Leveraging the objective intelligibility and noise estimation to improve Conformer-based MetricGAN

Chi-En Dai, Wan-Ling Zeng, Jia-Xuan Zeng, and Jeih-weih Hung

National Chi Nan University
No. 1, University Road
Puli Township, Nantou County, Taiwan
{s108323060, s109323001, s109323003}@mail1.ncnu.edu.tw, jwhung@ncnu.edu.tw

Abstract

Conformer-based MetricGAN (CMGAN) is a deep neural network (DNN)-based speech enhancement (SE) method that uses time-frequency (TF) domain features to learn a novel conformer-wise generative network, and it has demonstrated excellent SE performance in terms of various perceptual evaluation metrics.

In this study, we propose to revise CMGAN along three directions. To begin, we incorporate phone-fortified perceptual loss (PFPL) into its loss function. The PFPL is calculated using latent representations of speech from the wav2vec module. With PFPL as part of the loss function can effectively use perceptual and linguistic speech information to direct CMGAN model training. Next, we revise the discriminator output by adding the STOI values. The original discriminator is trained to estimate the enhanced PESQ score by taking both clean and enhanced spectrum as inputs as well as the associated PESQ label. In other words, the initial discriminator only takes into account the PESQ score. By further considering STOI, we expect to improve the discriminator. Finally, we add noise label estimation to the entire CMGAN framework. The original CMGAN only calculates the disparity between the estimated value provided by the model and the clean target with clean labels. Instead, we further take into account noise estimation loss, which can show the discrepancy between the predicted noise and the noise label.

The Voicebank-Demand dataset is used for the evaluation experiments. According to the experimental results, the revised CMGAN outperforms the original by gaining greater scores on objective perceptual metrics including PESQ and STOI. As a result, we confirm the success of the presented revisions in CMGAN.

Key words: speech enhancement, adaptive FullSubNet+, discrete wavelet transform, wavelet packet decomposition, PESQ, STOI, SI-SNR

I. Introduction

Speech is one of the most effective human communication tools. We can transmit speech messages to a variety of locations and environments thanks to highly effective telecommunication technology. However, during transmission, the speech signal is susceptible to noise and other types of interference, making it difficult to comprehend.

Speech advancements in recent decades have resulted in countermeasures to address the problem of speech distortion in communication. Speech enhancement (SE) is one such approach that attempts to increase the quality and intelligibility of speech signals. Conventional SE methods depend on statistical modeling of speech and noise, and they typically fail in non-stationary noise scenarios. With the rapid advancement of deep learning and deep neural networks (DNN) during the last decade, researchers have built SE frameworks based on DNN, with significant success compared to previous SE methods.

There are two types of DNN-based SE methods: mapping-based and masking-based [1]. Because of their capacity to adjust output dynamic range and faster convergence, masking-based approaches are more prevalent. Several time-frequency (T-F) masking techniques, including ideal binary mask [2] ideal ratio mask [3] and complex ideal ratio mask (cIRM) [4] have been proposed. cIRM, in particular, can deal with phase information implicitly without modeling its ambiguous structure. In particular, the generative adversarial network (GAN), a special DNN-based arrangement, has found considerable success in the SE field as well as many others. A GAN-based SE framework is made up of two networks: a generator and a discriminator. The generator network enhances the noisy utterance, whereas the discriminator network distinguishes between enhanced speech and real clean speech. The first SE method to use a GAN structure is SEGAN (Speech Enhancement Generative Adversarial Network) [5]. Furthermore, the novel GAN-wise architecture, conformer-based MetricGAN (CMGAN) [6], is proposed in particular, by leveraging two-stage conformer blocks in the time-frequency (TF)-domain speech characteristics for the generator. The conformer-based generator generates masks for processing input magnitude and complex spectrograms. The masked spectrograms from the two sources are then merged to form the final estimated spectrogram (masked magnitude plus noisy phase, masked real and imaginary components).

In this study, we build on CMGAN's excellent work and propose feasible modifications to improve its SE performance. The improvements provided here mostly concern the loss functions used in the learning of CMGAN's generator and discriminator networks. The preliminary evaluations on the Voicebank-Demand challenge show that the newly improved CMGAN framework outperforms the original version.

II. CMGAN

The CMGAN network uses an encoder-decoder architecture as the generator, and it specifically follows MetricGAN and employs a perceptual metric discriminator. One of the specialties 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

$$\mathcal{L}_D = E_{X_m}\{\|D(X_m, \hat{X}_m) - 1\|^2\} + E_{X_m,\hat{X}_m}\{\|D(X_m, \hat{X}_m) - Q_{PESQ}\|^2\} \quad (1)$$

with D denoting the discriminator and Q_{PESQ} referring to the normalized PESQ score [7], and

$$\mathcal{L}_G = \gamma_1\mathcal{L}_{TF} + \gamma_2\mathcal{L}_{GAN} + \gamma_3\mathcal{L}_{Time} \quad (2)$$

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

For the details of CMGAN, please refer to [6].

III. Proposed method

This study focuses on improving the SE performance of CMGAN, and the presented revision is threefold as stated in the following sub-sections:

A. Incorporating PFP loss into the loss function of the generator

According to [8], the phone-fortified perceptual loss (PFPL) is novel loss function for speech enhancement (SE) incorporates phonetic information when training SE models. To calculate the phone-fortified perceptual loss (PFPL), we use two types of losses: one based on the waveforms and another based on the phones. The phone-level loss uses the wav2vec model [9], which is a self-supervised encoder that can capture rich phonetic information. The PFPL measures the distance between the estimated and target distributions of phonetic representations using the Wasserstein distance. As such, we suggest incorporating PFPL into its loss function to further consider the phone-level loss in the CMGAN learning, given that the loss function of the generator of CMGAN is only concerned with the waveform-level information.

Accordingly, the updated loss function for the generator network of CMGAN is expressed by:

$$\tilde{\mathcal{L}}_G = \mathcal{L}_G + \gamma_4\mathcal{L}_{PFP} = \gamma_1\mathcal{L}_{TF} + \gamma_2\mathcal{L}_{GAN} + \gamma_3\mathcal{L}_{Time} + \gamma_4\mathcal{L}_{PFP}, \quad (3)$$

where $\mathcal{L}_{PFP}$ is the PFPL with weight γ_4 and is expressed by

$$\mathcal{L}_{PFP}(x, \hat{x}) = E_{x,\hat{x}}\{\|\Phi_{wave2vec}(\hat{x}) - \Phi_{wave2vec}(x)\|_1\}, \quad (4)$$

where $\Phi_{wave2vec}$ is the pre-trained $wav2vec$ encoder [9] used to extract low-dimensional feature vectors for calculating the L^1 distance between x and $\hat{x}$ in latent spaces.

B. Including STOI loss into the loss function of the discriminator

The original CMGAN discriminator is learned with the target PESQ score for each input utterance. PESQ is mostly used to assess the objective quality of an utterance. It is based on the human auditory model and computes a Mean Opinion Score-Listening Quality Objective (MOS-LQO) score by comparing the utterance with its clean version. Due to the non-differentiability of the PESQ function, the CMGAN discriminator attempts to emulate PESQ using a deep neural network structure that can serve as a loss function. Four convolution blocks make up this discriminator, which is then followed by instance normalization and a PReLU activation. The discriminator is trained to estimate the maximum normalized PESQ score by taking both clean and enhanced spectrograms as an input together with their label of PESQ score.

Here we suggest modifying the CMGAN discriminator to consider STOI [10] in addition to PESQ. STOI is an objective speech evaluation metric that estimates the intelligibility of an enhanced utterance by calculating the correlation between the enhanced utterance and its clean noise-free counterpart in STFT spectrogram. The loss function for the discriminator network is thus changed to:

$$\mathcal{L}_D = E_{X_m}\{\|D(X_m, \hat{X}_m) - 1\|^2\} + E_{X_m,\hat{X}_m}\{\|D(X_m, \hat{X}_m) - Q_{PESQ}\|^2\} + E_{X_m,\hat{X}_m}\{\|D(X_m, \hat{X}_m) - Q_{STOI}\|^2\} \quad (5)$$

where Q_{STOI} refers to the STOI score of $\hat{x}$ relative to x.

C. Incorporating the noise estimation into the loss function of the generator

The generator of the CMGAN framework is to predict the clean speech component in a noisy utterance y by minimizing the loss that measures the difference of enhanced (predicted) speech $\hat{x}$ and ground-truth clean speech x in the training set. Considering the fact that the noise component $n = y - x$ is also known for the utterances in the training set, we propose to incorporate the estimation of noise in the generator network. The resulting loss function for the generator is modified as:

$$\tilde{\mathcal{L}}_G = \beta\mathcal{L}_G(x, \hat{x}) + (1-\beta)\mathcal{L}_G(n, \hat{n}), \quad (6)$$

where $\hat{n} = y - \hat{x}$ is the estimated noise from the generator network, and we set the ratio $\beta = E_x\{\|x\|^2\}/E_y\{\|y\|^2\}$.

IV. Experimental Setup

The VoiceBank-DEMAND task [11] is used to evaluate the presented novel works, and the speech utterances and noise sample are taken from the VoiceBank and DEMAND data sets, respectively. The test set contains 824 utterances from 2 speakers other than those in the training set with five unseen

Table 1. Table 1: Averaged PESQ and STOI results of the unprocessed baseline, the CMGAN baseline, and the revised CMGAN that incorporates any one of noise estimation, STOI loss and PFPL, or the corresponding combinations. The marker "+" indicates that the revision is applied and the marker "−" indicates the revision is not applied.

			PESQ	STOI (%)
Unprocessed baseline			1.9700	92.10
CMGAN baseline			2.9884	95.15
Revised CMGAN			−	
PFPL	**STOI**	**noise estimation**		
+	−	−	**3.1191**	**95.21**
−	+	−	**3.0568**	**95.19**
−	−	+	**3.1509**	**95.15**
+	−	+	**3.0631**	**95.17**
+	+	−	**3.0901**	**95.21**
+	+	+	**3.0521**	**95.19**

2 speakers other than those in the training set with five unseen noise types at four SNRs. The training set consists of 11,572 utterances generated by 28 speakers that have been corrupted by ten noise types at four SNRs (0, 5, 10, and 15 dB). (2.5, 7.5, 12.5 and 17.5 dB). Notably, in our tests, all utterances are resampled at 16 kHz.

When it comes to the hyper-parameter settings for the CMGAN model training, it includes 150 epochs, a batch size of 4, a number of conformer blocks, 64 channels for the generator and discriminator, 200 frequency bins, and frame and hop sizes of 25 ms and 6.25 ms, respectively. The perceptual quality and intelligibility of the enhanced words in the test set are assessed using two objective measures, PESQ and STOI.

V. Experimental Results and Discussions

The PESQ and STOI results for the unprocessed baseline, the CMGAN baseline, and the CMGAN with some combinations of the presented revisions that incorporate non-zero PFPL loss, STOI adoption in the discriminator, and noise estimation are shown in Table 1. Here are some findings from this table:

1. Compared to the unprocessed baseline, CMGAN substantially improves PESQ and STOI scores, demonstrating excellent SE performance.
2. When further considering PFP loss, the corresponding CMGAN outperforms the CMGAN baseline in PESQ and STOI metrics. For PESQ and STOI, the greatest improvements are 0.1307 (from 2.9884 to 3.1191) and 0.0006 (from 0.9515 to 0.9521). The effectiveness of the revision with PFPL for CMGAN is thus demonstrated.
3. The second revision, which adds STOI to the discriminator, results in a modest boost in both PESQ (from 2.9884 to 3.0568) and STOI (from 0.9515 to 0.9519).
4. When the noise estimation loss is taken into consideration, PESQ significantly increases (from 2.9884 to 3.1509) while the STOI remains unchanged.
5. When choosing any combination of the presented revisions, PESQ and/or improvement over the CMGAN baseline can be also observed. However, it does not always outperform the single-component revision. One explanation for this could be that the used weights (γ_4, β) need to be fine-tuned.

VI. Conclusions and Future Works

The objective of this study is to revise the generator and discriminator networks' loss functions in the successful SE framework CMGAN. The modified CMGAN's higher SE performance, as shown by the evaluation results, reveals its promise. Regarding the potential course of action, we will look into how to combine the three revisions that have been offered to make them more additive.

References

[1] Y. Wang, A. Narayanan and D. Wang, "On training targets for supervised speech separation," IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2014.

[2] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in Speech separation by humans and machines, pp. 181–197. Springer,2005.

[3] Arun Narayanan and DeLiang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in Proc. ICASSP, 2013.

[4] Donald S Williamson et al., "Complex ratio masking for monaural speech separation," IEEE/ACM transactions on audio, speech, and language processing, 2015.

[5] S. S. Pascual, A. Bonafonte, and J. Serra', "Segan: Speech enhancement generative adversarial network," in Proc. Interspeech, 2017.

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

[7] 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

[8] 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.

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

[10] 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.

[11] C. V-Botinhao et al., "Investigating RNN-based speech enhance- ment methods for noise-robust text-to-speech," in Proc. SSW, 2016.

[12] J.Thiemannetal.,"Demand:a collection of multi-channel recordings of acoustic noise in diverse environments," in Proc. ICA, 2013.