
Proceedings Paper:
Close, G.L., Ravenscroft, W., Hain, T. orcid.org/0000-0003-0939-3464 et al. (1 more author) (2023) The University of Sheffield CHiME-7 UDASE challenge speech enhancement system. In: Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023). 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023), 25 Aug 2023, Dublin, Ireland. International Speech Communication Association (ISCA) , pp. 33-38.

https://doi.org/10.21437/chime.2023-7

7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023)
25 August 2023, Dublin, Ireland

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 domain adaptation to unlabelled speech data. This paper describes the University of Sheffield team's system submitted to the challenge. A generative adversarial network (GAN) methodology based on a conformer-based metric GAN (CMGAN) is employed as opposed to the unsupervised RemixIT strategy used in the CHiME-7 baseline system. The discriminator of the GAN is trained to predict the output score of a Deep Noise Suppression Mean Opinion Score (DNSMOS) metric. Additional data augmentation strategies are employed which provide the discriminator with historical training data outputs as well as more diverse training examples from an additional pseudo-generator. The proposed approach, denoted as CMGAN+/+, achieves significant improvement in DNSMOS evaluation metrics with the best proposed system achieving 3.51 OVR-MOS, a 24% improvement over the baseline.

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

1. Introduction

As work and lifestyle patterns shift towards more remote, online working, it is essential that voice and video communication software is able to reduce environmental distortion in transmitted audio. As such, speech enhancement techniques, especially those utilising neural networks are a high priority area of active research. The CHiME-7 unsupervised domain adaptation speech enhancement (UDASE) challenge [1] was proposed to improve speech enhancement research [2–5] using real-world training data in an unsupervised way. In supervised neural network based speech enhancement systems, there is often a mismatch between the synthetic data used to train the system and real-world recordings. This can lead to poor performance of such systems *in the wild* even if evaluation metrics on synthetic data are high [6]. To further compound this problem, metrics which are designed to measure the quality often do not correlate strongly with actual human assessment of speech audio in specific scenarios [7], and often require access to clean reference/label audio which may not be readily available for real-life recordings.

Recently, several new metrics [8–11] have been proposed which attempt to directly predict human quality assessment in a non-intrusive way, i.e. without need for a reference signal. These take the form of neural networks trained using datasets of distorted audio to predict a quality label assigned to the audio by the human assessors. Self-supervised speech representations (SSSRs) have been found to be useful feature representations for the prediction of audio quality [12, 13].

The baseline system for the CHiME-7 UDASE challenge addresses the speech enhancment task using a RemixIT [14] framework wherein a teacher network is trained using labelled data, and a student network trained on real data uses inference of the teacher network as pseudo-labels in its loss function. The speech enhancement problem is modelled as source separation task, using the 'Sudo rm -rf' [15] model structure. While both, student and teacher networks show good performance on synthetic labelled testsets in terms of Scale Invariant Signal-Distortion Ratio (SI-SDR), degradation in quality in terms of the DNSMOS non-intrusive quality metric is observed on the challenge evaluation sets, compared to the unprocessed input audio in real-world in-domain recordings [1].

This paper comprises a description and in-depth evaluation of the University of Sheffield UDASE challenge submission. Rather than using an unsupervised methodology, the proposed approach for this submission uses a supervised GAN-based methodology. Motivated by Mean Opinion Scores (MOSs) being the main ranking metrics of the challenge, the GAN discriminator is trained to predict a MOS-related metric, i.e. DNSMOS. Historical training data from a conventional generator and an additional pseudo-generator is used to augment the training data diversity.

The remainder of this paper is structured as follows. The input feature generation by the Hidden Unit BERT (HuBERT) [16] SSSR model as well as the DNSMOS [8] metric prediction network are described in Section 2 and Section 3, respectively. The proposed CMGAN+/+ model is described in Section 4. Experimental setup and results are discussed in Section 5 and Section 6, respectively. Finally, Section 7 draws some conclusions from the findings of the paper.

2. HuBERT Encoder Feature Representations

Recent work in metric prediction [12, 13] shows that SSSRs are useful as feature extractors for capturing quality-related information about speech audio. As such, the proposed system makes use of the HuBERT [16] SSSR as a feature extractor for the metric prediction component of the proposed framework. HuBERT, like most SSSRs which take time domain signals as input, consists of two distinct network stages, as shown in Figure 1. The first stage $\mathcal{H}_{\mathrm{FE}}(\cdot)$ comprises several 1D convolutional layers which map the input time domain audio $s[n]$ into a 2D representation. The second stage $\mathcal{H}_{\mathrm{OL}}(\cdot)$ consists of a number of transformer [17] layers, which takes the output of the first stage as input. For a input time domain signal $\mathbf{s}[n]$, two representations $\mathbf{S}_{\mathrm{FE}}$ (after the feature encoder (FE) stage) and $\mathbf{S}_{\mathrm{OL}}$ (at the final output (OL) layer) can thus be obtained from the

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. This work was also supported by TOSHIBA Cambridge Research Laboratory and 3M Health Information Systems Inc.

10.21437/CHiME.2023-7

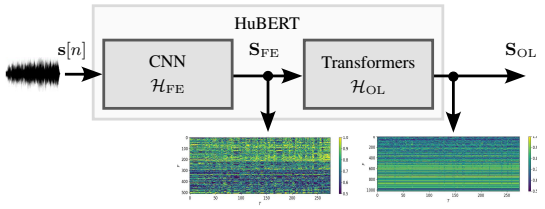

Figure 1: *Representations extracted from HuBERT model stages.*

HuBERT model:

$$\mathbf{S}_{\text{FE}} = \mathcal{H}_{\text{FE}}(s[n]) \quad (1)$$

$$\mathbf{S}_{\text{OL}} = \mathcal{H}_{\text{OL}}(\mathcal{H}_{\text{FE}}(s[n])) \quad (2)$$

Recent work in speech enhancement [13,18,19] have found that the outputs of HuBERT's $\mathcal{H}_{\text{FE}}(\cdot)$ stage are particularly useful for capturing quality-related information. The outputs of $\mathcal{H}_{\text{FE}}(\cdot)$ are 2D representations with dimensions $512 \times T$ where T depends on the length of the input audio in seconds. The HuBERT model used in this work is trained on 960 hours of audio-book recordings from the LibriSpeech [20] dataset, and is sourced from the FairSeq GitHub repository[1]. The HuBERT encoder representation $\mathbf{S}_{\text{FE}}$ in (1) is used as a feature extractor, and its parameters are not updated during in the proposed framework.

3. DNSMOS

The Deep Noise Suppression Mean Opinion Score (DNSMOS) [8] is a non-intrusive speech quality metric. It consists of a neural network which was trained to predict real human MOS ratings of input audio signals. As it is non-intrusive, it is particularly useful for assessing the quality of real recordings such as in the CHiME-7 UDASE challenge testset, and was one of the evaluation metrics used in assessing the entries to the challenge. For an input time domain speech signal $s[n]$ DNSMOS returns three values

$$Q_{\text{SIG}}, Q_{\text{BAK}}, Q_{\text{OVR}} = \text{DNSMOS}(s[n]), \quad (3)$$

where Q_{SIG}, Q_{BAK} and Q_{OVR} are each values between 1 and 5 which represent the estimated speech quality, background noise quality and overall quality, respectively (higher values indicating better quality). In the following Q is used to represent one of these values and Q' is the respective value normalized between 0 and 1.

While DNSMOS is a neural network meaning it is theoretically possible to backpropagate through it and use it directly in a loss function, it is not publicly available in this form. In order to incorporate DNSMOS as a loss function for speech enhancement in this work, a non-intrusive metric prediction discriminator [21] is trained to create a differentiable *copy* of the original implementation of DNSMOS provided in the CHIME-7 baseline system. This has the added benefit of allowing for an adversarial training of the metric prediction network in a GAN setting [22].

4. Speech Enhancement System Description

The overall architecture of the proposed system is largely based on the CMGAN framework proposed in [23], but with two extensions proposed in [24] and [25]. The first extension is to train the discriminator $\mathcal{D}$ on 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 $\mathcal{D}$. This work introduces a new structure for $\mathcal{D}$, as well as a new input feature which is derived from a pre-trained SSSR.

4.1. Conformer-based Generator

4.1.1. Conformer-based Generator Network Structure

The Conformer model generator $\mathcal{G}$ is based on the best performing CMGAN configuration in [23]. The network itself combines mapping and masking approaches for spectral speech enhancement, utilizing a conformer [26] based bottleneck. The model's input are short-time Fourier transform (STFT) components of the noisy audio $\mathbf{X}_{\text{Re}}$ and $\mathbf{X}_{\text{Im}}$ with a reasonably high temporal resolution (hop size of 6 ms) with a 50% overlap, and a fast Fourier transform (FFT) length of 400 samples at a sampling rate of $f_s = 16000$ Hz. The output of the model are the enhanced real and imaginary STFT components $\hat{\mathbf{S}}_{\text{Re}}$ and $\hat{\mathbf{S}}_{\text{Im}}$ from which the enhanced time domain audio $\hat{s}[n]$ is obtained by inverse short-time Fourier transform (ISTFT).

4.1.2. Generator Loss Function

The model is trained with a multi-term loss function

$$L_{\mathcal{G}} = \gamma_1 L_{\mathcal{G}_{\text{GAN}}} + \gamma_2 L_{\mathcal{G}_{\text{Time}}} + \gamma_3 L_{\mathcal{G}_{\text{TF}}}, \quad (4)$$

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

$$L_{\mathcal{G}_{\text{GAN}}} = \mathbb{E}\{(\mathcal{D}(\hat{\mathbf{S}}_{\text{FE}}) - 1)^2\}, \quad (5)$$

which represents an assessment of the enhanced signal by the metric Discriminator $\mathcal{D}$. The 1 in (5) represents the highest possible DNSMOS value of 5 after being normalized between 0 and 1.

$L_{\mathcal{G}_{\text{Time}}}$ is a mean absolute error between the enhanced and clean time domain mixtures

$$L_{\mathcal{G}_{\text{Time}}} = \mathbb{E}\{||s - \hat{s}||_1\}. \quad (6)$$

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

$$L_{\mathcal{G}_{TF}} = \alpha L_{\mathcal{G}_{\text{Mag}}} + (1 - \alpha) L_{\mathcal{G}_{\text{RI}}}, \quad (7)$$

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

$$L_{\mathcal{G}_{\text{Mag}}} = \mathbb{E}\{||\mathbf{S}_{\text{Mag}} - \hat{\mathbf{S}}_{\text{Mag}}||^2\}, \quad (8)$$

with $\hat{\mathbf{S}}_{\text{Mag}}$ defined as

$$\hat{\mathbf{S}}_{\text{Mag}} = \sqrt{\hat{\mathbf{S}}_{\text{Re}}^2 + \hat{\mathbf{S}}_{\text{Im}}^2}, \quad (9)$$

and $\mathbf{S}_{\text{Mag}}$ defined accordingly. $L_{\mathcal{G}_{\text{RI}}}$ represents a similar comparison between the enhanced and clean real and imaginary STFT components.

$$L_{\mathcal{G}_{\text{RI}}} = \mathbb{E}\{||\mathbf{S}_{\text{Re}} - \hat{\mathbf{S}}_{\text{Re}}||^2\} + \mathbb{E}\{||\mathbf{S}_{\text{Im}} - \hat{\mathbf{S}}_{\text{Im}}||^2\} \quad (10)$$

With the exception of (5), all terms of $L_{\mathcal{G}}$ require access to clean label/reference audio $s[n]$. The feature transformations and loss terms of $L_{\mathcal{G}}$ are visualised in Figure 2.

[1] https://github.com/facebookresearch/fairseq

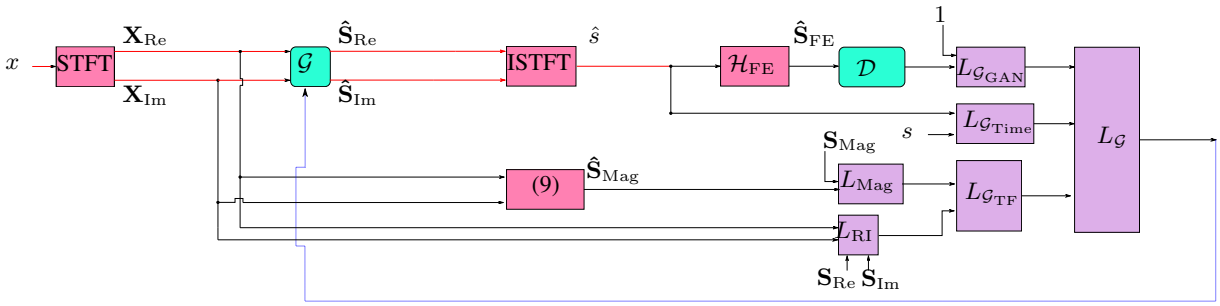

Figure 2: *Visualisation training of generator $\mathcal{G}$ and generator loss $L_{\mathcal{G}}$ in (4), (inference path shown in red, backpropagation in blue).*

4.1.3. Block Processing for Continuous Processing

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 [27]. 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 [27] similar to the baseline.

4.2. Metric Estimation Discriminator

The discriminator $\mathcal{D}$ part of the GAN structure is trained to predict a normalised DNSMOS [8] score for a given input signal. Inference of $\mathcal{D}$ is used in (5) as one of the loss terms of $\mathcal{G}$ and as the sole loss function of $\mathcal{N}$ in (12), enforcing an optimisation towards the target metric.
Experiments with training $\mathcal{D}$ to predict one of the outputs of DNSMOS (i.e $Q_{\mathrm{SIG}}, Q_{\mathrm{BAK}}$ or Q_{OVR}) are also conducted.

4.2.1. Discriminator Network Stucture

The discriminator network structure consists of 2 Bi-Directional Long Short-Term Memory (BLSTM) layers followed by a single attention feed-forward layer with a sigmoid activation, similar to the network proposed in [12]. The input to $\mathcal{D}$ is the output of the HuBERT feature encoder $\mathcal{H}_{\mathrm{FE}}(\cdot)$.

4.2.2. Discriminator Loss Function

Within each epoch, first the Discriminator $\mathcal{D}$ is trained on the current training elements:

$$\begin{aligned} L_{\mathcal{D},\mathrm{MG+}} = \mathbb{E}\{&(\mathcal{D}(\mathbf{S}_{\mathrm{FE}}) - Q'(s))^2 \\ &+ (\mathcal{D}(\hat{\mathbf{S}}_{\mathrm{FE}}) - Q'(\hat{s}))^2 \\ &+ (\mathcal{D}(\mathbf{X}_{\mathrm{FE}}) - Q'(x))^2 \\ &+ \mathcal{D}(\mathbf{Y}_{\mathrm{FE}}) - Q'(y))^2\} \end{aligned} \tag{11}$$

where $\mathbf{S}_{\mathrm{FE}}$, $\mathbf{X}_{\mathrm{FE}}$, $\hat{\mathbf{S}}_{\mathrm{FE}}$ and $\mathbf{Y}_{\mathrm{FE}}$ are HuBERT encoder representations, u.e. outputs of $\mathcal{H}_{\mathrm{FE}}(\cdot)$, 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 score of the input audio, normalized between 0 and 1.

4.2.3. Historical Training

The training procedure of $\mathcal{D}$ uses historical training data as it was first proposed in the MetricGAN+ framework [24]. In this stage, a sample of enhanced audio output from past epochs of $\mathcal{G}$ and $\mathcal{N}$ are used to train $\mathcal{D}$. The aim of this is to prevent $\mathcal{D}$ from '*forgetting* how to assess audio which is dissimilar to the current outputs of the enhancement network. In each epoch, $\mathcal{D}$ is trained using a randomly selected 10% of the outputs of the generator models from past epochs.

4.3. Metric Data Augmentation Pseudo-Generator

As first proposed in [25], an additional speech enhancement network $\mathcal{N}$ is trained, and its outputs y used to train the metric prediction discriminator $\mathcal{D}$ (last term in (11)) . This model is trained solely using the GAN loss in (5), similar to the original MetricGAN framework,

$$L_{\mathcal{N}_{\mathrm{GAN}}} = \mathbb{E}\{(\mathcal{D}(\mathbf{Y}_{\mathrm{FE}}) - w)^2\}, \tag{12}$$

where w is a hyperparameter value which corresponds to the target normalised DNSMOS score for which the output audio of $\mathcal{N}$ is being trained to reach.
Its network structure is based on the original MetricGAN enhancement model, consisting of a BLSTM which operates on 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 $\mathbf{Y}_{\mathrm{SPEC}}$. A time domain signal $y[n]$ is constructed by the overlap-add method using the original noisy phase.

5. Experiment Setup

The framework is trained on the simulated LibriMix dataset [28], using the same data loading configuration as the teacher network in the baseline system [1]. The labelled LibriMix training set consists of 33900 clean/noisy audio pairs, with the clean speech sourced from the LibriSpeech [20] dataset and the added noise from WHAM! [29] dataset. The framework is trained for 200 epochs, on a random sample of 100 training elements from the train set in each epoch. The Adam optimizer is used for all three networks, with learning rates of $0.005, 0.005$ and 0.001 for $\mathcal{G}, \mathcal{N}$ and $\mathcal{D}$ respectively. Frameworks are trained where $\mathcal{D}$ is trained to predict target metric $Q_{\mathrm{SIG}}, Q_{\mathrm{BAK}}$ and Q_{OVR}.
Following the configuration in the original CMGAN system, $\gamma_1, \gamma_2, \gamma_3$ in (4) are set to $1, 0.2$ and 0.05 respectively, while α in (7) is set to 0.9. An additional simulation completely disabling the GAN component of the framework, i.e. setting γ_3 to 0, as well as training solely using the GAN loss by setting γ_1 and γ_2 to 0 and γ_3 to 1 are performed. Additionally, we experiment with setting w, the hyperparameter which controls the

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

Model	w	Q	**SI-SDR (dB)**
unprocessed	–	–	*6.59*
Sudo rm -rf [15]	–	–	7.8
RemixIT [14]	–	–	9.44
RemixIT [14] w/ VAD	–	–	**10.05**
CMGAN+/+	1.00	SIG	4.71
fine-tuned			3.55
CMGAN+/+	0.80	SIG	4.53
fine-tuned			3.55
CMGAN+/+	0.45	SIG	5.98
fine-tuned			4.30
CMGAN+/+	1.00	BAK	6.95
fine-tuned			6.89
CMGAN+/+	0.80	BAK	6.31
fine-tuned			7.39
CMGAN+/+	0.45	BAK	6.42
fine-tuned			5.84
CMGAN+/+	1.00	OVR	7.41
fine-tuned			4.29
CMGAN+/+	0.80	OVR	1.19
fine-tuned			5.15
CMGAN+/+	0.45	OVR	4.75
fine-tuned			6.78
no GAN term	–	–	6.61
GAN only	1.00	SIG	-30.97
GAN only	1.00	BAK	-67.28
GAN only	1.00	OVR	-41.60

objective of $\mathcal{N}$ in (12), to 1.0, 0.8 and 0.45.

At evaluation time, the best-performing epoch in terms of the target metric on the LibriMix validation set is loaded. Note that only the labelled portion of the challenge training data is used in taining, unlike the baseline system. Additionally,results are reported for the best-performing epoch after further fine-tuning for 20 epochs on the labelled LibriCHiME dev set which consists is similar to LibriMix but with the noise sourced from the real CHiME recordings.

6. Results

Table 1 shows the results of the baseline systems and the proposed systems (for different w in (12) and different target metrics Q from (3)) on the simulated Reverberant LibriCHiME evaluation set in terms of Scale Invariant Signal-Distortion Ratio (SI-SDR) score. Here, the proposed system shows generally lower performance than the baselines, with the exception of the models which are trained with Q_{BAK} as their target metric. The model trained with a w value of 0.8 with Q_{BAK} as the objective when fine-tuned in the LibriCHiME dev set was able to achieve an average SI-SDR score of 7.41 dB. Similarly, the model trained with a w value of 1 and Q_{OVR} achieves an average SI-SDR score of 7.41 dB. The relatively poor overall performance by the proposed systems in terms of SI-SDR as evaluation metric can perhaps be explained by the fact that the baseline systems all explicitly use SI-SDR as a loss function during training; our system which incorporates SI-SDR loss directly outperforms the baseline in this measure as shown in the following.

Table 2 show results of the baseline systems and the proposed systems on the real CHiME evaluation set in terms of DNSMOS scores. Here, the proposed systems all show

Table 2: *DNSMOS results on CHiME5 eval set.*

Model	w	Q	**OVR**	**BAK**	**SIG**
unprocessed	-		*2.84*	*2.92*	*3.48*
Sudo rm -rf [15]	-		2.88	3.59	3.33
RemixIT [14]	-		2.82	3.64	3.26
RemixIT [14] w/ VAD	-		2.84	3.62	3.28
CMGAN+/+	1.00	SIG	3.29	3.85	3.76
fine-tuned			3.45	3.90	3.98
CMGAN+/+	0.80	SIG	3.20	3.70	3.68
fine-tuned			3.37	3.46	3.86
CMGAN+/+	0.45	SIG	3.33	3.81	3.80
fine-tuned			3.49	3.90	**3.98**
CMGAN+/+	1.00	BAK	3.12	3.90	3.39
fine-tuned			3.28	**4.08**	3.29
CMGAN+/+	0.80	BAK	3.06	3.82	3.32
fine-tuned			3.15	3.95	3.07
CMGAN+/+	0.45	BAK	2.87	3.74	3.18
fine-tuned			3.08	3.87	3.23
CMGAN+/+	1.00	OVR	**3.51**	3.99	3.78
fine-tuned			2.60	3.25	3.14
CMGAN+/+	0.80	OVR	3.37	3.87	3.56
fine-tuned			2.75	3.27	3.27
CMGAN+/+	0.45	OVR	3.23	3.94	3.33
fine-tuned			2.84	3.24	3.26
no GAN term	–	–	2.87	3.54	3.34
GAN only	1.00	SIG	2.66	1.58	3.72
GAN only	1.00	BAK	2.67	3.78	2.41
GAN only	1.00	OVR	2.70	3.68	3.00

a marked improvement over the baseline systems, with an improvement in terms of the target metric after fine-tuning in most cases. Furthermore, the inclusion of the GAN term in (4) also has a significant effect on this measure, as shown by the performance of the proposed system without the GAN term. Unlike Q_{SIG} and Q_{BAK} fine-tuning on the LibriCHiME dev set degrades performance on the models trained towards Q_{OVR}. Generally, the models trained with a w value of 1 perform better than the other values; this may be caused by the difficulty of the task of $\mathcal{N}$ to enhance or 'de-enhance' the input audio representation.

The results for the model trained solely using the GAN term towards Q_{SIG} are shown in the last row of Table 2. While this model shows good performance on its target metric, it scores rather poorly on the other two DNSMOS components. Furthermore, when played back, audio enhanced by this system is *significantly* distorted, with barely any of the original signal retained. The models trained only using the GAN term towards Q_{BAK} and Q_{SIG} are similarly distorted. Figure 3 shows exemplarily shows spectrograms for noisy (upper panel in Figure 3) and enhanced audio by the system with Q_{SIG} as target metric and a w of 1 (second panel), the system with no GAN term (3rd panel) and the system using the GAN term only (also with Q_{SIG}, w of 1, lower panel in Figure 3). In the lower panel of Figure 3, the significant distortion of the signal by the GAN-only model is visible, despite it achieving a similar DNSMOS SIG improvement relative to the noisy input as the other enhancement models. This suggests that the model has learned to 'enhance' the input audio in a way to trick the DNSMOS SIG metric into awarding it high scores. The reason as to why DNSMOS awards such high scores to significantly distorted audio remains unknown; it is possible that as DNSMOS is a data-driven system itself, the problem arises from its neural network not ever observing audio which has been distorted in such a way during its own training, resulting in it assigning an effectively meaningless score.

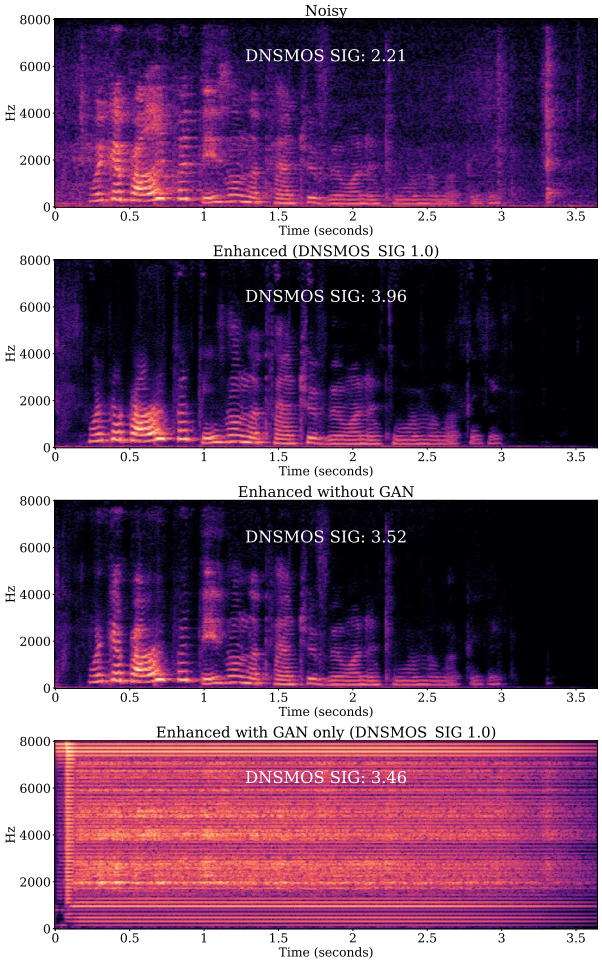

Figure 3: *Noisy and enhanced spectrograms of audio file* `S01_P01_0.wav` *from the CHiME-5 evaluation set.*

6.1. Challenge Results

Table 3 compares the challenge entries in terms of DNSMOS and SI-SDR on the sim challenge evaluation sets.

Table 3: *Comparison with other challenge entries ranked by DNSMOS OVR score.*

		CHiME-5 (DNSMOS)			Reverb Libri-CHiME-5
Rank	**System**	**OVRL**	**BAK**	**SIG**	**SI-SDR (dB)**
1	CMGAN+/+ fine	**3.55**	3.93	**3.92**	4.7
2	CMGAN+/+	3.40	**3.97**	3.76	7.8
3	NWPU/ByteAudio	3.07	3.93	3.39	**13.0**
4	Sogang ISDS1	2.90	3.60	3.39	12.4
5	Sogang ISDS2	2.88	3.70	3.32	12.4
6	*OOD teacher*	*2.88*	*3.59*	*3.33*	*7.8*
7	*RemixIT-VAD*	*2.84*	*3.62*	*3.28*	*10.1*
8	*Unprocessed*	*2.84*	*2.92*	*3.48*	*6.6*
9	*RemixIT*	*2.82*	*3.64*	*3.26*	*9.4*

The submitted system uses DNSMOS SIG as its target metric with a w value of 1. Note that the results shown here for our submitted systems differ slightly from those in the previous section, as they come from different runs of the model on a different random seed. Both our base and fine-tuned models significantly outperform all other entries in terms of DNSMOS on the real CHiME-5 evaluation set, but show lower performance for SI-SDR as target metric. After evaluation by the challenge organisers in terms of DNSMOS and SI-SDR as shown in Table 3, the two best-performing systems for each of the two target metrics (including the proposed system) were evaluated in listening tests.Table 4 shows the results listening-tests of audio enhanced by the top-performing systems, as well as the unprocessed audio. Interestingly, the proposed system shows lower performance in the listening tests than expected from the high scores in terms of DNSMOS in Table 4.

Table 4: *Comparison of top-performing challenge entries on listening tests with human participants, ranked by OVRL MOS.*

		CHiME-5 (Listening Tests)		
Rank	**System**	**OVRL**	**BAK**	**SIG**
1	NWPU/ByteAudio	**3.11**	**4.30**	3.41
2	Sogang ISDS1	2.75	3.08	3.43
3	*Unprocessed*	*2.68*	*2.20*	***3.97***
4	*RemixIT-VAD*	*2.45*	*2.97*	*3.02*
5	CMGAN+/+ fine	2.14	2.75	2.63

7. 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 discriminator input data augmentation strategies to improve metric prediction performance. Results on the unlabelled CHiME-5 evaluation set demonstrate improvements in DNSMOS evaluation metrics, significantly outperforming the baseline system in OVR, BAK and SIG measures. However, this does not directly translate to high ratings in listening tests with humans. By training solely using a metric optimisation loss, possible flaws in the metric being optimised towards have to be considered.

8. References

[1] S. Leglaive, L. Borne, E. Tzinis, M. Sadeghi, M. Fraticelli, S. Wisdom, M. Pariente, D. Pressnitzer, and J. R. Hershey, "The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement," 2023.

[2] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Berlin, Heidelberg: Springer, 2010.

[3] T. Rohdenburg, S. Goetze, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Combined Source Tracking and Noise Reduction for Application in Hearing Aids," in *8th ITG Conference on Speech Communication*, Aachen, Germany, Oct. 2008.

[4] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 Deep Noise Suppression Challenge," in *Proc. ICASSP'22*, 2022, pp. 9271–9275.

[5] G. Close, T. Hain, and S. Goetze, "PAMGAN+/-: Improving Phase-Aware Speech Enhancement Performance via Expanded Discriminator Training," in *AES 154th Conv.*, Espoo, Helsinki, Finland, May 2023.

[6] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, 2015.

[7] A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. Falk, "Performance Comparison of Intrusive and Non-Intrusive Instrumental Quality Measures for Microphone-Array Processed Speech," in *Proc. IWAENC 2016*, 2016.

[8] 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.

[9] A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "TorchAudio-Squim: Reference-less Speech Quality and Intelligibility measures in TorchAudio," in *ICASSP'23*, 2023.

[10] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1151–1163, Jul. 2019.

[11] 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*, Aug. 2021.

[12] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP'22*, 2022.

[13] 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.

[14] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "RemixIT: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, oct 2022.

[15] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2020.

[16] 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," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, Oct. 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[18] G. Close, T. Hain, and S. Goetze, "The Effect of Spoken Language on Speech Enhancement using Self-Supervised Speech Representation Loss Functions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'23)*, 2023.

[19] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP 2015*, 2015, pp. 5206–5210.

[21] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech," 2021.

[22] S.-W. Fu, C.-F. Liao, Y. Tsao, and S. de Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 2019, pp. 2031–2041.

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

[24] 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.

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

[26] 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.

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

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

[29] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," 2019.