Proceedings Paper:
Close, G., Ravenscroft, W., Hain, T. et al. (2024) Multi-CMGAN+/+: leveraging multi-objective speech quality metric prediction for speech enhancement. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024), 14-19 Apr 2024, Seoul, Korea. Institute of Electrical and Electronics Engineers (IEEE), pp. 351-355. ISBN: 979-8-3503-4486-8. ISSN: 1520-6149. EISSN: 2379-190X.

https://doi.org/10.1109/ICASSP48485.2024.10448343

MULTI-CMGAN+/+: LEVERAGING MULTI-OBJECTIVE SPEECH QUALITY METRIC PREDICTION FOR SPEECH ENHANCEMENT

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

Speech and Hearing Group, The University of Sheffield, Sheffield, UK

ABSTRACT

Neural network based approaches to speech enhancement have shown to be particularly powerful, being able to leverage a data-driven approach to result in a significant performance gain versus other approaches. Such approaches are reliant on artificially created labelled training data such that the neural model can be trained using intrusive loss functions which compare the output of the model with clean reference speech. Performance of such systems when enhancing real-world audio often suffers relative to their performance on simulated test data. In this work, a non-intrusive multi-metric prediction approach is introduced, wherein a model trained on artificial labelled data using inference of an adversarially trained metric prediction neural network. The proposed approach shows improved performance versus state-of-the-art systems on the recent CHiME-7 challenge unsupervised domain adaptation speech enhancement (UDASE) task evaluation sets.

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

1. INTRODUCTION

For training of 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 intrusive evaluation metrics on synthetic data are high. A compounding factor in this problem is that metrics which are designed to measure speech quality do not always correlate strongly with actual human assessment of speech audio quality in many scenarios [1, 2], and often require access to clean reference/label audio which may not be readily available for real-life recordings.

Recently, several new metrics [3, 4, 5] have been proposed which attempt to directly predict human quality assessment in a non-intrusive way, i.e. where the clean speech reference is not required. These take the form of neural networks which are trained using vast 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 [6].

This paper comprises a system which builds on the authors' entry [7] to the CHIME-7 challenge UDASE [8] track. It attempts to address the problem of model adaption to real world data via a metric prediction generative adversarial network (GAN) based methodology. A non-intrusive GAN discriminator is trained to predict multiple metrics including a MOS-related metric, as well as a traditional intrusive signal quality metric. Historical training data from a conventional generator and an additional *pseudo-generator* is used to augment the training data diversity. Then, during the training of the speech enhancement generator, inference of the multi-metric prediction discriminator is used to optimise the enhanced outputs towards the target metrics. In this way, metrics which are unable to be directly used as loss functions as well those which require access to a reference signal can be optimised.

The remainder of this paper is structured as follows. The target metrics are described in Section 2. A description of the proposed Multi-CMGAN+/+ model is given in Section 3. Experimental setup and results are discussed in Section 4 and Section 5, respectively. Finally, Section 6 draws some conclusions from the findings of the paper.

2. SPEECH QUALITY METRICS

Two speech quality metrics, Perceptual Evaluation of Speech Quality (PESQ) and Deep Noise Suppression Mean Opinion Score (DNSMOS), are used as target metrics which the speech enhancement generator in our proposed system is trained to optimise towards.

2.1. PESQ

Perceptual Evaluation of Speech Quality (PESQ) [9] is a well-known intrusive speech quality measure. It takes a time domain signal of the clean reference audio $s[n]$ and the time-domain audio of the signal to be evaluated, e.g. the noisy signal $x[n]$, and returns a value Q_{PESQ} between 1 and 4.5 which represents the quality of the test signal, higher meaning better

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 funded in part by TOSHIBA Cambridge Research Laboratory and 3M Health Information Systems Inc.

quality:

$$Q_{\mathrm{PESQ}} = \mathrm{PESQ}(s[n], x[n]) \tag{1}$$

The formulation of PESQ is non-differentiable, so direct use of it as a loss function for training enhancement models is not possible.

2.2. DNSMOS

Deep Noise Suppression Mean Opinion Score (DNSMOS) [3] is a non-intrusive speech quality metric. It consists of a neural network which was trained to predict human Mean Opinion Score (MOS) ratings for speech signals. As it is non-intrusive, it is particularly useful for assessing the quality of real-world 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 a input time domain speech signal $s[n]$ DNSMOS estimates three values, being estimates of the well-known composite measure [10]:

$$[Q_{\mathrm{SIG}}, Q_{\mathrm{BAK}}, Q_{\mathrm{OVR}}] = \mathrm{DNSMOS}(s[n]), \tag{2}$$

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 this work the non-neural implementation of DNSMOS provided in the CHIME-7 baseline system is used.

2.3. Non-intrusive Metric Prediction

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. Similarly, the computation of PESQ is non-differentiable, and requires access to a reference signal, meaning it cannot be used for most real-world scenarios. In order to incorporate DNSMOS and PESQ in loss functions for speech enhancement in this work, a non-intrusive metric prediction discriminator [11] is trained to create differentiable 'clones' of the metrics. This has the added benefit of allowing for an adversarial training of the metric prediction network in a GAN setting [12]. In the following, Q is used to represent one of these target metrics in (1) and (2) and Q' is the respective value normalised between 0 and 1.

3. SPEECH ENHANCEMENT SYSTEM

The overall architecture of the proposed system is based on the conformer-based metric GAN (CMGAN) framework proposed in [13], but with two extensions based on [14] and [15]. 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 $\mathcal{D}$ allowing it to predict multiple metrics at once, as well as a new input feature which is derived from a pre-trained SSSR model.

3.1. Conformer-based Speech Enhancement Generator

3.1.1. Conformer-based Generator Network Structure

The conformer model generator $\mathcal{G}$ is based on the best performing CMGAN configuration in [13]. The network itself combines mapping and masking approaches for spectral speech enhancement, utilizing a conformer [16] based bottleneck. The model's input are short-time Fourier transform (STFT) components of the complex-valued noisy audio, $\mathbf{X}_{\mathrm{Re}}, \mathbf{X}_{\mathrm{Im}}$, 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). The output of the model are the enhanced real and imaginary STFT components $\hat{\mathbf{S}}_{\mathrm{Re}}$ and $\hat{\mathbf{S}}_{\mathrm{Im}}$ from which the enhanced time domain audio $\hat{s}[n]$ is obtained by inverse short-time Fourier transform (ISTFT). Note that the time index n is omitted for clarity in the following.

3.1.2. Generator Loss Function

The generator model $\mathcal{G}$ is trained with a multi-term loss function:

$$L_{\mathcal{G}} = L_{\mathcal{G}_{\mathrm{GAN}}} + L_{\mathcal{G}_{\mathrm{Time}}} + L_{\mathcal{G}_{\mathrm{SI-SDR}}} \tag{3}$$

$L_{\mathcal{G}_{\mathrm{GAN}}}$ minimises the distance

$$L_{\mathcal{G}_{\mathrm{GAN}}} = \mathbb{E}\left\{\|\mathcal{D}(\hat{\mathbf{S}}_{\mathrm{FE}}) - \mathbf{1}\|_2^2\right\}, \tag{4}$$

which represents an assessment of the enhanced signal by the metric Discriminator $\mathcal{D}$. $\mathcal{D}(\hat{\mathbf{S}}_{\mathrm{FE}})$ is the inference of the metric prediction discriminator $\mathcal{D}$, given the enhanced signal as input, which has an output of dimension $N_Q \times 1$ representing the N_Q predicted normalised Q' values of the target metrics, i.e. N_Q equals 3 when using (2). The $\mathbf{1}$ vector in (4), also of length N_Q, represents the highest possible target metric values normalized between 0 and 1. Thus, the net effect of this loss term is to encourage $\mathcal{G}$ to maximise the predicted scores assigned to its outputs by $\mathcal{D}$.
$L_{\mathcal{G}_{\mathrm{Time}}}$ is a mean absolute error between the enhanced and clean time domain mixtures:

$$L_{\mathcal{G}_{\mathrm{Time}}} = \mathbb{E}\left\{||s - \hat{s}||_1\right\}. \tag{5}$$

Finally, $L_{\mathcal{G}_{\mathrm{SISDR}}}$ is the scale invariant signal-distortion ratio (SI-SDR) [17] loss

$$L_{\mathcal{G}_{\mathrm{SISDR}}} = -10\log_{10}\frac{\left\|\frac{\langle\hat{s},s\rangle s}{\|s\|^2}\right\|^2}{\left\|\hat{s} - \frac{\langle\hat{s},s\rangle s}{\|s\|^2}\right\|^2}. \tag{6}$$

With the exception of (4), all terms of $L_{\mathcal{G}}$ require access to clean label/reference audio s.

3.1.3. Block Processing for Longer Inputs

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 4s for evaluation and inference as this has been shown to be in an optimal signal length for attention-based enhancement models [18]. A 50% overlap with a Hann window is used to cross-fade each block with one an another. Models are trained with 4s signal length limits [18].

3.2. Metric Estimation Discriminator

The discriminator $\mathcal{D}$ part of the GAN structure is trained to predict three normalised speech quality metrics for a given input signal. Inference of $\mathcal{D}$ is used in (4) as one of the loss terms of $\mathcal{G}$ and as the sole loss function of $\mathcal{N}$ in (10), enforcing an optimisation towards the target metrics.
We experiment with training $\mathcal{D}$ to predict each outputs of DNSMOS (i.e $Q_{\mathrm{SIG}}, Q_{\mathrm{BAK}}$ or Q_{OVR}), as well as PESQ (Q_{PESQ}).

3.3. HuBERT Encoder Feature Representations

Recent work in metric prediction [19, 6] 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 Hidden Unit BERT (HuBERT) [20] 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. 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 $\mathbf{S}_{\mathrm{FE}}$. The second stage, $\mathcal{H}_{\mathrm{OL}}(\cdot)$, consists of a number of transformer layers, which takes the output of the first stage $\mathbf{S}_{\mathrm{FE}}$ as input. The two representations $\mathbf{S}_{\mathrm{FE}}$ and $\mathbf{S}_{\mathrm{OL}}$ can thus be obtained from the HuBERT model:

$$\mathbf{S}_{\mathrm{FE}} = \mathcal{H}_{\mathrm{FE}}(s[n]) \quad (7)$$

$$\mathbf{S}_{\mathrm{OL}} = \mathcal{H}_{\mathrm{OL}}(\mathcal{H}_{\mathrm{FE}}(s[n])) \quad (8)$$

Recent work in speech enhancement [6, 21, 22] have found that the outputs of HuBERT's encoder stage $\mathcal{H}_{\mathrm{FE}}(\cdot)$ are particularly useful for capturing quality-related information, outperforming the final transformer layer and weighted sums of each transformer output. The outputs of $\mathcal{H}_{\mathrm{FE}}(\cdot)$ 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 [23] dataset, sourced from the FairSeq GitHub repo[1]. This HuBERT encoder representation is used as a feature extractor, and its parameters are not updated during the training of the metric prediction network.

3.3.1. Discriminator Network Stucture

The discriminator network structure consists of 2 bi-directional long short-term memory (BLSTM) layers followed by three parallel attention feed-forward layers with sigmoid activations, similar to the network proposed in [19]. Each attention feed-forward layer outputs a single neuron which represents the prediction value of one of the three target metrics. The input to $\mathcal{D}$ is the output of the HuBERT feature encoder $\mathcal{H}_{\mathrm{FE}}(\cdot)$. The output of $\mathcal{D}$ has dimension $B \times N_Q$ where B is the batch size and each of N_Q values represents a normalised predicted metric value. Note that inference of $\mathcal{D}$ is always non-intrusive, even when if one of it's target metrics such as PESQ is intrusive.

3.3.2. Discriminator Loss Function

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

$$\begin{aligned} L_{\mathcal{D},\mathrm{MG+}} = \mathbb{E}\{&(\mathcal{D}(\mathbf{S}_{\mathrm{FE}}) - [Q'_1(s), ..., Q'_{N_Q}(s)])^2 \\ &+ (\mathcal{D}(\hat{\mathbf{S}}_{\mathrm{FE}}) - [Q'_1(\hat{s}), ..., Q'_{N_Q}(\hat{s})])^2 \\ &+ (\mathcal{D}(\mathbf{X}_{\mathrm{FE}}) - [Q'_1(x), ..., Q'_{N_Q}(x)])^2 \\ &+ (\mathcal{D}(\mathbf{Y}_{\mathrm{FE}}) - [Q'_1(y), ..., Q'_{N_Q}(y)])^2\} \end{aligned} \quad (9)$$

where $\mathbf{S}_{\mathrm{FE}}$, $\mathbf{X}_{\mathrm{FE}}$, $\hat{\mathbf{S}}_{\mathrm{FE}}$ and $\mathbf{Y}_{\mathrm{FE}}$ are HuBERT encoder representations, i.e. after $\mathcal{H}_{\mathrm{FE}}(\cdot)$, of the clean signal s, the noisy signal x, the signal enhanced by $\mathcal{G}$, $\hat{s}$, and the signal as enhanced by $\mathcal{N}$, y. $Q'_1(\cdot), Q'_2(\cdot)$ and $Q'_3(\cdot)$ are the true target metric scores of the input audio, normalized between 0 and 1. Please note that the Q' vectors in (9) can be shorter than 3 if less than $N_Q = 3$ metrics are considered. 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 $\mathcal{N}$.

3.3.3. Historical Training

The training procedure of $\mathcal{D}$ uses historical training data as first proposed in the MetricGAN+ framework [14]. In this stage, a sample of enhanced audio output from past epochs of $\mathcal{G}$ and $\mathcal{N}$ are used to train $\mathcal{D}$. This aim of this is to widen 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.

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

3.4. Metric Data Augmentation Pseudo-Generator

As first proposed in [15], a secondary speech enhancement network $\mathcal{N}$ is trained, and its outputs y used to train the metric prediction discriminator $\mathcal{D}$ (last term in (9)) . This model is trained solely using the GAN loss in (4), similar to the original MetricGAN framework:

$$L_{\mathcal{N}_{\text{GAN}}} = \mathbb{E}\{\|\mathcal{D}(\mathbf{Y}_{\text{FE}}) - \mathbf{1}w\|_2^2\} \tag{10}$$

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 obtain. Following on from prior work [7], here we fix the value of w at 1 meaning that $\mathcal{N}$ is trained to enhance relative to the target metrics, rather than to 'de-enhance' with a lower value of w.

$\mathcal{N}$s 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 $\mathbf{Y}_{\text{SPEC}}$. A time domain signal $y[n]$ is constructed by the overlap-add method using the original noisy phase.

4. EXPERIMENTS

4.1. Training Setup

The framework is trained on simulated labelled data from the LibriMix [24] for 200 epochs, following a similar dataloading system as in [8] generating mixtures of a single speaker with noise. The labelled LibriMix training set consists of 33900 clean/noisy audio pairs, with the clean speech sourced from the LibriSpeech [23] dataset and the added noise from WHAM! [25] dataset.

Each epoch, 300 samples from the training set are randomly selected. These are first used to train the metric prediction Discriminator $\mathcal{D}$ using (9). This is followed by the training of $\mathcal{D}$ on the historical set. Then the 300 random samples are used to train $\mathcal{N}$ using inference of $\mathcal{D}$ with (10), followed finally by the training of $\mathcal{G}$ using (3) which also uses inference of $\mathcal{D}$.

Different combinations of the DNSMOS terms and PESQ are experimented with as the three target metrics for $\mathcal{D}$ by setting each of Q_1, Q_2, Q_3 in (9) to be $Q_{\text{PESQ}}, Q_{\text{SIG}}, Q_{\text{BAK}}$ or Q_{OVR}.

The proposed models are evaluated on the CHiME7 UDASE task [8] evaluation sets. These are a real world unlabelled set consisting of CHIME5 recordings which are evaluated using DNSMOS and a simulated labelled set consisting of reverberant LibriMix audio which are evaluated using SI-SDR. The proposed system is compared to our prior entry to the CHiME7 UDASE challenge [7], as well as the challenge baselines [8]. Source code will be available at [2].

[2] https://github.com/leto19/MultiMetricGANplusplus

5. RESULTS

Table 1 shows the results of the proposed framework in terms of DNSMOS on the CHiME-7 UDASE task real evaluation set. The proposed systems significantly outperform the baseline systems in all measures, while also outperforming the author's prior work CMGAN+/+ in terms of OVR and BAK. However, CMGAN+/+ still outperforms the proposed system in terms of SIG, which is the only metric it is optimized towards.

Table 1. DNSMOS results on CHiME5 eval set.

Model	Q_1, Q_2, Q_3	**OVR**	**BAK**	**SIG**
unprocessed	–	*2.84*	*2.92*	*3.48*
Sudo rm -rf [26]	–	2.88	3.59	3.33
RemixIT [27] w/VAD	–	2.84	3.62	3.28
CMGAN+/+ [7]	SIG	3.29	3.85	**3.76**
Multi-CMGAN+/+	SIG/BAK/OVR	**3.42**	3.86	3.56
Multi-CMGAN+/+	SIG/BAK/PESQ	3.08	3.78	3.41
Multi-CMGAN+/+	SIG/OVR/PESQ	2.80	3.62	3.19
Multi-CMGAN+/+	BAK/OVR/PESQ	3.12	**3.86**	3.49

Table 2. SI-SDR results on the reverberant LibriCHiME eval set.

Model	Q_1, Q_2, Q_3	**SI-SDR (dB)**
unprocessed	–	*6.59*
Sudo rm -rf [26]	–	7.8
RemixIT [27] w/ VAD	–	**10.05**
CMGAN+/+	SIG	4.71
Multi-CMGAN+/+	SIG/BAK/OVR	3.36
Multi-CMGAN+/+	SIG/BAK/PESQ	4.47
Multi-CMGAN+/+	SIG/OVR/PESQ	0.09
Multi-CMGAN+/+	BAK/OVR/PESQ	6.95

Table 2 shows the results of the proposed framework in terms of SI-SDR on the CHiME-7 UDASE task simulated evaluation set. Here, the weaknesses of the proposed system relative to the CHiME-7 baseline systems is apparent, with our proposed framework significantly degrading the input with the exception of the model which does *not* optimise the SIG component of DNSMOS.

6. CONCLUSION

In this work a GAN framework utilising a multi-metric prediction discriminator is introduced. A number of combinations of target metric for this prediction network are experimented with, and improved performance on test set consisting of real data is shown. However a degradation in performance on a simulated testset is also shown, suggesting a significant distortion in the enhanced outputs of the proposed system.

7. REFERENCES

[1] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Objective Perceptual Quality Assessment for Self-Steering Binaural Hearing Aid Microphone Arrays," in *Proc. ICASSP*, 2008.

[2] S. Goetze, E. Albertin, J. Rennies, E. Habets, and K.-D. Kammeyer, "Speech Quality Assessment for Listening-Room Compensation," *J. Audio Eng. Soc.*, vol. 62, no. 6, 2014.

[3] 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.

[4] 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," 2023.

[5] 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.

[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] ——, "CMGAN+/+: The University of Sheffield CHiME-7 UDASE Challenge Speech Enhancement System," 2023.

[8] 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.

[9] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE ICASSP*, 2001.

[10] Z. Lin, L. Zhou, and X. Qiu, "A composite objective measure on subjective evaluation of speech enhancement algorithms," *Applied Acoustics*, vol. 145, 2019.

[11] 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.

[12] 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 *ICML Proc 2019*, 2019.

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

[14] 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.

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

[16] 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.

[17] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" 2018.

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

[19] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *Proc. ICASSP*, 2022.

[20] 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.

[21] G. Close, T. Hain, and S. Goetze, "The effect of spoken language on speech enhancement using self-supervised speech representation loss functions," in *Proc. WASPAA*, 2023.

[22] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *ICASSP 2023*, 2023.

[23] 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.

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

[25] 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.

[26] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient networks for universal audio source separation," in *MLSP 2020*. IEEE, sep 2020.

[27] 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*, 2022.