MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement

Szu-Wei Fu [1 2]   **Chien-Feng Liao** [1 2]   **Yu Tsao** [2]   **Shou-De Lin** [1]

Abstract

Adversarial loss in a conditional generative adversarial network (GAN) is not designed to directly optimize evaluation metrics of a target task, and thus, may not always guide the generator in a GAN to generate data with improved metric scores. To overcome this issue, we propose a novel MetricGAN approach with an aim to optimize the generator with respect to one or multiple evaluation metrics. Moreover, based on MetricGAN, the metric scores of the generated data can also be arbitrarily specified by users. We tested the proposed MetricGAN on a speech enhancement task, which is particularly suitable to verify the proposed approach because there are multiple metrics measuring different aspects of speech signals. Moreover, these metrics are generally complex and could not be fully optimized by L_p or conventional adversarial losses.

1. Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014) has shown its powerful generative ability in many different applications. In particular, for conditional GANs (CGANs) (Mirza & Osindero, 2014), in addition to the adversarial loss, there is an L_p loss, to guide the learning of generators. Ideally, the adversarial loss should make generated data indistinguishable from real (target) data. However, some applications of image (Ledig et al., 2017; Wang et al., 2018) and speech processing (Pandey & Wang, 2018; Wang & Chen, 2018; Donahue et al., 2018; Michelsanti & Tan, 2017) show that this loss term provides very marginal improvement (sometimes even degrade the performance) in terms of objective evaluation scores (in the case of image processing, the subjective score can be improved). For in-

stance, Donahue et al. (2018) applied CGAN on speech enhancement (SE) for automatic speech recognition (ASR); however, the following conclusion was obtained: "Our experiments indicate that, for ASR, simpler regression approaches may be preferable to GAN-based enhancement." This may be because the method that the discriminator uses to judge whether each sample is real or fake is not fully related to the metrics that we consider. In other words, similar to L_p loss, the way the adversarial loss guides the generator to generate data is still not matched to the evaluation metrics. We call this problem discriminator-evaluation mismatch (DEM). In this study, we propose a novel MetricGAN to solve this problem. We tested the proposed approach on the SE task because the metrics for SE are generally complex and difficult to directly optimize or adjust.

For human perception, the primary goal of SE is to improve the intelligibility and quality of noisy speech (Benesty et al., 2005). To evaluate an SE model in different aspects, several objective metrics have been proposed. Among the human perception-related objective metrics, the perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) and short-time objective intelligibility (STOI) (Taal et al., 2011) are two popular functions to evaluate speech quality and intelligibility, respectively. The design of these two metrics considers human auditory perception and has shown higher correlation to subjective listening tests than simple L_1 or L_2 distance between clean and degraded speech.

In recent years, various deep learning-based models have been developed for SE (Lu et al., 2013; Xu et al., 2014; Wang et al., 2014; Xu et al., 2015; Ochiai et al., 2017; Luo & Mesgarani, 2018; Grais et al., 2018; Germain et al., 2018; Chai et al., 2018). Most of these models were trained in a supervised fashion by preparing pairs of noisy and clean speeches. The deep models were then optimized by minimizing the distance between generated speech and clean speech. However, the distance (objective function) is usually based on simple L_p loss (where $p = 1$ or 2), which does not reflect human auditory perception or ASR accuracy (Bagchi et al., 2018) well. In fact, several researches have indicated that an enhanced speech with a smaller L_p distance, does not guarantee a higher quality or intelligibility score (Fu et al., 2018b; Koizumi et al., 2018).

---

[1]National Taiwan University, Taiwan [2]Academia Sinica, Taiwan. Correspondence to: Yu Tsao <yu.tsao@citi.sinica.edu.tw>, Shou-De Lin <sdlin@csie.ntu.edu.tw>.

Therefore, optimizing the evaluation metrics (i.e., STOI, PESQ, etc.) may be a reasonable direction to connect the model training with the goal of SE. Some latest studies (Fu et al., 2018b; Koizumi et al., 2018; Zhang et al., 2018; Zhao et al., 2018a; Naithani et al., 2018; Kolbæk et al., 2018; Venkataramani et al., 2018; Venkataramani & Smaragdis, 2018; Zhao et al., 2018b) have focused on STOI score optimization to improve speech intelligibility. A waveform based utterance-level enhancement manner is proposed to optimize the STOI score (Fu et al., 2018b). The results of a listening test showed that by combining STOI with MSE as an objective function, the speech intelligibility can be further increased. On the other hand, because the PESQ function is not fully differentiable and significantly more complex compared with STOI, only few (Koizumi et al., 2018; Zhang et al., 2018; Koizumi et al., 2017; Martín-Doñas et al., 2018) have considered it as an objective function. Reinforcement learning (RL) techniques such as deep Q-network (DQN) and policy gradient were employed to solve non-differentiable problems, as (Koizumi et al., 2017) and (Koizumi et al., 2018), respectively.

In summary, the abovementioned existing techniques can be categorized into two types depending on whether the details of evaluation metrics have to be obtained: (1) white-box: these methods approximate the complex evaluation metrics with a hand-crafted, simpler one; thus, it is differentiable and easy to be applied as a loss function. However, the details of the metrics have to be known; (2) black-box: these methods mainly treat the metrics as a reward and apply RL-based techniques to increase the scores. However, because of less efficiency in training, most of them have to be pre-trained by conventional supervised learning.

In this study, to solve the drawbacks of the abovementioned methods and the DEM problem, the discriminator in GAN is associated with the evaluation metrics of interest (Although these evaluation functions are complex, Fu et al. (2018a) showed that they can be approximated by neural networks). In particular, when training the discriminator, instead of always giving a false label (e.g., "0") to the generated speech, the labels of MetricGAN are given according to the evaluation metrics. Therefore, the target space of discriminator transforms from discrete (1 (true) or 0 (false)) to continuous (evaluation scores). Through this modification, the discriminator can be treated as a learned surrogate of the evaluation metrics. In other words, the discriminator iteratively estimates a surrogate loss that approximates the sophisticated metric surface, and the generator uses this surrogate to decide a gradient direction for optimization. Compared with previous existing methods, the main advantages of MetricGAN are as follows:

(1) The surrogate function (discriminator) of the complex evaluation metrics is learned from data. In other words, it is still in a black-box setting and no computational details of the metric function have to be known.

(2) Experiment result shows that the training efficiency of MetricGAN to increase metric score is even higher than conventional supervised learning with L_p loss.

(3) Because the label space of the discriminator is now continuous, any desired metric scores can be assigned to the generator. Therefore, MetricGAN has the flexibility to generate speech with specific evaluation scores.

(4) Under some non-extreme conditions, MetricGAN can even achieve multi-metrics assignments by employing multiple discriminators.

2. CGAN for SE

GAN has recently attracted a significant amount of attention in the community. By employing an alternative mini-max training scheme between a generator network (G) and a discriminator network (D), adversarial training can model the distribution of real data. One of its applications is to serve as a trainable objective function for a regression task. Instead of explicitly minimizing the L_p losses, which may cause over smoothing problems, D provides a high-level abstract measurement of realness (Liao et al., 2018).

In the applications of GAN on SE, CGAN is usually employed to generate enhanced speech. To achieve this, G is trained to map noisy speech x to its corresponding clean speech y by minimizing the following loss function (as in (Pascual et al., 2017). The least-squares GAN (LSGAN) approach (Mao et al., 2017) is used with binary coding (1 for real, 0 for fake)):

$$L_{G(CGAN)} = \mathbb{E}_x[\lambda(D(G(x), x) - 1)^2] + ||G(x) - y||_1 \quad (1)$$

Because G usually simply learned to ignore the noise prior z in the CGAN (Isola et al., 2017), we directly neglected it here. The first term in Eq. (1) is called adversarial loss for cheating D with a weighting factor λ. The goal of D is to distinguish between real data and generated data by minimizing the following loss function:

$$L_{D(CGAN)} = \mathbb{E}_{x,y}[(D(y, x) - 1)^2 + (D(G(x), x) - 0)^2] \quad (2)$$

We argue that to optimize the metric scores, the training of D should be associated with the metric.

3. MetricGAN

3.1. Associating the Discriminator with the Metrics

The main difference between the proposed MetricGAN and the conventional CGAN is how the discriminator is trained. Here, we first introduce a function $Q(I)$ to represent the

evaluation metric to be optimized, where I is the input of the metric. For example, for PESQ and STOI, I is the pair of speech that we want to evaluate and the corresponding clean speech y. Therefore, to ensure that D behaves similar to Q, we simply modify the objective function of D:

$$L_{D(MetricGAN)} = \mathbb{E}_{x,y}[(D(y,y) - Q(y,y))^2 \\ + (D(G(x),y) - Q(G(x),y))^2] \quad (3)$$

Because we can always map Q to Q', which is between 0 and 1 (here, 1 represents the best evaluation score), Eq. (3) can be reformulated as

$$L_{D(MetricGAN)} = \mathbb{E}_{x,y}[(D(y,y) - 1)^2 \\ + (D(G(x),y) - Q'(G(x),y))^2] \quad (4)$$

where $0 \leq Q'(G(x),y) \leq 1$. There are two main differences between Eq. (4) and Eq. (2):

1.) In CGAN, as long as the data is generated, its label for D is always a constant 0. However, the target label of the generated data in our MetricGAN is based on its metric score. Therefore, D can evaluate the degree of realness (clean speech), instead of just distinguishing real and fake. (Therefore, maybe "D" should be called an evaluator; however, here we just follow the convention of GAN.)

2.) The condition used in the D of CGAN is the noisy speech x, which is different from the condition used in the proposed MetricGAN (clean speech y). This is because we want D and Q to have similar behavior. Therefore, the input argument of D is chosen to be the same as Q.

3.2. Continuous Space of the Discriminator Label

The training of G is similar to Eq. (1). However, we found that the gradient provided by D in our MetricGAN is more efficient than the L_p loss. Therefore, the training of G can completely rely on the adversarial loss :

$$L_{G(MetricGAN)} = \mathbb{E}_x[(D(G(x),y) - s)^2] \quad (5)$$

where s is the desired assigned score. For example, to generate clean speech, we can simply assign s to be 1. On the contrary, we can also generate more noisy speech by assigning a smaller s. This flexibility is caused by the label of the generated speech in D, which is now continuous and related to the metric. Unlike surrogate loss learning in the multi-class classification (Hsieh et al., 2018), because the output space of our G is continuous, the local neighbors need not be explicitly selected to learn the behavior of metric surface.

3.3. Explanation of MetricGAN

In MetricGAN, the target of G is to cheat D to reach specified score, and D tries to not be cheated by learning the
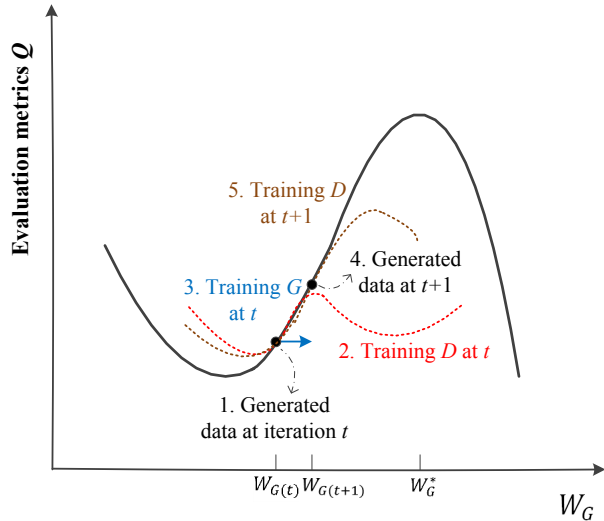

Figure 1. Learning process of MetricGAN to optimize the evaluation metric Q (the horizontal axis represents the weights of G). In each iteration, there are three steps. First, some data are generated by G with weights $W_{G(t)}$. Second, training D to have similar behavior as metric Q in those points. Third, training G according to the gradient provided by D (cheat D).

true score. Here, we also explain the learning process of MetricGAN in a different manner. As shown in Figure 1, training of D can be treated as learning a local surrogate of Q; and training of G is to adjust its weights W_G toward the optimum value of D. Because D may only approximate Q well in the observed region (Fu et al., 2019), this learning framework should be alternatively trained until convergence.

4. Experiments

4.1. Network Architecture

The input features x for G is the normalized noisy magnitude spectrogram utterance. The generator used in this experiment is a BLSTM (Weninger et al., 2015) with two bidirectional LSTM layers, each with 200 nodes, followed by two fully connected layers, each with 300 LeakyReLU nodes and 257 sigmoid nodes for mask estimation, respectively. When this mask (between 0 to 1) is multiplied with the noisy magnitude spectrogram, the noise components should be removed. In addition, as reported in (Koizumi et al., 2018), to prevent musical noise, flooring was applied to the estimated mask before T-F-mask processing. Here, we used the lower threshold of the T-F mask as 0.05.

The discriminator herein is a CNN with four two-dimensional (2-D) convolutional layers with the number of filters and kernel size as follows: [15, (5, 5)], [25, (7, 7)], [40, (9, 9)], and [50, (11, 11)]. To handle the variable-length input (different speech utterance has different length),

Table 1. Performance comparisons of different loss functions in terms of PESQ and STOI (* represents pre-trained from another model).

SNR (dB)	Noisy		IRM (L1)		IRM (CGAN)		PE policy grad*(P)		MetricGAN (P)		MetricGAN (S)	
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
12	2.375	0.919	2.913	0.935	2.879	0.936	**2.995**	0.927	2.967	0.936	2.864	**0.939**
6	1.963	0.831	2.52	0.878	2.479	0.876	2.595	0.869	**2.616**	0.881	2.486	**0.885**
0	1.589	0.709	2.086	0.787	2.053	0.786	2.144	0.776	**2.200**	0.796	2.086	**0.802**
-6	1.242	0.576	1.583	0.655	1.551	0.653	1.634	0.644	**1.711**	0.668	1.599	**0.679**
-12	0.971	0.473	1.061	0.508	1.046	0.507	1.124	0.500	**1.169**	0.521	1.090	**0.533**
Avg.	1.628	0.702	2.033	0.753	2.002	0.751	2.098	0.743	**2.133**	0.760	2.025	**0.768**

a 2-D global average pooling layer was added such that the features can be fixed at 50 dimensions (50 is the number of feature maps in the previous layer). Three fully connected layers were added subsequently, each with 50 and 10 LeakyReLU nodes, and 1 linear node. In addition, to make D a smooth function (we do not want a small change in the input spectrogram can result in a significant difference to the estimated score), it is constrained to be 1-Lipschitz continuous by spectral normalization (Miyato et al., 2018). Our preliminary experiments found that adding this constraint can stabilize the training of D. All models are trained using Adam (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

4.2. Experiment on the TIMIT Dataset

In this section, we show the experiments about PESQ and STOI scores. PESQ was designed to evaluate the quality of processed speech, and the score ranges from -0.5 to 4.5. STOI was designed to compute the speech intelligibility, and the score ranges from 0 to 1. Both the two metrics are the higher the better.

4.2.1. DATASET

In this experiments, the TIMIT corpus (Garofolo et al., 1988) was used to prepare the training, validation, and test sets. 300 utterances were randomly selected from the training set of the TIMIT database for training in this experiment. These utterances were further corrupted with 10 noise types (crowd, 2 machines, alarm and siren, traffic and car, animal sound, water sound, wind, bell, and laugh noise) from (Hu), at five SNR levels (from -8 dB to 8 dB with steps of 4 dB) to form 15000 training utterances. To monitor the training process and choose the hyperparameters, we randomly selected another clean 100 utterances from the TIMIT training set to form our validation set. Each utterance was further corrupted with one of the noise types (different from those already used in the training set) from (Hu) at five different SNR levels (from -10 dB to 10 dB with steps of 5 dB). To evaluate the performance of different training methods, 100 utterances from the TIMIT test set were randomly selected as our test set. These utterances were mixed with four un-

seen noise types (engine, white, street, and baby cry), at five SNR levels (-12 dB, -6 dB, 0 dB, 6 dB, and 12 dB). In summary, 2000 utterances exist in the test set.

4.2.2. OBJECTIVE EVALUATION WITH DIFFERENT LOSS FUNCTIONS

In this experiment, to evaluate the performance of different objective functions, the structure of G is fixed and trained with different losses. As one of our baseline models, we adopt ideal ratio mask (IRM) (Narayanan & Wang, 2013) based mask estimation with L_1 loss (denoted as **IRM (L1)**). The other baseline (denoted as **IRM (CGAN)**) is the CGAN with the loss function of G shown in Eq. (1). Compared to **IRM (L1)**, **IRM (CGAN)** has an additional adversarial loss term with $\lambda = 0.01$ as in (Bagchi et al., 2018; Pascual et al., 2017). A parameter exploring policy gradients (Sehnke et al., 2010) based black-box optimization, which is similar to the one used in (Zhang et al., 2018), is also compared. However, we found that this method is very sensitive to the hyperparameters (e.g., weight initialization, step size of jitter, etc.). We could only obtain improved results for PESQ optimization (denoted as **PE policy grad (P)**). In addition, because of the lower training efficiency, its generator was first pre-trained from **IRM (L1)**. The proposed MetricGAN with PESQ or STOI metric as Q, is indicated as **MetricGAN (P)** and **MetricGAN (S)**, respectively.

Table 1 presents the results of the average PESQ and STOI scores on the test set for the baselines and proposed methods. From this table, we can first observe that the performance of **IRM (CGAN)** is similar to or slightly worse than the simple **IRM (L1)**, which is in agreement with the results presented in previous papers. (Pandey & Wang, 2018; Donahue et al., 2018). This implies that the adversarial loss term used to cheat D is not helpful in this application. One possible reason for this result may be that the decision boundary of D is very different from the metrics we consider. We also attempted to train **IRM (CGAN)** with larger λ; however, their evaluation scores were worse than the reported scores. Although **PE policy grad (P)** can obtain some PESQ scores improvements, the STOI scores decreased compared to its
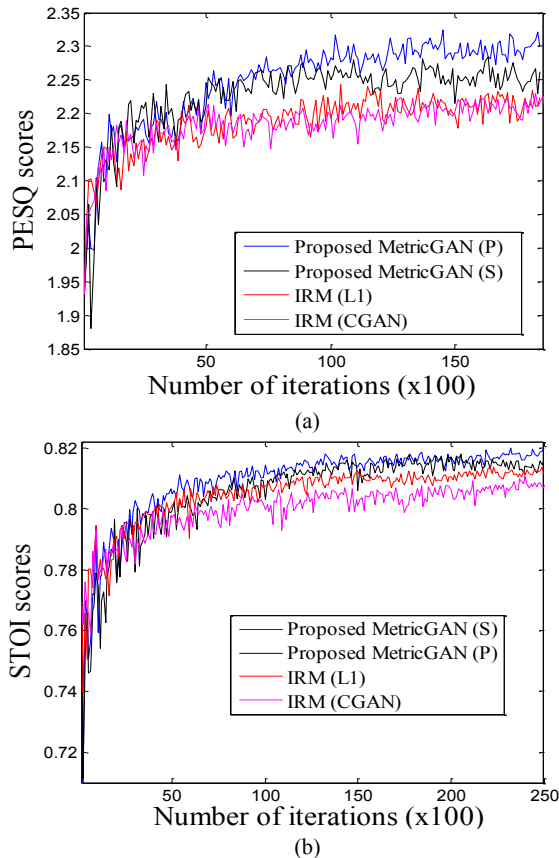
Figure 2. Learning curve of different objective functions evaluated on the validation set (structure of G is fixed). In terms of: (a) PESQ score and (b) STOI score.
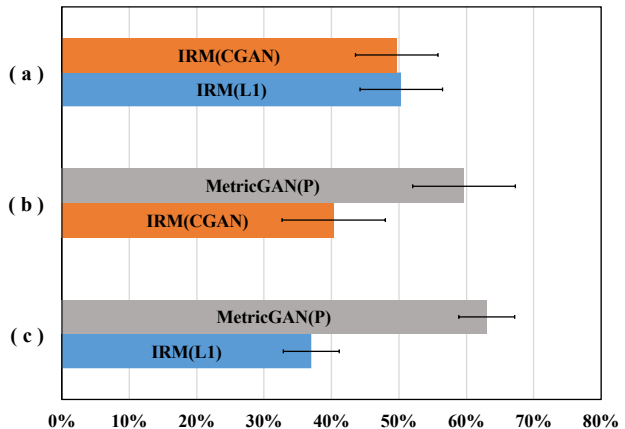

Figure 3. Results of AB preference test (with 95% confidence intervals) on speech quality compared between proposed **MetricGAN(P)** and the two baseline models.

IRM (CGAN) is not helpful for improving the scores and training efficiency.

Finally, an example of the enhanced spectrograms by different training objective functions are shown in Figure 4. The spectrogram generated by **IRM (CGAN)** is similar to that of **IRM (L1)**. If we simply increase the weight λ of the adversarial loss term in Eq.(1), some unpleasant artifacts begin to appear (this is not shown here, owing to limited space). Interestingly, in comparison to others, the spectrogram (f) generated by **MetricGAN (S)** can best recover the speech components with clear structures (as shown by the black-dashed rectangles) and hence, obtain the highest STOI score.

4.2.3. SUBJECTIVE EVALUATION

To evaluate the perceptual quality of the enhanced speech, we conducted AB preference tests to compare the proposed method with the baseline models. Three pairs of listening tests were conducted: **IRM(CGAN)** versus **IRM(L1)**, **MetricGAN(P)** versus **IRM(CGAN)**, and **MetricGAN(P)** versus **IRM(L1)**. Each pair of samples are presented in a randomized order. For each listening test, 20 sample pairs were randomly selected from the test set; 15 listeners participated. Listeners were instructed to select the sample with the better quality. The stimuli were played to the subjects in a quiet environment through a set of Sennheiser HD headphones at a comfortable listening level. In Figure 3 (a), we can observe that the preference score between **IRM (L1)** and **IRM (CGAN)** overlap in the confidence interval, which is in agreement with the result of the objective evaluation. Further, as shown in Figure 3 (b) and Figure 3 (c), **MetricGAN(P)** significantly outperforms both baseline systems, without an overlap in the confidence intervals.

initialization, **IRM (L1)**. On the contrary, when we employed PESQ as Q in our MetricGAN, it could achieve the highest PESQ scores among all the models with the second highest STOI score. Note that unlike L_p loss, the loss function of G in MetricGAN is Eq. (5), and there is no specific target for each T-F bin. In terms of the STOI score, **MetricGAN (S)** outperforms the other models, and the improvement is most evident for the low SNR conditions (where speech intelligibility improvement is most critical).

In addition to the final results of the test set, the learning process of different loss functions evaluated on the validation set are also presented in Figure 2. For both the scores, we can observe that the learning efficiency (in terms of the number of iterations) of MetricGAN is higher than the others. This implies that the gradient provided by D (surrogate of Q) is the most accurate toward the maximum value of Q. However, if the Q used to train MetricGAN does not match the evaluation metric, the performance is sub-optimal. Therefore, the information from Q is important; our preliminary experiment also shows that without Q, the learning cannot converge. The conventional adversarial loss term in

PESQ=4.500, STOI=1.000

PESQ=1.080, STOI=0.600

PESQ=1.966, STOI=0.721

(a) clean target

(b) noisy input

(c) **IRM (L1)**

PESQ=1.950, STOI=0.709

PESQ=**2.137**, STOI= 0.735

PESQ=2.049, STOI=**0.744**

(d) **IRM (CGAN)**

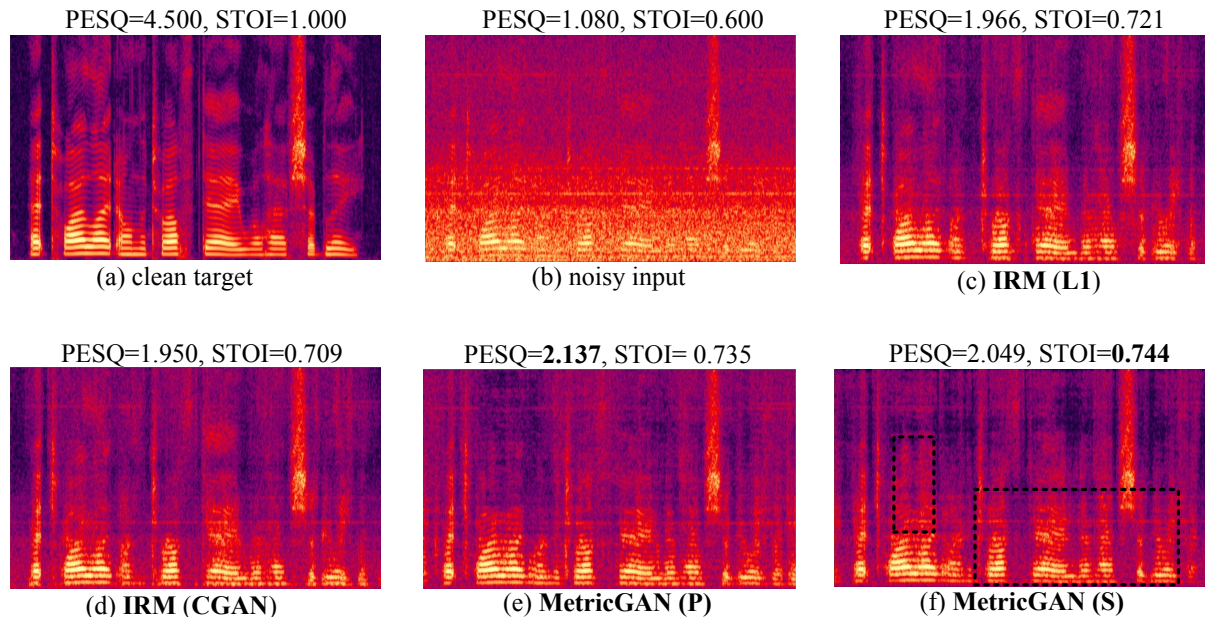(e) **MetricGAN (P)**

(f) **MetricGAN (S)**

Figure 4. Spectrograms of a TIMIT utterance in the teset set: (a) clean target, (b) noisy speech (engine noise at 0 dB). (c) to (f): enhanced speech with different loss functions.

4.2.4. ASSIGNING ANY DESIRED SCORE TO THE GENERATOR

Because the label of D in the conventional GAN is in a discrete space, there is no guarantee that the generated data can gradually improve toward real data when the label assigned to G (i.e., the constant in the first term of Eq.(1)) increases from 0 (fake) toward 1 (real). For example, the generated data from label 0.9 is not necessarily better (more like real data) than that from label 0.8. However, as pointed out in section 3.2, because the output of D in MetricGAN is continuous according to Q, we can assign any desired score during the training of G as in Eq. (5). Therefore, different s in Eq. (5) correspond to generated speech with different qualities. Interestingly, setting s as a small value can convert the generator from a speech enhancement model to a speech degradation model. This provides us with another method to understand the factors that affect the metric. To achieve this, a uniform mask constraint (penalize estimated mask away from 0.5) was also applied to G so that G has to choose the most efficient way to attain the assigned score s without significantly changing the initialized mask. (Owing to the sigmoid activation used in the output layer of G, all the initially estimated mask values were close to 0.5). Figure 5 shows an example of assigning different s to G, and the learning process evaluated on the validation set is also illustrated in Figure 5 (c) and (g). Compared to the generation of clean speech (the entire learning process for generating clean speech is presented in Figure 2), Metric-

GAN can attain the desired score more easily when s is small. This phenomenon is because the number of solutions decreases gradually when s increases (it is easier to obtain noisy speech than a clean speech). Therefore, the solution for a large s is considerably difficult to obtain. Figures 5 (d) to (f) and (h) to (j) present the generated speech by assigning different s with STOI and PESQ as Q, respectively. Intriguingly, the speech components gradually disappear when we attempt to generate a speech with low STOI score (the speech components are almost removed as shown by the black rectangle in Figure 5 (f)). Because STOI measures the intelligibility of speech, it is reasonable that the speech component is most crucial in this metric. On the contrary, because PESQ measures the quality of speech, the generated speech with lower s seems to become more noisy (for extremely low s values (Figure 5 (j)), in spite of not as serious as the STOI case, there is also some speech components being removed). These results verify that the MetricGAN can generate data according to the designate metric score and make the label space of D continuous.

4.2.5. MULTI-METRIC SCORES ASSIGNMENT

In this section, we further explore the assignment of scores for multiple metrics simultaneously. Compared with single metric assignment, this is a more difficult task because the requirement to achieve other metrics can be treated as adding constraints.
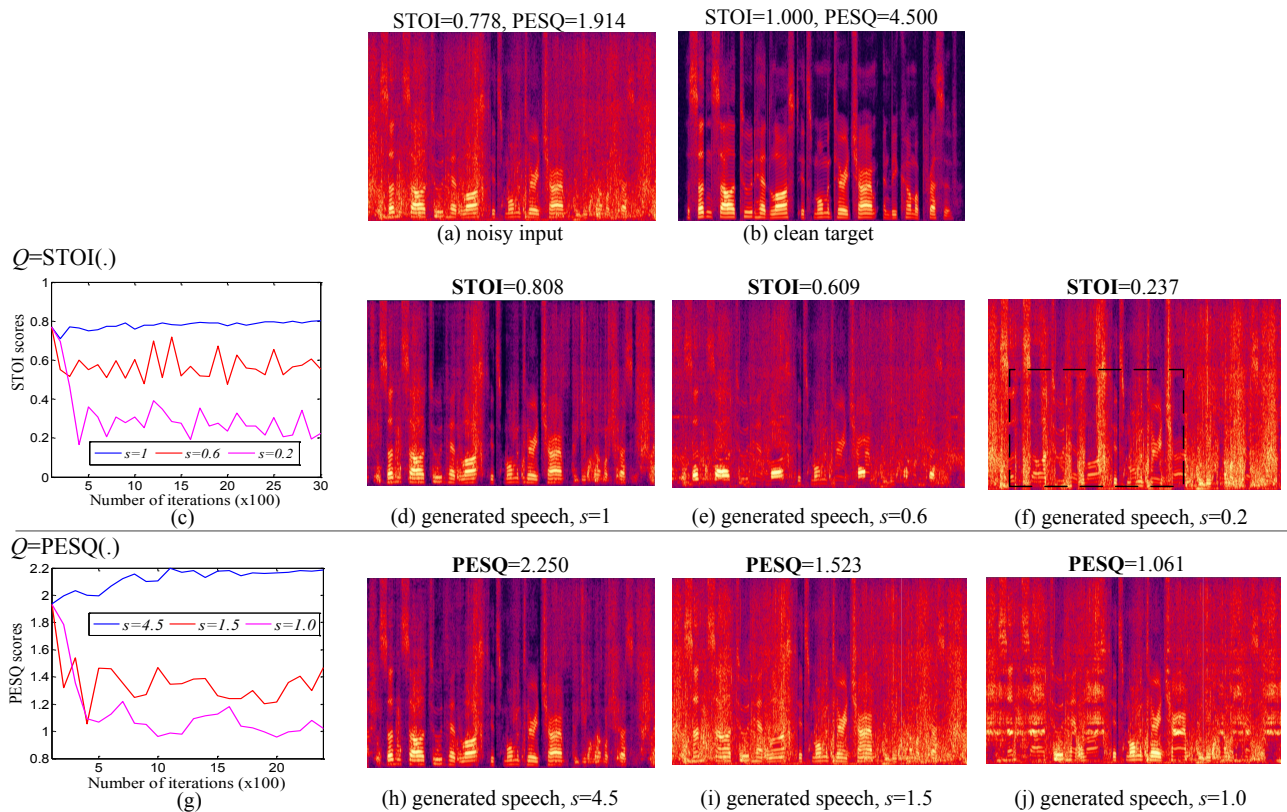
STOI=0.778, PESQ=1.914   STOI=1.000, PESQ=4.500

(a) noisy input   (b) clean target

Q=STOI(.)

STOI=0.808   STOI=0.609   STOI=0.237

(c)   (d) generated speech, s=1   (e) generated speech, s=0.6   (f) generated speech, s=0.2

Q=PESQ(.)

PESQ=2.250   PESQ=1.523   PESQ=1.061

(g)   (h) generated speech, s=4.5   (i) generated speech, s=1.5   (j) generated speech, s=1.0

Figure 5. Results of assigning different s to Eq. (5) for the generator training. Note that the learning curves of generating clean speech in (c) and (g) are not yet converged. For more complete learning processes, please refer to Figure 2.

Algorithm 1 Multi-Metric Scores Assignment

Input: desired score s_1 for metric $Q'_1(.)$ to s_N for metric $Q'_N(.)$ (assume there are N different metrics).
repeat
    1) Find metric index i with the largest distance between achieved and assigned score:
$$i = \operatorname{argmax}_n |Q'_n(G(x), y) - s_n|$$
    2) Train G to minimize the loss from D_i:
$$L_{G(MetricGAN)} = \mathbb{E}_x[(D_i(G(x), y) - s_i)^2]$$
    3) Train all D_n to minimize the distance from Q_n:
$$L_{D(MetricGAN)} = \mathbb{E}_{x,y}[(D_n(y, y) - 1)^2$$
$$+ (D_n(G(x), y) - Q'_n(G(x), y))^2]$$
until converge

Algorithm 1 shows the proposed training method for multi-metric scores assignment. Assuming that there are N different metrics, we have to employee N discriminators. In each iteration, only D with the largest distance between achieved score, $Q'_n(G(x), y)$, and assigned score, s_n, would guide the learning of G (steps 1 and 2). However, in the training of D, all the discriminators D_n are updated, irrespective of

whether it is used to provide loss to G (step 3).

Figure 6 shows the learning curves for the case of N=2. To explore more possible combinations, these results are based on the subset (top 10% metric score) of the original validation set. To clearly illustrate the results of multi-metric learning, in each column of this figure, the assignment of STOI score is fixed with different PESQ scores. Because different metrics may have some positive correlation between each other, MetricGAN is difficult to converge when the score assignments are too extreme (in this case, the solution may not even exist). However, we still obtain some flexibility to generate speech with desired multiple scores. This experiment verifies that MetricGAN can approximate and distinguish different metrics well.

4.3. Comparison with Other State-of-the-Art SE Models

To further compare the proposed MetricGAN with other state-of-the-art methods, we use a publicly available dataset released by (Valentini-Botinhao et al., 2016). This dataset contains a large amount of pre-mixed noisy-clean paired data and is already used by several SE models. By using the
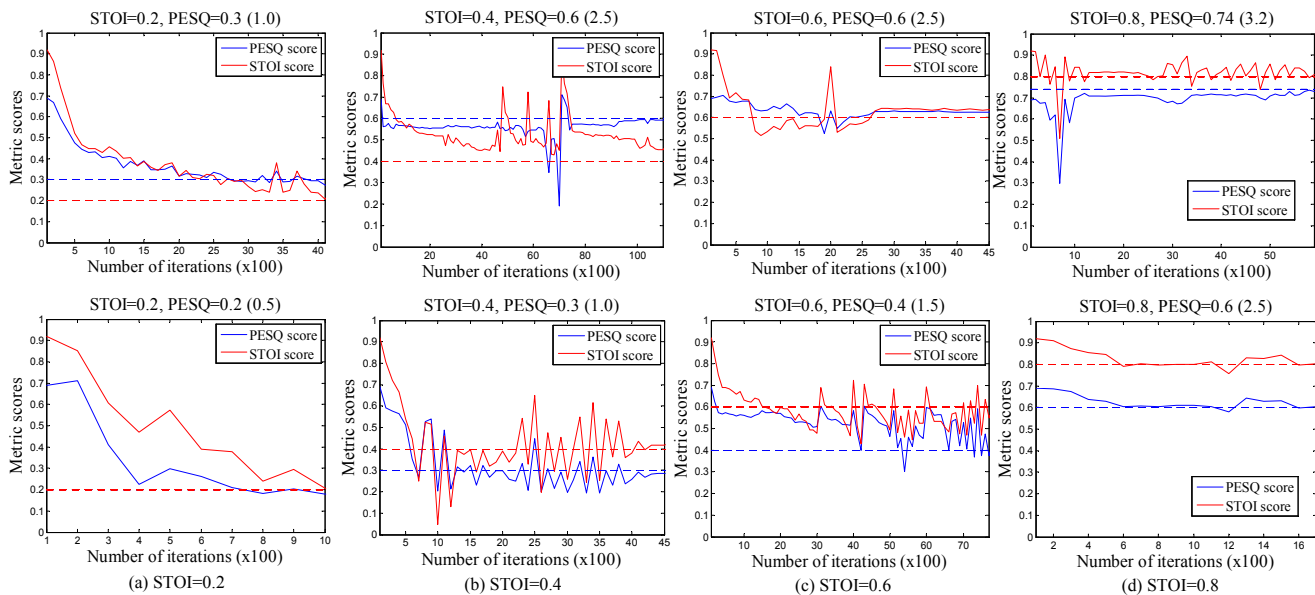
Figure 6. Learning curves of assigning different pairs of (STOI, PESQ) scores (shown in the title of each figure). Given a specified STOI score, the upper row and lower row is the maximum and minimum PESQ scores MetricGAN can reach, respectively. Note that the PESQ score is normalized ((original score+0.5)/5) between 0 to 1 with the original score shown in the parentheses.

Table 2. Compared MetricGAN with other state-of-the-art methods. The highest score per metric is highlighted with bold text.

	PESQ	CSIG	CBAK	COVL
Noisy	1.97	3.35	2.44	2.63
SEGAN	2.16	3.48	2.94	2.80
MMSE-GAN	2.53	3.80	3.12	3.14
WGAN-GP	2.54	-	-	-
Deep Feature Loss	-	3.86	**3.33**	3.22
SERGAN	2.62	-	-	-
MetricGAN (P)	**2.86**	**3.99**	3.18	**3.42**

exact same training and test dataset split, we can establish a fair comparison with them easily.

Experimental Setup and Results: Details about the data can be found in the original paper. Except for input features and activation functions, the network architecture and training strategy are the same as described in the previous section. In addition to the PESQ score, we also report another three metrics over the test set to compare with previous works: **CSIG** predicts the mean opinion score (MOS) of the signal distortion, **CBAK** predicts the MOS of the background noise interferences, and **COVL** predicts the MOS of the overall speech quality, these three metrics range from 1 to 5.

Five baseline models that rely on another network to provide loss information are compared with the proposed **MetricGAN (P)**. We briefly explain these models as follows: **SEGAN** (Pascual et al., 2017) directly operates on the raw

waveform and the model is trained to minimize the combination of adversarial and L_1 losses. **MMSE-GAN** (Soni et al., 2018) is a time-frequency masking-based method that uses a GAN objective along with L_2 loss. Similar to the structure of **SEGAN**, **WGAN-GP** and **SERGAN** (Baby & Verhulst, 2019) introduced Wasserstein loss and relativistic least-square loss for GAN training, respectively. Finally, **Deep Feature Loss** (Germain et al., 2018) also operates on the raw waveform and is trained with a deep feature loss from another network that classifies acoustic environments. Table 2 summarizes that our proposed method outperforms all previous works with respect to three metrics. This implies that although **MetricGAN** is only trained to optimize a certain score (PESQ), it also has a great generalization ability to other metrics.

5. Conclusion

In this paper, we proposed a novel MetricGAN approach to directly optimize generators based on one or multiple evaluation metric scores. By associating a discriminator with the metrics of interest, MetricGAN can be treated as an iterative process between surrogate loss learning and generator learning. This surrogate can successfully capture the behavior of the metrics and provides accurate gradients guiding the generator updates. In addition to outperforming other loss functions and state-of-the-art models in SE, MetricGAN can also be trained to generate data according to the designate metric scores. To the best of our knowledge, this is the first work that employs GAN to directly train the generator with respect to multiple evaluation metrics.

References

Baby, D. and Verhulst, S. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 106–110, 2019.

Bagchi, D., Plantinga, P., Stiff, A., and Fosler-Lussier, E. Spectral feature mapping with mimic loss for robust speech recognition. *arXiv preprint arXiv:1803.09816*, 2018.

Benesty, J., Makino, S., and Chen, J. *Speech Enhancement*. Berlin, Germany: Springer, 2005.

Chai, L., Du, J., and Lee, C.-H. Error modeling via asymmetric laplace distribution for deep neural network based single-channel speech enhancement. In *Interspeech*, pp. 3269–3273, 2018.

Donahue, C., Li, B., and Prabhavalkar, R. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5024–5028, 2018.

Fu, S.-W., 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 *Interspeech*, 2018a.

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):1570–1584, 2018b.

Fu, S.-W., Liao, C.-F., and Tsao, Y. Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality. *arXiv preprint arXiv:1905.01898*, 2019.

Garofolo, J. S. et al. Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database. *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, 107:16, 1988.

Germain, F. G., Chen, Q., and Koltun, V. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Grais, E. M., Ward, D., and Plumbley, M. D. Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders. *arXiv preprint arXiv:1803.00702*, 2018.

Hsieh, C.-Y., Lin, Y.-A., and Lin, H.-T. A deep model with local surrogate loss for general cost-sensitive multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Hu, G. 100 nonspeech environmental sounds, 2004.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koizumi, Y., Niwa, K., Hioka, Y., Kobayashi, K., and Haneda, Y. Dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 81–85, 2017.

Koizumi, Y., Niwa, K., Hioka, Y., Koabayashi, K., and Haneda, Y. Dnn-based source enhancement to increase objective sound quality assessment score. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

Kolbæk, M., Tan, Z.-H., and Jensen, J. Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure. *arXiv preprint arXiv:1802.00604*, 2018.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

Liao, C.-F., Tsao, Y., Lee, H.-Y., and Wang, H.-M. Noise adaptive speech enhancement using domain adversarial training. *arXiv preprint arXiv:1807.07501*, 2018.

Lu, X., Tsao, Y., Matsuda, S., and Hori, C. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pp. 436–440, 2013.

Luo, Y. and Mesgarani, N. Tasnet: Surpassing ideal time-frequency masking for speech separation. *arXiv preprint arXiv:1809.07454*, 2018.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Smolley, S. P. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821, 2017.

Martín-Doñas, J. M., Gomez, A. M., Gonzalez, J. A., and Peinado, A. M. A deep learning loss function based on the perceptual evaluation of the speech quality. *IEEE Signal Processing Letters*, 25(11):1680–1684, 2018.

Michelsanti, D. and Tan, Z.-H. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. *arXiv preprint arXiv:1709.01703*, 2017.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Naithani, G., Nikunen, J., Bramslow, L., and Virtanen, T. Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 386–390, 2018.

Narayanan, A. and Wang, D. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7092–7096, 2013.

Ochiai, T., Watanabe, S., Hori, T., and Hershey, J. R. Multi-channel end-to-end speech recognition. In *International Conference on Machine Learning (ICML)*, pp. 2632–2641, 2017.

Pandey, A. and Wang, D. On adversarial training and loss functions for speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5414–5418, 2018.

Pascual, S., Bonafonte, A., and Serra, J. Segan: Speech enhancement generative adversarial network. In *Interspeech*, 2017.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pp. 749–752, 2001.

Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., and Schmidhuber, J. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.

Soni, M. H., Shah, N., and Patil, H. A. Time-frequency masking-based speech enhancement using generative adversarial network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

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):2125–2136, 2011.

Valentini-Botinhao, C., Wang, X., Takaki, S., and Yamagishi, J. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA Speech Synthesis Workshop*, pp. 146–152, 2016.

Venkataramani, S. and Smaragdis, P. End-to-end networks for supervised single-channel speech separation. *arXiv preprint arXiv:1810.02568*, 2018.

Venkataramani, S., Higa, R., and Smaragdis, P. Performance based cost functions for end-to-end speech separation. *arXiv preprint arXiv:1806.00511*, 2018.

Wang, D. and Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., and Tang, X. Esrgan: Enhanced super-resolution generative adversarial networks. *arXiv preprint arXiv:1809.00219*, 2018.

Wang, Y., Narayanan, A., and Wang, D. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 22(12):1849–1858, 2014.

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 *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 91–99, 2015.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1):65–68, 2014.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19, 2015.

Zhang, H., Zhang, X., and Gao, G. Training supervised speech separation system to improve stoi and pesq directly. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5374–5378, 2018.

Zhao, Y., Xu, B., Giri, R., and Zhang, T. Perceptually guided speech enhancement using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5074–5078, 2018a.

Zhao, Z., Liu, H., and Fingscheidt, T. Convolutional neural networks to enhance coded speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018b.