Audio Engineering Society

Convention Paper 10656

Presented at the 154th Convention
2023 May 13–15, Espoo, Helsinki, Finland

PAMGAN+/-: Improving Phase-Aware Speech Enhancement Performance via Expanded Discriminator Training

George Close[1], Thomas Hain[1], and Stefan Goetze[1]

[1]*Speech and Hearing (SpandH), Dept. of Computer Science, The University Of Sheffield, United Kingdom*

Correspondence should be addressed to George Close (`glclose1@sheffield.ac.uk`)

ABSTRACT

Recent speech enhancement work, which makes use of neural networks trained with a loss derived in part using an adversarial metric prediction network, has shown to be very effective. However, by limiting the data used to train this metric prediction network to only the clean reference and the output of the speech enhancement network, only a limited range of the metric is learnt. Additionally, such speech enhancement systems are limited because they typically operate solely over magnitude spectrogram representations so they do not encode phase information. In this work, recent developments for phase-aware speech enhancement in such an adversarial framework are expanded in two ways to enable the metric prediction network to learn a full range of metric scores. Firstly, the metric predictor is also exposed to unenhanced 'noisy' data during training. Furthermore, an additional network is introduced and trained alongside which attempts to produce outputs with a fixed 'lower' target metric score, and expose the metric predictor to these 'de-enhanced' outputs. It is found that performance increases versus a baseline system utilising a magnitude spectrogram speech enhancement network.

1 Introduction

Speech enhancement (SE) has been an active research topic for some decades now, given its myriad applications in human-to-human (h2h) communication (in video or voice calls) [1, 2, 3], as well as in human-to-machine (h2m) communication in home, car, industry, mobile devices or smart assistants [4, 5]. The use of neural network systems to perform speech enhancement has shown great success in recent years [6, 7]. In the training of neural networks to accomplish speech enhancement selection of an objective function that is appropriate for this task is important [**?** 8]. 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 uncorrelated with objective measures (metrics) of intelligibility, quality and performance for both forms of speech communication [9, 10]. Recent publications [11, 12] have proposed the use of an objective function that better represents one or more of these objective measures. However, these objective functions must be carefully designed as many signal assessment measures have calculations that are non-differentiable. A commonly used signal assessment measure in such systems is the Perceptual Evaluation of Speech Quality (PESQ) [13] score, which has been shown to correlate with human perception [12, 14]. Many systems [15, 16, 17, 18] circumvent the limitation of non-differentiable losses by training an additional network alongside the speech enhancement network to mimic the behaviour of the objective metric. This network, the discriminator, is

then used to formulate a loss for the training of the speech enhancement network, the generator. The two networks are trained in a generative adversarial network (GAN) style. In such systems, the data used to train the discriminator network is of particular importance, as the generator network's training is entirely reliant upon it.
A limitation of the above is that the speech enhancement network is trained solely using the metric estimation network in all cases. In [19], two additional losses derived from a distance between a reference signal and the model output in the time and complex frequency domains respectively are introduced. Also in [19], a more complex speech enhancement network which is able to directly encode phase information is proposed. This work further uses a phase-aware speech enhancement network structure inspired by [19], but the whole system is trained as described in [16], including the use of a tertiary 'de-generator' network, as first proposed in [18].
The remainder of this paper is structured as follows: in Section 2 the proposed system and baselines are detailed including their model structures, loss functions and training setups. The experimental setup is described in Section 3, performance of the proposed method is analysed in Section 4 and Section 5 concludes the paper.

2 PAMGAN+/-

2.1 Signal Model and Feature Computation (FC)

Single channel speech enhancement can be defined as the recovery of the clean speech signal $s[n]$ from the noisy mixture

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

corrupted by a disturbance $v[n]$ for discrete time index n (omitted in the following to increase readability), i.e. estimating a clean speech signal $\hat{s}$ from x.

Complex tempo-spectral features are calculated as in [19]: spectrogram matrices $\mathbf{P} \in \mathbb{C}^{L_{\text{DFT}} \times L_{\mathbf{P}}}$ are calculated using the short time Fourier transform (STFT) of length L_{DFT} for each $L_{\mathbf{P}}$ frames of a time domain signal p, which are compressed by the power law compression [20] to obtain feature matrices $\mathbf{P}_f = \mathbf{P}^c$. From this, magnitude, phase, real and imaginary components are calculated, denoted as $\mathbf{P}_{\text{M}} = |\mathbf{P}_f|$, $\mathbf{P}_{\text{P}} = \angle \mathbf{P}_f$, $\mathbf{P}_{\text{Re}} = \text{Re}\{\mathbf{P}_f\}$, and $\mathbf{P}_{\text{Im}} = \text{Im}\{\mathbf{P}_f\}$, respectively. Please note that p and $\mathbf{P}$ are considered as placeholders for the signals, i.e. $p \in \{x, s, \hat{s}, y\}$ and $\mathbf{P}_{(\cdot)} \in \{\mathbf{X}_{(\cdot)}, \mathbf{S}_{(\cdot)}, \hat{\mathbf{S}}_{(\cdot)}, \mathbf{Y}_{(\cdot)}\}$, respectively [18, 16]. Note also that the magnitude representation $\mathbf{P}_{\text{M}}$, by definition, does not encode any phase information.

2.2 Framework Overview

In this work, the enhancement generator $\mathscr{G}$ is inspired by the model proposed in [19], including two additional loss terms when compared to the MetricGAN+ [16] baseline. This is trained alongside the discriminator $\mathscr{D}$ in the manner proposed in MetricGAN+ [16] and a 'de-generator' $\mathscr{N}$ as introduced in MetricGAN+/- [18]. We call this proposed system PAMGAN+/- (Phase-Aware-Metric-GAN) since it extends the MetricGAN+/- framework by phase information. Table 1 analyses differences of related frameworks which will be detailed in the following.

2.3 MetricGAN+ Baseline

The MetricGAN+ baseline framework [16], extending ideas from [15], consists of two networks, (i) a speech enhancement model, the generator $\mathscr{G}$, which aims to remove the undesired signal parts v from the noisy signal x in Eq. 1 and (ii) a metric estimation network, the discriminator $\mathscr{D}$, which provides an estimate $\widehat{Q}'(\cdot)$ for the normalised performance metric $Q'(\cdot)$, providing a target to optimise the signal enhancement.

2.3.1 Loss of Discriminator $\mathscr{D}$

In this work, a normalised version of the time domain intrusive speech quality metric PESQ [13] is used as $Q'(\cdot)$. PESQ is defined between 1 and 4.5, higher being better, and frequently used to assess speech enhancement [21]. The discriminator $\mathscr{D}$ is trained to reproduce the normalised target metric $Q'(\cdot)$ by minimising the distance from its output $\widehat{Q}'(\cdot)$ and the 'true' normalised metric score $Q'(\cdot)$ used as its objective function, i.e. the discriminator loss function is

$$\begin{aligned} L_{\mathscr{D},\text{MG+}} \;=\; & \mathbb{E}\{(\mathscr{D}(\mathbf{S}_{\text{M}}, \mathbf{S}_{\text{M}}) - 1)^2 \\ & + (\mathscr{D}(\hat{\mathbf{S}}_{\text{M}}, \mathbf{S}_{\text{M}}) - Q'(\hat{s}, s))^2 \\ & + (\mathscr{D}(\mathbf{X}_{\text{M}}, \mathbf{S}_{\text{M}}) - Q'(x, s))^2\} \end{aligned} \tag{2}$$

where each term represents $\mathscr{D}$'s ability to reproduce the $Q'(\cdot)$ score of time domain signals s, $\hat{s}$, and x, respectively. Figure 1 visualises the discriminator structure. The value of $Q'(s, s)$, i.e. comparing the reference

Table 1: Properties of different MetricGAN (MG) derived frameworks.

	Generator $\mathscr{G}$			Discriminator $\mathscr{D}$		De-generator $\mathscr{N}$
System	**Features**, Eq. 2.1	**Structure**, Section 2.6.1	$L_{\mathscr{G}}$, Section 2.4.1	**Input**, Section 2.3.1	**Historical**, Section 2.7.1	**Component**, Section 2.5
MG [15]	$\mathbf{X}_{\mathrm{M}}$	BLSTM	Eq. 3	$s,\hat{s}$	x	x
MG+ [16]	$\mathbf{X}_{\mathrm{M}}$	BLSTM	Eq. 3	$s,\hat{s},x$	✓	x
MG+/- [18]	$\mathbf{X}_{\mathrm{M}}$	BLSTM	Eq. 3	$s,\hat{s},x,y$	✓	✓
PAMGAN+/- (prop.)	$\mathbf{X}_{\mathrm{M}},\mathbf{X}_{\mathrm{Re}},\mathbf{X}_{\mathrm{Im}}$	Conformer	Eq. 8	$s,\hat{s},x,y$	✓	✓

signal s to itself, is always 1 (cf. first term in Eq. 2, while the values of $Q'(x,s)$, $Q'(\hat{s},s)$, etc. usually vary depending on corpus and generator training success, respectively, typically from 0.3 to 1 for common corpora, higher being better.

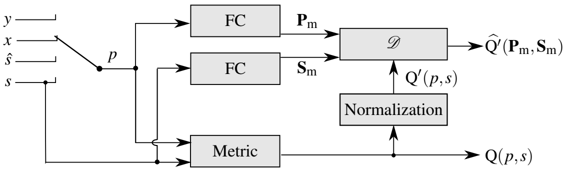

Fig. 1: Training of discriminator $\mathscr{D}$.

2.3.2 Loss of MetricGAN Generator $\mathscr{G}$

The metric score of the enhanced output of $\mathscr{G}$, $\hat{s}$ as predicted by inference of $\mathscr{D}$ is used to train $\mathscr{G}$ based on the loss

$$L_{\mathscr{G}_{\mathrm{MG}}} = \mathbb{E}\{(\mathscr{D}(\hat{\mathbf{S}}_{\mathrm{m}},\mathbf{S}_{\mathrm{m}})-1)^2\}, \tag{3}$$

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

2.4 Phase-Aware Enhancement Losses

2.4.1 Phase-Aware Generator $\mathscr{G}$

Inspired by [19], two additional losses can be used additionally to Eq. 3 to train the generator $\mathscr{G}$. These losses are intended to train the generator structure to fully utilise the phase information encoded in its inputs. Firstly a time domain loss [22]

$$L_{\mathscr{G}_{\mathrm{Time}}} = \mathbb{E}\{||s-\hat{s}||_1\} \tag{4}$$

is minimised which directly compares the enhanced time domain signal $\hat{s}$ with the clean reference signal s. Secondly, a time-frequency (TF) domain loss $L_{\mathscr{G}_{TF}}$ [20] is considered, which makes explicit use of the component outputs of the signal enhancement network $\mathscr{G}$, i.e. $\hat{\mathbf{S}}_{\mathrm{m}},\hat{\mathbf{S}}_{\mathrm{Im}}$ and $\hat{\mathbf{S}}_{\mathrm{Re}}$. For this, the distance between the enhanced and the reference magnitude

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

and the complex (real and imaginary) components

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

are computed. The two loss terms in Eq. 5 and Eq. 6 are finally combined using a weighing hyperparameter α to result in

$$L_{\mathscr{G}_{TF}} = \alpha L_{\mathscr{G}_{\mathrm{M}}} + (1-\alpha) L_{\mathscr{G}_{\mathrm{RI}}}. \tag{7}$$

The final loss for $\mathscr{G}$ is then given as in [19]

$$L_{\mathscr{G}} = \gamma_1 L_{\mathscr{G}_{\mathrm{GAN}}} + \gamma_2 L_{\mathscr{G}_{\mathrm{Time}}} + \gamma_3 L_{\mathscr{G}_{\mathrm{TF}}} \tag{8}$$

where γ_1, γ_2, and γ_3 are hyperparameter weights to control the influence of each loss term. The model structure has three component outputs, a magnitude mask $\mathbf{M}_{\mathscr{G}}$, a real and imaginary component, $\hat{\mathbf{S}}'_{\mathrm{Re}}$ and $\hat{\mathbf{S}}'_{\mathrm{Im}}$, respectively. The magnitude mask $\mathbf{M}_{\mathscr{G}}$ is multiplied with the noisy signal magnitude $\mathbf{X}_{\mathrm{M}}$ to produce the enhanced signal estimate $\hat{\mathbf{S}}_{\mathrm{M}}$. Then, the combination of the enhanced magnitude $\hat{\mathbf{S}}_{\mathrm{M}}$ with the original noisy phase $\mathbf{X}_{\mathrm{P}}$ is added with the other two outputs $\hat{\mathbf{S}}'_{\mathrm{Re}}$ and $\hat{\mathbf{S}}'_{\mathrm{Im}}$ as follows:

$$\hat{\mathbf{S}}_{\mathrm{Re}} = \hat{\mathbf{S}}_{\mathrm{M}}\cos(\mathbf{X}_{\mathrm{P}}) + \hat{\mathbf{S}}'_{\mathrm{Re}} \tag{9a}$$

$$\hat{\mathbf{S}}_{\mathrm{Im}} = \hat{\mathbf{S}}_{\mathrm{M}}\sin(\mathbf{X}_{\mathrm{P}}) + \hat{\mathbf{S}}'_{\mathrm{Im}} \tag{9b}$$

The generator, thus, combined masking-based and mapping-based signal enhancement. The power law compression is then inverted and an inverse short time Fourier transform (ISTFT) is taken to obtain $\hat{s}$ in the Inv-FC block as visualised in Figure 2 which visualises the generator training.

2.5 Degenerator Extension $\mathscr{N}$

Based on findings from previous work [18], adding an 'de-generator' network $\mathscr{N}$ to the PAMGAN framework as depicted in Figure 2 is proposed to generate

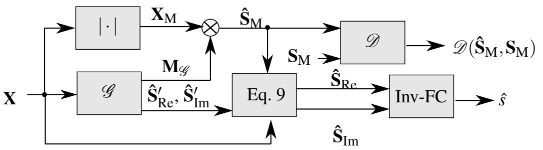
$|\cdot|$
$\mathbf{X}_\mathrm{M}$
$\hat{\mathbf{S}}_\mathrm{M}$
$\mathscr{D}$
$\mathscr{D}(\hat{\mathbf{S}}_\mathrm{M},\mathbf{S}_\mathrm{M})$
$\mathbf{S}_\mathrm{M}$
$\mathbf{M}_\mathscr{G}$
$\mathbf{X}$
$\mathscr{G}$
$\hat{\mathbf{S}}'_\mathrm{Re}, \hat{\mathbf{S}}'_\mathrm{Im}$
Eq. 9
$\hat{\mathbf{S}}_\mathrm{Re}$
Inv-FC
$\hat{s}$
$\hat{\mathbf{S}}_\mathrm{Im}$

Fig. 2: Input and output of PAMGAN+/- $\mathscr{G}$, showing inference of $\mathscr{D}$ and computation of $\hat{s}$.

outputs with a target score $Q'(\cdot)$ which is lower than 1, i.e. which have lower quality. These outputs, denoted as y, are then used to augment the training of the discriminator $\mathscr{D}$ by extending its loss from Eq. 2:

$$\begin{aligned} L_{\mathscr{D},\mathrm{PAMGAN+/-}} &= L_{\mathscr{D},\mathrm{MG+}} \\ &+ \mathbb{E}\left\{(\mathscr{D}(\mathbf{Y}_\mathrm{m},\mathbf{S}_\mathrm{m}) - Q'(y,s))^2\right\} \end{aligned} \quad (10)$$

Figure 3 shows the input and training to the degenerator network $\mathscr{N}$, where inference of $\mathscr{D}$ is used to train the network to produce outputs with $Q'(\cdot)$ scores of w. The proposed hyperparameter w orresponds to the normalised 'lowered' target metric $Q'(\cdot)$ that $\mathscr{N}$ is being trained to output audio with. The loss of $\mathscr{N}$ is given similarly to Eq. 3:

$$L_\mathscr{N} = \mathbb{E}\{(\mathscr{D}(\mathbf{Y}_\mathrm{m},\mathbf{S}_\mathrm{m}) - w)^2\}, \quad 0 < w < 1 \quad (11)$$

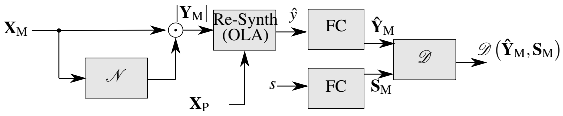

Fig. 3: Feature Computation (FC) and training of PAMGAN+/- degenerator network $\mathscr{N}$ via inference of $\mathscr{D}$.

Note that we also experiment with having the input to $\mathscr{N}$ be the clean reference magnitude $\mathbf{S}_\mathrm{m}$ and train it using a modified version of Eq. 4:

$$L_{\mathscr{N}_\mathrm{Clean}} = \mathbb{E}\{\mathscr{D}(\bar{\mathbf{S}}_\mathrm{m},\mathbf{S}_\mathrm{m}) - w)^2\} + \mathbb{E}\{||x - \bar{s}||_1\} \quad (12)$$

where $\bar{\mathbf{S}}_\mathrm{m}$ is the magnitude of the de-enhanced time domain signal $\bar{s}$. The time loss is used to encourage $\mathscr{N}$ to recreate the distortion caused by the noisy v in x.

2.6 Network Structures

2.6.1 Generator Network Structure

The generator $\mathscr{G}$ is inspired by that proposed in [19] and has a loss function Eq. 8 as described in Section 2.4.1. It takes noisy signal parts $\mathbf{X}_\mathrm{m}$, $\mathbf{X}_\mathrm{Re}$, $\mathbf{X}_\mathrm{Im}$ as input. The network consists of four main components, an encoder, a Conformer [23, 24] based bottleneck and two decoder structures, one producing a magnitude mask which is applied to $\mathbf{X}_\mathrm{m}$ to produce the enhanced magnitude $\hat{\mathbf{S}}_\mathrm{m}$ and the other the complex and real components $\hat{\mathbf{S}}'_\mathrm{Re}$ and $\hat{\mathbf{S}}'_\mathrm{Im}$. The encoder structure consists of two convolutional layers with a dilated DenseNet [25] in between.

Each block of the bottleneck structure consists of two sequential Conformer blocks, with residual connections and reshape operations to allow the first Conformer to operate over the time dimension and the second over the frequency. The Conformer blocks are designed to capture long-term dependency in the spectogram inputs [26].

The output of the bottleneck is fed in parallel to two decoders, which use a dilated DenseNet followed by two convolutional layers. The mask decoder and the complex decoder are identical with the exception of the final convolutional layer. In the mask decoder the final convolutional layer outputs a single layer (the magnitude mask) with a ReLU activation, while the complex decoder outputs two representations (the real $\hat{\mathbf{S}}'_\mathrm{Re}$ and imaginary $\hat{\mathbf{S}}'_\mathrm{Im}$) with its final layer having no activation.

2.6.2 Discriminator Network Structure

The discriminator network $\mathscr{D}$ takes the magnitude of the clean reference signal $\mathbf{S}_\mathrm{m}$ as input, stacked with that of the signal being assessed ($\mathbf{S}_\mathrm{m}$, $\mathbf{X}_\mathrm{m}$, $\hat{\mathbf{S}}_\mathrm{m}$ or $\mathbf{Y}_\mathrm{m}$) and returns an estimation of that signal's $Q'(.)$ score. It consists of four convolutional layers with 32, 64, 128, and 256 filters, respectively, each with an instance normalisation and LeakyReLU [27] activation. These are followed by a global average pooling layer, and then three feed-forward layers each with a LeakyReLU activation aside from the final layer which has a sigmoid activation. These feed-forward layers have 50, 10, and 1 output neurons, respectively.

2.6.3 Degenerator Network Structure

The degenerator network $\mathscr{N}$ takes as input the noisy magnitude $\mathbf{X}_{\text{M}}$ and returns a magnitude mask which is multiplied with $\mathbf{X}_{\text{M}}$ to produce a degraded magnitude $\mathbf{Y}_{\text{M}}$. It follows the structure detailed in [18], consisting of a Bidirectional Long Short-Term Memory (BLSTM) [28] 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 activation, and the second with 257 output neurons and a 'Learnable' Sigmoid activation function. The degraded waveform y is computed using the overlap-add resynthesis method, using the original noisy phase $\mathbf{X}_{\text{P}}$.

2.7 MetricGAN+ Training

2.7.1 Historical Training

First proposed in [16], this is a technique where $\mathscr{D}$ is trained using a 'replay buffer' where saved outputs of the generator $\mathscr{G}$ from past epochs. The size of this replay buffer is decided by a '`history_portion`' hyper-parameter H, which corresponds to the replay buffer growing by a fixed percentage of the audio segments observed each epoch. This is done to prevent $\mathscr{D}$ from 'forgetting' too much about the behaviour of $Q'(\cdot)$ on previously enhanced speech.

2.7.2 Training Cycle

Each training epoch consists of four steps, the first three representing the training of $\mathscr{D}$ and the final one the training of $\mathscr{G}$. At the start of each epoch, I audio segments are randomly picked from the training set. Firstly, $\mathscr{D}$ is trained as given in Eq. 2 on these I random audio segments. These audio segments are 2 seconds in length due to technical constraints, but the system uses the longer audio at validation and test. Next, $\mathscr{D}$ is trained using the historical set as described above. Then the first step is repeated with $\mathscr{D}$ again being trained using the I random samples. Finally, $\mathscr{G}$ is trained using Eq. 8. $\mathscr{G}$ is trained also using the I samples.
While training the discriminator $\mathscr{D}$, the generator $\mathscr{G}$ is 'frozen', i.e. its parameters are not updated, and the opposite is true during training the generator $\mathscr{G}$. Note that samples are added to the replay buffer during the first step of $\mathscr{D}$'s training, meaning that 20% of the current epoch's samples are always present in the replay buffer. As $\mathscr{D}$ is trained before $\mathscr{G}$, the $\hat{s}$ in Eq. 2 actually represents the output of the previous epoch's generator $\mathscr{G}$.

2.8 MetricGAN+/- and PAMGAN+/- Training

Due to the introduction of $\mathscr{N}$, some changes are made to the training described above. Firstly, outputs of $\mathscr{N}$, i.e. y, are also used to populate the replay buffer, thus doubling its size as each enhanced sample with now have a corresponding 'de-enhanced' version. Additionally, the training of $\mathscr{N}$ occurs immediately before that of $\mathscr{G}$ in each epoch.

3 Experiments

3.1 Dataset

The dataset used in the following experiments is VoiceBank-DEMAND [29], a popular and commonly used dataset for single-channel speech enhancement. Its training set consists of 11572 clean and noisy speech audio file pairs (s,x), mixed at four different signal-to-noise ratios (SNRs) of {0, 5, 10, 15} dB. Eight noise files are sourced from the DEMAND [30] noise dataset - a cafeteria, a car interior, a kitchen, a meeting, a metro station, a restaurant, a train station and heavy traffic noise. Two others, a babble noise and a speech-shaped noise, were also used. The utterances in the set vary in length from around 2 seconds to 10, but are segmented into 2 second blocks for training. The training set contains speech from 28 different speakers (14 male, 14 female), with 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.

3.2 Experiment Setup

The aim of the following experiments is to compare the performance of the baseline systems MetricGAN+, which is available as part of the SpeechBrain [31] toolkit, and the CMGAN with the proposed system, denoted as ConformerMetricGAN+/-. The framework is trained for 300 epochs with a sample size t of 100 and a batch size of 1. The historical training hyperparameter H is set to 0.2, such that the historical set grows by 40 entries each epoch (20 enhanced, 20, 'de-enhanced'). The Adam optimiser [32] with a learning rate of 0.0005 is used for all three networks $\mathscr{G}$, $\mathscr{D}$ and $\mathscr{N}$. Two TS-Conformer blocks are used in $\mathscr{G}$. The STFT has a DFT length of $L_{\text{DFT}} = 400$, a window length of 25 ms at sampling frequency of $f_s = 16$ kHz and a hop (overlap) length of 6.25 ms, resulting in a 75% overlap between frames.The feature compression factor is $c = 0.3$ [20]. The hyperparameter α in the calcualtion of $L_{G_{TF}}$ is set to 0.7, while $\gamma_1, \gamma_2, \gamma_3$ in L_G are all set to 1. The value of w in L_N is set to 0.45 (corresponding to a PESQ score of 2.5).

Two proposed models are trained, one of which has $\mathscr{N}$ trained using Eq. 11, denoted '$\mathscr{N}$ noisy' in the following and the other using Eq. 12, denoted as '$\mathscr{N}$ clean'. In addition to the proposed system PAMGAN+/-, models using MetricGAN+[16] and MetricGAN+/-[18] are also trained as baselines. The MetricGAN+ and MetricGAN+/- baseline models are trained for the same number of epochs (300 rather than 600 in their originally published versions) as the proposed system to ensure comparability. Models are evaluated using PESQ [13], short-time objective intelligibility (STOI) [33] and the Composite [34] measure. STOI is a measure of speech intelligibility valued between 0 and 100%, while the three Composite scores Csig, Cbak and Covl are valued between 0 and 5 and represent the speech signal quality, the background noise reduction and the overall quality of the speech, respectively.

4 Results

Table 2: Performance of PAMGAN+/- on VoiceBank-DEMAND test set

Model Name	**PESQ**	**STOI**	**Csig**	**Cbak**	**Covl**
Noisy	*1.97*	*92.0*	*3.35*	*2.44*	*2.63*
MetricGAN+ [16]	2.93	93.0	3.99	2.81	3.44
MetricGAN+/- [18]	**3.05**	92.0	3.95	2.87	3.49
PAMGAN+/- ($\mathscr{N}$ noisy)	2.97	93.1	4.09	2.89	3.53
PAMGAN+/- ($\mathscr{N}$ clean)	3.04	**93.4**	**4.16**	**2.93**	**3.61**

Table 2 shows the results of the PAMGAN+/- models trained on the VoiceBank-DEMAND test set. Both proposed models outperform the baseline systems in terms of STOI and the Composite measure and perform similarly in terms of PESQ. This slight decrease in PESQ score can perhaps be explained by the difference in $\mathscr{G}$'s loss function; in MetricGAN+/-, $\mathscr{G}$ is trained using Eq. 3, i.e. solely to maximise the PESQ score, while in the proposed PAMGAN+/- Eq. 3 is but one component of Eq. 8 used to train $\mathscr{G}$. The model where $\mathscr{N}$ is trained using Eq. 12, '$\mathscr{N}$ clean', outperforms '$\mathscr{N}$ noisy' in all tests. This is interesting, given that a similar experiment in [18] found the opposite; however in that case all $\mathscr{N}$ were trained only via inference of $\mathscr{D}$ as in Eq. 11. This indicates that the additional loss term which represents the distance between $\mathscr{N}$'s output and the noisy signal x helps $\mathscr{N}$ to produce outputs which are more useful for the training of $\mathscr{D}$.

4.1 Spectrogram Analysis

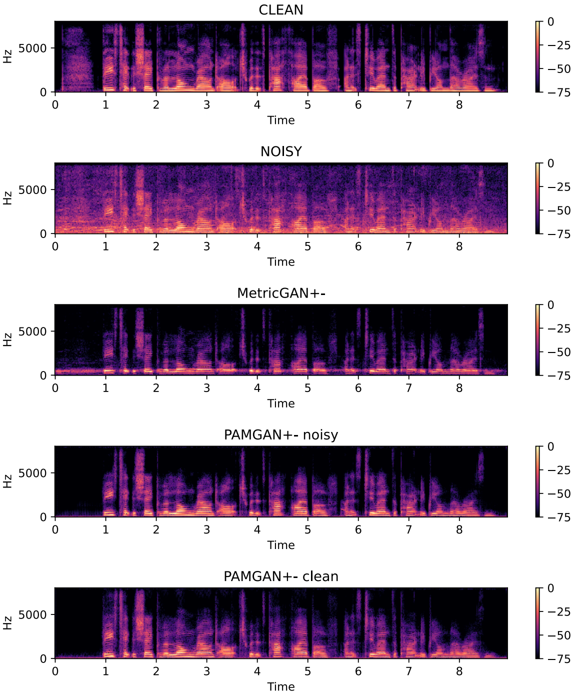

Fig. 4: Spectrogram plots of `p257_008.wav` from VoiceBank-DEMAND testset s, x, $\hat{s}$ outputs of MetricGAN+/-, PAMGAN+/- with x as input to $\mathscr{N}$ and PAMGAN+/- with s as input to $\mathscr{N}$

Figure 4 show spectrogram plots of clean signal s, noisy signal x and the enhanced signal $\hat{s}$ for the baseline system MetricGAN+/- and for the two proposed PAMGAN+/- models. From these, it can be observed that an artefact in the low-frequency region of the MetricGAN+/- spectrogram is not present in either PAMGAN+/- spectrogram. This indicates that the phase-aware Generator structure is more robust to such artefacts compared to the baseline system.

5 Conclusion

In this work, an extension to the MetricGAN+/- framework incorporating a phase-aware, Conformer based network structure leads to increased performance and reduced artefacts while utilising more input features as well as a more nuanced loss function for the speech enhancement network $\mathscr{G}$.

References

[1] *Speech enhancement / [edited by] Benesty, J., Makino, S., and Chen, J.*, Signals and communication technology, Springer, Berlin, 2005, ISBN 354024039X.

[2] Goetze, S., Mildner, V., and Kammeyer, K.-D., "A Psychoacoustic Noise Reduction Approach for Stereo Hands-Free Systems," in *Audio Engineering Society (AES), 120th Convention*, Paris, France, 2006.

[3] Reddy, C. K., Dubey, H., Koishida, K., Nair, A., Gopal, V., Cutler, R., Braun, S., Gamper, H., Aichner, R., and Srinivasan, S., "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *INTERSPEECH*, 2021.

[4] Xiong, F., Meyer, B., Moritz, N., Rehr, R., Anemüller, J., Gerkmann, T., Doclo, S., and Goetze, S., "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*, 2015(1), 70, 2015, doi:10.1186/s13634-015-0256-4.

[5] Haeb-Umbach, R., Heymann, J., Drude, L., Watanabe, S., Delcroix, M., and Nakatani, T., "Far-Field Automatic Speech Recognition," *Proceedings of the IEEE*, 109(2), pp. 124–148, 2021, doi:10.1109/JPROC.2020.3018668.

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

[7] Tammen, M. and Doclo, S., "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*, pp. 8443–8447, 2021, doi:10.1109/ICASSP39728.2021.9413775.

[8] Close, G., Ravenscroft, W., Hain, T., and Goetze, S., "Perceive and predict: self-supervised speech representation based loss functions for speech enhancement," in *Proc. 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, 2023.

[9] Bagchi, D., Plantinga, P. W. V., Stiff, A., and Fosler-Lussier, E., "Spectral Feature Mapping with MIMIC Loss for Robust Speech Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5609–5613, 2018.

[10] Goetze, S., Albertin, E., Rennies, J., Habets, E., and Kammeyer, K.-D., "Speech Quality Assessment for Listening-Room Compensation," *J. Audio Eng. Soc.*, 62(6), pp. 386–399, 2014.

[11] wei Fu, S., Tsao, Y., Hwang, H.-T., and Wang, H.-M., "Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model Based on BLSTM," in *Proc. Interspeech 2018*, pp. 1873–1877, 2018, doi:10.21437/Interspeech.2018-1802.

[12] Fu, S.-W., Wang, T.-W., Tsao, Y., Lu, X., and Kawai, H., "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*, 26(9), pp. 1570–1584, 2018, doi:10.1109/TASLP.2018.2821903.

[13] Rix, A., Beerends, J., Hollier, M., and Hekstra, A., "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*, pp. 749–752 vol.2, 2001, doi:10.1109/ICASSP.2001.941023.

[14] K. A. Reddy, C., Gopal, V., and Cutler, R., "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors," in *2020 International Conference on Acoustics, Speech, and Signal Processing*, pp. 6493–6497, IEEE, 2020.

[15] Fu, S.-W., Liao, C.-F., Tsao, Y., and Lin, S.-D., “MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement,” in K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2031–2041, PMLR, 2019.

[16] Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., and Tsao, Y., “MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement,” in *Proc. Interspeech 2021*, pp. 201–205, 2021, doi:10.21437/Interspeech.2021-599.

[17] Fu, S.-W., Yu, C., Hung, K.-H., Ravanelli, M., and Tsao, Y., “MetricGAN-U: Unsupervised Speech Enhancement/ Dereverberation Based Only on Noisy/ Reverberated Speech,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7412–7416, 2022, doi:10.1109/ICASSP43922.2022.9747180.

[18] Close, G., Hain, T., and Goetze, S., “MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data,” in *30th European Signal Processing Conference, EUSIPCO 2022*, pp. 165–169, Belgrade, Serbia, 2022, doi:10.23919/EUSIPCO55093.2022.9909682.

[19] Cao, R., Abdulatif, S., and Yang, B., “CMGAN: Conformer-based Metric GAN for Speech Enhancement,” in *Proc. Interspeech 2022*, pp. 936–940, 2022, doi:10.21437/Interspeech.2022-517.

[20] Braun, S. and Tashev, I., “A consolidated view of loss functions for supervised deep learning-based speech enhancement,” in *44th Int. Conf. on Telecommunications and Signal Processing (TSP)*, pp. 72–76, 2021, doi:10.48550/ARXIV.2009.12286.

[21] Avila, A., Cauchi, B., Goetze, S., Doclo, S., and Falk, T., “Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech,” in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, doi:10.1109/IWAENC.2016.7602907.

[22] Abdulatif, S., Armanious, K., Sajeev, J. T., Guirguis, K., and Yang, B., “Investigating Cross-Domain Losses for Speech Enhancement,” in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 411–415, 2021, doi:10.23919/EUSIPCO54536.2021.9616267.

[23] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R., “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *Proc. Interspeech 2020*, pp. 5036–5040, 2020, doi:10.21437/Interspeech.2020-3015.

[24] Chen, S., Wu, Y., Chen, Z., Wu, J., Li, J., Yoshioka, T., Wang, C., Liu, S., and Zhou, M., “Continuous Speech Separation with Conformer,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5749–5753, 2021, doi:10.1109/ICASSP39728.2021.9413423.

[25] Pandey, A. and Wang, D., “Densely Connected Neural Network with Dilated Convolutions for Real-Time Speech Enhancement in The Time Domain,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6629–6633, 2020, doi:10.1109/ICASSP40776.2020.9054536.

[26] Dang, F., Chen, H., and Zhang, P., “DPT-FSNet: Dual-Path Transformer Based Full-Band and Sub-Band Fusion Network for Speech Enhancement,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6857–6861, 2022, doi:10.1109/ICASSP43922.2022.9746171.

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

[28] Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., and Schuller, B., “Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR,” in E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, editors, *Latent Variable Analysis and Signal Separation*, 2015, ISBN 978-3-319-22482-4.

[29] Valentini-Botinhao, C., “Noisy speech database for training speech enhancement algorithms and TTS models,” University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017, doi: 10.7488/ds/2117.

[30] Thiemann, J., Ito, N., and Vincent, E., “DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments,” 2013, doi:10.5281/zenodo.1227121, Supported by Inria under the Associate Team Program VERSAMUS.

[31] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y., “SpeechBrain: A General-Purpose Speech Toolkit,” 2021.

[32] Kingma, D. and Ba, J., “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations*, 2014.

[33] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J., “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), pp. 2125–2136, 2011, doi:10.1109/TASL.2011.2114881.

[34] Lin, Z., Zhou, L., and Qiu, X., “A composite objective measure on subjective evaluation of speech enhancement algorithms,” *Applied Acoustics*, 145, pp. 144–148, 2019, doi:10.1016/j.apacoust.2018.10.002.