoto, Song, and Kim 2020), continue to remain mainstream. These findings lead us to raise the first question: **(i) Why does SNR fail in measuring the audio quality?**

To answer this question, we carry out detailed mathematical formula derivation and visualize the intermediate results to analyze the inherent characteristics of SNR. The results point to one possible factor, which is the inaccurate measurement of the distance of phase. Specifically, SNR can

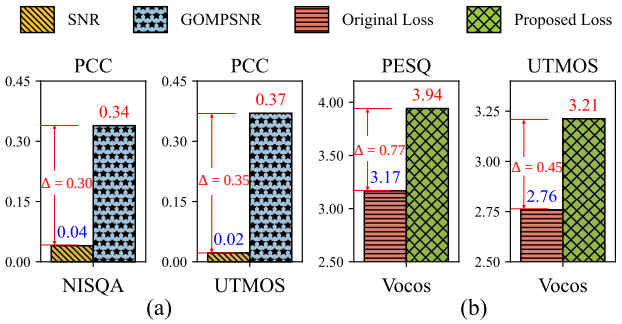

Figure 1: Illustration of the performance improvement brought by our proposed methods. (a) The correlation of SNR and our proposed GOMPSNR with other perceptual metrics. (b) The performance comparison between the original loss and our proposed loss.

*Corresponding authors are Andong Li and Chengshi Zheng.

be seamlessly transformed from the time domain into the time-frequency (T-F) domain, where quantitative assessment of the raw waveform is transformed into the assessment of magnitude and phase in a coupled manner. As the phase spectrum exhibits a more irregular structure compared to the magnitude spectrum, measuring the distance of the phase component remains less dependable than that of the magnitude (Yin et al. 2020). The absence of an effective method for measuring the distance of phase is thus confirmed as the critical factor. Naturally, this leads to the second question: **(ii) How to improve the reliability of SNR as an objective metric?**

As the partial derivatives of the phase, *i.e.*, instantaneous frequency (IF) and group delay (GD), exhibit clearer structures than the instantaneous phase (IP), calculating the distance of IF and GD between the estimated audio and the reference audio rather than IP can provide more reliable measurements (Masuyama et al. 2022). Recent studies have leveraged the intrinsic properties of the phase derivatives for both phase optimization and phase discrepancies quantification. In (Li et al. 2025), the phase derivatives are further developed from the omnidirectional perspective and consolidated into a more convenient and unified form. Based on that, we reformulate the SNR metric by replacing IP with the omnidirectional phase derivatives and correct the unstable terms in the formula, introducing **G**eneralized **OM**nidirectional **P**hase-oriented **S**NR, abbreviated as **GOMPSNR**, as an effective substitute metric of SNR. Meanwhile, we reinvestigate the formulation of several commonly adopted loss functions and propose novel forms of magnitude-guided phase refinement and joint magnitude–phase optimization by leveraging the property of phase derivatives. Experimental results on several state-of-the-art vocoders show that our proposed GOMPSNR exhibits higher relevance with perceptual metrics compared with SNR, and our reformulated loss functions significantly outperform their original loss functions. Additionally, comprehensive experiments are conducted to evaluate the optimal combination of different types of loss functions, and our well-chosen combination further improves the model performance.

Related Works

Audio Generation Tasks

In recent years, the development of deep learning techniques has led to significant advancements in audio generation fields, which encompass a wide range of applications, such as text-to-speech (TTS) (Chen et al. 2024; Du et al. 2024), voice conversion (VC) (Yao et al. 2024), and singing voice synthesis (SVS) (Zhang et al. 2024b). Given an audio clip, text, or other modalities as input, these tasks aim to produce high-fidelity and diverse audio signals with high relevance, where generative models, including flow-based models, Generative Adversarial Networks (GANs), and diffusion probabilistic models are commonly employed (Miao et al. 2020; Popov et al. 2021; Chengyi Wang 2023). Despite the form of generative model varies, when designing the audio synthesis pipeline, a prevalent approach is to decompose the sophisticated audio generation process into several stages, with each focusing on a specific aspect of the task. Typically, a pretrained neural vocoder is employed to convert the mel-spectrogram or latent representations into a raw waveform in the last stage of such a paradigm, where the vocoder plays a crucial role in reconstructing the audio signals (Siuzdak 2024).

Objective Metrics for Audio Quality Assessment

Despite subjective evaluation provides gold standards for evaluating the quality of audio signals, it is often laborious and time-consuming to collect massive subjective mean opinion scores (MOS). Moreover, MOS is sensitive to factors such as listener preferences and equipments. Instead, objective metrics provide quantitative assessments of audio quality and further facilitate fair comparisons between different methods from diverse acoustic aspects. One category of objective metrics measures raw waveform discrepancy or spectral distance in a point-wise manner, including SNR, Log Spectral Distance (LSD), and M-STFT. Moreover, some assess the audio quality by measuring the difference in terms of acoustic features. For instance, Periodicity Root Mean Square Error (RMSE), V/UV F1 score, and pitch RMSE (Morrison et al. 2022) are developed to evaluate the periodicity, voicing, and pitch accuracy. Another category focuses on the perceptual quality, such as PESQ (Rec. 2005), UTMOS (Saeki et al. 2022), and Scoreq (Ragano, Skoglund, and Hines 2024), where the metrics are designed to align with human auditory perception and are often regarded as the economical and practical substitutes for subjective evaluations.

Loss Functions for Enhancing the Audio Quality

Loss functions are essential for guiding the training process and optimizing the model's performance, where the choice of loss function can significantly impact the quality of the generated audio. In the early stages of neural audio generation and signal processing development, the reconstruction losses were primarily composed of simple L1 or L2 losses of the raw waveform or the complex spectrum. However, these loss functions illustrate insufficient ability in enhancing the auditory quality, leading to further exploration of advanced loss functions. In (Kong, Kim, and Bae 2020), a mel-spectrum loss is introduced for improving the perceptual quality inherent to the characteristics of the human auditory system. Similarly, (Chao et al. 2022) designs a critical band importance function for perceptual enhancement of the training target. Furthermore, objective metrics and pretrained large-scale models are also integrated for model optimization (Richter, De Oliveira, and Gerkmann 2025; Babaev et al. 2024). Aside from the focus on perceptual loss design, several other studies concentrate on the point-wise optimization. For instance, (Ai and Ling 2023b) and (Li et al. 2025) propose phase-related loss functions to improve the continuity of phase spectra, which has been shown to further boost the audio quality.

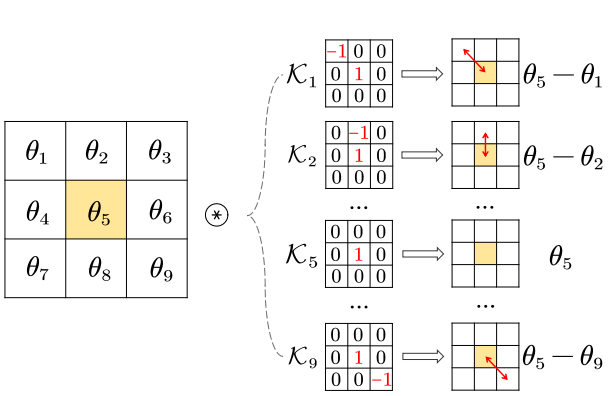

Figure 2: Illustration of obtaining the omnidirectional phase derivatives.

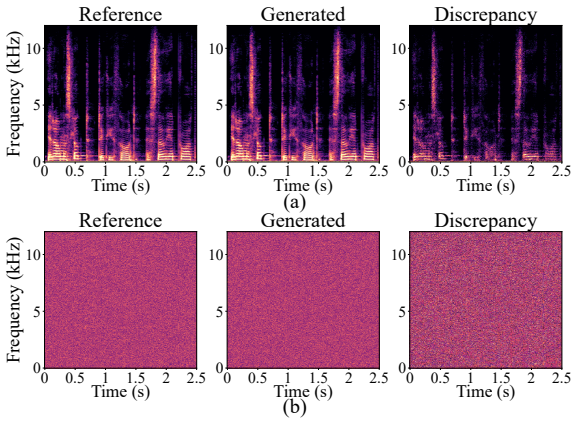

Figure 3: The visualization of the spectrogram discrepancy in terms of (a) magnitude and (b) phase.

Metric	Definition
SNR	$-2\|Y\|\|\widehat{Y}\|\cos\left(\theta-\widehat{\theta}\right)$
OMPSNR	$-\frac{2}{9}\|Y\|\|\widehat{Y}\|\sum_i\cos\left(\nabla_i\theta-\nabla_i\widehat{\theta}\right)$
GOMPSNR	$\frac{2}{9}\|Y\|\|\widehat{Y}\|\sum_i\left(\frac{1}{\pi}f_{AW}\left(\nabla_i\theta-\nabla_i\widehat{\theta}\right)-1\right)$

Table 1: The definition of the correlation component C in SNR and our proposed OMPSNR and GOMPSNR.

Methodology

For audio signal processing and audio generation tasks, SNR and its variants have been previously widely applied in measuring the discrepancy between the estimated signal $\widehat{y} \in \mathbb{R}^n$ and the target $y \in \mathbb{R}^n$, where n denotes the length of the raw waveform. By calculating the ratio of the power of the target signal to that of the residual error over the entire waveform, SNR reflects the energy discrepancy with a quite simple and intuitive mathematical formulation. However, recent researches have repeatedly confirmed that SNR illustrates a low correlation with auditory perception. To isolate the key factor of such misalignment, we shift the formulation from the time domain to the T-F domain, where the raw waveform is decomposed into the representation of magnitude and phase. We use $\{Y, \widehat{Y}\} \in \mathbb{C}^{L\times K}$, $\{\theta, \widehat{\theta}\} \in \mathbb{R}^{L\times K}$ to present the complex spectrum and the phase spectrum of $\{y, \widehat{y}\}$, where the L and K denote the frame and frequency size, respectively. Figure 1 illustrates the performance improvement brought by our proposed methods.

Omnidirectional Phase Derivatives

Compared with magnitude, the direct prediction of phase has remained a formidable challenge (Yin et al. 2020). Specifically, the phase spectrogram exhibits an irregular structure due to the wrapping property, which restricts the phase value to a limited range of $[-\pi, \pi)$. Moreover, phase is also highly sensitive to waveform shifts, which further brings challenges into the prediction process (Zhang et al. 2024a). As computing phase derivatives and applying basic anti-wrapping functions yields more structural representations, researchers now design phase-aware loss functions and dedicated metrics to quantify phase discrepancies based on these unwrapped phase derivatives (Ai and Ling 2023b; Liu et al. 2024; Choi et al. 2018; Lu et al. 2024).

In (Li et al. 2025), a novel omnidirectional phase (OP) representation is proposed to unify and complement the phase derivatives from omni directions. Specifically, as presented in Figure 2, nine 3×3 kernels with fixed parameters $\mathcal{K} = Cat(\mathcal{K}_1, ..., \mathcal{K}_9) \in \mathbb{R}^{9\times3\times3}$ are designed to obtain the derivatives from eight adjacent T-F bins, as well as the IP. Then the omnidirectional phase derivatives $\{\nabla\theta, \nabla\widehat{\theta}\}$ are formulated as:

$$\nabla\theta = \theta \circledast \mathcal{K},\ \nabla\widehat{\theta} = \widehat{\theta} \circledast \mathcal{K}, \tag{1}$$

where $\circledast$ denotes the convolution operation.

Moreover, an OP loss function is further derived to optimize the phase by minimizing the difference between the estimated phase and the target one:

$$\mathcal{L}_{OP} = \frac{1}{9KL}\sum_{i,k,l} f_{AW}\left(\nabla_i\theta - \nabla_i\widehat{\theta}\right), \tag{2}$$

where an anti-wrapping function f_{AW} is applied to unwrap the phase derivatives:

$$f_{AW}(x) = \left|x - 2\pi \cdot round\left(\frac{x}{2\pi}\right)\right|. \tag{3}$$

Generalized Omnidirectional Phase-oriented SNR

In the T-F domain, SNR is formulated as the ratio of two components, where the numerator measures the energy of the target spectrogram, and the denominator aggregates the total squared deviation between the target and the estimated spectrograms:

$$SNR = 10\log_{10}\frac{\sum_{k,l}|Y|^2}{\sum_{k,l}|Y-\widehat{Y}|^2}. \tag{4}$$

Phase	RI	PESQ↑	UTMOS↑	MCD↓	M-STFT↓	V/UV F1↑	Periodicity RMSE↓	Pitch RMSE↓	GOMPSNR↑
-	-	3.749	4.128	2.451	0.990	0.963	0.104	26.104	4.299
P	-	3.711	4.082	2.500	0.994	0.962	0.107	25.323	4.282
OP	-	3.752	4.111	2.449	0.981	0.964	0.102	24.647	4.395
WOP	-	**3.928**	**4.168**	**2.256**	**0.964**	**0.969**	**0.088**	**20.698**	**5.232**
WOP	RI (L1)	3.891	4.180	2.394	1.015	0.966	0.098	19.557	5.718
WOP	ORI (L2)	3.791	4.123	2.614	1.051	0.965	0.097	19.488	5.593
WOP	ORI (L1)	3.998	**4.207**	2.218	0.957	**0.971**	**0.083**	**19.138**	**5.818**
WOP	CORI (L2)	**4.001**	4.164	2.238	**0.935**	**0.971**	**0.083**	19.176	5.674
WOP	CORI (L1)	3.992	4.186	**2.212**	0.944	**0.971**	0.085	19.476	5.622

Table 2: Experimental results of phase-oriented loss functions and co-optimized phase and magnitude loss functions on the LJSpeech Dataset. Inside the parentheses is the adopted type of the point-wise distance. The **optimal** results are marked in **bold**.

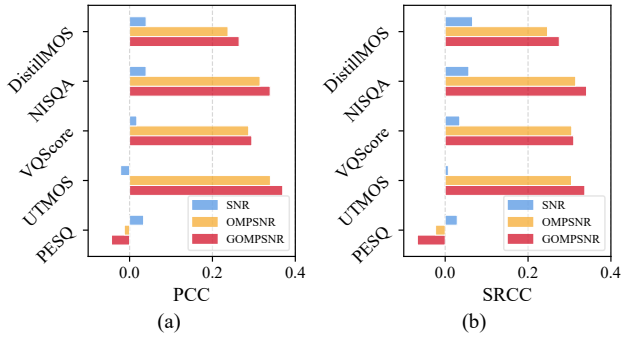

Figure 4: The correlation of SNR and GOMPSNR with several commly adopted perceptual metrics in terms of PCC and SRCC on the LibriTTS dataset.

The denominator can be unfolded into three additive components: the energy of the target spectrogram, the energy of the estimated spectrogram, and a summation of the signed correlation component C:

$$SNR = 10\log_{10}\frac{\sum_{k,l}|Y|^2}{\sum_{k,l}\left(|Y|^2+|\widehat{Y}|^2+C\right)}, \tag{5}$$

where C is presented in Table 1.

In the above equation, SNR is further transformed into a coupled manner of magnitude and phase. To disentangle the influence of each component, we visualize the residuals of the magnitude and phase spectrogram between the vocoder output and its ground-truth reference in Figure 3. As one can observe, the discrepancy of magnitude spectrogram is structurally salient and clearly highlights the regions where the model fails to recover the spectral details. Instead, the visualization of the phase discrepancy yields uninformative results, indicating that conventional metrics, whether explicitly or implicitly calculating the phase distance, may have miscalculated the phase discrepancy. To mitigate the deviation in measuring the distance of phase, we propose to use the omnidirectional phase derivatives to replace the IP and upgrade SNR into **OM**nidirectional **P**hase-oriented **SNR**, dubbed **OMPSNR**, that simultaneously accounts for the magnitude discrepancy and the misalignment of omnidirectional phase derivatives.

Moreover, we find that the sign of the correlation component C changes around the value of $\pm\pi/2$ for $\theta-\widehat{\theta}$, which probably leads to the numerical oscillation in the summation of C. Such a pattern intensifies the influence of phase, making SNR highly sensitive to perturbations in phase errors. Therefore, we further propose the generalized version of OMPSNR (**GOMPSNR**), by modifying the correlation component C into a non-positive item to alleviate the instability caused by the phase error. Additionally, the nonlinear transformation is replaced with a strictly linear mapping function. The final formulation is presented in Table 1.

Coupling Magnitude and Phase in Loss Functions

Apart from metrics that suffer from the assessment deviation brought by the incorrect measurement of phase distance, existing loss functions may also face the inherent difficulties of implicit or direct phase optimization. While the OP loss has been proven effective (Li et al. 2025), we further explore more generalized designs of loss function that cooperate phase derivatives and magnitude to innovate the classical ones. Specifically, we reformulate the loss function from two different aspects: magnitude-guided phase refinement and joint magnitude–phase optimization.

Firstly, following (Zhang et al. 2024a; Dai et al. 2025), we apply a magnitude-based weighting term to the OP loss function, assigning region-specific importance according to the relative value of the magnitude spectrogram. The weighted OP (WOP) loss is defined as:

$$\mathcal{L}_{WOP} = \frac{1}{9KL}\sum_{i,k,l}\frac{|Y|\cdot f_{AW}\left(\nabla_i\theta-\nabla_i\widehat{\theta}\right)}{\max(|Y|)}. \tag{6}$$

Then we reformulate the commonly adopted RI loss, which shares a similar formulation with SNR. Aligned with OMPSNR, one straightforward operation is to substitute the IP with the omnidirectional phase derivatives. Then we propose the OmniRI (ORI) loss, given by:

Model	Mag	Phase	RI	PESQ↑	UTMOS↑	MCD↓	M-STFT↓	V/UV F1↑	Periodicity RMSE↓	Pitch RMSE↓	GOMPSNR↑
Vocos	-	-	-	3.749	4.128	2.451	0.990	0.963	0.104	26.104	4.299
	Lin	-	-	3.775	4.096	2.428	0.972	0.965	0.099	23.975	4.482
	-	WOP	-	3.928	4.168	2.500	0.964	0.969	0.088	20.698	5.232
	-	-	CORI	3.806	4.146	2.424	0.963	0.964	0.101	24.238	4.653
	Lin	WOP	CORI	**4.035**	**4.190**	**2.162**	**0.922**	**0.971**	**0.082**	**19.657**	**5.749**
APNet	Log	P	RI	2.836	2.950	3.716	1.134	0.952	0.132	27.502	3.335
	Lin	P	RI	3.124	3.235	5.924	1.437	0.936	0.153	26.278	-0.957
	Log	WOP	RI	3.503	3.648	**3.190**	**1.032**	**0.959**	**0.112**	**22.310**	**4.318**
	Log	P	CORI	3.293	3.211	3.696	1.083	0.953	0.129	26.656	2.674
	Lin	WOP	CORI	**3.569**	**3.676**	7.144	1.461	0.936	0.149	23.811	-3.381
APNet2	Log	P	RI	3.643	3.764	2.769	1.041	0.960	0.111	22.812	4.961
	Lin	P	RI	3.731	3.844	2.667	1.034	0.964	0.100	22.490	5.163
	Log	WOP	RI	3.803	**4.073**	2.518	1.002	0.967	0.093	**19.543**	**5.629**
	Log	P	CORI	3.702	3.784	2.596	0.972	0.964	0.102	23.055	4.889
	Lin	WOP	CORI	**3.901**	4.056	**2.369**	**0.949**	**0.970**	**0.086**	19.729	5.533
RNDVoc	Log	P	RI	4.033	4.085	1.946	0.877	0.974	0.080	19.241	5.655
	Lin	P	RI	4.029	4.047	1.959	0.910	0.973	0.082	19.236	5.636
	Log	WOP	RI	4.105	**4.231**	1.872	0.884	0.974	0.076	**18.000**	**5.870**
	Log	P	CORI	4.064	4.135	1.896	0.871	0.973	0.079	19.173	5.588
	Lin	WOP	CORI	**4.121**	4.223	**1.844**	**0.875**	**0.975**	**0.074**	18.026	5.822

Table 3: Objective results of different vocoder models with various combinations of loss function on the LJSpeech Dataset.

$$\begin{aligned}\mathcal{L}_{ORI} = &\frac{1}{9}\sum_i h\left(|Y|\cos\left(\nabla_i\theta\right), |\widehat{Y}|\cos\left(\nabla_i\widehat{\theta}\right)\right) \\ &+ \frac{1}{9}\sum_i h\left(|Y|\sin\left(\nabla_i\theta\right), |\widehat{Y}|\sin\left(\nabla_i\widehat{\theta}\right)\right),\end{aligned} \tag{7}$$

where $h(\cdot,\cdot)$ denotes the point-wise distance, *i.e.*, L1 and L2 distance. Note that the point-wise averaging operator is omitted for clarity. Additionally, we further combine the phase derivatives and the magnitude in a coupled manner to obtain the Coupled OmniRI (CORI) loss:

$$\mathcal{L}_{CORI} = \frac{2}{9\pi}\sum_i h\left(|Y|,\ |\widehat{Y}|\right) f_{AW}\left(\nabla_i\theta - \nabla_i\widehat{\theta}\right). \tag{8}$$

Implementation Details

Data Preparation

We conduct experiments on two commonly adopted benchmarks in neural vocoders, namely LJSpeech (Ito and Johnson 2017) and LibriTTS (Zen et al. 2019). The LJSpeech dataset is a relatively small benchmark with 13,100 clean speech clips by a single female speaker, and the sampling rate is 22.05 kHz. Aligned with previous works, we use the division in the open-sourced VITS repository[1] for training, validation, and testing, respectively. The LibriTTS dataset contains approximately 960 hours of speech with a sampling rate of 24 kHz. Following (Lee et al. 2022), we use the entire training subsets (train-clean and train-other) for training and adopt the subsets test-clean and test-other for testing. For LJSpeech, the mel-spectrogram dimension is set to 80, with a hop size of 256 and a window size of 1024, and an effective frequency range from 0 to 8 kHz. For LibriTTS, the mel-spectrogram dimension is set to 100 at a full frequency range of 0 to 12 kHz, with identical hop size and window size settings as LJSpeech. Additionally, we use a hanning window with a window size of 1024 and a hop size of 256 to transform the raw waveform into the complex spectrum when calculating GOMPSNR.

Training Settings

In this paper, several state-of-the-art neural vocoders, including Vocos (Siuzdak 2024), APNet (Ai and Ling 2023a), APNet2 (Du et al. 2023), and RNDVoc (Li et al. 2025), are implemented in the experiments. For training, we follow the official training pipeline of APNet2[2] for all vocoders, except that the learning rate is set to 5e-4. Specifically, the multi-period discriminator (MPD) (Kong, Kim, and Bae 2020) and multi-resolution spectrogram discriminator (MRSD) (Jang et al. 2021) are adopted for adversarial training, with the hinge GAN losses used as the adversarial training loss. Additionally, the feature matching loss and the mel-spectrogram loss are also included as the basic loss functions. We retain all other loss functions adopted in the original paper for each vocoder and each is trained for 2 million steps. [3] Additionally, Neural Audio Codecs (NACs) including WavTokenizer (Ji et al. 2025) and Vocos (Siuzdak 2024) are implemented for auxiliary evaluations. We adopt the of-

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

[2] https://github.com/redmist328/APNet2

[3] Note that the results may be different from the original papers due to different training settings.

Model	Mag	Phase	RI	PESQ↑	UTMOS↑	MCD↓	M-STFT↓	V/UV F1↑	Periodicity RMSE↓	Pitch RMSE↓	GOMPSNR↑
Vocos	-	-	-	3.167	2.758	3.928	1.074	0.920	0.166	43.456	3.909
	Lin	-	-	3.298	2.910	3.641	0.995	0.940	0.136	33.961	4.314
	-	WOP	-	3.886	3.120	3.028	0.927	0.957	0.103	27.835	5.306
	-	-	CORI	3.227	2.830	3.752	1.010	0.929	0.148	38.231	4.122
	Lin	WOP	CORI	**3.942**	**3.212**	**2.866**	**0.887**	**0.959**	**0.101**	**22.585**	**5.777**
APNet	Log	P	RI	2.676	2.034	4.453	1.307	0.926	0.159	42.059	3.532
	Lin	P	RI	3.030	2.376	4.018	1.241	0.932	0.152	43.897	4.262
	Log	WOP	RI	3.288	2.536	3.837	1.144	0.946	0.128	**22.879**	4.792
	Log	P	CORI	3.083	2.292	3.819	1.074	0.938	0.142	38.061	4.048
	Lin	WOP	CORI	**3.654**	**2.725**	**3.344**	**0.915**	**0.957**	**0.105**	27.180	**5.297**
APNet2	Log	P	RI	1.685	1.327	6.305	1.998	0.707	0.360	220.525	1.615
	Lin	P	RI	1.819	1.356	5.728	1.777	0.772	0.320	230.402	1.969
	Log	WOP	RI	2.932	2.454	4.110	1.162	0.934	0.146	33.382	4.343
	Log	P	CORI	3.347	2.605	3.607	0.978	0.946	0.125	27.798	4.532
	Lin	WOP	CORI	**3.789**	**2.974**	**3.187**	**0.901**	**0.957**	**0.107**	**21.601**	**5.269**
RNDVoc	Log	P	RI	4.071	3.106	2.347	0.779	0.964	0.088	36.525	6.056
	Lin	P	RI	4.103	3.175	2.253	0.775	0.966	0.085	26.900	6.203
	Log	WOP	RI	**4.162**	3.285	**2.203**	**0.764**	0.968	0.080	26.540	**6.646**
	Log	P	CORI	4.102	3.200	2.293	0.774	0.967	0.085	**25.916**	6.051
	Lin	WOP	CORI	4.159	**3.291**	2.237	0.767	**0.969**	0.079	28.257	6.559

Table 4: Objective results of different vocoder models with various combinations of loss function on the LibriTTS Dataset.

ficial implementations of Vocos[4] and WavTokenizer[5]. Following (Siuzdak 2024), we only optimize the decoder with frozen encoder and codebooks. Models are trained on the LibriTTS dataset for 2 million steps.

Results and Discussion

Validation on GOMPSNR

In this section, we validate the performance of OMPSNR and GOMPSNR as an intrusive metric in comparison with SNR. Several perceptual objective metrics, including PESQ (Rec. 2005), UTMOS (Saeki et al. 2022), VQScore (Fu et al. 2024), NISQA (Mittag et al. 2021), and DistillMOS (Mittag et al. 2021), are employed as perceptual references for calculating the Pearson Correlation Coefficient (PCC) and the Spearman's Rank Correlation Coefficient (SRCC), with higher absolute values indicating stronger relevance. Figure 4 presents the results of the officially pretrained Vocos on LibriTTS. As one can observe, SNR exhibits weak correlation with all perceptual metrics, with PCC and SRCC scores not exceeding 0.1, suggesting its inefficacy as an objective metric. In contrast, by simply substituting the IP counterpart with omnidirectional phase derivatives in SNR, the correlation with these perceptual metrics improves significantly, further confirming that the inaccurate estimation of phase distance is a primary reason for the diminished efficacy of SNR. Moreover, the rectification of the correlation component in GOMPSNR further improves the effectiveness, where GOMPSNR outperforms OMPSNR and exhibits a comparatively strong correlation with most of the perceptual metrics in terms of PCC and SRCC, indicating its superior performance as an intrusive metric for audio quality assessment. With a straightforward and interpretable mathematical formulation, GOMPSNR is convenient to be applied in any audio generation or signal processing tasks that require frame-wise alignment and is expected to become a new standard metric for audio quality assessment.

Phase-oriented Loss Functions

As the under-modeling of phase distance leads to insufficient modeling of SNR as an objective metric, the incorrect phase optimization in training process may also hinder the overall model performance. In this section, we evaluate the impact of phase-oriented loss functions on audio generation tasks by conducting experiments on Vocos. The objective results are presented in the upper bound of Table 2. Note that we use P to denote the vanilla phase loss proposed in (Ai and Ling 2023b). As illustrated in Table 2, trained with hinge GAN loss, feature matching loss, and mel-spectrogram loss, Vocos is able to achieve satisfying performance on the LJSpeech dataset, and the two other phase-oriented loss functions including the vanilla phase loss and the OP loss provide no further improvement. However, the proposed WOP loss, which is a simple magnitude-weighted version of the OP loss, brings significant improvements to Vocos across all objective metrics. Such finding highlights the benefits of integrating magnitude as prior information and further underlines the need to investigate how phase and magnitude should be coupled during the phase optimization process.

Joint Magnitude and Phase Optimization

In addition to incorporating magnitude as an auxiliary guidance in phase-oriented loss functions, another common prac-

[4]https://github.com/gemelo-ai/vocos

[5]https://github.com/jishengpeng/WavTokenizer

tice in loss function design is to explore the joint optimization of magnitude and phase. To validate whether the co-optimized magnitude and phase can lead to better performance, we conduct experiments on Vocos with different RI losses. The results are given in the lower bound of Table 2, and we use RI to denote the vanilla RI loss. It is noticeable that the vanilla RI loss even slightly degrades the model performance on most of the objective metrics. In contrast, our proposed OmniRI loss and the Coupled OmniRI loss in L1 distance further enhance the model performance, even assisted with the WOP loss that provides direct supervision on phase. These results indicate that disentangling the explicit phase component from the complex spectrum and reforming it benefits both magnitude and phase retrieval, despite the magnitude counterpart remaining unchanged. Furthermore, co-optimizing the phase and magnitude with properly designed forms is demonstrated to further enhance the overall performance and the Coupled OmniRI loss exhibits robust performance on point-wise distance choice.

Comprehensive Evaluation on Loss Functions

In this section, we conduct comprehensive experiments on advanced vocoders to validate an optimal combination of different loss functions, covering magnitude-oriented, phase-oriented, and the co-optimized form of both magnitude and phase, by replacing the corresponding losses in each vocoder with our well-chosen alternatives. Table 3 and Table 4 present the results of different vocoders on the LJSpeech dataset and LibriTTS dataset, respectively. Note that we use the Coupled OmniRI loss in L1 distance by default. Remarkably, each alternative loss function outperforms the original one used in the paper, demonstrating that the original loss function did not fully exploit the model's potential. One may notice that the linear form of the magnitude loss leads to the degradation on the LJSpeech dataset in terms of MCD, M-STFT, and GOMPSNR, which can likely be attributed to the overfitting of speech energy on such a limited dataset. However, it is more effective in improving the perceptual quality compared with the logarithmic one. Furthermore, the combination of our selected loss functions consistently outperforms the original loss function settings across all the vocoders, including RNDVoc, which already achieves outstanding results without elaborately designed losses. The experimental results further reveal that poorly designed loss functions tend to yield sub-optimal performance, whereas our proposed loss functions exploit the model's full potential and significantly enhance the audio quality. Meanwhile, our proposed GOMPSNR also exhibits a strong correlation with other metrics, indicating its effectiveness for providing a quantitative assessment between the estimated signal and the reference.

Evaluation on Neural Audio Codecs

To verify the applicability of our well-chosen combination of loss functions on other audio generation tasks, we further conduct extensive experiments on NACs, which compress the raw waveform into discrete acoustic codec representations and then reconstruct the audio from the latent representations. As presented in Figure 5, after employing our well-chosen losses, both WavTokenizer and Vocos demonstrate superior performance across different bandwidths compared with their original configurations. Notably, the improvement is more pronounced at lower bandwidths, indicating that this set of loss functions is especially beneficial under higher compression levels. The appealing results further confirm the effectiveness of our proposed loss functions in enhancing the audio quality in audio generation tasks, and also demonstrate the applicability of our proposed GOMPSNR as an objective metric for evaluating the audio quality.

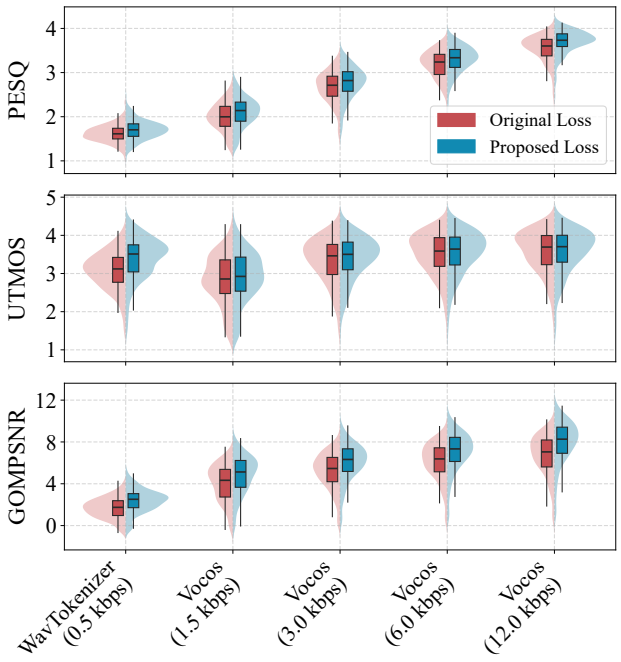

Figure 5: The objective results of different codec models on the LibriTTS Dataset.

Conclusion

This paper proposes GOMPSNR, an effective alternative to SNR as an objective metric. By reformulating SNR with modified phase-distance terms, GOMPSNR exhibits improved measurement rationality and a stronger correlation with auditory perception. Furthermore, we extend GOMPSNR to develop novel loss functions, including a phase-oriented type and a co-optimized form of magnitude and phase, and explore a feasible combination of different loss functions. Quantitative experiments validate the effectiveness of GOMPSNR in audio generation tasks, highlighting its potential as a reliable objective metric for evaluating audio quality. Additionally, extensive experimental results demonstrate that our proposed loss functions yield significant performance improvement on advanced neural vocoders, and our well-chosen combination of loss functions manages to further enhance the overall audio quality.

Acknowledgments

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

References

Ai, Y.; and Ling, Z.-H. 2023a. APNet: An all-frame-level neural vocoder incorporating direct prediction of amplitude and phase spectra. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2145–2157.

Ai, Y.; and Ling, Z.-H. 2023b. Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses. In *Proc. ICASSP*, 1–5.

Babaev, N.; Tamogashev, K.; Saginbaev, A.; Shchekotov, I.; Bae, H.; Sung, H.; Lee, W.; Cho, H.-Y.; and Andreev, P. 2024. FINALLY: fast and universal speech enhancement with studio-like quality. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 934–965. Curran Associates, Inc.

Chao, R.; Yu, C.; wei Fu, S.; Lu, X.; and Tsao, Y. 2022. Perceptual Contrast Stretching on Target Feature for Speech Enhancement. In *Proc. Interspeech*, 5448–5452.

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. In *arXiv:2410.06885*.

Chengyi Wang, Y. W. Z. Z. L. Z. S. L. Z. C. Y. L. H. W. J. L. L. H. S. Z. F. W., Sanyuan Chen. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. In *arXiv:2301.02111*.

Choi, H.-S.; Kim, J.-H.; Huh, J.; Kim, A.; Ha, J.-W.; and Lee, K. 2018. Phase-aware speech enhancement with deep complex u-net. In *ICLR*.

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

Du, H.-P.; Lu, Y.-X.; Ai, Y.; and Ling, Z.-H. 2023. APNet2: High-quality and High-efficiency Neural Vocoder with Direct Prediction of Amplitude and Phase Spectra. In *Proc. NCMSC*, 66–80.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; Gao, Z.; and Yan, Z. 2024. CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. In *arXiv:2407.05407*.

Erdogan, H.; Wisdom, S.; Chang, X.; Borsos, Z.; Tagliasacchi, M.; Zeghidour, N.; and Hershey, J. R. 2023. TokenSplit: Using Discrete Speech Representations for Direct, Refined, and Transcript-Conditioned Speech Separation and Recognition. In *Proc. Interspeech*, 3462–3466.

Fu, S.-W.; Hung, K.-H.; Tsao, Y.; and Wang, Y.-C. F. 2024. Self-Supervised Speech Quality Estimation and Enhancement Using Only Clean Speech. In *ICLR*.

Hao, X.; Su, X.; Horaud, R.; and Li, X. 2021. Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement. In *Proc. ICASSP*, 6633–6637.

Hauret, J.; Joubaud, T.; Zimpfer, V.; and Bavu, 2023. EBEN: Extreme Bandwidth Extension Network Applied To Speech Signals Captured With Noise-Resilient Body-Conduction Microphones. In *Proc, ICASSP*, 1–5.

Ito, K.; and Johnson, L. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/. Accessed: 2025-12-10.

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

Ji, S.; Jiang, Z.; Wang, W.; Chen, Y.; Fang, M.; Zuo, J.; Yang, Q.; Cheng, X.; Wang, Z.; Li, R.; et al. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *ICLR*.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33: 17022–17033.

Kubichek, R. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, 125–128.

Kumar, R.; Kumar, K.; Anand, V.; Bengio, Y.; and Courville, A. 2020. NU-GAN: High resolution neural upsampling with GAN. In *arXiv:2010.11362*.

Lee, S.-g.; Ping, W.; Ginsburg, B.; Catanzaro, B.; and Yoon, S. 2022. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR*.

Li, A.; Lei, T.; Sun, Z.; Chen, R.; Yin, E.; Li, X.; and Zheng, C. 2025. Learning neural vocoder from range-null space decomposition. In *Proc. IJCAI*.

Li, Y.; Tagliasacchi, M.; Rybakov, O.; Ungureanu, V.; and Roblek, D. 2021. Real-Time Speech Frequency Bandwidth Extension. In *Proc. ICASSP*, 691–695.

Liu, F.; Ai, Y.; Du, H.-P.; Lu, Y.-X.; Zheng, R.-C.; and Ling, Z.-H. 2024. Stage-Wise and Prior-Aware Neural Speech Phase Prediction. In *Proc. SLT*, 638–644.

Lu, Y.-X.; Ai, Y.; Du, H.-P.; and Ling, Z.-H. 2024. Towards High-Quality and Efficient Speech Bandwidth Extension with Parallel Amplitude and Phase Prediction. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Luo, Y.; and Mesgarani, N. 2019. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8): 1256–1266.

Masuyama, Y.; Yatabe, K.; Nagatomo, K.; and Oikawa, Y. 2022. Online phase reconstruction via DNN-based phase differences estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 163–176.

Miao, C.; Liang, S.; Chen, M.; Ma, J.; Wang, S.; and Xiao, J. 2020. Flow-TTS: A non-autoregressive network for text to speech based on flow. In *Proc. ICASSP*, 7209–7213.

Mittag, G.; Naderi, B.; Chehadi, A.; and Möller, S. 2021. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Proc. Interspeech*, 2127–2131.

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

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *Proc. ICML*, 8599–8608.

Ragano, A.; Skoglund, J.; and Hines, A. 2024. SCOREQ: Speech Quality Assessment with Contrastive Regression. In *Proc. NeurIPS*, volume 37, 105702–105729.

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

Richards, D. 1965. Speech-transmission performance of pcm systems. *Electronics Letters*, 1(2): 40–41.

Richter, J.; De Oliveira, D.; and Gerkmann, T. 2025. Investigating Training Objectives for Generative Speech Enhancement. In *Proc, ICASSP*, 1–5.

Roux, J. L.; Wisdom, S.; Erdogan, H.; and Hershey, J. R. 2019. SDR – Half-baked or Well Done? In *Proc. ICASSP*, 626–630.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Proc. Interspeech*, 4521–4525.

Siuzdak, H. 2024. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *Proc. 12th Int. Conf. Learn. Representations*, 1–15.

Tribolet, J.; Noll, P.; McDermott, B.; and Crochiere, R. 1978. A study of complexity and quality of speech waveform coders. In *Proc. ICASSP*, volume 3, 586–590.

Wang, Z.; Zhu, X.; Zhang, Z.; Lv, Y.; Jiang, N.; Zhao, G.; and Xie, L. 2024. SELM: Speech Enhancement using Discrete Tokens and Language Models. In *Proc. ICASSP*, 11561–11565.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In *Proc. ICASSP*, 6199–6203.

Yao, J.; Yang, Y.; Lei, Y.; Ning, Z.; Hu, Y.; Pan, Y.; Yin, J.; Zhou, H.; Lu, H.; and Xie, L. 2024. Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts. In *Proc. ICASSP*, 10571–10575.

Yin, D.; Luo, C.; Xiong, Z.; and Zeng, W. 2020. PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network. In *Proc. AAAI Conf. Artif. Intell.*, volume 34, 9458–9465.

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

Zhang, K.; Ren, Y.; Xu, C.; and Zhao, Z. 2021. WSRGlow: A Glow-Based Waveform Generative Model for Audio Super-Resolution. In *Proc. Interspeech*, 1649–1653.

Zhang, S.; Qiu, Z.; Takeuchi, D.; Harada, N.; and Makino, S. 2024a. Unrestricted Global Phase Bias-Aware Single-Channel Speech Enhancement with Conformer-Based Metric Gan. In *Proc. ICASSP*, 1026–1030.

Zhang, W.; Saijo, K.; Cornell, S.; Scheibler, R.; Li, C.; Ni, Z.; Kumar, A.; Sach, M.; Wang, W.; Fu, Y.; Watanabe, S.; Fingscheidt, T.; and Qian, Y. 2025. Lessons Learned from the URGENT 2024 Speech Enhancement Challenge. In *arXiv:2506.01611*.

Zhang, Y.; Huang, R.; Li, R.; He, J.; Xia, Y.; Chen, F.; Duan, X.; Huai, B.; and Zhao, Z. 2024b. Stylesinger: Style transfer for out-of-domain singing voice synthesis. In *Proc. AAAI*, volume 38, 19597–19605.