UNRESTRICTED GLOBAL PHASE BIAS-AWARE SINGLE-CHANNEL SPEECH ENHANCEMENT WITH CONFORMER-BASED METRIC GAN

Shiqi Zhang[*], *Zheng Qiu*[*], *Daiki Takeuchi*[†], *Noboru Harada*[†], *Shoji Makino*[*]

[*] Waseda University, Japan and [†] NTT Coporation, Japan

ABSTRACT

With the rapid development of neural networks in recent years, the ability of various networks to enhance the magnitude spectrum of noisy speech in the single-channel speech enhancement domain has become exceptionally outstanding. However, enhancing the phase spectrum using neural networks is often ineffective, which remains a challenging problem. In this paper, we found that the human ear cannot sensitively perceive the difference between a precise phase spectrum and a biased phase (BP) spectrum. Therefore, we propose an optimization method of phase reconstruction, allowing freedom on the global-phase bias instead of reconstructing the precise phase spectrum. We applied it to a Conformer-based Metric Generative Adversarial Networks (CMGAN) baseline model, which relaxes the existing constraints of precise phase and gives the neural network a broader learning space. Results show that this method achieves a new state-of-the-art performance without incurring additional computational overhead.

Index Terms— Single-channel, speech enhancement, biased phase spectrum, phase derivative

1. INTRODUCTION

Speech enhancement [1] is a technique that involves processing noisy speech signals to output relatively clean speech signals. It has a wide range of applications, spanning communication devices, intelligent interactive devices, hearing aids, and more. A good speech enhancement front-end technology can not only provide a better communication experience but also lead to improved accuracy in automatic speech recognition (ASR) [2, 3].

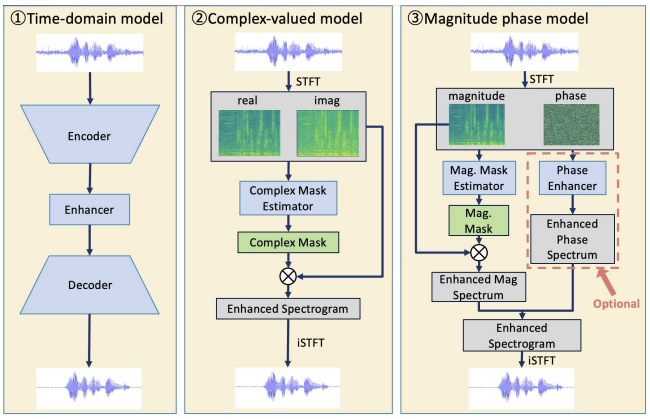

Fig. 1. Three different types of speech enhancement models

With the rapid development of neural networks in recent years, a large number of excellent single-channel speech enhancement neural network models have emerged in the field of acoustic signal processing, including the classic magnitude-phase models [4–10] and the recently popular complex-valued models [11–16] and time-domain models [17–20] as shown in Fig. 1. Among these, the magnitude-phase models and complex-valued models both process noisy speech signals in the time-frequency domain. In this processing flow, the noisy speech is converted to the time-frequency domain using the short-time Fourier transform (STFT), the spectrogram is enhanced, and the enhanced spectrogram is converted back to the time-domain signal and output using the inverse STFT (iSTFT). The difference between these two models is that the former estimates a real-valued mask matrix applied to the magnitude and a cleaner phase spectrum (optional), while the latter estimates a complex-valued mask [11, 12] matrix applied to the entire complex spectrum. The time-domain model usually consists of an encoder, a decoder, and a network in the middle for enhancement. After inputting the signal, it is directly encoded in the time domain, and then decoded after enhancement in the encoding domain to obtain a cleaner audio signal. Although the implementation details of the above three methods are different, they all aim at reconstructing precise speech signals that have accurate magnitude and phase.

However, due to the complexity of the phase, accurately reconstructing the phase spectrum [8, 21, 22] has posed significant challenges for existing neural networks [23–27]. Through our own experiments, we discovered that it is difficult for the human ear to distinguish between an accurate phase and a globally biased phase spectrum [28, 29]. Utilizing this characteristic, we propose a new optimization method based on unrestricted globally biased phase reconstruction that improves the performance of speech enhancement without increasing the number of model parameters or the computational cost, and achieves new state-of-the-art (SoTA) results.

2. BIASED PHASE PERCEPTUAL EXPERIMENT

The experiment conducted in this section reveals that the human auditory system is not sensitive to precise phase. A perfect reconstruction of the phase spectrum is not necessary, and global phase bias can therefore be ignored. This provides one more dimension to the solution space that does not affect the perceptual quality, thereby facilitating easier convergence of neural networks.

First, we processed 100 audio samples using STFT with a Hamming window h and a shift value θ as shown in (1). The modification only involves adding a random global angle value, $\theta \in [-\pi, \pi)$, to the phase part of the original spectrum without affecting the magnitude spectrum at all. Therefore, (1) merely produces a time-frequency spectrogram with a global phase bias. Subsequently, the two types of spectrograms obtained are ultilized for time-domain signal reconstruction using iSTFT. (Note that (1) can also be written in the form of (2) if we want to show that this processing is not a simple time shift for the waveform. For different frequency components ω, the time shift amount $\frac{\theta}{\omega}$ is also different. The reconstructed signal is inconsistent with the original signal in terms of the time-domain waveform, as shown in Fig. 2.)

This work was supported by JSPS KAKENHI Grant Number 23H03423.

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j(\omega\tau+\theta)}d\tau \quad (1)$$

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j\omega(\tau+\frac{\theta}{\omega})}d\tau \quad (2)$$

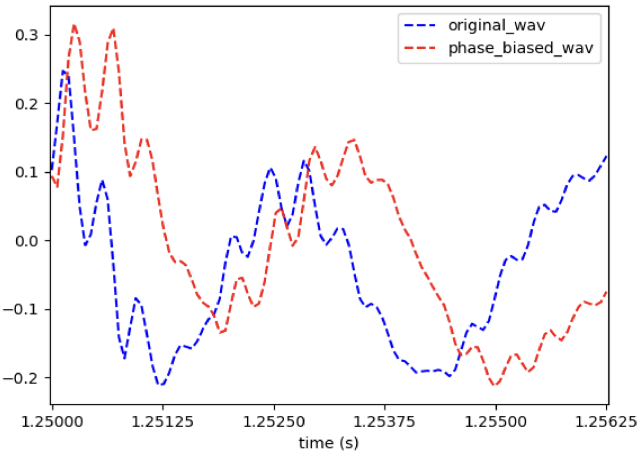

Fig. 2. Comparison of original phase reconstructed signal and biased phase reconstructed signal (partial)

Next, we conducted a hearing test with four listeners. They first listened to the original audio and then to the signals reconstructed using iSTFT after transformation with STFT and (1), and judged and recorded which of the two latter signals was closer to the original audio. If the judgment was correct, it was counted as 1 for "True"; otherwise, it was counted as 1 for "False". The results, as listed in Table 1, indicate that the ratio of True to False counts is close to 50%. This indicates a similar phenomenon to [28, 29], namely, that the human ear has little ability to accurately distinguish between the phase spectrum and the phase spectrum with a global bias.

Table 1. Hearing results of phase unbiased and biased signals

Result	P_1	P_2	P_3	P_4	Total
True	46	58	55	50	209 (52.25%)
False	54	42	45	50	191 (47.75%)

We also calculated the average Perceptual Evaluation of Speech Quality (PESQ) [30], Segment Signal-to-Noise Ratio (SegSNR) [31], and Scale-invariant Source-to-Noise Ratio (SiSNR) [19] for the signals reconstructed using STFT-iSTFT and those reconstructed using (1)-iSTFT. (SegSNR and SiSNR are limited to the range of [-10.000, 35.000] dB.) Table 2 shows results consistent with the perception experiment in Table 1 and Figure 2. PESQ is mostly unaffected by the biased phase, whereas the time-domain metrics SegSNR and SiSNR are significantly impacted. These findings also indicate that time-domain metrics in the field of signal processing do not directly represent the perception effects of audio.

Table 2. Metrics on phase unbiased and biased signals

Signal	PESQ	SegSNR	SiSNR
Unbiased	4.644	35.000	35.000
Biased	4.636	0.561	1.627

3. METHODOLOGY

In this paper, we selected CMGAN [32] as the baseline and improved the optimization method using the phenomenon discussed in Sec. 2.

3.1. Baseline: CMGAN

CMGAN is a generalized magnitude-phase model that generates a real-valued mask matrix and a complex-valued residual matrix through a stacked network structure. The mask is multiplied by the original magnitude spectrum to obtain a relatively clean magnitude spectrum, which is then reconstructed into a spectrogram together with the original noisy phase spectrum. The complex-valued residual matrix is then added to this spectrogram, achieving the effect of simultaneously estimating the magnitude and phase. The original optimization function of CMGAN ($\mathcal{L}_{ori}$) is a composite loss function (3) consisting of the following weighted ($\lambda_{1\sim4}$) components: magnitude Mean Square Error (MSE) Loss ($\mathcal{L}_{mag}$), spectrum real and imaginary part MSE Loss ($\mathcal{L}_{ri}$), time-domain L1 Loss ($\mathcal{L}_{time}$), and Adversarial Loss ($\mathcal{L}_{adv}$). These are jointly used for optimizing the generator network.

$$\mathcal{L}_{ori} = \lambda_1\mathcal{L}_{mag} + \lambda_2\mathcal{L}_{ri} + \lambda_3\mathcal{L}_{time} + \lambda_4\mathcal{L}_{adv} \quad (3)$$

As for the discriminator part, it can be viewed as a normalized PESQ predictor. It undergoes joint training consisting of two loss components. The first is to narrow the gap between the real normalized PESQ and the discriminator's prediction for the output of the generator and the clean signal. The second is to make the predictions for the same clean signal closer to 1, ensuring accuracy in assessing signal quality.

3.2. Proposed methods

As shown in Sec. 2, a precise phase is unnecessary for human ear perception. Therefore, we propose the following methods, hoping to ignore any global bias on the phase spectrum, relax the existing constraints of the precise phase, and give the neural network a broader learning space, thereby more easily achieving better enhancement effects.

3.2.1. Unrestricted phase bias (UPB)-aware loss function

To estimate a globally biased phase spectrum, we naturally thought of using the MSE loss of its partial derivatives for model optimization, due to the property that the derivative of a constant is zero. At the same time, to avoid the impact of meaningless accurately estimating the phase, we removed the $\mathcal{L}_{ri}$ and $\mathcal{L}_{time}$ parts from the original CMGAN optimization function.

Specifically, we first use (4) and (5) to compute derivatives of the phase spectrum's time and frequency, referred to as Time Phase Derivative (TPD [33]) $\delta_\tau \in \mathbb{R}^{(T-1)\times F}$ and Frequency Phase Derivative (FPD [33]) $\delta_\omega \in \mathbb{R}^{T\times(F-1)}$, respectively.

$$\boldsymbol{\delta}_\tau(t,f) = \boldsymbol{\phi}(t+1,f) - \boldsymbol{\phi}(t,f) \quad (4)$$

$$\boldsymbol{\delta}_\omega(t,f) = \boldsymbol{\phi}(t,f+1) - \boldsymbol{\phi}(t,f) \quad (5)$$

Here, t and f represent the time and frequency coordinates of the elements in the spectrum matrix, $\boldsymbol{\phi} \in \mathbb{R}^{T\times F}$ represents the phase angle, T is the total number of time frames, and F is the total number of frequency bins. However, this method has a wrapping problem during the calculation process. Specifically, when the absolute sum of two phase angles exceeds π, for example, $\theta_1 = \frac{3}{4}\pi$, $\theta_2 = -\frac{3}{4}\pi$, their algebraic difference is $\theta_2 - \theta_1 = -\frac{3}{2}\pi$. However, $\theta_1 + \frac{1}{2}\pi = \frac{5}{4}\pi = -\frac{3}{4}\pi + 2\pi$ and $\theta_2 = -\frac{3}{4}\pi$ are equivalent, i.e., increasing θ_1 by $\frac{1}{2}\pi$ results in θ_2.

Therefore, we use a common method (6) similar to the one in [34] to calculate the angle difference and solve the wrapping issue.

$$\mathcal{S}(\theta_1,\theta_2) = arctan\big(tan(\theta_1 - \theta_2)\big) \quad (6)$$

As a result, the wrapped TPD $\delta_{w\tau} \in \mathbb{R}^{(T-1)\times F}$ and FPD $\delta_{w\omega} \in \mathbb{R}^{T\times(F-1)}$ can be represented by (7) and (8), respectively.

$$\boldsymbol{\delta}_{w\tau}(t,f) = \mathcal{S}\big(\boldsymbol{\phi}(t+1,f), \boldsymbol{\phi}(t,f)\big) \tag{7}$$

$$\boldsymbol{\delta}_{w\omega}(t,f) = \mathcal{S}\big(\boldsymbol{\phi}(t,f+1), \boldsymbol{\phi}(t,f)\big) \tag{8}$$

Hence, the loss function $\mathcal{L}_{upb}$ for estimating unrestricted biased phase can be calculated using (9).

$$\begin{aligned}\mathcal{L}_{upb} = &\frac{1}{2}\mathbb{E}_{\boldsymbol{\delta}_{w\tau}(t,f),\hat{\boldsymbol{\delta}}_{w\tau}(t,f)}[\|\mathcal{S}\big(\boldsymbol{\delta}_{w\tau}(t,f), \hat{\boldsymbol{\delta}}_{w\tau}(t,f)\big)\|^2] \\ &+ \frac{1}{2}\mathbb{E}_{\boldsymbol{\delta}_{w\omega}(t,f),\hat{\boldsymbol{\delta}}_{w\omega}(t,f)}[\|\mathcal{S}\big(\boldsymbol{\delta}_{w\omega}(t,f), \hat{\boldsymbol{\delta}}_{w\omega}(t,f)\big)\|^2]\end{aligned} \tag{9}$$

Finally, the model loss function of CMGAN is modified here to (10).

$$\mathcal{L}_1 = \lambda_1\mathcal{L}_{mag} + \lambda_4\mathcal{L}_{adv} + \lambda_5\mathcal{L}_{upb} \tag{10}$$

The loss factors $\lambda_2 L_{ri}$ and $\lambda_3 L_{time}$ have been removed from (3). This method directly unrestricts the phase bias instead of reconstructing the phase spectrum from TPD and FPD through an additional model like [35].

3.2.2. Magnitude-based Weighted UPB-aware loss function

In Sec. 3.2.1, we did not weight the individual time-frequency bins in the TPD and FPD spectra, which implies that all TPD and FPD were considered to have the same perceptual impact on the final enhanced signal. However, this contradicts common sense. Consider the following scenario: there are two time-frequency bins, and if the magnitude of one is very small (i.e., close to 0) while the magnitude of the other is very large, will the corresponding TPD and FPD have the same perceptual impact? Clearly, they will not, because when converting back to the time domain, the time-frequency bin with a larger magnitude will produce stronger sinusoidal signals, whereas the signals from the time-frequency bin close to 0 can be almost ignored. At this point, its phase offset value can also be ignored.

On the basis of the above insight, we propose the magnitude-based weighted UPB loss function on the basis of (10). First, we use (11) to obtain the power-compressed magnitude spectrum $\mathbf{M}_{cmp}(t,f) \in \mathbb{R}^{T\times F}$.

$$\mathbf{M}_{cmp}(t,f) = \big(\mathbf{M}(t,f)\big)^c = \big(|\mathbf{X}(t,f)|\big)^c \tag{11}$$

Here, $\mathbf{X}$ represents the spectrogram of the audio processed by STFT.

After obtaining $\mathbf{M}_{cmp}$, we use (12) and (13) to compute the magnitude-weighted wrapped TPD $\delta_{ww\tau} \in \mathbb{R}^{(T-1)\times F}$ and FPD $\delta_{ww\omega} \in \mathbb{R}^{T\times(F-1)}$. Σ here means to sum up.

$$\boldsymbol{\delta}_{ww\tau}(t,f) = \frac{\mathbf{M}_{cmp}(t+1,f) + \mathbf{M}_{cmp}(t,f)}{\Sigma\big(\mathbf{M}_{cmp}(t+1,f) + \mathbf{M}_{cmp}(t,f)\big)}\boldsymbol{\delta}_{w\tau}(t,f) \tag{12}$$

$$\boldsymbol{\delta}_{ww\omega}(t,f) = \frac{\mathbf{M}_{cmp}(t,f+1) + \mathbf{M}_{cmp}(t,f)}{\Sigma\big(\mathbf{M}_{cmp}(t,f+1) + \mathbf{M}_{cmp}(t,f)\big)}\boldsymbol{\delta}_{w\omega}(t,f) \tag{13}$$

Consequently, the magnitude-weighted UPB loss $\mathcal{L}_{wupb}$ can be expressed as (14).

$$\begin{aligned}\mathcal{L}_{wupb} = &\frac{1}{2}\mathbb{E}_{\boldsymbol{\delta}_{wu\tau}(t,f),\hat{\boldsymbol{\delta}}_{wu\tau}(t,f)}[\|\mathcal{S}\big(\boldsymbol{\delta}_{wu\tau}(t,f), \hat{\boldsymbol{\delta}}_{wu\tau}(t,f)\big)\|^2] \\ &+ \frac{1}{2}\mathbb{E}_{\boldsymbol{\delta}_{wu\omega}(t,f),\hat{\boldsymbol{\delta}}_{wu\omega}(t,f)}[\|\mathcal{S}\big(\boldsymbol{\delta}_{wu\omega}(t,f), \hat{\boldsymbol{\delta}}_{wu\omega}(t,f)\big)\|^2]\end{aligned} \tag{14}$$

It is important to note that this equation is the loss function during training optimization. The weights should be calculated using the clean magnitude as the standard, so even in $\hat{\boldsymbol{\delta}}_{ww\tau}(t,f)$ and $\hat{\boldsymbol{\delta}}_{ww\omega}(t,f)$, the weighted portion still uses the clean audio magnitude spectrum $\mathbf{M}_{cmp}$ instead of the estimated magnitude spectrum $\hat{\mathbf{M}}_{cmp}$.

Finally, the overall loss function $\mathcal{L}_2$ of CMGAN optimized by the magnitude-weighted biased phase loss function can be expressed as (15).

$$\mathcal{L}_2 = \lambda_1\mathcal{L}_{mag} + \lambda_4\mathcal{L}_{adv} + \lambda_6\mathcal{L}_{wupb} \tag{15}$$

3.2.3. UPB-aware discriminator

We have also incorporated the aforementioned properties of the biased phase into the Discriminator of CMGAN. The original discriminator input consists only of the magnitude spectra of clean and estimated speech, $(\mathbf{M}, \hat{\mathbf{M}})$, to which we add four additional inputs: wrapped TPD and FPD of clean and estimated speech $(\boldsymbol{\delta}_{w\tau}, \boldsymbol{\delta}_{w\omega}, \hat{\boldsymbol{\delta}}_{w\tau}, \hat{\boldsymbol{\delta}}_{w\omega})$.

Initially, we concatenate the above spectra together, using (16) to generate the corresponding input matrices $\mathbf{A}$ and $\hat{\mathbf{A}}$ for the discriminator. Here, $\mathbf{A} \in \mathbb{R}^{T\times F}$ represents the input matrix corresponding to the clean speech signal, and $\hat{\mathbf{A}} \in \mathbb{R}^{T\times F}$ corresponds to the estimated signal.

$$\mathbf{A} = [\boldsymbol{\delta}_{w\tau} : \boldsymbol{\delta}_{w\omega} : \mathbf{M}] \tag{16}$$

In this formula, ':' represents the concatenation operation. Since the shapes of $\delta_{w\tau} \in \mathbb{R}^{(T-1)\times F}$, $\delta_{w\omega} \in \mathbb{R}^{T\times(F-1)}$, and $\mathbf{M} \in \mathbb{R}^{T\times F}$ do not match, we first need to pre-perform a one-length zero padding on the boundaries of $\delta_{w\tau}$ and $\delta_{w\omega}$ to make the shapes consistent before performing the concatenation operation.

Then, the prediction process of the normalized PESQ score by the discriminator $\mathcal{D}(\cdot)$, $Q_{PESQ} = \frac{PESQ-1}{3.65}$, can be represented as in (17).

$$\hat{Q}_{PESQ} = \mathcal{D}(\mathbf{A}, \hat{\mathbf{A}}) \tag{17}$$

Here, $\hat{Q}_{PESQ}$ represents the predicted value of the normalized PESQ score from the discriminator. The optimization function of the discriminator is shown in (18).

$$\begin{aligned}\mathcal{L}_D = &\mathbb{E}_{\mathbf{A},\mathbf{A}}\big[\|\mathcal{D}(\mathbf{A},\mathbf{A}) - 1\|^2\big] \\ &+ \mathbb{E}_{\mathbf{A},\hat{\mathbf{A}}}\big[\|\mathcal{D}(\mathbf{A},\hat{\mathbf{A}}) - Q_{PESQ}\|^2\big]\end{aligned} \tag{18}$$

At this point, the new adversarial loss of the generator based on the UPB discriminator, $\mathcal{L}_{upb-adv}$, is given by (19).

$$\mathcal{L}_{upb-adv} = \mathbb{E}_{\mathbf{A},\hat{\mathbf{A}}}\big[\|\mathcal{D}(\mathbf{A},\hat{\mathbf{A}}) - 1\|^2\big] \tag{19}$$

Ultimately, the overall loss function of the optimized CMGAN, $\mathcal{L}_3$, can be expressed as in (20).

$$\mathcal{L}_3 = \lambda_1\mathcal{L}_{mag} + \lambda_6\mathcal{L}_{wupb} + \lambda_7\mathcal{L}_{upb-adv} \tag{20}$$

3.2.4. UPB-based data augmentation

Finally, we developed two data augmentation methods, one based on a global biased phase and the other on another frequency-based angle biased phase, and combined them with methods on the magnitude spectrum to augment input noisy speech. The global biased phase-based data augmentation is consistent with the method in Sec. 2 (1), so it is not detailed here. We focus on the linear biased phase-based data augmentation, which is similar to (1), but we replaced the θ with $\omega\tau'$, as shown in (21).

$$\begin{aligned}X(t,\omega) &= \int_{-\infty}^{\infty} x(\tau)h(\tau-t)e^{-j(\omega\tau+\omega\tau')}d\tau \\ &= \int_{-\infty}^{\infty} x(\tau)h(\tau-t)e^{-j\omega(\tau+\tau')}d\tau\end{aligned} \tag{21}$$

Table 3. Performance comparison on the Voice Bank+DEMAND dataset. "/" denotes that the current model does not have this hyper-parameter. "-" denotes that the result is not provided. ('T' stands for True, 'F' stands for False.)

Model	Input type	UPB				No. of paras. (M)	PESQ	CSIG	COVL	STOI
		Loss	Weight	Disc.	Data aug.					
Noisy	/	/	/	/	/	–	1.97	3.35	2.63	0.91
PHASEN [8]	Magnitude + Phase	/	/	/	/	–	2.99	4.21	3.62	–
MetricGAN+ [5]	Magnitude	/	/	/	/	–	3.15	4.14	3.64	–
MANNER [36]	Time domain	/	/	/	/	–	3.21	4.53	3.91	0.95
D^2Net [37]	Complex	/	/	/	/	–	3.27	4.63	3.92	**0.96**
DPT-FSNet [15]	Complex	/	/	/	/	0.91	3.33	4.58	4.00	**0.96**
D2Former [38]	Complex	/	/	/	/	0.87	3.43	4.66	4.22	**0.96**
MP-SENet [7]	Magnitude + Phase	/	/	/	/	2.05	3.50	4.73	4.22	**0.96**
CMGAN (baseline) [32]	Magnitude + Complex	/	/	/	/	1.83	3.41	4.63	4.12	**0.96**
UPB-CMGAN (proposed)	Magnitude + Complex	T	F	F	F	1.83	3.46	4.73	4.21	**0.96**
	Magnitude + Complex	T	T	F	F	1.83	3.49	4.75	4.25	**0.96**
	Magnitude + Complex	T	T	T	F	1.83	3.52	4.75	4.24	**0.96**
	Magnitude + Complex	T	T	T	T	1.83	**3.55**	**4.78**	**4.28**	**0.96**

Here, we set the value of $\tau' \in \frac{rand(0,2\pi)}{f_s}$, where f_s means the sampling rate. From the relationship in the formula, we can see that this data augmentation method is equivalent to a time shift of τ' length in the time domain.

We also utilized a method of adding random real values on the magnitude spectrum, which follows the distribution of $\mathcal{N}(0, 4 \times 10^{-6})$.

Note that the probability of using each of the above methods on each training audio sample is set to 50%.

4. EXPERIMENTS

4.1. Dataset

We conducted experiments on the VoiceBank-DEMAND dataset [39]. The train set consists of 11,572 utterances (from 28 speakers) mixed with noise data with four signal-to-noise ratios (SNRs) (15, 10, 5, and 0 dB). The test set consists of 824 utterances (from two unseen speakers) mixed with unseen noise data with four SNRs (17.5, 12.5, 7.5, and 2.5 dB). All the utterances have been resampled to 16 kHz for a fair comparison.

4.2. Evaluation metrics

Considering Table 2 in Sec. 2, which indicates the lack of correlation between time-domain metrics and perceived auditory quality, we abandoned the use of such metrics and only utilized the following four non-time-domain objective metrics to evaluate the performance of the proposed method and compare it with previous models. These included PESQ [30], which assesses speech quality on a scale from –0.5 to 4.5, and the Short-time Objective Intelligibility (STOI) [40], which evaluates speech intelligibility on a scale from 0 to 100%. Two mean opinion score (MOS)-based measures were also considered, both of which operate on a scale from 1 to 5. These include CSIG [41], which predicts signal distortion, and COVL [41], which forecasts overall signal quality. For all these metrics, higher scores indicate a superior speech enhancement performance.

4.3. Experimental configuration

The training process involved segmenting the signals into 2-second chunks, with shorter signals being padded to meet this duration. The batch size was set to 32. All the models underwent training for a total of 180 epochs. To ensure optimal model selection, we saved the checkpoint after each epoch and recorded the best model on the validation set during the training progress. The optimizer we used was AdamW, with distinct initial learning rates for the generator (4×10^{-3}) and discriminator (8×10^{-3}). A StepLR scheduler with a gamma of 0.6 was utilized to adjust the learning rate every 30 epochs, facilitating a balanced learning pace. The $\lambda_{1\sim7}$ weight parameters mentioned above are $[0.9,\ 0.1,\ 0.2,\ 0.05,\ 0.05,\ 0.05,\ 0.05]$. All the experiments were conducted using PyTorch (version 2.0.1) on 8 Nvidia RTX 3090Ti GPUs endowed with 24 GB of memory.

5. RESULTS AND ABLATION STUDY

The results of the speech enhancement effects of the UPB-CMGAN model on the VoiceBank-DEMAND dataset, along with its ablation study results and comparisons with other models, are presented in Table 3. The results indicate that replacing the original ri loss and time loss of CMGAN with UPB Loss led to an increase in PESQ, CSIG, and COVL. Furthermore, when we incorporated magnitude-based weights into the UPB Loss, these objective evaluation metrics showed additional improvement. A similar trend was observed when the phase derivative input was added to the original discriminator, which was solely based on magnitude. With an informal listening of generated samples, we observed a slight hum noise recognized in some worse-case samples. However, the objective metrics used to evaluate the performance of the methods on VoiceBank-DEMAND indicated that leveraging UPB-based data augmentation had a synergistic effect that led to the aforementioned improvements. Overall, our model achieved the PESQ of 3.55, CSIG of 4.78, and COVL of 4.28, which are SoTA results.

6. CONCLUSION

In this paper, we propose four mutually compatible improvements for the estimation of phase spectra based on the CMGAN framework. We abandon the traditional approach of precise phase reconstruction and instead estimate phases with unrestricted global-phase bias to reduce the training burden of the model. We conducted thorough ablation experiments on the proposed UPB-CMGAN model and compared it with various existing models. The results indicate that all four proposed methods are reasonable and effective. Furthermore, when these methods were implemented collaboratively, the UPB-CMGAN achieved a SoTA performance on the VoiceBank-DEMAND dataset without incurring additional computational costs.

References

[1] Jacob Benesty, Shoji Makino, and Jingdong Chen, *Speech enhancement*, Springer Science & Business Media, 2006.

[2] Dong Yu and Lin Deng, *Automatic speech recognition*, vol. 1, Springer, 2016.

[3] Sebastian Braun and Hannes Gamper, "Effect of noise suppression losses on speech distortion and asr performance," in *Proc. ICASSP*, 2022, pp. 996–1000.

[4] Szu-Wei Fu *et al.*, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019, pp. 2031–2041.

[5] Szu-Wei Fu *et al.*, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.

[6] Guochen Yu *et al.*, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *Proc. ICASSP*, 2022, pp. 7847–7851.

[7] Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling, "MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra," in *Proc. Interspeech*, 2023, pp. 3834–3838.

[8] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI*, 2020, vol. 34, no. 05, pp. 9458–9465.

[9] Ziyi Xu, Samy Elshamy, and Tim Fingscheidt, "Using separate losses for speech and noise in mask-based speech enhancement," in *Proc. ICASSP*, 2020, pp. 7519–7523.

[10] Meet H Soni, Neil Shah, and Hemant A Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP*, 2018, pp. 5039–5043.

[11] Ke Tan and DeLiang Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. ICASSP*, 2019, pp. 6865–6869.

[12] Ke Tan and DeLiang Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE Trans. ASLP*, vol. 28, pp. 380–390, 2019.

[13] Hsieh Tsun-An *et al.*, "Improving Perceptual Quality by Phone-Fortified Perceptual Loss Using Wasserstein Distance for Speech Enhancement," in *Proc. Interspeech*, 2021, pp. 196–200.

[14] Hu Yanxin *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[15] Feng Dang, Hangting Chen, and Pengyuan Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *Proc. ICASSP*, 2022, pp. 6857–6861.

[16] Yin Dacheng *et al.*, "TridentSE: Guiding Speech Enhancement with 32 Global Tokens," in *Proc. Interspeech*, 2023, pp. 3839–3843.

[17] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[18] Eesung Kim and Hyeji Seo, "Se-conformer: Time-domain speech enhancement using conformer.," in *Proc. Interspeech*, 2021, pp. 2736–2740.

[19] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2018, pp. 696–700.

[20] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. ASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[21] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. ASLP*, vol. 32, no. 2, pp. 236–243, 1984.

[22] Yang Ai and Zhen-Hua Ling, "Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses," in *Proc. ICASSP*, 2023, pp. 1–5.

[23] Doyeon Kim *et al.*, "Phase continuity: Learning derivatives of phase spectrum for speech enhancement," in *Proc. ICASSP*, 2022, pp. 6942–6946.

[24] Yoshiki Masuyama *et al.*, "Deep griffin–lim iteration," in *Proc. ICASSP*, 2019, pp. 61–65.

[25] Nguyen Binh Thien *et al.*, "Inter-frequency phase difference for phase reconstruction using deep neural networks and maximum likelihood," *IEEE Trans. ASLP*, 2023.

[26] Nguyen Binh Thien *et al.*, "Two-stage phase reconstruction using dnn and von mises distribution-based maximum likelihood," in *Proc. APSIPA*, 2021, pp. 995–999.

[27] Yoshiki Masuyama *et al.*, "Deep griffin–lim iteration: Trainable iterative phase reconstruction using neural network," *JSTSP*, vol. 15, no. 1, pp. 37–50, 2020.

[28] Mikko-Ville Laitinen, Sascha Disch, and Ville Pulkki, "Sensitivity of human hearing to changes in phase spectrum," *JAES*, vol. 61, no. 11, pp. 860–877, 2013.

[29] Roy D Patterson, "A pulse ribbon model of monaural phase perception," *JASA*, vol. 82, no. 5, pp. 1560–1586, 1987.

[30] Antony W Rix *et al.*, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752.

[31] John HL Hansen and Bryan L Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP*, 1998.

[32] Ruizhe Cao, Sherif Abdulatif, and Bin Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. Interspeech*, 2022, pp. 936–940.

[33] Yoshiki Masuyama *et al.*, "Online phase reconstruction via dnn-based phase differences estimation," *IEEE Trans. ASLP*, vol. 31, pp. 163–176, 2022.

[34] Anthony P Stark and Kuldip K Paliwal, "Speech analysis using instantaneous frequency deviation," in *Proc. Interspeech*, 2008.

[35] Lars Thieling, Daniel Wilhelm, and Peter Jax, "Recurrent phase reconstruction using estimated phase derivatives from deep neural networks," in *Proc. ICASSP*, 2021, pp. 7088–7092.

[36] Hyun Joon Park *et al.*, "Manner: Multi-view attention network for noise erasure," in *Proc. ICASSP*, 2022, pp. 7842–7846.

[37] Liusong Wang *et al.*, "D 2 net: A denoising and dereverberation network based on two-branch encoder and dual-path transformer," in *Proc. APSIPA*, 2022, pp. 1649–1654.

[38] Shengkui Zhao and Bin Ma, "D2former: A fully complex dual-path dual-decoder conformer network using joint complex masking and complex spectral mapping for monaural speech enhancement," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[39] Cassia Valentini-Botinhao *et al.*, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *Proc. SSW*, 2016, pp. 146–152.

[40] Cees H Taal *et al.*, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[41] Yi Hu and Philipos C Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. ASLP*, vol. 16, no. 1, pp. 229–238, 2007.

Appendix

This paper was accepted and presented at ICASSP 2024. After presenting it at ICASSP 2024, we recognized that a data augmentation scheme [42] similar to the ”frequency-based angle biased phase” part (21) of our method in Section 3.2.4 has been proposed in ICASSP 2023. Therefore, we added a reference to the paper here.

References

[42] Junhyeok Lee, Seungu Han, Hyunjae Cho, and Wonbin Jung, “PHASEAUG: a differentiable augmentation for speech synthesis to simulate one-to-many mapping,” in *Proc. ICASSP*, 2023, pp. 1–5.