Improving the performance of CMGAN in speech enhancement with the phone fortified perceptual loss

Chi-En Dai[1]*, Jia-Xuan Zeng*[1]*, Wan-Ling Zeng*[1]*, Eric S. Li*[2] *and Jeih-weih Hung*[1]

[1]National Chi Nan University, Taiwan

[2]National Taipei University of Technology, Taiwan

s108323060@mail1.ncnu.edu.tw, s109323003@mail1.ncnu.edu.tw, s109323001@mail1.ncnu.edu.tw, ericli@ntut.edu.tw, jwhung@ncnu.edu.tw

Abstract

This study presents to promote the behaviour of a celebrated speech enhancement framework, the Conformer-based Metric Generative Adversarial Network (CMGAN). CMGAN adopts a generative adversarial network (GAN) structure and employs a Conformer-based metric to evaluate the synthetic data from the generator work. In CMGAN, the loss function for the generator network is mainly from three sources: the time-frequency domain loss, the waveform-domain loss and the loss fed by the discriminator network. We argue that this loss function can be further revised by adding the phone-fortified perceptual loss (PFPL). PFPL reflects the mismatch between the synthetic data and the real data in the respective latent phone-level representation created by the wave2vec encoder. The preliminary experiments conducted on the VoiceBank-DEMAND task indicate that when integrating the PFPL in the learning of CMGAN, the objective metrics (PESQ and STOI) scores for the test data can be significantly promoted.

Index Terms: speech enhancement, CMGAN, perceptual speech quality, short-time objective intelligibility, phone-fortifed perceptual loss

1. Introduction

Speech enhancement (SE) is the process of alleviating noise, reverberation, or other types of interference in speech signals to improve the corresponding quality and intelligibility. Novel SE methods usually consist of a deep neural network (DNN) structure, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs). Compared to conventional SE methods that rely on estimating the statistics regarding speech or interference, DNN-based SE methods can learn complex features of speech and interference. They can also adapt to different interference scenarios. The GAN-wise SE framework consists of two networks, a generator and a discriminator. The generator network enhances the noisy utterance, and the discriminator network highlights the difference between enhanced speech and real clean speech. SEGAN (Speech Enhancement Generative Adversarial Network) [1] is the first SE method using a GAN structure. What follows include WSEGAN [2], HiFi-GAN [3], and MetricGAN [4], to name a few.

In particular, a novel GAN-wise framework, conformer-based MetricGAN (CMGAN) is presented by employing two-stage conformer blocks in the time-frequency (TF)-domain speech features for the generator. The conformer-based generator processes input magnitude and complex spectrograms by creating the respective masks. The masked spectrograms from the two sources (masked magnitude plus noisy phase, masked real and imaginary components) are then added to constitute the final estimated spectrogram.

In this study, we follow the excellent work of CMGAN and provide possible revisions to improve its SE performance. In the original CMGAN, the loss function of its generator consists of three parts: the TF-domain loss, the waveform (time-domain) loss, and the loss regarding the output of the discriminator part. The first two parts are directly associated with the plain signals irrelevant to human perception and comprehension. The third part is implicitly correlated with the perceptual quality due to using PESQ in the discriminator. Partially inspired by the work [6], we argue that CMGAN can be further improved in SE if it considers the human comprehension loss, and present to add phone fortified perceptual loss (PFPL) to learn the generator of the CMGAN. Briefly speaking, PFPL considers the phonetic information dwelled in speech and employs the powerful $wav2vec$ [7] model to capture it.

The rest of this paper is organized as follows. Section 2 briefly introduces the CMGAN system and the presented revision. Sections 3 and 4 reveal the experimental setup, results, and corresponding findings and analyses. Finally, a brief concluding remark is given in Section 5.

2. Proposed method

The CMGAN network inherits most advanced DNN-based SE methods that contain an encoder-decoder architecture as the generator, and it specifically follows MetricGAN, which employs a perceptual metric discriminator. One of the specialities of CMGAN is that its generator network employs the time-frequency (TF)-domain features for the following encoder with two-stage conformer blocks, a magnitude mask decoder and a complex decoder. The dual-path transformer is applied to the two-stage conformer blocks to reduce the computation load. Let (x, X_m, X_r, X_i) denote the target time-domain waveform, magnitude, real and imaginary spectrograms and $(\hat{x}, \hat{X}_m, \hat{X}_r, \hat{X}_i)$ denote the respective enhanced version from the CMGAN. The loss functions for the discriminator and generator of CMGAN are set to be

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{X_m}\{\|D(X_m, X_m) - 1\|^2\} \\ &+ \mathbb{E}_{X_m, \hat{X}_m}\{\|D(X_m, X_m) - Q_{PESQ}\|^2\},\end{aligned} \tag{1}$$

with D denoting the discriminator and Q_{PESQ} referring to the normalized PESQ score, and

$$\mathcal{L}_G = \gamma_1 \mathcal{L}_{TF} + \gamma_2 \mathcal{L}_{GAN} + \gamma_3 \mathcal{L}_{Time}, \tag{2}$$

where $\mathcal{L}_{TF}$, $\mathcal{L}_{GAN}$ and $\mathcal{L}_{Time}$ are three loss types with the corresponding weights γ_1, γ_2, and γ_3. They are defined as follows:

- TF loss: $\mathcal{L}_{TF} = \alpha \mathcal{L}_{mag} + (1-\alpha)\mathcal{L}_{RI}$, with $\mathcal{L}_{mag} = \mathbb{E}_{X_m, \hat{X}_m}\{\|X_m, \hat{X}_m|^2\}$, and $\mathcal{L}_{RI} = \mathbb{E}_{X_r, \hat{X}_r}\{\|X_r - \hat{X}_r|^2\} + \mathbb{E}_{X_i, \hat{X}_i}\{\|X_i - \hat{X}_i|^2\}$.
- GAN loss: $\mathcal{L}_{GAN} = \mathbb{E}_{X_m, \hat{X}_m}\{\|D(X_m, \hat{X}_m) - 1\|^2\}$.
- Time loss: $\mathcal{L}_{Time} = \mathbb{E}_{x, \hat{x}}\{\|x - \hat{x}\|_1\}$.

Observing the loss for the generator network shown in Eq. (2), its two components $\mathcal{L}_{TF}$ and $\mathcal{L}_{Time}$ are directly related to the physical attributes of speech signals, and the term $\mathcal{L}_{GAN}$ is somehow implicitly correlated with the signals' perceptual nature through the discriminator D that contains perceptual metric Q_{PESQ}. Accordingly, we wonder whether CMGAN can behave even better in perceptual quality and intelligibility if its loss function is explicitly associated with phonetic information in speech signals. Here, we present using the phone-fortified perceptual (PFP) [9] loss to reflect perceptual distance between enhanced speech $\hat{x}$ and clean target speech x, which is expressed by

$$\mathcal{L}_{PFP}(x, \hat{x}) = \mathbb{E}_{x, \hat{x}}\{\|\Phi_{wave2vec}(\hat{x}) - \Phi_{wave2vec}(x)\|_1\}, \tag{3}$$

where $\Phi_{wave2vec}$ is the pre-trained $wav2vec$ encoder [7] used to extract low-dimensional feature vectors for calculating the L^1 distance between x and $\hat{s}$ in latent spaces.

As a result, the updated loss function for the generator network of CMGAN is expressed by:

$$\begin{aligned}\tilde{\mathcal{L}}_G &= \mathcal{L}_G + \gamma_4 \mathcal{L}_{PFP} \\ &= \gamma_1 \mathcal{L}_{TF} + \gamma_2 \mathcal{L}_{GAN} + \gamma_3 \mathcal{L}_{Time} + \gamma_4 \mathcal{L}_{PFP},\end{aligned} \tag{4}$$

3. Experimental Setup

We validate our novel work by experimenting with the VoiceBank-DEMAND task—the speech utterances and the noise sample are from the VoiceBank [8] and DEMAND [9] data sets, respectively. The training set includes 11,572 utterances produced by 28 speakers and corrupted by ten noise types at four SNRs (0, 5, 10, and 15 dB), and the test set contains 824 utterances from 2 speakers other than those in the training set with five unseen noise types at four SNRs (2.5, 7.5, 12.5 and 17.5 dB). Notably, all utterances are re-sampled at 16 kHz in our experiments.

Regarding the hyper-parameter settings for the CMGAN model training, there are 150 epochs with the batch size $B = 4$, the number of conformer blocks is $N = 4$, the channel numbers for both the generator and discriminator are $C = 64$, the number of frequency bins is $F = 200$, and the frame length and hop size are 25 ms and 6.25 ms, respectively. We use two objective metrics, PESQ [10] and STOI [11], to evaluate the perceptual quality and intelligibility of the enhanced utterances in the test set.

4. Experimental results and discussions

Table 1 reveals the PESQ and STOI results for the unprocessed baseline, the CMGAN baseline, and the revised various non-zero assignments of the PFPL loss weight $\gamma_4 = 0$ in Eq. (4). From this table, we have some observations:

1. CMGAN promotes PESQ and STOI scores significantly compared with the unprocessed baseline, indicating its excellent SE performance.
2. When adopting PFP loss with $\gamma_4 > 0$, the corresponding revised CMGAN outperforms the CMGAN baseline in PESQ and STOI metrics at almost all cases. The best possible improvements for PESQ and STOI are respectively 0.1307 (from 2.9884 to 3.1191 with $\gamma_4 = 2.0$) and 0.0023 (from 0.9515 to 0.9538 $\gamma_4 = 1.0$). Accordingly, the effectiveness of the presented revision for CMGAN is evidenced.
3. With the revised CMGAN, PESQ increases more significantly than STOI, which is possibly due to the fact the achieved STOI has been quite close to its ideal value 1.
4. Increasing the weight γ_4 from 0.001 to 0.01 improves PESQ steadily, while it is not always the case by assigning γ_4 higher than 0.01. Setting $\gamma_4 = 2.0$ gives the optimal PESQ score (3.1191), which is close to 3.1025 obtained by the case $\gamma_4 = 0.01$. Since the PFP loss might be partially correlated with the other three losses in Eq. (4), enlarging its portion does not necessarily influence the performance substantially.

Figure 1 depicts the magnitude spectrogram of an utterance in the test set at four situations. Comparing Figs. 1(a)(b), we see that noise causes a significant distortion, and this distortion can effectively reduced by the original and revised CMGAN as shown in Figs. 1(b)(c)(d). Observing Figs. 1(c)(d), the revised CMGAN is shown to alleviate the distortion further than the original CMGAN, especially in non-speech portions at the start and the end of the utterance.

Table 1: *Averaged PESQ and STOI results of and the test set for the network learned with the loss function in Eq. (4) at different assignments of the PFPL loss weight* γ_4

		PESQ	STOI
unprocessed baseline		1.9700	0.9210
CMGAN baseline ($\gamma_4 = 0$)		2.9884	0.9515
revised CMGAN	$\gamma_4 = 0.001$	**3.0323**	**0.9524**
	$\gamma_4 = 0.005$	**3.0550**	0.9515
	$\gamma_4 = 0.01$	**3.1025**	**0.9522**
	$\gamma_4 = 0.05$	**3.0607**	0.9510
	$\gamma_4 = 0.1$	**3.0891**	**0.9517**
	$\gamma_4 = 1.0$	**3.0739**	**0.9538**
	$\gamma_4 = 1.5$	**3.0715**	**0.9527**
	$\gamma_4 = 2.0$	**3.1191**	**0.9521**

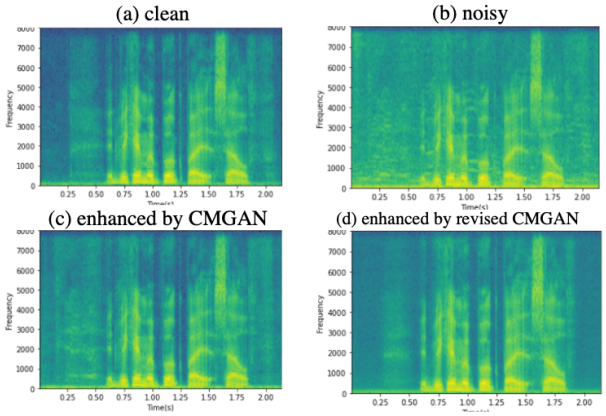

Figure 1: *The different forms of spectrograms of an utterance: (a) clean, (b) noisy, (c) noisy and enhanced by original CMGAN, (d) noisy and enhanced by the revised CMGAN*

5. Conclusions

This study examines if the phone-fortified perceptual loss (PFPL) can benefit the learning for the generator network of CMGAN to promote its speech enhancement capability. The preliminary experiments with the VoiceBank-DEMAND task confirms this supposition, showing that the PESQ and STOI metric scores for the test data can be further improved by the revised CMGAN. In the future avenue, we will investigate if CMGAN can be downsized by simply exploiting the complex spectrogram as the input features and maintain its superior performance with the help of PFPL.

6. References

[1] S. S. Pascual, A. Bonafonte, and J. Serra‘, “Segan: Speech enhancement generative adversarial network,” *in Proc. Interspeech*, 2017.

[2] S Pascual, A Bonafonte, J Serrà, JA Gonzalez, “Whispered-to-voiced alaryngeal speech conversion with generative adversarial networks,” *arXiv:1808.10687v2*, 2018. hifigan

[3] J. Su, Z. Jin, and A. Finkelstein, “Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” *in Proc. Interspeech*, 2020.

[4] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, “Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” *in Proc. ICML*, 2019.

[5] R. Cao, S. Abdulatif and B. Yang, “CMGAN: Conformer-based Metric GAN for speech enhancement,” *arXiv:2203.15149v3*, 2022

[6] T. Hsieh, C. Yu, S. Fu, X. Lu, and Y. Tsao, “Improving perceptual quality by phone-fortified perceptual loss for speech enhancement,” *arXiv preprint arXiv:2010.15174*, 2020.

[7] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” *in Proc. Interspeech*, 2019.

[8] C. V-Botinhao *et al.*, “Investigating RNN-based speech enhancement methods for noise-robust text-to-speech,” *in Proc. SSW*, 2016.

[9] J. Thiemann *et al.*, “Demand: a collection of multi-channel recordings of acoustic noise in diverse environments,” *in Proc. ICA*, 2013.

[10] A. 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.

[11] C. H. Taal *et al.*, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, 2011.