The University of Sheffield CHiME-7 UDASE Challenge Speech Enhancement System

George Close, William Ravenscroft, Thomas Hain, and Stefan Goetze

Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom
{glclose1, jwravenscroft1, t.hain, s.goetze}@sheffield.ac.uk

Abstract

The CHiME-7 unsupervised domain adaptation speech enhancement (UDASE) challenge targets in-domain adaptation to unlabelled speech data. This paper describes the University of Sheffield team's system submitted for the challenge. A generative adversarial network (GAN) structure is employed as opposed to the unsupervised RemixIT method proposed in the baseline system. The system uses a conformer-based metric GAN (CMGAN) structure. The discriminator part of the GAN is trained to predict the output of a DNSMOS model. Data augmentation strategies are employed which enable training on historical training data as well as miscellaneous data from an additional generator. The proposed approach, referred to as CMGAN+/+, achieves significant improvement in DNSMOS evaluation metrics with the best proposed system achieving 3.40 OVR-MOS, an 18% improvement over the baselines.

Index Terms: speech enhancement, model generalisation, generative adversarial networks, conformer, metric prediction

1. Introduction

The CHiME-7 UDASE challenge was proposed for improving speech enhancement research using real training data in an unsupervised fashion. This paper comprises the system description for the University of Sheffield challenge submission. Rather than using an unsupervised methodology, the submission described here uses as supervised GAN based approach. The GAN discriminator is trained to learn MOS-related metrics, motivated by MOS being the main ranking metric of the challenge. Historical training data from a conventional generator and an additional pseudo-generator is used to augment the training data diversity.

2. System Description

The overall architecture of this submission is largely based on the CMGAN framework proposed in [1], but with two extensions proposed in [2] and [3]. The first extension is to train the discriminator $\mathcal{D}$ with a historical set of past generator outputs every epoch. The second extension is to train $\mathcal{D}$ to predict the metric score of noisy, clean and enhanced audio, as well as the output of a secondary pseudo-generator network $\mathcal{N}$ which is designed to increase the range of metric values observed by $\mathcal{D}$.

2.1. HuBERT Encoder Feature Representations

Recent works in metric prediction [4, 5] show that self-supervised speech representations (SSSRs) are useful as feature extractors for capturing quality of audio. Furthermore, recent work in speech enhancement [6] found that specifically the encoder outputs of SSSRs are particularly useful in this regard. As such, we make use of the feature encoder of the HuBERT [7] SSSR as a feature extractor for the metric prediction component of the proposed framework. We purely use this representation as a feature extractor, and do not update the parameters in the training of the metric prediction network.

2.2. Conformer-Based MetricGAN+ With Data Augmentation Generator (CMGAN+/+)

2.2.1. Conformer-Based Generator Description & Training

The Conformer model generator is based on the best performing CMGAN configuration in [1]. The network itself combines mapping and masking approaches to spectral speech enhancement, utilizing a conformer [8] based bottleneck. The model uses short-time Fourier transforms (STFTs) with a reasonably high temporal resolution (hop size of 6 ms) with a 50% overlap, and a fast Fourier transform (FFT) size of 400. The model is trained with a multi term loss function:

$$L_{\mathcal{G}} = \gamma_1 L_{\mathcal{G}_{\mathrm{GAN}}} + \gamma_2 L_{\mathcal{G}_{\mathrm{Time}}} + \gamma_3 L_{\mathcal{G}_{\mathrm{TF}}} \tag{1}$$

where $\gamma_1, \gamma_2, \gamma_3$ are hyperparameter weights. $L_{\mathcal{G}_{\mathrm{GAN}}}$ is defined as:

$$L_{\mathcal{G}_{\mathrm{GAN}}} = \mathbb{E}\{(\mathcal{D}(\hat{\mathbf{S}}_{\mathrm{HuBERTenc}}) - 1)^2\}, \tag{2}$$

which represents an assessment of the enhanced signal by the metric Discriminator $\mathcal{D}$. $L_{\mathcal{G}_{\mathrm{Time}}}$ is a mean absolute error between the enhanced and clean time domain mixtures:

$$L_{\mathcal{G}_{\mathrm{Time}}} = \mathbb{E}\{||s - \hat{s}||_1\}. \tag{3}$$

Finally $L_{\mathcal{G}_{\mathrm{TF}}}$ itself consists of two weighted components:

$$L_{\mathcal{G}_{TF}} = \alpha L_{\mathcal{G}_{\mathrm{Mag}}} + (1-\alpha) L_{\mathcal{G}_{\mathrm{RI}}}. \tag{4}$$

where α is a hyperparameter weight between the terms. $L_{\mathcal{G}_{\mathrm{Mag}}}$ represents the distance between magnitude spectrogram representations of the enhanced and clean mixtures:

$$L_{\mathcal{G}_{\mathrm{Mag}}} = \mathbb{E}\{||\mathbf{S}_{\mathrm{Mag}} - \hat{\mathbf{S}}_{\mathrm{Mag}}||^2\}, \tag{5}$$

while $L_{\mathcal{G}_{\mathrm{RI}}}$ represents a a similar comparison between the enhanced and clean real and imaginary STFT components:

$$L_{\mathcal{G}_{\mathrm{RI}}} = \mathbb{E}\{||\mathbf{S}_{\mathrm{Re}} - \hat{\mathbf{S}}_{\mathrm{Re}}||^2\} + \mathbb{E}\{||\mathbf{S}_{\mathrm{Im}} - \hat{\mathbf{S}}_{\mathrm{Im}}||^2\}. \tag{6}$$

Following the configuration in the original CMGAN system, $\gamma_1, \gamma_2, \gamma_3$ are set to $1, 0.2$ and 0.05 respectively, while α in (4) is set to 0.9. We also experiment with completely disabling the GAN component of the framework, i.e setting γ_3 to 0.

Due to the quadratic time-complexity of the transformer layers in the Conformer models, processing long sequences can be unfeasible due to high memory requirements. Transformers are also typically unsuitable for continuous processing as the entire sequence is required to compute self-attention. To address these issues input signals are processed in overlapping blocks of 4 s for evaluation and inference as this has been shown to be in an optimal signal length range for attention based enhancement models [9]. A 50% overlap with a Hann window is used to cross-fade each block with one another. Models are trained with 4 s signal length limits [9] similar to the baseline.

2.2.2. DNSMOS Metric Estimation Discriminator

The discriminator part of the GAN structure is trained to predict a normalised DNSMOS [10] score for a given input signal. The 'SIG' component of the DNSMOS model is used. The training procedure of $\mathcal{D}$ uses historical training data as was first proposed in the MetricGAN+ framework [2]. Within each epoch, first the Discriminator $\mathcal{D}$ is trained on the current training elements:

$$\begin{aligned} L_{\mathcal{D},\mathrm{MG+}} = \mathbb{E}\{&(\mathcal{D}(\mathbf{S}_{\mathrm{HuBERTenc}}) - Q'(s))^2 \\ &+ (\mathcal{D}(\hat{\mathbf{S}}_{\mathrm{HuBERTenc}}) - Q'(\hat{s}))^2 \\ &+ (\mathcal{D}(\mathbf{X}_{\mathrm{HuBERTenc}}) - Q'(x))^2 \\ &+ \mathcal{D}(\mathbf{Y}_{\mathrm{HuBERTenc}}) - Q'(y))^2\} \end{aligned} \tag{7}$$

where $\mathbf{S}_{\mathrm{HuBERTenc}}$, $\mathbf{X}_{\mathrm{HuBERTenc}}$, $\hat{\mathbf{S}}_{\mathrm{HuBERTenc}}$ and $\mathbf{Y}_{\mathrm{HuBERTenc}}$ are HuBERT encoder representations [6] of the clean audio mixture s, the noisy mixture x, the mixture as enhanced by $\mathcal{G}$, $\hat{s}$, and the mixture as enhanced by $\mathcal{N}$, y. This is followed by a historical training stage, where $\mathcal{D}$ is trained to predict the metric scores from past outputs of the generative networks $\mathcal{G}$ and $\mathcal{N}$. $Q'(\cdot)$ is the true DNSMOS SIG score of the input audio, normalized between 0 and 1. The discriminator network structure consists of a BLSTM followed by a single attention feed-forward layer with a sigmoid activation, similar to the network proposed in [5].

2.2.3. Metric Data Augmentation Pseudo-Generator

As first proposed in [3], an additional speech enhancement network $\mathcal{N}$ is trained, and its outputs y used to train the metric prediction discriminator $\mathcal{D}$. This model is trained solely using the GAN loss in (2), similar to the original MetricGAN framework. Its network structure is also based on the original MetricGAN enhancement model, consisting of a BLSTM which operates over a magnitude spectrogram representation of the input, followed by 3 linear layers. Its output is a magnitude mask which is multiplied by the input noisy spectrogram to produce an enhanced spectrogram.

3. Experiment Setup

The framework is trained on the simulated LibriMix dataset [11], using the same data loading configuration as the teacher network in the baseline system. The framework is trained for 200 epochs, on a random sample of 100 training elements from the train set each epoch. The Adam optimizer is used for all three networks, with learning rates of $0.005, 0.005$ and 0.001 for the $\mathcal{G}, \mathcal{N}$ and $\mathcal{D}$ respectively. At evaluation time, the best performing epoch in terms of DNSMOS SIG-MOS on the LibriMix validation set is loaded. Note that we only make use of the labeled portion of the challenge training data, unlike the baseline system. We additionally report the results of the best performing epoch after further fine-tuning for 20 epochs on the LibriCHiME dev set.

4. Results

Model	SI-SDR (dB)
unprocessed	*6.59*
Sudo rm -rf	7.80
RemixIT	9.44
RemixIT w/ VAD	**10.05**
CMGAN+/+	7.81
CMGAN+/+ fine-tuned	4.73
CMGAN+/+ (no GAN term)	6.61

Table 1: *SI-SDR Results on the reverberant LibriCHiME eval set*

Table 1 show the performance of the baseline and proposed system on the challenge reverberant LibriCHiME evaluation set in terms of SI-SDR. Our proposed system performs similarly to the baseline supervised teacher system on this evaluation set.

Table 2 shows the results of both the baseline and proposed systems on the CHiME 5 eval set in terms of DNSMOS. Our proposed system significantly outperforms all baseline systems in this measure, with a slight improvement in OVR and SIG measures by our fine-tuned model. Interestingly, there is a significant difference between our proposed framework which incorporates the GAN metric prediction loss component versus the version which did not, suggesting that this loss term is particularly useful.

Model	OVR	BAK	SIG
unprocessed	*2.84*	*2.92*	*3.48*
Sudo rm -rf	2.88	3.59	3.33
RemixIT	2.82	3.64	3.26
RemixIT w/ VAD	2.84	3.62	3.28
CMGAN+/+	3.40	**3.97**	3.76
CMGAN+/+ fine-tuned	**3.55**	3.93	**3.92**
CMGAN+/+ (no GAN term)	2.88	3.47	3.40

Table 2: *DNSMOS results on CHiME5 eval set*

Model	SI-SDR (dB)
unprocessed	*4.91*
Sudo rm -rf	**13.23**
RemixIT	11.47
RemixIT w/ VAD	12.15
CMGAN+/+	9.49
CMGAN+/+ fine-tuned	5.27
CMGAN+/+ (no GAN term)	11.11

Table 3: *SI-SDR Results on LibriMix eval set*

5. Conclusions

In this paper the University of Sheffield's CMGAN+/+ speech enhancement system for the CHiME-7 UDASE challenge is described. The system uses a GAN based model with data augmentation strategies to improve generalisation. Results on the unlabelled CHiME-5 evaluation set demonstrate significant improvements in DNSMOS evaluation metrics, far outperforming the baseline system in OVR, BAK and SIG measures. Other metrics were reported for labelled datasets showing comparable results with the baseline system in SI-SDR, a measure which the proposed model had not been trained explicitly to optimise for. We submit results from our base and fine-tuned systems as our entries 1 and 2 respectively.

6. References

[1] R. Cao, S. Abdulatif, and B. Yang, “CMGAN: Conformer-based Metric GAN for Speech Enhancement,” in *Proc. Interspeech 2022*, 2022, pp. 936–940.

[2] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, “MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement,” in *Proc. Interspeech 2021*, 2021, pp. 201–205.

[3] G. Close, T. Hain, and S. Goetze, “MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data,” in *EUSIPCO 2022*, Belgrade, Serbia, Aug. 2022.

[4] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Interspeech 2021*. ISCA, aug 2021.

[5] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, “Generalization ability of mos prediction networks,” 2021. [Online]. Available: https://arxiv.org/abs/2110.02635

[6] G. Close, W. Ravenscroft, T. Hain, and S. Goetze, “Perceive and predict: self-supervised speech representation based loss functions for speech enhancement,” in *Proc. ICASSP 2023*, 2023.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[8] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” 2020.

[9] W. Ravenscroft, S. Goetze, and T. Hain, “On data sampling strategies for training neural network speech separation models,” in *EUSIPCO 2023*, Sep 2023.

[10] C. K. A. Reddy, V. Gopal, and R. Cutler, “Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” 2022.

[11] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, “Librimix: An open-source dataset for generalizable speech separation,” 2020.