Proceedings Paper:
Close, G. orcid.org/0000-0002-9478-5421, Hain, T. orcid.org/0000-0003-0939-3464 and Goetze, S. orcid.org/0000-0003-1044-7343 (Accepted: 2022) MetricGAN+/- : increasing robustness of noise reduction on unseen data. In: Proc. 30th European Signal Processing Conference, EUSIPCO 2022. 30th European Signal Processing Conference, EUSIPCO 2022, 29 Aug - 02 Sep 2022, Belgrade, Serbia. European Association for Signal Processing (EURASIP) . (In Press)

https://doi.org/10.48550/arXiv.2203.12369

MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data

George Close, Thomas Hain, and Stefan Goetze
Dept. of Computer Science, The University of Sheffield, Sheffield, United Kingdom
{glclose1, t.hain, s.goetze}@sheffield.ac.uk

***Abstract*—Training of speech enhancement systems often does not incorporate knowledge of human perception and thus can lead to unnatural sounding results. Incorporating psychoacoustically motivated speech perception metrics as part of model training via a predictor network has recently gained interest. However, the performance of such predictors is limited by the distribution of metric scores that appear in the training data. In this work, we propose MetricGAN+/- (an extension of MetricGAN+, one such metric-motivated system) which introduces an additional network - a "de-generator" to improve the robustness of the prediction network (and by extension of the generator) by ensuring observation of a wider range of metric scores in training. Experimental results on the VoiceBank-DEMAND dataset show relative improvement in PESQ score of 3.8% (3.05 vs. 3.22 PESQ score), as well as better generalisation to unseen noise and speech signals.**

***Index Terms*—speech enhancement, noise reduction, speech quality metrics, neural networks, GAN, metric prediction**

I. Introduction

Speech enhancement (SE) has been an active research topic for decades now, given its myriad applications in human-to-human (h2h) communication in video or voice calls as well as in human-to-machine (h2m) communication in home, industry and mobile device assistant products [1], [2]. Use of neural network (NN) systems to perform speech enhancement has shown great success in recent years [3]–[7]. Training of NNs for speech enhancement requires selection of an objective function appropriate for the task. Direct comparison between 'clean' audio and the output of a neural network given an artificially corrupted version of that audio has been found to be only weakly correlated with objective measures (metrics) of intelligibility, quality and performance for both forms of speech communication [8]–[10]. A recent publication [11] proposed a loss function that corresponds to one of these psychoacoustically motivated metrics. However, such objective functions must be carefully designed as many objective measures contain calculations that are non-differentiable. Several systems circumvent this limitation via use of an additional model that mimics the behaviour of the metric [12]–[14], with this network being used as a surrogate of the metric used as an objective function in training of the speech enhancement model. The baseline system that this work builds upon is one such system, MetricGAN+ [15] (itself an extension of previous work MetricGAN [16]). Two popular objective measures the Perceptual Evaluation of Speech Quality (PESQ) [17] and the Short-Time Objective Intelligibility (STOI) [18] for speech quality and intelligibility respectively are used. Both measures account for human perception and are often highly correlated with Mean Option Score (MOS) of human evaluators [8], [11], [19]. The computation of STOI is relatively simple, and a version of it suitable for use as an objective function is detailed in [11]. Calculation of PESQ is more complex, and thus cannot be formulated in a differentiable way to be used as objective functions. To handle this problem, a secondary "discriminator" network is introduced that, given a representation of the reference and the degraded signal, predicts the metric score corresponding to those two signals. Such a metric prediction network is sometimes referred to as a QualityNet [12]. The output of this discriminator network is then used to train the speech enhancement (generator) network. The two networks are trained in a Generative Adversarial Network (GAN) style strategy. In this work we introduce a further network, a 'de-generator' which attempts to produce outputs with a set lower metric score, aiming to improve the ability of the discriminator to predict the metric on a more complete range of metric scores.

The remainder of this paper is structured as follows: Section II presents the baseline system, its model structure and training setup. The proposed extension is introduced in Section III, followed by a comparison to the baseline in Section IV and a brief conclusion in Section V.

II. Baseline System - MetricGAN+

The MetricGAN+ framework [15] consists of two networks: a speech enhancement model $\mathcal{G}$, which aims to remove the undesired signal parts, i.e the noise $v[n]$ from a noisy signal

$$x[n] = s[n] + v[n] \tag{1}$$

to produce an estimate of a clean signal $s[n]$, denoted in the following by $\hat{s}[n]$ and a metric discriminator (more correctly an evaluator) $\mathcal{D}$, which predicts the possibly psychoacoustically motivated performance metrics(s) providing a target to optimise the signal enhancement.

1) Input Features: Features are calculated from different time domain signals denoted here by $p[n]$ and which is a placeholder for the signals $x[n], s[n], \hat{s}[n]$ for the time index n which we will omit in the following. First a spectral magnitude $P_{k,\ell}$ for frequency k and frame ℓ is calculated of the time

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.

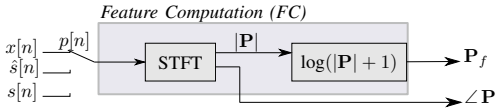

Fig. 1. Illustration of feature computation (FC), $p[n]$ can be any signal $x[n], \hat{s}[n], s[n]$

domain audio signal p using the Short Time Fourier Transform (STFT), followed by transformation to the feature space by adding 1 to and taking the logarithm of each element to give the feature representation $\mathbf{P}_f$ as shown in Fig. 1. The phase of the spectral bins $\angle p_{k,\ell}$ will be used later to resythesize the time domain signal using the Overlap-Add (OLA) method.

2) Generator Network for Signal Enhancement: Fig. 2 shows the training of $\mathcal{G}$. The dotted blue arrows and processes show the objective function and loss calculation back-propagated to the model. In order to obtain the enhanced signal $\hat{s}$ from the noisy features $\mathbf{X}_f$ in the generator $\mathcal{G}$'s training and inference, the transform is reversed by subtracting 1 from each element and taking the exponential of each element in the feature representation. The output of $\mathcal{G}$ is a time-frequency

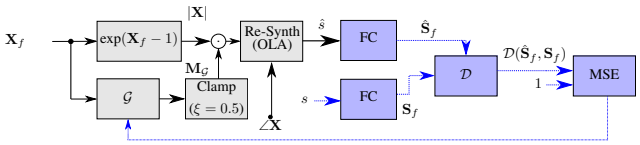

Fig. 2. Training and inference of MetricGAN+ Generator.

(T-F) mask matrix $\mathbf{M}_\mathcal{G}$, which is then multiplied with the noisy magnitude spectogram $|\mathbf{X}|$ to result in the enhanced signal matrix $|\hat{\mathbf{S}}|$. The enhanced time domain audio signal $\hat{s}[n]$ is calculated using OLA resythesis. Note that each element in $\mathbf{M}_\mathcal{G}$ is 'clamped' in order to reduce residual musical tones caused by the mask, i.e. it is limited to element wise values $\xi \leq \mathbf{M}_\mathcal{G} \leq 1$. The objective function of the speech enhancement network $\mathcal{G}$ is dependent entirely on the metric score of its output $\hat{s}$ (in its feature space representation $\hat{\mathbf{S}}_f$) as predicted by discriminator $\mathcal{D}$

$$L_{\mathcal{G},\text{MG+}} = \mathbb{E}[(\mathcal{D}(\hat{\mathbf{S}}_f, \mathbf{S}_f) - 1)^2] \quad (2)$$

where 1 represents a 'perfect' score in the normalised metric $Q'(\cdot)$.

3) Discriminator Network for Metric Prediction: The discriminator $\mathcal{D}$ is trained to reproduce the normalised target metric $Q'(\cdot)$ minimising the distance from its output and the 'true' normalised metric score used as its objective function, as visualised in Fig. 3. Arrows and processes marked blue denote those which occur only during training. The loss of the discriminator comprises three mean squared error (MSE) terms depending on the clean reference signal s, or $\mathbf{S}_f$, the degraded noisy signal x, or $\mathbf{X}_f$, and the enhanced signal $\hat{s}$, or $\hat{\mathbf{S}}_f$. More specifically, its objective function is given as:

$$L_{D,MG+} = \mathbb{E}[(\mathcal{D}(\mathbf{S}_f, \mathbf{S}_f) - 1)^2 + (\mathcal{D}(\hat{\mathbf{S}}_\mathbf{f}, \mathbf{S}_f) - Q'(\hat{s}, s))^2 + (\mathcal{D}(\mathbf{X}_f, \mathbf{S}_f) - Q'(x, s))^2] \quad (3)$$

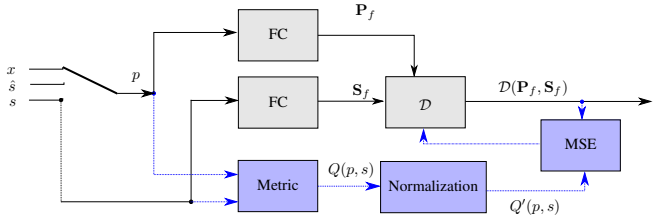

Fig. 3. Training and inference of MetricGAN+ Discriminator.

The 1 in the first term of (3) represents the fact that $Q'(s, s) = 1$. In the second term, the scores of signals enhanced by $\mathcal{G}$, $\hat{s}$ are considered and compared to the ground truth score for the enhanced signal. In the final term, the scores of noisy signals x are considered and compared to the true score for the noisy signal. Note that in the case of the metrics investigated in this work the input to the function that defines the metric are the time domain signals x, $\hat{s}$ and s, but this may not always be the case.

4) MetricGAN+ Training: Each epoch of training consists of four steps, the first three representing the training of $\mathcal{D}$ and the final step the training of $\mathcal{G}$. At the start of each epoch, I audio segments are randomly picked out from the training set. Firstly $\mathcal{D}$ is trained as given in (3) on these I random audio segments. The audio segments are time domain signals of varying length. Then, in the second step, $\mathcal{D}$ is trained using a 'replay buffer' where saved enhanced outputs of the generator $\mathcal{G}$ from past epochs are used to train $\mathcal{D}$. The size of this replay buffer is decided by a '`history_portion`' H hyper-parameter, which corresponds to the replay buffer growing by a set percentage of the audio segments observed each epoch. This is done to prevent $\mathcal{D}$ from 'forgetting' too much about the behaviour of $Q'(\cdot)$ on previously enhanced speech.
Then the first step is repeated with $\mathcal{D}$ again being trained using the t random samples. Finally, $\mathcal{G}$ is trained also using these t samples as in (2). During training of $\mathcal{D}$, $\mathcal{G}$ is 'frozen' and its parameters are not updated; the opposite is true during $\mathcal{G}$'s training. Note that samples are added to the replay buffer during the first step of $\mathcal{D}$'s training, meaning that 20% of the 'current' epoch data are always present in the replay buffer. As $\mathcal{D}$ is trained before $\mathcal{G}$, the $\hat{s}$ in (3) actually represents the output of the previous epoch's $\mathcal{G}$.

5) Discriminator Model Structure: The discriminator $\mathcal{D}$'s structure is a Convolutional Neural Network (CNN) with four 2D convolutional layers with 15 filters of a kernel size of $(5, 5)$. To account for the variable length of input data, a global 2-D average pooling layer is placed immediately after the input, fixing the feature representation at 15 dimensions. After the convolutional layers, a mean is taken over the 2nd and 3rd dimensions, and this representation is fed into three sequential linear layers, with 50, 10 and 1 output neurons, respectively. The first two of these layers have a LeakyReLU activation while the final layer has no activation.

The generator $\mathcal{G}$'s network structure consists of a Bidirectional Long Short-Term Memory (BLSTM) [20] with two LSTM layers with 200 neurons each. This is followed by two

fully connected layers, the first with 300 output neurons and a LeakyReLU [21] activation and the second 257 output neurons with a 'Learnable' Sigmoid activation function. This Learnable Sigmoid is given as:

$$y_{\text{learnable}-\text{sigmoid}} = \frac{\beta}{1 + e^{-\alpha x}} \tag{4}$$

where β is a hyper-parameter (default to 1.2) and α is a learnable parameter. In the original work the authors found that allowing β to be learnable did not increase performance.

III. Proposed System - MetricGAN+/-

A. MetricGAN+/- Framework

The framework proposed in this work, MetricGAN+/-, expands on MetricGAN+ in one major way - we introduce an additional network, a 'de-generator' $\mathcal{N}$ which, given an input signal x, will attempt to output a signal with a non-perfect score of metric Q'. The key idea of this extension is to allow $\mathcal{D}$ to observe a wider range of metrics scores outside of those present in the training data. The output audio of $\mathcal{N}$'s mask $\mathbf{M}_{\mathcal{N}}$ applied to noisy magnitude spectogram $|\mathbf{X}|$ is defined as y and its feature space representation as $\mathbf{Y}_f$. An extra term is appended to the objective function of $\mathcal{D}$ that accounts for the prediction of the Q' scores of these 'de-enhanced' signals:

$$\begin{aligned} L_{D,\text{MG}+/-} = \mathbb{E}[(\mathcal{D}(\mathbf{S}_f, \mathbf{S}_f) - 1)^2 + (\mathcal{D}(\hat{\mathbf{S}}_f, \mathbf{S}_f) - Q'(\hat{s}, s))^2 \\ + (\mathcal{D}(\mathbf{X}_f, \mathbf{S}_f) - Q'(x, s))^2 + (\mathcal{D}(\mathbf{Y}_f, \mathbf{S}_f) - Q'(y, s))^2] \end{aligned} \tag{5}$$

where y represents the output of the de-generator network on the noisy signal x. The objective function of $\mathcal{N}$ is given as

$$L_{\mathcal{N},\text{MG}+/-} = \mathbb{E}[(\mathcal{D}(\mathbf{Y}_f, \mathbf{S}_f) - w)^2], \text{ for } 0 < w < 1, \tag{6}$$

where w is a hyper-parameter corresponding to the value of Q' we train $\mathcal{N}$ to output signals with. The objective function of $\mathcal{G}$ is the same as for MetricGAN+, as given in (2). The training of $\mathcal{N}$ is the same as the training of $\mathcal{G}$ depicted in Fig. 2 except that $\mathcal{G}$ is replaced by $\mathcal{N}$, $\hat{s}$, $\hat{\mathbf{S}}_f$ by y, $\mathbf{Y}_f$ and the 1 in the MSE by w. This means that the training of $\mathcal{N}$ is influenced entirely by its performance as assessed by $\mathcal{D}$, in the same manner as $\mathcal{G}$. We use an identical network structure to $\mathcal{G}$ for $\mathcal{N}$ - We leave to future work to change this structure, as well as related hyper-parameters such as the clamp threshold.

The training of MetricGAN+/- is similar to that that of MetricGAN+ given above with slight differences. Firstly $\mathcal{D}$ is trained using (5); as a result the replay buffer now contains both enhanced and de-enhanced data, effectively doubling its size. After $\mathcal{D}$'s training, $\mathcal{N}$ is trained using (6). Then $\mathcal{G}$ is trained as usual.

IV. Experiments

A. Dataset

The dataset used in the following experiments is VoiceBank-DEMAND [22]. This is a popular and commonly used dataset for single channel speech enhancement. Its training set consists of 11572 clean $s[n]$ and noisy $x[n]$ speech audio file pairs, mixed at 4 Signal to Noise Ratios (SNRs) 0, 5, 10, 15 dB. Eight noise files are sourced from the DEMAND [23] noise dataset - a cafeteria, a car interior, a kitchen, a meeting, a metro station, a restaurant, a train station and heavy traffic, and two others a babble noise and a speech-shaped noise. The utterances in the set vary in length from around 10 seconds to 1. The training set contains speech from 28 different speakers (14 male, 14 female), English or Scottish accents. The testset containing 824 utterances is mixed at SNRs of 2.5, 7.5, 12.5 and 17.5 dB, with five different noises which do not appear in the training set from the DEMAND corpus (bus, cafe, office, public square and living room) and contains speech from two (one male, one female) speakers who do not appear in the training set.

In order to better assess the system's ability to generalise to unseen noise types and recording scenarios as well as real recordings, we also assess performance of the models trained on VoiceBank-DEMAND training set on the test set of the CHiME3 [3] challenge dataset. This test set consists of 1320 real and 1320 simulated noisy clean/speech pairs. For the real recordings the clean 'reference' is a close-talk headset microphone which may also capture some of the background noise from the recording environment. There are 6 channels of noisy recordings; we select the 5th channel as input to the single channel systems as it has the most direct energy to the speaker and is the one used in the baseline system of the CHiME3 challenge. The recording environments of the real data and background noise of the simulated are a bus, a cafe, a pedestrian area and a street junction. The simulated data is not mixed at any fixed SNR, instead an ideal mixing SNR is calculated from analysis of the clean reference and the background recording.

B. Experiment Setup

The aim of the following experiments is to compare the performance of the baseline system MetricGAN+ which is available as part of the SpeechBrain [24] toolkit with our extension, MetricGAN+/-. The Adam optimiser [25] with a learning rate of 0.0005 is used. The STFT is used with a DFT length of L_{DFT} =512, a window length of 512 (32 ms) at sampling frequency of $f_s = 16$ kHz and a hop (overlap) length 256 (16 ms), resulting in a 50% overlap between frames. The minimum value in the time frequency masks $\mathbf{M}_{\mathcal{G}}$ and $\mathbf{M}_{\mathcal{N}}$ is set to $\xi = 0.05$.

We experiment with both PESQ and STOI as objective Q and different values of w. The values of w are selected such that they correspond to sparely populated values of Q' in the dataset. We also performed one experiment (denoted by * in Table I) where the value of β in $\mathcal{N}$'s Learnable Sigmoid activation as given in (4) to also learned (in addition to α). Additionally, we experiment with reducing the size of the replay buffer training step for $\mathcal{D}$, via modifying H. In order to ensure that our performance gain does not come entirely from the larger H in MetricGAN+/-, we report also the baseline MetricGAN+ performance with H set to 0.4.

C. Experiment Results

Table I shows the performance of MetricGAN+/- relative to the MetricGAN+ baseline and the unprocessed noisy audio on the VoiceBank-DEMAND testset. We also compare performance with a second baseline system SEGAN [26], a state-of-the-art speech enhancement system. For more comparison baseline performances the interested reader is referred to Table 3 in [15], which shows that MetricGAN+ with a PESQ objective outperforms all systems listed terms of PESQ score. We assess this performance using PESQ and STOI and also using the Composite [27] Measure, where Csig, Cbak and Covl are intrusive measures of speech signal quality, background noise reduction quality, and overall quality respectively.

TABLE I
PERFORMANCE OF METRICGAN+ (MG+) AND METRICGAN+/- (MG+/-) ON VOICEBANK-DEMAND TEST SET FOR OBJECTIVE PESQ (P) OR STOI (S), * DENOTES THE SIMULATION WHERE β IS MADE LEARNABLE

Model Name	Obj.	w	H	P	S	Csig	Cbak	Covl
Noisy	-	-	-	*1.97*	*92*	*3.35*	*2.44*	*2.63*
MG+ (P) [15]	P	-	0.2	3.05	93	4.03	2.87	3.52
MG+ (S)	S	-	0.2	2.42	**93.4**	3.56	2.58	2.97
SEGAN [26]	-	-	-	2.42	92.5	3.61	2.61	3.01
MG+	P	-	0.4	3.17	92.3	4.05	2.91	3.59
MG+/-	P	1.0	0.2	3.20	93.0	4.08	2.94	3.62
MG+/-	P	0.50	0.2	**3.22**	91.3	4.05	2.94	3.62
MG+/-	P	0.45	0.2	3.21	91.9	4.09	**2.95**	3.64
MG+/-*	P	0.45	0.2	3.17	93.0	**4.16**	2.93	**3.65**
MG+/-	P	0.45	0.1	3.13	92.1	4.05	2.91	3.58
MG+/-	P	0.30	0.2	3.04	93.0	4.07	2.88	3.55
MG+/-	S	0.45	0.1	2.13	93.2	3.04	2.42	2.56
MG+/-	S	0.30	0.2	2.31	93.3	3.19	2.49	2.72

The first four rows in Table I present the results the unenhanced noisy data and of different baselines. The results for the baseline MetricGAN+ models shown here are obtained using the implementation in SpeechBrain [24]. Further simulations are conducted for various values of hyperparameter w used in the training of $\mathcal{N}$. Table I shows a clear improvement in PESQ score for PESQ objective MetricGAN+/- models over the baseline MetricGAN+ (3.05 vs 3.22 PESQ), and also versus the PESQ value reported in [15] of 3.15. We also observe increase in the composite measure scores. Interestingly, there is an improvement even when $w = 1$, which is the case where $\mathcal{N}$ and $\mathcal{G}$ have the same objective, and thus $\mathcal{N}$ also learns to enhance. We hypothesise that this is due to slight variations in the outputs of $\mathcal{N}$ and $\mathcal{G}$ during training, as well as the increased replay buffer size compared to the baseline. Highest performance in terms of PESQ score is obtained with a w value set to 0.5, which means that $\mathcal{N}$ attempts to produce signals with a PESQ score of 3. We speculate that this performance increase is due to there being few clean/noisy pairs in the training set with a PESQ score around this value.

By making the β parameter in $\mathcal{N}$'s activation function learnable, we observe a slight improvement against the baseline, as well as increased Csig and Covl scores versus all other simulations. We find also that increasing H in the baseline MetricGAN+ from 0.2 to 0.4 such that its size is comparable to MetricGAN+/-'s does slightly improve PESQ score. This is contrary to the findings in [15] where they report no improvement for values larger than 0.2. Larger values of H will drastically increase the training time requirement of the system. We speculate that a better understanding of what $\mathcal{D}$ leans from the replay buffer training and better curation of its contents is the key to further performance gains, as well as reduced training time required.

D. Spectogram Analysis

Fig. 4 shows the spectrograms of the clean reference $|\mathbf{S}|$, noisy input $|\mathbf{X}|$, generator output mask $\mathbf{M}_{\mathcal{G}}$ and this mask applied $|\hat{\mathbf{S}}|$ for baseline MetricGAN+, MetricGAN+/- ($w = 0.45$) PESQ objective models. The mask in Fig. 4 (e) attempts to remove low frequency signal content while boosting the area corresponding to the frequency curve of the fundamental speech frequencies. Futhermore, the baseline MetricGAN+ PESQ model in Fig. 4 (c, d) attenuates the signal in the initial non speech region, while the MetricGAN+/- model in Fig. 4 (e, f) suppresses less energy around 400 Hz over the whole utterance. This artefact can already be observed in the

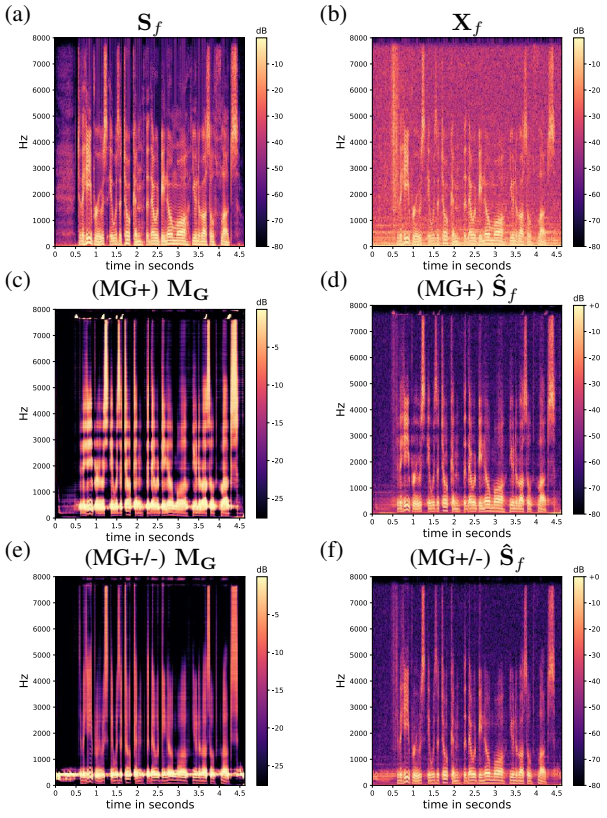

Fig. 4. Spectograms of: (a) clean reference features $\mathbf{S}_f$, (b) noisy features $\mathbf{X}_f$, (c) Mask $\mathbf{M}_{\mathcal{G}}$ and (d) enhanced output $\hat{S}_f$ for MetricGAN+ baseline PESQ objective model, (e) Mask $\mathbf{M}_{\mathcal{G}}$ and (f) enhanced output $\hat{S}_f$ for MetricGAN+/- PESQ objective model. Source audio file is `p232_014.wav` of VoiceBank-DEMAND testset.

baseline MetricGAN+ but is more prominent for the proposed method, which could explains the relatively low Cbak score for this method. We hypothesise that this is a results of $\mathcal{D}$ not learning to properly penalise errors in this region, perhaps due to the additional influence of $\mathcal{N}$'s outputs on it's training.

Further experimentation is required to fully understand this artefact.

E. Generalisation To Unseen Data

TABLE II
PERFORMANCE ON REAL COMPONENT OF CHIME3 TEST SET

Model Type	PESQ	STOI	Csig	Cbak	Covl
Noisy	*1.37*	*44.0*	***2.96***	*1.42*	*2.09*
MG+ PESQ	1.54	45.8	2.67	2.09	2.00
MG+ STOI	1.24	44.7	2.45	1.84	1.76
MG+/- PESQ	**1.76**	44.3	2.86	**2.03**	**2.20**
MG+/- STOI	1.22	**45.3**	2.31	1.81	1.67

TABLE III
PERFORMANCE ON SIMULATED COMPONENT OF CHIME3 TEST SET

Model Type	PESQ	STOI	Csig	Cbak	Covl
Noisy	*1.27*	*87.0*	*2.61*	*1.39*	*1.88*
MG+ PESQ	2.14	87.4	3.05	2.31	2.53
MG+ STOI	1.52	**88.9**	2.75	2.07	2.08
MG+/- PESQ	**2.38**	86.1	**3.17**	**2.41**	**2.70**
MG+/- STOI	1.47	88.5	2.62	2.02	1.99

Tables II and III shows the performance of the baseline MetricGAN+ and the best performing proposed MetricGAN+/- systems on this test set. We observe an increased performance in terms of PESQ, Csig, Cbak and Covl between PESQ objective MetricGAN+/- and the baseline, as well as a slight improvement in STOI score for STOI objective MetricGAN+/- . This suggests that $\mathcal{D}$'s access to the de-generated signals produced by $\mathcal{N}$ allows $\mathcal{G}$ in MetricGAN+/- systems to generalise better to unseen environments.

V. CONCLUSION

In this work, we present an extension to the MetricGAN+ baseline framework, which improves its performance in terms of PESQ score and related measures, as well as improving its generalisation to unseen data. We find that training the discriminator network on a wider range of metric scores and with a larger replay buffer achieves greater performance that the baseline system.

REFERENCES

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[2] F. Xiong, B. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-end technologies for robust ASR in reverberant environments - spectral enhancement-based dereverberation and auditory modulation filterbank features," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

[3] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: dataset, task and baselines," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2015*, Scottsdale, Arizona, USA, 2015, pp. 504–511.

[4] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

[5] W. Ravenscroft, S. Goetze, and T. Hain, "Att-TasNet: Attending to Encodings in Time-Domain Audio Speech Separation of Noisy, Reverberant Speech Mixtures," *Frontiers in Signal Processing*, vol. 2, 2022.

[6] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada*, 2021, pp. 8443–8447.

[7] N. Moritz, K. Adiloğlu, J. Anemüller, S. Goetze, and B. Kollmeier, "Multi-channel speech enhancement and amplitude modulation analysis for noise robust automatic speech recognition," *Computer Speech & Language*, vol. 46, pp. 558–573, Nov 2017.

[8] S. Goetze, A. Warzybok, I. Kodrasi, J. Jungmann, B. Cauchi, J. Rennies, E. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A Study on Speech Quality and Speech Intelligibility Measures for Quality Assessment of Single-Channel Dereverberation Algorithms," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Antibes, France, Sep. 2014.

[9] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," 2018.

[10] L. Chai, J. Du, and C.-H. Lee, "Acoustics-guided evaluation (age): a new measure for estimating performance of speech enhancement algorithms for robust asr," *ArXiv*, vol. abs/1811.11517, 2018.

[11] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, 2018.

[12] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-m. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Proc of Interspeech 2018*, 09 2018, pp. 1873–1877.

[13] T. F. Ziyi Xu, Maximilian Strake, "Deep noise suppression with non-intrusive PESQNet supervision enabling the use of real training data," in *Proc. of INTERSPEECH*, Brno, Czechia, 2021, pp. 1–5.

[14] Z. Xu, M. Strake, and T. Fingscheidt, "Deep Noise Suppression Maximizing Non-Differentiable PESQ Mediated by a Non-Intrusive PESQNet," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1572–1585, 2022.

[15] 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.

[16] 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.

[17] 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 International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, 2001.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[19] C. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors," in *ICASSP 2020*, October 2020, pp. 6493–6497.

[20] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds., 2015.

[21] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[22] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and tts models," University of Edinburgh. Centre for Speech Technology Research (CSTR), 2017, doi: 10.7488/ds/2117.

[23] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," Jun. 2013, Supported by Inria under the Associate Team Program VERSAMUS.

[24] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[26] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017.

[27] Z. Lin, L. Zhou, and X. Qiu, "A composite objective measure on subjective evaluation of speech enhancement algorithms," *Applied Acoustics*, vol. 145, pp. 144–148, 02 2019.