MetricGAN-OKD: Multi-Metric Optimization of MetricGAN via Online Knowledge Distillation for Speech Enhancement

Wooseok Shin [1] **Byung Hoon Lee** [1] **Jin Sob Kim** [1] **Hyun Joon Park** [1] **Sung Won Han** [1]

Abstract

In speech enhancement, MetricGAN-based approaches reduce the discrepancy between the L_p loss and evaluation metrics by utilizing a non-differentiable evaluation metric as the objective function. However, optimizing multiple metrics simultaneously remains challenging owing to the problem of confusing gradient directions. In this paper, we propose an effective multi-metric optimization method in MetricGAN via online knowledge distillation—MetricGAN-OKD. MetricGAN-OKD, which consists of multiple generators and target metrics, related by a one-to-one correspondence, enables generators to learn with respect to a single metric reliably while improving performance with respect to other metrics by mimicking other generators. Experimental results on speech enhancement and listening enhancement tasks reveal that the proposed method significantly improves performance in terms of multiple metrics compared to existing multi-metric optimization methods. Further, the good performance of MetricGAN-OKD is explained in terms of network generalizability and correlation between metrics.

1. Introduction

Speech enhancement (SE) involves the improvement of the intelligibility and perceptual quality of human speech in noisy environments. SE is a crucial task in various speech-related applications, such as automatic speech recognition (ASR), teleconference systems, and hearing aids. Although recent deep learning-based models (Tan & Wang, 2018; Yin et al., 2020; Defossez et al., 2020; Hao et al., 2021; Park et al., 2022) have significantly improved SE performance, dissonance continues to exist between the evaluation metrics and the L_p distance loss (e.g., L_1 or L_2), typically used as the objective function. Existing studies have demonstrated that this problem arises because the L_p loss is not highly correlated with the evaluation metrics designed to consider human auditory perception (Bagchi et al., 2018; Fu et al., 2018; Zhao et al., 2018; Manocha et al., 2020). To reduce this dissonance, several studies (Fu et al., 2018; Zhao et al., 2018; Martin-Donas et al., 2018; Fu et al., 2019a; Kim et al., 2019) have directly optimized the evaluation metrics, such as the perceptual evaluation of speech quality (PESQ; Rix et al. 2001) and short-time objective intelligibility (STOI; Taal et al. 2011). Some (Fu et al., 2018; Zhao et al., 2018) have demonstrated that adopting the STOI metric as an objective function improves speech intelligibility. Others (Martin-Donas et al., 2018; Fu et al., 2019a; Kim et al., 2019) have utilized the metric function approximation or policy gradient approaches owing to the profound complexity and not completely differentiable nature of the PESQ metric (Fu et al., 2019b). However, the disadvantages of these methods include the following: (i) Lack of flexibility and scalability for application to other metrics. (ii) Inefficient training costs and limited performance improvements (Fu et al., 2019b; 2021).

Alternatively, generative adversarial network (GAN; Goodfellow et al. 2014)-based approaches (Pascual et al., 2017; Donahue et al., 2018), which train a generator using adversarial loss, have been proposed in SE. However, as in the case of the L_p loss, the conventional adversarial loss that considers the authenticity of a given sample is not consonant with the optimization of the evaluation metric (Fu et al., 2019b).

To overcome the aforementioned problems, MetricGAN (Fu et al., 2019b), which is a GAN-based architecture that utilizes non-differentiable evaluation metrics as objective functions with efficient cost, was proposed. MetricGAN consists of a surrogate function (discriminator) that learns the behavior of the metric function and a generator that generates enhanced speech based on the guidance of the discriminator. Following demonstrations verifying that MetricGAN optimizes metric scores directly, optimization of multiple metrics to improve different metrics representing

[1]School of Industrial and Management Engineering, Korea University, Seoul, Republic of Korea. Correspondence to: Sung Won Han <swhan@korea.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

various aspects of human auditory perception has been attempted. Some studies (Li et al., 2020; Li & Yamagishi, 2021) have proposed multi-metric optimization versions of MetricGAN in listening enhancement (LE) tasks that improve speech intelligibility on the listener's side. They employed multiple nodes in the final layer of the discriminator corresponding to the number of target metrics to be used for optimization. Although these studies have demonstrated the potential of multi-metric optimization, simultaneous performance improvements are still limited. For example, a single-target metric model often performs better in terms of a particular metric than a model trained on multiple metrics.

This phenomenon is attributed to the one-to-many matching between a generator and multiple metrics. In the context of SE tasks, several metrics, such as PESQ, STOI, speech intelligibility in bits (SIIB; Van Kuyk et al. 2017), hearing-aid speech perception index (HASPI; Kates & Arehart 2014), and extended STOI (ESTOI; Jensen & Taal 2016), have been proposed to reflect various aspects of human auditory perception. As these metrics have inherently different properties, discriminators trained using multiple metrics may provide a single generator with confusing gradient directions that limit stable optimization and network generalizability.

From another perspective, as the discriminator of MetricGAN serves as the loss function, multi-metric optimization can be regarded as a multi-task or multi-objective optimization problem. In general, multi-objective problems involve conflicting objectives (Sener & Koltun, 2018). A standard compromise is to optimize a linear combination of multiple losses. However, although this strategy functions well when multiple objectives do not compete with each other (Sener & Koltun, 2018), in our case, multiple objectives can conflict owing to the differences between the properties of various metrics when a generator is trained in terms of multiple metrics directly.

In this paper, we propose an effective multi-metric optimization method for MetricGAN via online knowledge distillation (MetricGAN-OKD) to improve the performance in terms of all target metrics. To mitigate the aforementioned problem involving confusing gradient directions, we design a special OKD learning scheme, which consists of a one-to-one correspondence between generators and target metrics. In particular, each generator learns from the gradient of each discriminator trained using a single target metric for stability. Subsequently, other metrics are improved by transferring knowledge of other generators trained on different metrics to the target generator. This strategy enables stable multi-metric optimization, where the generator learns the target metric from a single discriminator easily and improves multiple metrics by mimicking other generators.

Extensive experiments on SE and LE tasks reveal that the proposed MetricGAN-OKD outperforms existing single- and multi-metric optimization methods significantly. In particular, compared to causal models, the models trained using the proposed method outperform state-of-the-art models in terms of PESQ and COVL using 23 and 82 times fewer parameters and floating point operations (FLOPs), respectively. Besides quantitative evaluation, we explain the success of MetricGAN-OKD in terms of high network generalizability and the correlation between different metrics.

2. Related Work

2.1. MetricGAN

GAN is a representative generative method and has been adopted in various domains. In SE, SEGAN (Pascual et al., 2017) and FSEGAN (Donahue et al., 2018) utilize the GAN structure, which provides real or fake labels (i.e., 0 or 1) while training the discriminator. Although these methods improve SE performance, the use of discrete labels to train discriminators may not be completely relevant to the optimization of evaluation metrics (Fu et al., 2019b). In addition, previous studies (Fu et al., 2018; Zhao et al., 2018) have demonstrated that L_p loss, which is widely used as a loss function, does not optimize evaluation metrics designed to reflect human auditory perception satisfactorily. To resolve these problems, MetricGAN utilizes evaluation metric scores (continuous) as labels during discriminator training. In this way, the discriminator acts as a surrogate function similar to the evaluation metric, while the generator learns to optimize the evaluation metric directly using a gradient direction of the discriminator. In a follow-up study, Kawanaka et al. (2020) proposed an improved cost function that includes noisy speech during discriminator training to train MetricGAN stably. Additionally, MetricGAN+ (Fu et al., 2021) was proposed by improving MetricGAN using several training techniques.

2.2. Multi-MetricGAN Methods

In MetricGAN (Fu et al., 2019b), the authors also explored the optimization of multiple metrics. They employed a single generator and N discriminators corresponding to N distinct target metrics. Each discriminator learns from a single metric, and the generator is updated by the discriminator based on the largest loss in each iteration. Although the conclusions of the aforementioned paper imply that multi-metric optimization is possible in SE, it indicates the difficulty of simultaneous optimization under extreme conditions (e.g., in the absence of high positive correlation between the evaluation metrics; Fu et al. 2019b).

Li et al. (2020) proposed iMetricGAN, a modified version of MetricGAN, for LE. Similar to MetricGAN, iMetricGAN consists of a generator and a discriminator; but its discriminator has multiple output nodes corresponding to

the number of target metrics. In addition, Li & Yamagishi (2021) proposed a new framework for LE consisting of a generator, a multi-node intelligibility discriminator, and a multi-node quality discriminator to improve performance in terms of both intelligibility and quality metrics.

2.3. Online Knowledge Distillation

Knowledge distillation (KD; Hinton et al. 2015), which was first proposed in the field of image recognition, improves compact student networks using knowledge gained from large teacher networks. KD has also been widely used to improve the performance of student networks in SE tasks (Nakaoka et al., 2021; Shin et al., 2022). Online knowledge distillation (OKD; Zhang et al. 2018; Guo et al. 2020), a practical variant of KD, performs mutual learning among student models during the training phase, instead of a one-sided knowledge transfer from a pre-trained teacher network to a student network. Specifically, OKD begins training with initialized student networks, and each student is trained using standard cross entropy (CE) loss as well as a mimicry loss (e.g., KL loss–Kullback-Leibler divergence loss) that penalizes the difference between predictions of student networks. In terms of efficacy of KD and OKD, several studies (Pereyra et al., 2017; Zhang et al., 2018) have demonstrated that KD can help a network converge to flat minima, improving its generalizability.

3. Methodology

This section describes the overall training flow and the formulation of general MetricGAN comprising a discriminator (D) and a generator (G). The MetricGAN-OKD method is also introduced here.

3.1. Overall Workflow

The goal of the standard SE model is to remove background noise n from noisy speech x. The noisy speech x is defined as a mixture of clean speech y and background noise n. In MetricGAN-based method, the SE model G receives noisy speech (spectrogram) x as input and outputs enhanced speech $\hat{y} \approx y$. This can be formulated as follows:

$$x = y + n, \qquad \hat{y} = G(x) \tag{1}$$

1. **Discriminator Flow** MetricGAN (Fu et al., 2019b) introduces a metric function (Q) to train D, and D is learned to approximate Q. As depicted in Figure 1b, to this end, D receives inputs in the same way that Q receives predictions and labels for metric calculations (i.e., $D(\hat{y}, y)$ and $Q(\hat{y}, y)$). D is then trained to minimize the difference between D and Q outputs.

2. **Generator – Standard Flow** The dotted blocks of Figure 1a illustrate the standard flow of the training process of G (red arrow). Given noisy speech, G produces enhanced speech from it. The enhanced speech and the corresponding clean speech are then transmitted to D to obtain prediction scores. G is trained to reduce the difference between the prediction scores and the allocated scores, s. To produce clean speech, s is set to the maximum score of the evaluation metrics.

3. **Generator – Distillation Flow** Given multiple Gs trained in terms of different metrics, enhanced speeches (spectrograms) are obtained from each G based on a singular input noisy speech. Each G mimics the enhanced speeches of other Gs to improve performance in terms of other metrics, as depicted in the OKD flow presented in Figure 1a (blue arrow). To match the enhanced speeches, we use standard L_2 loss.

All Gs and Ds are optimized alternately in each epoch until convergence. Our training algorithm is presented in Appendix A.

3.2. MetricGAN

Discriminator In MetricGAN, the goal of D is to serve as a surrogate function of Q by imitating it. Following previous studies (Kawanaka et al., 2020; Fu et al., 2021), we formulate the objective function of D as follows:

$$L_D = \mathbb{E}_{x,y}[(D(y,y) - Q'(y,y))^2 + (D(x,y) - Q'(x,y))^2 + (D(G(x),y) - Q'(G(x),y))^2], \tag{2}$$

where $Q'(\cdot)$ refers to the target metric function normalized to [0, 1]. The three terms in Equation (2) are used to minimize the difference between $D(\cdot)$ and $Q'(\cdot)$ based on clean, noisy, enhanced speech, respectively. In this way, the three terms enable D to learn the quality distribution of the signal.

Generator In MetricGAN, G, i.e., the enhancement network, receives noisy speech and generates enhanced speech. Then, D acts as a differentiable surrogate function by receiving the enhanced and clean speeches as input and generating scores that mimic metric scores. G only utilizes a gradient direction provided by D because the adversarial loss of D is more effective than L_p loss. The objective function is defined as follows:

$$L_G = \mathbb{E}_x[(D(G(x),y) - s)^2], \tag{3}$$

where s denotes the score that determines the level of enhanced speech produced by G. s is simply set to 1 to generate clean speech. In this way, G is trained to achieve the highest possible metric scores.

3.3. MetricGAN via Online Knowledge Distillation

To avoid the problem of confusing gradient directions, we propose MetricGAN-OKD, which consists of multiple Gs

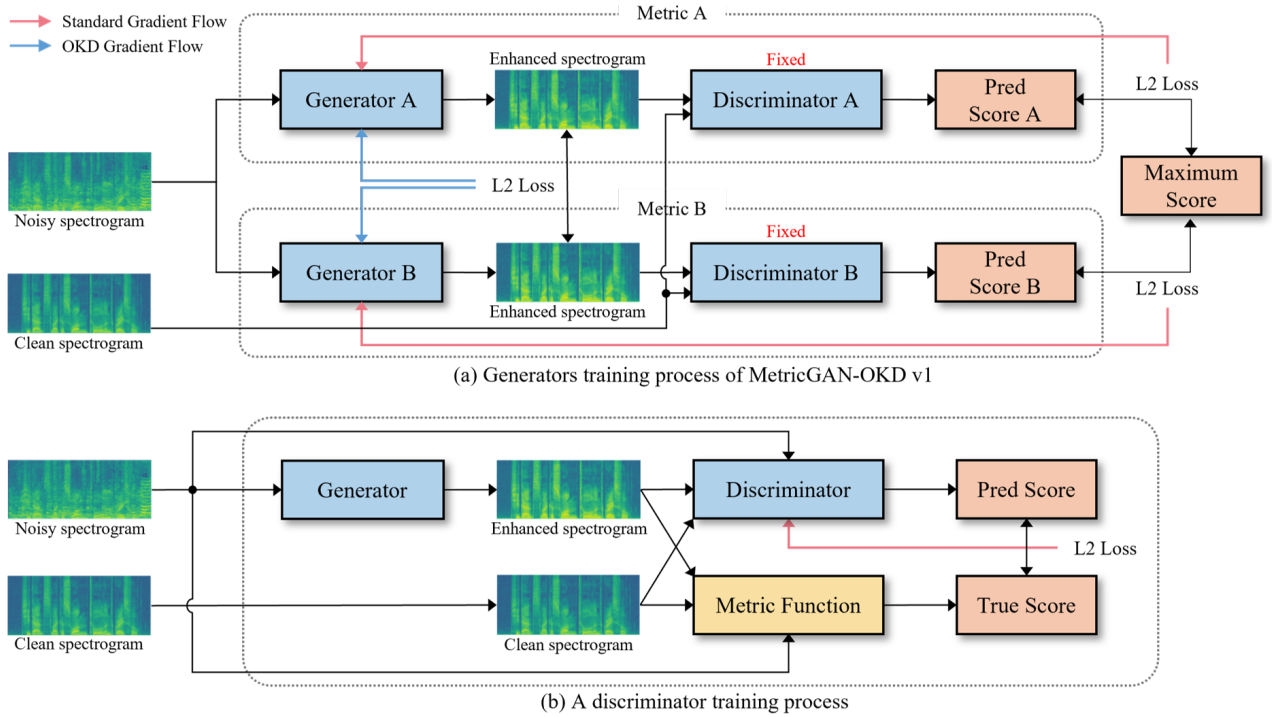


Figure 1. The overall framework of MetricGAN-OKD v1. For simplicity, (a) represents the generator training process with respect to two target metrics, and (b) represents the training process of each discriminator.

Table 1. The number of generators, discriminators, nodes in the discriminator, and target metrics corresponding to different methods. MGAN denotes MetricGAN.

Method	# Gs	# Ds	# D nodes	# target metrics
MGAN(+) (Fu et al., 2019b; 2021)	1	1	1	1
Multi-D MGAN (Fu et al., 2019b)	1	n	1	n
Multi-node MGAN (Li et al., 2020)	1	1	n	n
MGAN-OKD v1 (ours)	n	n	1	n
MGAN-OKD v2 (ours)	n	1	n	n

and target metrics related by a one-to-one correspondence. Table 1 lists the differences between the proposed framework and existing methods. The first row corresponds to MetricGAN(+), a single-metric optimization method. The other rows correspond to multi-metric optimization methods. Existing multi-metric methods, Multi-D and Multi-node MetricGAN, use a one-to-many correspondence, consisting of a single G and multiple target metrics. In particular, Multi-D MetricGAN adopts a single G and multiple Ds, and each D learns from each target metric. Multi-node MetricGAN adopts a single G and a single D; but, in this case, D contains multiple output nodes corresponding to the number of target metrics. Thus, in summary, in existing multi-metric methods, a single G is directly trained in terms of multiple target metrics, which may induce unstable optimization due to conflicting directions of multiple metrics gradients. We hypothesize that this can lead to sharpness of the loss landscape and degrade the generalization performance of the network (which will be explored in Section 5.2).

We propose the strategy that enables indirect training with respect to multiple metrics addressing the aforementioned issues by performing multi-metric optimization using KD. To apply KD to a multi-metric optimization pipeline, the following typical KD approach may be considered: a network is trained in terms of a single metric to be used as a pre-trained teacher network, and, subsequently, a student network is trained with respect to another metric while mimicking the predictions of the teacher network (Refer to Appendix D.1 for related experiments). However, this approach suffers from the disadvantage of being at least a two-stage training process. Thus, its complexity increases rapidly as the number of target metrics to be optimized increases.

To achieve stable multi-metric optimization and an efficient training process simultaneously, we design an OKD learning scheme with a one-stage training process and comparable efficacy to KD. Based on the design of D in previously proposed multi-metric methods, we propose two versions of MetricGAN-OKD: **(i) Multiple *G*s–Multiple *D*s** (Refer to Figure 1). **(ii) Multiple *G*s–Multi-node *D*** (Refer to Figure 4 in Appendix B). We assume that there are N different target metrics. The first version (v1) consists of N different Gs and Ds related by a complete one-to-one correspondence. Therefore, each G and D are trained in the

same way as in the case of a single MetricGAN, which is given by Equations (2) and (3), except for the target metric type. The second version (v2) is composed of N different Gs and a single D with N output nodes. The single D is simultaneously trained in terms of multiple metrics, and each node in the final output learns from each target metric score. To construct a one-to-one correspondence, we design the framework in such a way that the single G learns from the gradient of one node representing one metric. In both versions, the objective function of each D is defined as:

$$L_{D_i} = \mathbb{E}_{x,y}[(D_i(y,y) - Q'_i(y,y))^2 + (D_i(x,y) - Q'_i(x,y))^2 + (D_i(G_i(x),y) - Q'_i(G_i(x),y))^2], \quad (4)$$

where $1 \leq i \leq N$, and D_i denotes the i^{th} D in version 1 and the i^{th} node of the output in version 2.

In terms of G, the training process is divided into the standard and OKD flows, as depicted in Figure 1a. In the standard flow, G_i learns from the gradient of D_i (red arrow in Figure 1a) trained on i^{th} single metric as follows:

$$L_{G_i(standard)} = \mathbb{E}_x[(D_i(G_i(x),y) - s_i)^2] \quad (5)$$

In the OKD flow, G_i mimics the G_j trained using different target metrics by minimizing the element-wise difference between enhanced spectrograms, $G_i(x)$ and $G_j(x)$, as indicated by the blue arrow in Figure 1a. Unlike the classification tasks that use logits for distillation, we use spectrogram-level outputs. The objective function is defined as follows:

$$L_{G_i(OKD)} = \mathbb{E}_x\left[\sum_{j=1}^{N} |G_i(x) - G_j(x)|^2\right], \quad \text{if } i \neq j \quad (6)$$

The total loss of each generator G_i is defined by combining the standard and OKD flow losses as follows:

$$L_{G_i} = L_{G_i(standard)} + \alpha \times L_{G_i(OKD)}, \quad (7)$$

where α denotes the weight of the OKD flow. In this way, our Gs perform multi-metric optimization stably without succumbing to the problem of confusing gradient directions.

The efficacy of MetricGAN-OKD can be explained intuitively as follows. In the standard flow, the gradient of D leads G along the direction in which only the target metric loss is minimized—the improvement of other metrics is not considered. On the other hand, spectrograms of OKD flow contain information related to various metrics in addition to the target metric (Refer to Table 5 for related experiments). Therefore, mimicking the spectrogram can be considered to be similar to KL loss that increases the posterior entropy by matching class probabilities between teacher and student networks. In image classification, several studies (Pereyra et al., 2017; Zhang et al., 2018; Chaudhari et al., 2019) have demonstrated that high posterior entropy aids convergence of networks to flat minima, improving their generalizability. Based on this fact, we explore the generalizability of MetricGAN-OKD in Section 5.2.

4. Experiments

To validate the proposed method, experiments were conducted on SE and LE tasks.

4.1. Common Experimental Configuration

Baselines To validate the effect of MetricGAN-OKD, the performance of the proposed method is compared with those of the existing single- and multi-metric optimization methods listed in Table 1. For simplicity, single-target MetricGAN defined with respect to a single metric is denoted by metric name-GAN (e.g., PESQ-GAN or SIIB-GAN). To ensure fair comparison, all baselines and training schemes are reimplemented in PyTorch and all experiments are performed using the same environment and hyperparameters.

Implementation details For architecture, the G and D structures proposed in MetricGAN+ (Fu et al., 2021) are adopted in both SE and LE. G consists of two bidirectional-LSTM (Weninger et al., 2015) layers and two fully-connected layers. D is composed of four 2D convolution layers, an average-pooling layer, and three fully-connected layers. We refer to Appendix E for the architecture details. For other setups, we use Adam optimizer with a learning rate of 0.0005 for G and D. The number of samples per epoch and history portion of the replay buffer (Refer to Appendix A.1) are set to 100 and 0.2, respectively, as in MetricGAN+ (Fu et al., 2021). The batch size is set to one, and the signal is downsampled to 16 kHz. An input spectrogram with 257 frequency bins is generated using window and hop lengths of 512 and 256, respectively. Moreover, the total number of epochs of SE and LE are set to 750 and 100, respectively. Finally, the value of the loss function weight, α, is set to ten and five in SE and LE tasks, respectively.

4.2. Speech Enhancement

The goal of the SE task is to remove background noise from noisy speech. MetricGAN was introduced in the context of this task. However, to the best of our knowledge, multi-metric optimization has not been extensively investigated in SE. Thus, the following extensive experiments are performed in SE.

Dataset The VoiceBank-DEMAND dataset (Valentini-Botinhao et al., 2017) is used for validation. The VoiceBank-DEMAND dataset consists of 30 speakers, of which 28 are used to comprise the training/validation dataset and the remaining two are used to comprise the test dataset. The training/validation and test datasets contain 11,572 utter-

Table 2. Evaluation results on VoiceBank-DEMAND dataset. PESQ and CSIG are selected as target metrics for optimization. For all metrics, higher values are better.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Single MGAN	PESQ (PE)	-	3.13	4.13	3.03	3.61
	CSIG (CS)	-	3.11	4.23	3.07	3.68
Multi-D MGAN	PE+CS	-	3.14	4.21	3.09	3.67
Multi-node MGAN	PE+CS	-	3.13	4.22	**3.15**	3.68
MGAN-OKD v1 (ours)	PE	CS	3.20	4.23	3.11	3.71
	CS	PE	3.12	4.25	3.10	3.68
MGAN-OKD v2 (ours)	PE	CS	**3.24**	4.23	3.07	**3.73**
	CS	PE	3.19	**4.26**	3.12	3.72

ances with four signal-to-noise ratio (SNR) (15, 10, 5, and 0 dB) levels and 824 utterances with four SNR (17.5, 12.5, 7.5, and 2.5 dB) levels, respectively.

Target and evaluation metrics Identical target and evaluation metrics are adopted for training and evaluation. PESQ and composite measures (CSIG, CBAK, and COVL; Hu & Loizou 2007), which are widely used in SE, are selected for this purpose. CSIG, CBAK, and COVL are used to verify the signal distortion, intrusiveness of noise, and overall quality of the signal, respectively. The PESQ score ranges from -0.5 to 4.5. Values of the three composite measures range from 1 to 5. Note that all scores are normalized to [0, 1] during training.

Results Two main experiments are conducted with [PESQ, CSIG] and [PESQ, CSIG, CBAK, COVL] as the target metrics. As reported in Table 2, Multi-D and Multi-node MetricGAN exhibit higher CSIG scores and comparable PESQ scores compared to the PESQ-GAN. However, compared to the CSIG-GAN, they exhibit lower CSIG scores, which indicates that the multi-metric optimization performed by the existing methods is insufficient. On the other hand, the proposed methods are observed to outperform PESQ- and CSIG-GAN significantly in terms of both PESQ and CSIG. In particular, models of OKD v2 achieve performance improvements with large margins in the PESQ score. Moreover, Table 3 presents the results in the case involving four target metrics. Similar to the results listed in Table 2, previously proposed multi-metric methods exhibit limited improvement in multi-metric performance, whereas the proposed methods achieve significant improvement. Finally, the proposed method generates as many models as the number of target metrics, enabling the selection of a single preferred model depending on the task during the inference phase. We refer to Appendix D for further experiments.

Comparison with State-of-the-art Causal Methods In SE, causality is an essential condition for real-time operation, and causality-based models should rely only on inputs on steps prior to the current step, instead of relying on future information. The efficacy of the proposed method for causality-based models is thus experimentally verified. To this end, bidirectional-LSTM layers are modified to unidirectional-LSTM layers in the G architecture described in Section 4.1. Table 4 lists the experimental results obtained using MetricGAN-based and state-of-the-art non-MetricGAN methods. The proposed method exhibits significant improvement in overall performance compared to existing multi-MetricGAN methods. Further, the models trained using the proposed method are observed to outperform DEMUCS and CleanUNet in terms of PESQ and COVL with significantly fewer parameters and FLOPs.

Table 3. Evaluation results on VoiceBank-DEMAND dataset. PESQ, CSIG, CBAK, and COVL are selected as target metrics for optimization.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Single MGAN	PESQ (PE)	-	3.13	4.13	3.03	3.61
	CSIG (CS)	-	3.11	4.23	3.07	3.68
	CBAK (CB)	-	2.97	4.10	**3.32**	3.53
	COVL (CO)	-	3.12	4.20	3.14	3.69
Multi-D MGAN	PE+CS+CB+CO	-	3.12	4.18	3.25	3.65
Multi-node MGAN	PE+CS+CB+CO	-	3.09	4.22	3.24	3.67
MGAN-OKD v1 (ours)	PE	CS+CB+CO	3.14	4.22	3.24	3.67
	CS	PE+CB+CO	3.11	4.25	3.25	3.67
	CB	PE+CS+CO	2.98	4.12	**3.32**	3.53
	CO	PE+CS+CB	3.13	**4.29**	3.23	**3.71**
MGAN-OKD v2 (ours)	PE	CS+CB+CO	**3.15**	4.26	3.25	**3.71**
	CS	PE+CB+CO	3.12	4.26	3.22	3.69
	CB	PE+CS+CO	3.10	4.26	3.30	3.69
	CO	PE+CS+CB	3.13	4.27	3.24	**3.71**

Table 4. Comparison with state-of-the-art causality-based methods on VoiceBank-DEMAND dataset. MetricGAN-based methods are trained using PESQ, CSIG, and CBAK as target metrics. The three rows of MGAN-OKD represent models trained with PESQ, CSIG, and CBAK as the main metric, respectively. FLOPs are computed using a duration of four seconds at 16 kHz.

Method	Params	FLOPs	PESQ	CSIG	CBAK	COVL
Non-MetricGAN methods						
DEMUCS (Defossez et al., 2020)	18.87M	34.56G	2.93	4.22	3.25	3.52
CleanUNet (Kong et al., 2022)	39.77M	104.7G	2.88	**4.32**	**3.41**	3.63
MetricGAN-based methods						
Multi-D MGAN	0.82M	0.42G	2.98	4.14	3.21	3.56
Multi-node MGAN	0.82M	0.42G	2.98	4.13	3.18	3.55
MGAN-OKD v2 (ours)	0.82M	0.42G	**3.12**	4.17	3.13	**3.64**
			3.07	4.21	3.14	**3.64**
			3.00	4.12	3.22	3.56

Ablation Study To investigate the reasons why the proposed method enables stable optimization despite the presence of multiple loss terms (i.e., standard and OKD terms), additional experiments are conducted by employing identical metrics as multiple target metrics. In other words, the source of additional knowledge is investigated in the absence of information about different metrics. This configuration can be considered to be similar to the OKD of the image classification problem in that multiple models solve identical main tasks. The blocks in Table 5 present the results obtained using PESQ and CSIG as singular target metrics. As evidenced by both blocks, Multi-node MGAN performs comparably to single-metric MGAN, whereas the proposed method im-

Table 5. Ablation study on efficacy of identical target metrics on VoiceBank-DEMAND dataset. In this experiment, both output nodes of Multi-node MGAN are trained with respect to identical PESQ or CSIG metrics. In the proposed method, both models are trained using the identical PESQ or CSIG metric using the OKD scheme.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Single MGAN	PESQ (PE)	-	3.13	4.13	3.03	3.61
Multi-node MGAN	PE+PE	-	3.15	4.05	3.07	3.58
MGAN-OKD v2 (ours)	PE	PE	3.16	**4.16**	**3.17**	**3.66**
	PE	PE	**3.17**	4.15	**3.17**	3.65
Single MGAN	CSIG (CS)	-	3.11	4.23	3.07	3.68
Multi-node MGAN	CS+CS	-	3.11	4.25	3.12	3.69
MGAN-OKD v2 (ours)	CS	CS	**3.17**	4.26	3.13	**3.71**
	CS	CS	**3.17**	**4.27**	**3.16**	**3.71**

Table 6. Evaluation results on the Harvard Sentences dataset. SIIB, HASPI, and ESTOI are selected as target metrics for optimization. For all metrics, higher values are better. All values are averaged over five runs.

Method	Target Metric	KD Metric	SIIB	HASPI	ESTOI
Unmodified	-	-	25.42	2.51	0.313
Single MGAN	SIIB (S)	-	69.78	4.43	0.269
	HASPI (H)	-	65.50	4.49	0.291
	ESTOI (E)	-	58.57	4.05	0.355
Multi-D MGAN	S+H+E	-	67.71	4.42	0.332
Multi-node MGAN	S+H+E	-	71.07	4.53	0.328
MGAN-OKD v1 (ours)	S	H, E	**79.40**	**4.78**	0.382
	H	S, E	78.79	4.75	0.386
	E	S, H	79.17	4.76	**0.387**
MGAN-OKD v2 (ours)	S	H, E	76.32	4.68	0.374
	H	S, E	76.55	4.68	0.376
	E	S, H	74.50	4.60	0.373

proves performance with respect to different metrics, as well as the target metric, significantly. In particular, even in the absence of additional information about other metrics, a significant improvement in the performance of other metrics is observed. This suggests that spectrogram-level distillation transfers information related to various metrics. Therefore, thanks to this property, MetricGAN-OKD induces stable optimization even with multiple loss terms during training.

Table 7. Ablation study for the weight α of the OKD flow using MGAN-OKD v1 on Harvard Sentences dataset. Range denotes the maximum difference between the scores of the three models.

Target	KD	$\alpha = 1.0$			$\alpha = 5.0$			$\alpha = 10.0$		
		S	H	E	S	H	E	S	H	E
S	H, E	77.7	4.63	0.345	79.4	4.78	0.382	79.1	4.79	0.405
H	S, E	76.7	4.65	0.356	78.8	4.75	0.386	78.9	4.77	0.402
E	S, H	76.2	4.48	0.372	79.2	4.76	0.387	78.6	4.75	0.405
Range		1.5	0.17	0.027	0.6	0.03	0.005	0.5	0.04	0.003

4.3. Listening Enhancement

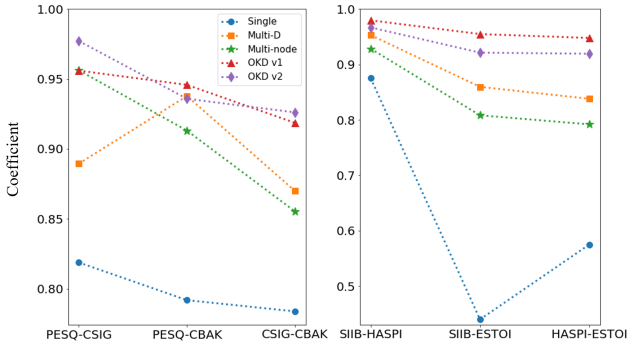
The goal of the LE task is to improve intelligibility by changing the input speech in situations where the noise on the listener side is stationary. We refer to Appendix C for formulating the LE problem.

Dataset In the LE experiment, two public English speeches (one male speaker; Valentini-Botinhao et al. 2019 and one female speaker; Demonte 2019), created using Harvard Sentences (Rothauser, 1969), are used. Harvard Sentences consists of 720 sentences, and sentences 1-600, 601-660, and 661-720 are used to comprise the training, validation, and test datasets, respectively. For the training and validation utterances, five types of noise (babble, station, restaurant, metro, and traffic) provided by the MS-SNSD (Reddy et al., 2019) dataset and three levels of SNRs (-11, -7, and -3 dB) are used. To evaluate the robustness of the proposed method, three unseen types of noise (cafeteria, neighbor, and airport announcement; Reddy et al. 2019) and unseen levels of SNRs (-13, -9, -5, and -1 dB) are used as the test set. In summary, 18,000, 1,800, and 1,440 utterances are used as training, validation, and test datasets, respectively.

Target and evaluation metrics For LE, we selected three intelligibility measures—SIIB, HASPI, and ESTOI. SIIB (Van Kuyk et al., 2017) is calculated based on the information shared between clean and degraded speech signals in bits per second. HASPI (Kates & Arehart, 2014) is based on the cepstral correlation and auditory model that integrates changes induced by hearing loss. HASPI v2 (Kates & Arehart, 2021), which is an improved version of HASPI, is used in this paper. ESTOI focuses on the spectral correlation between clean and degraded speech. The SIIB and HASPI scores range from 0 to $+\infty$, and the ESTOI score ranges from 0 to 1. For score normalization during training, logistic compression parameters used in existing study (Li & Yamagishi, 2021) are used.

Results Experimental results for LE are presented in Table 6. Although existing multi-metric optimization methods show performance improvements to a certain extent, the single-target MetricGAN outperforms these methods with regard to certain metrics. For example, the ESTOI-GAN is outperformed by the Multi-D and Multi-node MGANs in terms of SIIB and HASPI; however, the ESTOI-GAN significantly outperforms the two methods in terms of ESTOI by 2.3–2.7%. In our methods, all models of v1 and v2 achieve performance gains with large margins in terms of all metrics compared to existing multi-metric methods. Moreover, the models outperform the three single-target models in terms of all metrics. This suggests that the OKD learning approach synergizes between models trained in terms of different metrics as well as leads to stable multi-metric optimization by preventing the problem of confusing gradient directions.

Ablation Study The effect corresponding to the OKD flow weight α is evaluated. In Table 7, large overall performance improvements on all three α values are observed, as compared to existing methods listed in Table 6. In addition,



(a) Speech Enhancement (b) Listening Enhancement

Figure 2. Visualization of the Pearson correlation coefficient between evaluation metric scores of the five methods. Correlation coefficients are calculated using the validation scores of the metrics per epoch during training. The single (blue) model is trained using PESQ or SIIB, and the others using three metrics. For the OKD models, the model trained in terms of the main target metric of the PESQ or SIIB is selected from among the three models generated. The X-axis represents the combination of the three metrics.

the results obtained using $\alpha = 5$ and 10 correspond to performance gains with large margins in the main as well as other KD metrics. This suggests that powerful mimicking of the features of other generators also provides synergy for main metric learning. In terms of the "Range", the difference between the scores of the three models decreases with the increase in α. This implies that the three models are trained to predict similarly as the influence of the OKD flow increases.

5. Discussion

In this section, we attempt to explain the performance of the MetricGAN-OKD in terms of its high model generalizability and the correlation between the different metrics.

5.1. Correlation Analysis between Target Metrics

The effect of the proposed method on the changes between the validation scores of multiple metrics during multi-metric optimization is depicted in Figure 2. A lower correlation coefficient indicates that the performances of the two metrics improve or decrease together relatively infrequently. The single-target MetricGAN (blue) exhibits the lowest correlation coefficients in both the SE and LE tasks. Although existing multi-metric methods (orange and green) achieve an overall improvement, they exhibit lower correlation coefficients compared to the proposed methods (red and purple). This implies that the proposed framework optimizes multiple metrics easily without significant conflicts between the gradient directions of different metrics.

Table 8. Evaluation results on the unseen set mixed with LibriSpeech test-clean set (Panayotov et al., 2015) and new noise set (Hu & Wang, 2010). Each result is inferred without re-training using the weights of the model trained on VoiceBANK-DEMAND in Table 2 and Figure 3. (a)–(e) correspond to (a)–(e) in Figure 3.

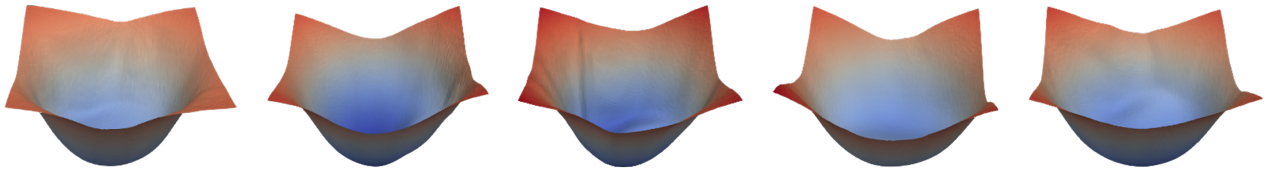
Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Single MGAN (a)	PESQ (PE)	-	2.62	3.80	3.04	3.21
Multi-D MGAN (b)	PE+CS	-	2.62	3.90	3.07	3.26
Multi-node MGAN (c)	PE+CS	-	2.61	3.91	3.07	3.26
MGAN-OKD v1 (ours) (d)	PE	CS	2.70	4.01	3.12	3.36
MGAN-OKD v2 (ours) (e)	PE	CS	**2.80**	**4.02**	**3.13**	**3.42**

5.2. Loss Landscape and Generalization

Although KD is observed to be effective in improving generalizability of image classification, it has not been investigated in the context of SE. Therefore, in this study, we investigate the effect of spectrogram-level distillation on the generalizability of a model. To verify the generalizability of the proposed method, we utilize the landscape geometry of loss. Several studies (Hochreiter & Schmidhuber, 1997; Li et al., 2018; Foret et al., 2020; Chen et al., 2021) have investigated the relationship between landscape geometry (sharpness/flatness) and generalization, demonstrating that visually flatter landscapes consistently correspond to good generalization performance. In this section, we visualize the loss function curvatures of five models trained in different ways using the method proposed by Li et al. (2018).

As depicted in Figure 3, the loss landscape of single MetricGAN presented in (a) is flat, whereas that of existing multi-metric methods presented in (b)–(c) exhibit sharper regions and more fluctuations in comparison. In contrast, the proposed methods exhibit smoother landscapes as depicted in (d)–(e) despite optimizing multiple metrics simultaneously. These results suggest the following: (1) Optimizing multiple metrics leads to greater non-convexity of loss landscapes, making multi-metric optimization difficult. (2) Our training procedure aids model convergence to flat local minima without significant fluctuations. (3) Even in SE, the flatness of the loss landscape correlates with better test scores, i.e., better generalizability. Note that the landscape of (a) is flatter than those in (b)–(c), but a direct performance comparison is difficult because of differences in additional information about an additional metric (Refer to Appendix F for further discussion and setups).

To further verify the relationship between loss landscape flatness and generalizability, we conduct experiments using disparate speech and noise datasets, which are datasets other than the one used for training and testing, as presented in Table 8. Comparisons depicted in (b)–(c) and (d)–(e) reveal a larger difference in the score on the new test dataset as compared to that on VoiceBANK-DEMAND, indicating that the proposed method improves model generalizability



(a) Single 3.13 / 4.13 (b) Multi-D 3.14 / 4.21 (c) Multi-node 3.13 / 4.22 (d) OKD v1 3.20 / 4.23 (e) OKD v2 3.24 / 4.23

Figure 3. Visualization of the loss landscapes of the models trained using different methods. The (a) is trained using PESQ, and (b)–(e) are trained using PESQ and CSIG (Refer to the results presented in Table 2). Each loss in the landscape is computed to be equal to the loss of D used for training. In other words, the losses corresponding to (b)–(e) are calculated by averaging the PESQ and CSIG losses, while the loss corresponding to (a) is calculated as the PESQ loss. Test PESQ and CSIG scores are listed below each figure.

by smoothing the loss landscapes.

6. Conclusion

In this work, we proposed a MetricGAN-based multi-metric optimization method called MetricGAN-OKD using online knowledge distillation to address the limited simultaneous improvement of multiple target metrics. MetricGAN-OKD, which consists of multiple generators and target metrics, related by a one-to-one correspondence, enables generators to learn with respect to a single target metric reliably while improving performance with respect to other metrics by mimicking other generators. Extensive experiments were conducted on SE and LE, revealing that the proposed method achieves significant performance improvement with respect to multiple target metrics by avoiding the problem of confusing gradient directions. In addition to conducting a quantitative evaluation, we also explained the desirable performance of the MetricGAN-OKD in terms of network generalizability and the correlation between different metrics. The limitations of the proposed method and directions of future work are discussed in Appendix G.

Acknowledgements

This research was supported by Brain Korea 21 FOUR.

References

Bagchi, D., Plantinga, P., Stiff, A., and Fosler-Lussier, E. Spectral feature mapping with mimic loss for robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5609–5613. IEEE, 2018.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Chen, X., Hsieh, C.-J., and Gong, B. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. Intelligibility-enhancing speech modifications: the hurricane challenge. In *Interspeech*, pp. 3552–3556, 2013.

Defossez, A., Synnaeve, G., and Adi, Y. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*, 2020.

Demonte, P. Harvard speech corpus–audio recording 2019. *University of Salford. Collection*, 2019.

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

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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, 2018.

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. *IEEE Signal Processing Letters*, 27:26–30, 2019a.

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 *International Conference on Machine Learning*, pp. 2031–2041. PMLR, 2019b.

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. *arXiv preprint arXiv:2104.03538*, 2021.

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.

Guo, Q., Wang, X., Wu, Y., Yu, Z., Liang, D., Hu, X., and Luo, P. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11020–11029, 2020.

Hao, X., Su, X., Horaud, R., and Li, X. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6633–6637. IEEE, 2021.

Hinton, G., Vinyals, O., Dean, J., et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Hu, G. and Wang, D. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2067–2079, 2010.

Hu, Y. and Loizou, P. C. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.

Jensen, J. and Taal, C. H. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

Kates, J. M. and Arehart, K. H. The hearing-aid speech perception index (haspi). *Speech Communication*, 65: 75–93, 2014.

Kates, J. M. and Arehart, K. H. The hearing-aid speech perception index (haspi) version 2. *Speech Communication*, 131:35–46, 2021.

Kawanaka, M., Koizumi, Y., Miyazaki, R., and Yatabe, K. Stable training of dnn for speech enhancement based on perceptually-motivated black-box cost function. In *Proc. ICASSP*, pp. 7524–7528. IEEE, 2020.

Kim, J., El-Kharmy, M., and Lee, J. End-to-end multi-task denoising for joint sdr and pesq optimization. *arXiv preprint arXiv:1901.09146*, 2019.

Kong, Z., Ping, W., Dantrey, A., and Catanzaro, B. Speech denoising in the waveform domain with self-attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7867–7871. IEEE, 2022.

Li, H. and Yamagishi, J. Multi-metric optimization using generative adversarial networks for near-end speech intelligibility enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3000–3011, 2021.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Li, H., Fu, S.-W., Tsao, Y., and Yamagishi, J. imetricgan: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning. *arXiv preprint arXiv:2004.00932*, 2020.

Lu, Y.-J., Wang, Z.-Q., Watanabe, S., Richard, A., Yu, C., and Tsao, Y. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7402–7406. IEEE, 2022.

Manocha, P., Finkelstein, A., Zhang, R., Bryan, N. J., Mysore, G. J., and Jin, Z. A differentiable perceptual audio metric learned from just noticeable differences. *arXiv preprint arXiv:2001.04460*, 2020.

Martin-Donas, 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.

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

Nakaoka, S., Li, L., Inoue, S., and Makino, S. Teacher-student learning for low-latency online speech enhancement using wave-u-net. In *Proc. ICASSP*, pp. 661–665. IEEE, 2021.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Park, H. J., Kang, B. H., Shin, W., Kim, J. S., and Han, S. W. Manner: Multi-view attention network for noise erasure. In *Proc. ICASSP*, pp. 7842–7846. IEEE, 2022.

Pascual, S., Bonafonte, A., and Serra, J. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Reddy, C. K., Beyrami, E., Pool, J., Cutler, R., Srinivasan, S., and Gehrke, J. A scalable noisy speech dataset and online subjective test framework. *arXiv preprint arXiv:1909.08050*, 2019.

Rennies, J., Schepker, H. F., Valentini-Botinhao, C., and Cooke, M. Intelligibility-enhancing speech modifications-the hurricane challenge 2.0. In *INTERSPEECH*, pp. 1341–1345, 2020.

Richter, J., Welker, S., Lemercier, J.-M., Lay, B., and Gerkmann, T. Speech enhancement and dereverberation with diffusion-based generative models. *arXiv preprint arXiv:2208.05830*, 2022.

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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Rothauser, E. Ieee recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246, 1969.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Shin, W., Park, H. J., Kim, J. S., Lee, B. H., and Han, S. W. Multi-view attention transfer for efficient speech enhancement. *arXiv preprint arXiv:2208.10367*, 2022.

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.

Tan, K. and Wang, D. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, volume 2018, pp. 3229–3233, 2018.

Valentini-Botinhao, C., Mayo, C., Cooke, M., et al. Hurricane natural speech corpus-higher quality version. 2019.

Valentini-Botinhao, C. et al. Noisy speech database for training speech enhancement algorithms and tts models. 2017.

Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. An instrumental intelligibility metric based on information theory. *IEEE Signal Processing Letters*, 25(1):115–119, 2017.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J. L., 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*, pp. 91–99. Springer, 2015.

Xu, B., Wang, N., Chen, T., and Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Yen, H., Germain, F. G., Wichern, G., and Roux, J. L. Cold diffusion for speech enhancement. *arXiv preprint arXiv:2211.02527*, 2022.

Yin, D., Luo, C., Xiong, Z., and Zeng, W. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9458–9465, 2020.

Zhang, Y., Xiang, T., Hospedales, T. M., and Lu, H. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4320–4328, 2018.

Zhao, Y., Xu, B., Giri, R., and Zhang, T. Perceptually guided speech enhancement using deep neural networks. In *Proc. ICASSP*, pp. 5074–5078. IEEE, 2018.

A. Algorithm

Algorithm 1 How to train MetricGAN-OKD

Input: The number of samples M, The number of target metrics and networks N, Normalized metric functions Q_i, Loss weight α, Generators G_i, Discriminators D_i
Initialize: G_i and D_i to different initial conditions, $\forall i \in \{1, 2, \ldots, N\}$
for epoch=1 to Total epochs **do**
  # Generators training step
  for m=1 to M **do**
    Sample a pair of noisy and clean speeches (x, y)
    for j=1 to N **do**
      Store enhanced speeches of other generators for the OKD step:
      $\hat{y_j} \leftarrow G_j(x)$
    end for
    for i=1 to N **do**
      $\hat{y} \leftarrow G_i(x)$
      $\hat{s} \leftarrow D_i(\hat{y}, y)$
      Update G_i with $L = \|\hat{s} - 1\|^2 + \alpha \times \sum_{j=1}^{N} \|\hat{y} - \hat{y_j}\|^2$ (Equation (7))
    end for
  end for
  # Discriminators training step
  for m=1 to M **do**
    Sample a pair of noisy and clean speech (x, y)
    for i=1 to N **do**
      $\hat{y} \leftarrow G_i(x)$
      Calculate prediction and real scores from D and Q, respectively.
      $\hat{s_c}, \hat{s_e}, \hat{s_n} \leftarrow D_i(y, y), D_i(\hat{y}, y), D_i(x, y)$
      $s_c, s_e, s_n \leftarrow Q_i(y, y), Q_i(\hat{y}, y), Q_i(x, y)$
      Update D_i with $L = \|s_c - \hat{s_c}\|^2 + \|s_n - \hat{s_n}\|^2 + \|s_e - \hat{s_e}\|^2$ (Equation (4))
    end for
  end for
end for

We discuss the training method in Algorithm 1. In each epoch, the training phase consists of training generators and discriminators. In the generator training phase, enhanced speeches to be used in the OKD phase are first stored in a list, followed by the update step of each generator, which includes the standard and OKD flows. Then, each discriminator is trained with respect to different evaluation metrics using the three terms given by Equation (4). Although a replay buffer technique is adopted that stores enhanced speech samples during training and reuses them in discriminator training, as in the case of MetricGAN+ (Fu et al., 2021), we describe a simple discriminator training procedure for simplicity. The replay buffer is described below.

A.1. Replay buffer in discriminator training

A replay buffer technique proposed in MetricGAN+ (Fu et al., 2021) is adopted for discriminator training in this study. The replay buffer stores a fraction of the enhanced speech samples per epoch during training and then reuses them during discriminator training. For example, if the number of samples per epoch is 100 and the history portion of the replay buffer is 0.2, then 20 samples are stored. Subsequently, the discriminator is trained using 120 (100 + 20) samples in the following epoch. In this way, the total number of samples, N_T, used to train the discriminator in the epoch, T, is given by:

$$N_T = 100 + 100 \times 0.2 \times T \tag{8}$$

Although a replay buffer is significantly effective for discriminator training, its operation becomes time-consuming with the accumulation of samples. As presented in Table 9, the training time of the replay buffer accounts for approximately 90% of the total training time. Therefore, we intend to make the replay buffer more efficient in future work. The generator training time is discussed in Appendix G.

Table 9. The time spent on each component of the five training methods is measured. All methods are trained for 750 epochs in SE, and multi-metric methods are trained with two target metrics. Experiments are conducted in the following environments—UBUNTU 20.04, PYTHON 3.8.13, PyTorch 1.10.1, CUDA 11.3, and NVIDIA RTX A6000.

Method	Training Time (hour – h)				GPU Memory Usage
	Generator	Discriminator	Replay buffer of D	Total	
Single MGAN	0.6 h (1.9%)	2.1 h (6.5%)	29.5 h (91.6%)	32.2 h	2.2 G
Multi-D MGAN	0.7 h (1.1%)	4.9 h (7.6%)	59.1 h (91.3%)	64.7 h	2.5 G
Multi-node MGAN	0.6 h (1.8%)	3.0 h (8.9%)	30.0 h (89.3%)	33.6 h	2.3 G
MGAN-OKD v1 (ours)	1.4 h (2.1%)	4.9 h (7.5%)	59.1 h (90.4%)	65.4 h	2.6 G
MGAN-OKD v2 (ours)	1.4 h (4.1%)	2.9 h (8.5%)	29.9 h (87.4%)	34.2 h	2.3 G

B. Figure of MetricGAN-OKD v2

The figure of MetricGAN-OKD v2 is presented in Figure 4. Unlike MetricGAN-OKD v1, the discriminators of the two generator paths in Figure 4a are identical networks sharing weights. To achieve a one-to-one correspondence, only one metric score from the discriminator's several metric nodes is used during the training of each generator.

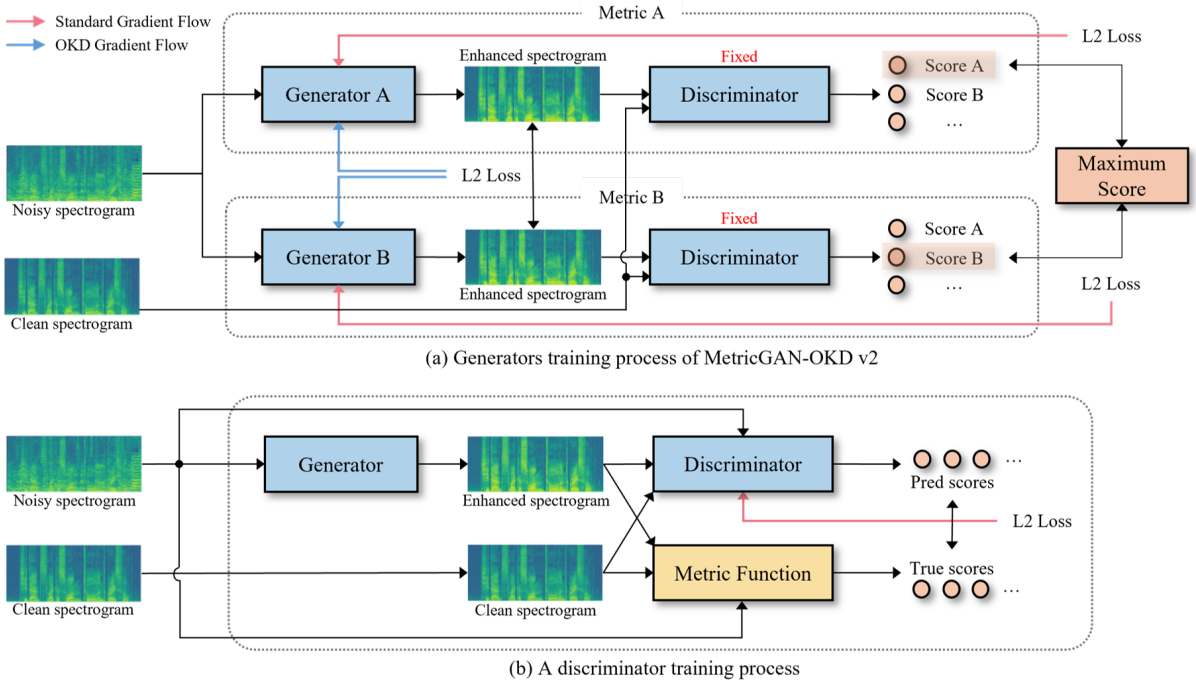


Figure 4. The overall framework of MetricGAN-OKD v2. For simplicity, (a) represents the generator training process with respect to two target metrics, and (b) represents the training process of a multi-node discriminator.

C. Listening Enhancement

The goal of the LE task, also known as near-end SE, is to improve speech intelligibility on the listener's side under the equal-level signal power constraint as follows:

$$\hat{y} = G(x, n), \qquad o = \hat{y} + n, \tag{9}$$

where x, n, and $\hat{y}$ denote input speech, listener-side background noise, and enhanced speech, respectively. LE is useful in environments where accurate speech reception is difficult due to adverse listening conditions. Some challenge contests have been organized to improve performance in this task (Cooke et al., 2013; Rennies et al., 2020). In contrast to SE, the

generator receives two inputs: (i) input speech (noisy, denoised, or clean speech), (ii) noise observed by the listener. The generator in LE improves the intelligibility of the input speech in a noisy environment. The final output, o, which is the sound heard by the listeners, is obtained by combining the enhanced speech, $\hat{y}$, obtained via G and the noise, n (including reverberation). For discriminator training, the objective function is defined as follows:

$$L_{D_i} = \mathbb{E}_{x,n}[(D_i(G_i(x,n),x,n) - Q_i^{'}(G_i(x,n),x,n))^2] \tag{10}$$

where $1 \leq i \leq N$. By optimizing this objective function, the discriminator acts as a surrogate function similar to the target metric function, while the generator learns to improve intelligibility using surrogate guidance from the discriminator as follows:

$$L_{G_i(standard)} = \mathbb{E}_{x,n}[(D_i(G_i(x,n),x,n) - s_i)^2] \tag{11}$$

In the OKD flow, G_i mimics G_j trained on different target metrics by minimizing the difference between enhanced speech $\hat{y}_i$ and $\hat{y}_j$ as follows:

$$L_{G_i(OKD)} = \mathbb{E}_{x,n}\left[\sum_{j=1}^{N} |G_i(x,n) - G_j(x,n)|^2\right], \quad \text{if } i \neq j \tag{12}$$

The total loss of G_i is defined by combining the standard and OKD flow losses, where α denotes the weight of the OKD flow.

$$L_{G_i} = L_{G_i(standard)} + \alpha \times L_{G_i(OKD)} \tag{13}$$

D. Additional Experiments

D.1. MetricGAN with Standard KD Approach

As described in Section 3.3, although utilizing the typical KD approach for multi-metric optimization suffers from the complexity of the training process, it can be used to resolve the problems raised in this study. Therefore, we conduct experiments to validate the effect of MetricGAN trained using the standard KD approach, as presented in Table 10.

Table 10. Evaluation results on efficacy of standard KD approach (at least a two-stage training process) on VoiceBank-DEMAND dataset. PESQ and CSIG are selected as target metrics for optimization. Specifically, in the case of the PESQ (target) and CSIG (KD) combination, the following process is performed: (1) Train the single CSIG-GAN and obtain the pre-trained generator. (2) Train the single PESQ-GAN while transferring the knowledge obtained from the pre-trained generator to the generator of PESQ-GAN.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Single MGAN	PESQ (PE)	-	3.13	4.13	3.03	3.61
	CSIG (CS)	-	3.11	4.23	3.07	3.68
Multi-D MGAN	PE+CS	-	3.14	4.21	3.09	3.67
Multi-node MGAN	PE+CS	-	3.13	4.22	**3.15**	3.68
MGAN-OKD v1 (ours)	PE	CS	3.20	4.23	3.11	3.71
	CS	PE	3.12	4.25	3.10	3.68
MGAN-OKD v2 (ours)	PE	CS	**3.24**	4.23	3.07	**3.73**
	CS	PE	3.19	**4.26**	3.12	3.72
MGAN-Standard KD (ours)	PE	CS	3.17	4.23	3.09	3.69
	CS	PE	3.15	4.24	3.09	3.68

MetricGAN with the standard KD approach is observed to outperform existing single- and multi-metric methods significantly with respect to target metrics, i.e., PESQ and CSIG. This reveals that the proposed distillation scheme is the effective method for the multi-metric optimization of MetricGAN. Although the standard KD approach functions well, it exhibits limited performance improvements compared to the proposed method, MetricGAN-OKD. This implies that the OKD training scheme synergizes between multiple models trained in terms of different metrics during training.

D.2. Another design of the proposed method

To allow flexibility and scalability of the proposed method, we executed additional experiments for another design as follows: its framework consists of one generator and two discriminators. One discriminator is trained with respect to two target

metrics (e.g. SIIB and HASPI), and another discriminator is trained in terms of another target metric (e.g., ESTOI). The single generator learns from the gradients of multiple discriminators. The results are shown in Table 11.

Table 11. Evaluation results on the Harvard Sentences dataset. SIIB, HASPI, and ESTOI are selected as target metrics for optimization.

Method	# Gs	# Ds	# D nodes	# Target metrics	Target Metric	KD Metric	SIIB	HASPI	ESTOI
Multi-D MGAN	1	3	1/1/1	3	S/H/E	-	67.7	4.42	0.332
Multi-node MGAN	1	1	3	3	S+H+E	-	71.1	4.53	0.328
MGAN-OKD v1	3	3	1/1/1	3	S	H, E	**79.4**	**4.78**	**0.382**
MGAN-OKD v2	3	1	3	3	S	H, E	76.3	4.68	0.374
MGAN-OKD another design	2	2	2/1	3	S+H	E	79.1	4.67	0.373
	2	2	2/1	3	E	S+H	74.8	4.56	0.375
MGAN-OKD another design	2	2	2/1	3	E+S	H	75.9	4.57	0.367
	2	2	2/1	3	H	E+S	73.8	4.51	0.379

The results show that the various combinations of the proposed method still outperform existing multi-metric methods, indicating that the proposed method is flexible and scalable with respect to design. We dealt with combinations involving three target metrics, but if there are more than four target metrics, more combinations are also possible.

D.3. Efficacy of ensemble

The proposed method creates as many generators as the number of target metrics. Therefore, we verify the efficacy of the ensemble of multiple models, as shown in Tables 12 and 13. As universally expected, the results reveal that the ensemble of models improves overall performance. Note that ensembles are performed with equal weights.

Table 12. Ensemble results for VoiceBank-DEMAND dataset based on the experiments in Table 2.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Multi-D MGAN	PE+CS	-	3.14	4.21	3.09	3.67
Multi-node MGAN	PE+CS	-	3.13	4.22	**3.15**	3.68
MGAN-OKD v1 (ours)	PE	CS	3.20	4.23	3.11	3.71
	CS	PE	3.12	4.25	3.10	3.68
	Ensemble		3.19	**4.29**	3.13	3.74
MGAN-OKD v2 (ours)	PE	CS	3.24	4.23	3.07	3.73
	CS	PE	3.19	4.26	3.12	3.72
	Ensemble		**3.25**	**4.29**	3.11	**3.77**

Table 13. Ensemble results for VoiceBank-DEMAND dataset based on the experiments in Table 3.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Multi-D MGAN	PE+CS+CB+CO	-	3.12	4.18	3.25	3.65
Multi-node MGAN	PE+CS+CB+CO	-	3.09	4.22	3.24	3.67
MGAN-OKD v1 (ours)	PE	CS+CB+CO	3.14	4.22	3.24	3.67
	CS	PE+CB+CO	3.11	4.25	3.25	3.67
	CB	PE+CS+CO	2.98	4.12	**3.32**	3.53
	CO	PE+CS+CB	3.13	4.29	3.23	3.71
	Ensemble		3.14	4.30	3.29	3.73
MGAN-OKD v2 (ours)	PE	CS+CB+CO	3.15	4.26	3.25	3.71
	CS	PE+CB+CO	3.12	4.26	3.22	3.69
	CB	PE+CS+CO	3.10	4.26	3.30	3.69
	CO	PE+CS+CB	3.13	4.27	3.24	3.71
	Ensemble		**3.17**	**4.31**	3.28	**3.75**

D.4. Various target metric combinations

In Tables 14 to 18, the experiments with different combinations of evaluation metrics, which are omitted from the main body of the paper, are conducted.

Table 14. Evaluation results on VoiceBank-DEMAND dataset. PESQ and CBAK are selected as target metrics for optimization.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Single MGAN	PESQ (PE)	-	3.13	4.13	3.03	3.61
	CBAK (CB)	-	2.97	4.10	3.32	3.53
Multi-D MGAN	PE+CB	-	3.06	4.13	3.29	3.59
Multi-node MGAN	PE+CB	-	3.05	4.19	3.27	3.62
MGAN-OKD v1 (ours)	PE	CB	3.13	4.21	3.26	3.67
	CB	PE	2.95	4.12	3.30	3.54
	Ensemble		3.09	**4.23**	3.32	3.67
MGAN-OKD v2 (ours)	PE	CB	**3.17**	4.18	3.24	3.67
	CB	PE	3.05	4.15	**3.33**	3.60
	Ensemble		3.15	4.22	3.31	**3.69**

Table 15. Evaluation results on VoiceBank-DEMAND dataset. PESQ and COVL are selected as target metrics for optimization.

Method	Target Metric	KD Metric	PESQ	CSIG	CBAK	COVL
Single MGAN	PESQ (PE)	-	3.13	4.13	3.03	3.61
	COVL (CO)	-	3.12	4.20	3.14	3.69
Multi-D MGAN	PE+CO	-	3.09	4.20	3.14	3.64
Multi-node MGAN	PE+CO	-	3.09	4.16	3.13	3.61
MGAN-OKD v1 (ours)	PE	CO	3.17	4.19	3.13	3.67
	CO	PE	3.15	4.23	3.13	3.69
	Ensemble		**3.19**	4.26	3.15	**3.72**
MGAN-OKD v2 (ours)	PE	CO	3.13	4.23	3.11	3.68
	CO	PE	3.15	4.24	3.16	3.69
	Ensemble		3.17	**4.27**	**3.16**	**3.72**

Table 16. Evaluation results on the Harvard Sentences dataset. SIIB and HASPI are selected as target metrics for optimization.

Method	Target Metric	KD Metric	SIIB	HASPI	ESTOI
Unmodified	-	-	25.42	2.51	0.313
Single MGAN	SIIB (S)	-	69.78	4.43	0.269
	HASPI (H)	-	65.50	4.49	0.291
Multi-D MGAN	S+H	-	74.01	4.59	0.303
Multi-node MGAN	S+H	-	78.00	4.60	0.306
MGAN-OKD v1 (ours)	S	H	79.26	4.70	**0.362**
	H	S	**81.30**	**4.72**	0.347
MGAN-OKD v2 (ours)	S	H	80.25	4.64	0.351
	H	S	81.19	4.63	0.341

Table 17. Evaluation results on the Harvard Sentences dataset. SIIB and ESTOI are selected as target metrics for optimization.

Method	Target Metric	KD Metric	SIIB	HASPI	ESTOI
Single MGAN	SIIB (S)	-	69.78	4.43	0.269
	ESTOI (E)	-	58.57	4.05	0.355
Multi-D MGAN	S+E	-	73.99	4.35	0.363
Multi-node MGAN	S+E	-	71.55	4.37	0.356
MGAN-OKD v1 (ours)	S	E	**80.96**	**4.59**	0.374
	E	S	78.91	4.57	**0.388**
MGAN-OKD v2 (ours)	S	E	75.29	4.52	0.384
	E	S	78.15	4.53	0.369

Table 18. Evaluation results on the Harvard Sentences dataset. HASPI and ESTOI are selected as target metrics for optimization.

Method	Target Metric	KD Metric	SIIB	HASPI	ESTOI
Single MGAN	HASPI (H)	-	65.50	4.49	0.291
	ESTOI (E)	-	58.57	4.05	0.355
Multi-D MGAN	H+E	-	68.03	4.23	0.336
Multi-node MGAN	H+E	-	69.82	4.28	0.344
MGAN-OKD v1 (ours)	H	E	75.62	4.60	**0.388**
	E	H	76.71	**4.66**	0.376
MGAN-OKD v2 (ours)	H	E	77.57	4.59	0.380
	E	H	**77.79**	4.58	0.379

Table 19. Comparison with state-of-the-art methods on VoiceBank-DEMAND dataset.

Method	Params	PESQ
Encoder-Decoder structure		
DEMUCS-L (Defossez et al., 2020)	33.5M	3.07
SE-Conformer (Kong et al., 2022)	-	3.13
MANNER (Park et al., 2022)	24.1M	3.21
Diffusion-based methods		
CDiffuSE-Base (Lu et al., 2022)	2.6M	2.44
CDiffuSE-Large (Lu et al., 2022)	-	2.52
SGMSE+ (Richter et al., 2022)	65M	2.93
Unfolded CD (Yen et al., 2022)	5.6M	2.77
MetricGAN-based methods		
Single MGAN (PE)	1.9M	3.13
MGAN-OKD v1-PE,CS (ours)	1.9M	3.20
MGAN-OKD v2-PE,CS (ours)	1.9M	**3.24**

D.5. Comparison with State-of-the-art SE Methods

We compared the proposed method with recent encoder-decoder and diffusion-based methods. The results are shown in Table 19. The results indicate that the proposed method outperforms different methods while requiring significantly fewer model parameters.

E. Network Architecture

In the SE and LE experiments, the G and D structures proposed in MetricGAN+ (Fu et al., 2021) are adopted. G consists of two bidirectional-LSTM (Weninger et al., 2015) layers with 200 hidden nodes and two fully-connected layers with 300 and 257 nodes for mask estimation. The input $x \in \mathbb{R}^{1\times T\times F}$ of G is the magnitude spectrogram of noisy speech, which is generated using window and hop lengths of 512 and 256, respectively. Here, T and F denote the time length and frequency bin of the spectrogram. The enhanced spectrogram is obtained by applying the mask to the input spectrogram. Note that short-time Fourier transform (STFT) and inverse STFT are used to transform between speech signal and spectrogram. D is composed of four 2D convolution layers with a 5×5 kernel, an average-pooling layer, and three fully-connected layers. All layers of D are followed by spectral normalization (Miyato et al., 2018) and LeakyReLU (Xu et al., 2015), except for the pooling and the final fully-connected layer.

F. Discussion and Configurations for Loss Landscape and Generalization

As discussed in Section 5.2, the loss landscapes of existing multi-metric methods exhibit sharper regions and more fluctuations than those of the single-metric method and the proposed methods. However, the loss landscapes of existing multi-metric methods are not as severely non-convex as the loss landscapes of ResNet without skip connections shown in the study (Li et al., 2018). This can be attributed to the simple structure of the generator, comprising two bidirectional-LSTM layers and two fully-connected layers. Li et al. (2018) demonstrated that the loss landscape becomes more chaotic as the complexity of the network increases. Therefore, we intend to apply MetricGAN-OKD to more complex generators and investigate changes in the loss landscapes with respect to the five training schemes.

For visualization of loss landscapes depicted in Figure 3, we used the official code provided by Li et al. (2018) and the ParaView tool for rendering. For the configuration of loss landscapes, the resolutions are set to 100×100, creating 10,000 points on 2D projection space in aggregate, and the losses are calculated using 10% of the training samples of the VoiceBank-DEMAND dataset. For LibriSpeech (Panayotov et al., 2015) test-clean sets with added new noises, the 20 types of noise provided by Hu & Wang (2010) and 20 dB of SNR are used.

G. Limitations and Future Work

The proposed method does not affect the inference time of the model as it is a training method. However, it suffers from some limitations in terms of training costs. As MetricGAN-OKD contains multiple generators, the cost of the generator training process increases linearly with the number of target metrics. Although the proposed method increases generator training time, it still accounts for a small portion of the total training time because of the overhead of the discriminator training process, as presented in Table 9. Nevertheless, as the complexity of the generator increases, the total training time of the proposed method can increase significantly.

In future work, we intend to address the aforementioned limitations by designing an efficient training scheme for multi-metric optimization based on the concept of self-knowledge distillation or a siamese network, involving a single generator. Further, MetricGAN-OKD v2 is observed to outperform v1 on SE, whereas the reverse is true on LE. The main difference between v1 and v2 is as follows: each discriminator of v1 mimics a single metric function, and the multi-node discriminator of v2 learns from multiple metrics. This difference and experimental results imply that the discriminator has good and bad cases to learn simultaneously different metrics depending on the properties of evaluation metrics. Therefore, we intend to investigate these causes in terms of discriminator training.

Finally, this study demonstrated that the OKD training scheme is an effective method to mitigate conflicts between multiple objectives when optimizing them. Therefore, this approach can be utilized for various tasks that combine multiple losses, such as depth estimation, pose estimation, and semantic segmentation.