

Perceptual Loss Function for Speech Enhancement Based on Generative Adversarial Learning

Xin Bai, Xueliang Zhang, Hui Zhang and Haifeng Huang

College of Computer Science, Inner Mongolia University

National & Local Joint Engineering Research Center

of Intelligent Information Processing Technology for Mongolian

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology

E-mail: 31909068@mail.imu.edu.cn, cszxl@imu.edu.cn, cszh@imu.edu.cn, Huanghf1@lenovo.com

Abstract— Loss function is important to current deep learning-based speech enhancement. Although the commonly used loss function is minimum mean square error (MSE) between the enhanced speech and the target speech, it cannot accurately reflect the speech quality after noise reduction. Perceptual metrics of speech quality (PESQ) and short-time objective intelligibility (STOI) are usually employed to evaluate speech enhancement algorithms. This paper proposes a generative adversarial architecture that can directly uses PESQ and STOI as training targets for deep learning-based speech enhancement. Specifically, the enhancement network and the assessment network are used as the generator and discriminator in the generative adversarial network (GAN), respectively. In each epoch of adversarial training, we first fix the generator and make the discriminator estimate the true PESQ and STOI of enhanced speech. Then, the discriminator is fixed, and the generator is trained to make the PESQ and STOI of enhanced speech close to those of the target speech. The assessment network is based on Multi-gate Mixture-of-Experts (MMoE), which is suitable for multi-task learning. Compared with using the clean speech spectrum as training target, the proposed method is more effective on speech enhancement.

I. INTRODUCTION

In real environment, interfering noises are inevitable and severely affects many applications, such as, mobile speech communication, teleconference and automatic speech recognition. Speech enhancement aims to remove the noise and improve the speech quality and intelligibility.

Compared with traditional speech enhancement methods, speech enhancement based on deep learning achieved great success for over the past decades, especially under challenging test conditions (non-stationary noise and low signal-to-noise ratio) [1]-[14]. Generally, most of these methods regard the speech enhancement as a supervised learning, which trains deep neural networks to estimate the clean speech from noisy speech by using mean square error (MSE) as loss. But the reduction of MSE loss cannot always represent better speech enhancement effect [15], [16]. The discrepancy between the loss function used for training a speech enhancement model and human auditory perception usually makes the quality of enhanced speech unsatisfactory.

We aim to find a perfect loss function to replace the MSE loss function to serve speech enhancement more effectively. For speech enhancement, we need to focus on two points: 1) *noise reduction* and 2) *speech distortion*. A perfect loss function should reduce noise while maintaining the quality of speech.

PESQ and STOI are two commonly used metrics for evaluating the quality of the enhanced speech. They evaluate speech quality by comparing the gap between speech signals under test and clean speech signals. The higher the score, the better the speech enhancement effect of the model. Therefore, optimizing the enhancement models directly using these two functions is a reasonable direction. In recent years, some studies have used neural networks to estimate the PESQ score of speech without reference, which called Quality-Net [17]. Fu et al. proved that speech enhancement networks using Quality-Net to calculate the loss outperform which using MSE of spectrum as loss [18]. This new framework was named MetricGAN. It is based on GAN. The generator of MetricGAN is a speech enhancement network and the discriminator is Quality-Net.

We argue that speech quality evaluation networks will be more conducive to speech enhancement by learning more speech quality evaluation metrics. So, we propose a new speech quality assessment network that can learn both PESQ and STOI simultaneously. We name it PESQ-STOI-Quality-Net (PSQN). PSQN is based on multi-task learning. The goal of multi-task learning is to use the useful information contained in multiple learning tasks to help each task learn a more accurate learner. Pareto optimality provides theoretical guidance for multi-target learning [19]. When multiple targets are jointly optimized, one of them can reach a better convergence point than training it alone. Ma et al. proposed Multi-gate Mixture-of-Experts (MMoE) [20]. Different from the traditional multi-task learning using the shared bottom layer [21], MMoE divides the shared bottom layer into multiple experts to provide different weights for each task. Individualized learning for each task is more conducive to multi-task learning. Multi-task learning has been successfully applied to many large-scale applications, such as recommender systems. In e-commerce, predictions of

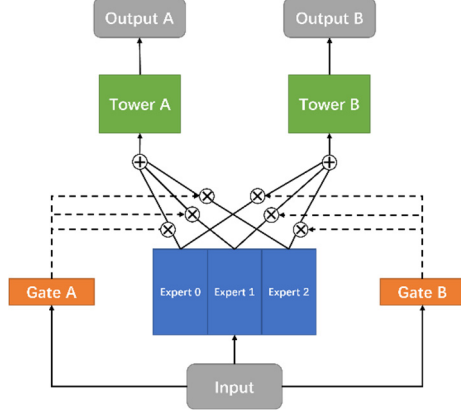


Fig. 1 Structure of Multi-gate Mixture-of-Experts

click-through rate and conversion rate are jointly optimized by it. Besides, in video sites, predictions of video through rate, share rate and comment rate are jointly optimized by it. Due to the characteristics of MMoE, we believe that compared with commonly predicting a single speech quality evaluation metric, PSQN based on MMoE can not only predict more accurately, but also constrain the convergence direction of speech enhancement network by simultaneously predicting multiple indicators. We name this speech enhancement framework Multi-MetricGAN.

The generator of Multi-MetricGAN is a speech enhancement network based on Convolutional Recurrent Neural Network (CRNN) [22] and the discriminator is a speech quality assessment network based on MMoE. After inputting the speech enhanced by the generator into the discriminator, the discriminator will output the predicted PESQ and STOI of the speech. Multi-MetricGAN aims to optimize the speech enhancement network by minimizing the MSE loss between the metric scores predicted by discriminator and the actual scores. The experimental results show that the speech enhancement network which is based on Multi-MetricGAN is not only perform better than that based on CRNN, but also better than that based on MetricGAN.

A. MMoE

One of the widely used multi-task learning models is proposed by Caruana [20], [23], which has a shared-bottom model structure, where the bottom hidden layers are shared across tasks. This structure substantially reduces the risk of overfitting, but can suffer from optimization conflicts caused by task differences, because all tasks need to use the same set of parameters on shared-bottom layers.

MMoE is designed to capture the task differences without requiring significantly more model parameters compared to the shared-bottom multi-task model. The improvement of MMoE is achieved by dividing the shared bottom layer into multiple experts and adding a separate gating network for each task.

Fig.1 shows the structure of MMoE. Specifically, given K tasks, where $k=1, 2, \dots, K$ for each task respectively. The

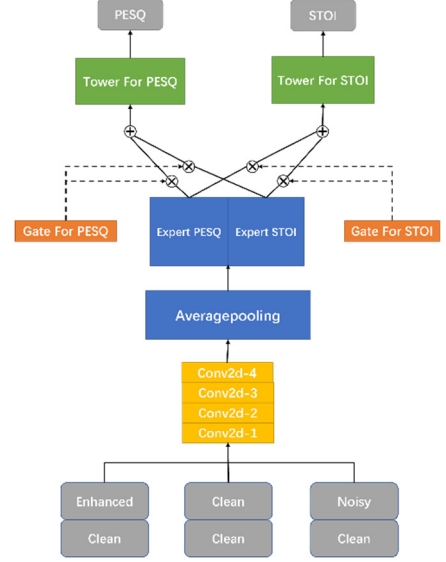


Fig. 2 Structure of PESQ-STOI-Quality-Net

model consists of K tower networks h^k and a personalize shared-bottom network represented as function f . When x is the input of the network, y_k , the output of task k can be denoted as the input of the network, y_k , the output of task k can be denoted as

$$y_k = h^k(f^k(x)) \quad (1)$$

The shared-bottom consists of expert networks. Here, n denotes the number of expert networks. We record the i th expert network as $E_i(x)$. The output of the k th gating network is shown as $g^k(x)$. The i th component of $g^k(x)$ is denoted as $g^k(x)_i$. After multiplied and added with each component of the gating networks, the output of each expert network will be input into the corresponding tower network for each task. So, $f^k(x)$ can be denoted as

$$f^k(x) = \sum_{i=1}^n g^k(x)_i E_i(x) \quad (2)$$

The gating networks are linear transformations of the input with a softmax layer:

$$g^k(x) = \text{softmax}(W_{gk}x) \quad (3)$$

where W_{gk} denotes the weight parameter of the k th gating network.

B. PESQ-STOI-Quality-Net

Our proposed speech quality assessment network PSQN is based on MMoE. Fig.2 shows the network structure of PSQN. To provide more visible information and to stabilize the learning process of Multi-MetricGAN [24], we deliberately divide the input into three items, namely clean speech, noisy speech and enhanced speech, where the enha-

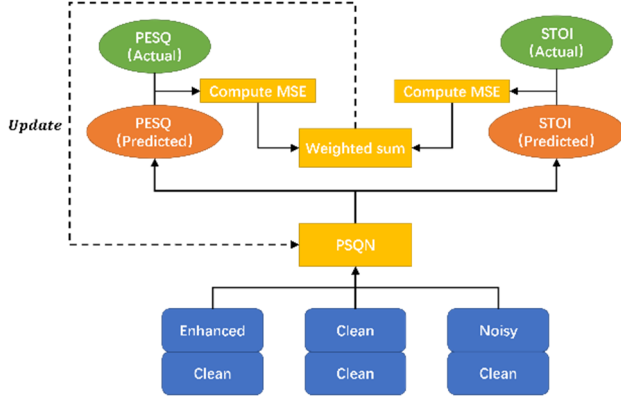


Fig. 3 Training process of PESQ-STOI-Quality-Net

ned speech is generated by a pre-trained speech enhancement network. Each of these three items forms a feature map with clean speech and input into PSQN together. PSQN is aim to make the predicted speech quality evaluation metric scores close to the actual metric scores. So, its loss function is the MSE between predicted scores and actual scores. Since PSQN has two learning objectives, PESQ and STOI, the loss function of PSQN, L_{PSQN} , can be denotes as

$$L_{PSQN} = \mathbb{E}_x(\alpha_1 \cdot L_{PESQ} + \alpha_2 \cdot L_{STOI}) \quad (4)$$

where α denotes the weight of each metric during optimization. This is because each task converges at a different rate during training. Properly adjusting weight of the loss for each task can effectively improve the overall convergence speed. We found that when $\alpha_1, \alpha_2 = 1$, the convergence speed of PESQ learning is far less than that of STOI. After extensive experimentation, we found a perfect set of weights to achieve simultaneous convergence of PESQ learning and STOI learning in a very short time, where $\alpha_1 = 3, \alpha_2 = 1$.

Let $Q'(I)$ be a function that represents the target evaluation metric, where I denotes the input of the metric. For example, for PESQ and STOI, I denotes a pair of enhanced speech, y' (or noisy speech, x , clean speech, y) that we want to evaluate, and its corresponding clean speech, y . To ensure that PSQN behaves similar to Q' , the objective function of PESQ and STOI can be denotes as

$$L_{PESQ} = (Q'_{PESQ}(x, y) - Q_{PESQ}(x, y))^2 + (Q'_{PESQ}(y, y) - Q_{PESQ}(y, y))^2 + (Q'_{PESQ}(y', y) - Q_{PESQ}(y', y))^2 \quad (5)$$

$$L_{STOI} = (Q'_{STOI}(x, y) - Q_{STOI}(x, y))^2 + (Q'_{STOI}(y, y) - Q_{STOI}(y, y))^2 + (Q'_{STOI}(y', y) - Q_{STOI}(y', y))^2 \quad (6)$$

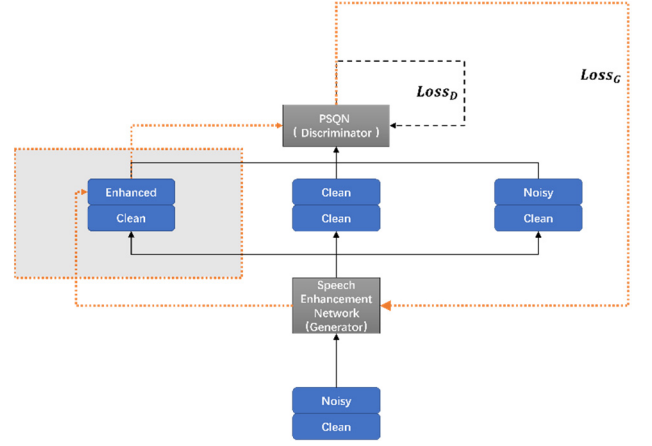


Fig. 4 Training process of Multi-MetricGAN

It is worth mentioning that to reduce the gradient updated by the network and prevent oscillation, we normalized the PESQ score when computing the label metric scores. Fig.3 shows the training process of PSQN.

C. Multi-MetricGAN

The main idea of MetricGAN is to mimic the behavior of a target evaluation function (e.g., PESQ function) with a neural network (e.g., Quality-Net). The surrogate evaluation function is learned from raw scores, MetricGAN treat the target evaluation function as a black box. Once the surrogate evaluation is trained, it can be used as a loss function for the speech enhancement model. The whole training process of MetricGAN is based on the idea of GAN, where the surrogate loss and enhancement model are alternately updated.

MetricGAN also mentions the use of multiple metrics to optimize. But it only uses a simple heuristic method. In each iteration, metric with the largest loss is selected as the optimization goal. This approach has two drawbacks. First, there are multiple discriminators. When the number of discriminator parameters is too large, it is easy to overfit. Second, the correlation between speech evaluation metrics is unconsidered so that the optimization of each metric is isolated. Therefore, we propose Multi-MetricGAN.

Fig.4 shows the whole training process of Multi-MetricGAN. The main training idea of Multi-MetricGAN is the same as that of MetricGAN. The difference of them is that Multi-MetricGAN replaces the single-target discriminator (Quality-Net, which can only predict PESQ) of MetricGAN with a multi-target discriminator (PSQN, which can predict both PESQ and STOI simultaneously).

Similar to training PSQN alone, the loss function of the discriminator (D) of Multi-MetricGAN can be denotes as

$$L_D = \mathbb{E}_x(\beta_1 \cdot L_{PESQ} + \beta_2 \cdot L_{STOI}) \quad (7)$$

where β denotes the weight of each metric during optimization. Here $\beta_1 = 3, \beta_2 = 1$. Loss function for

PESQ learning, L_{PESQ} , and that of STOI, L_{STOI} can be denoted as

$$L_{PESQ} = (D_{PESQ}(y, y) - Q_{PESQ}(y, y))^2 + (D_{PESQ}(G(x), y) - Q_{PESQ}(G(x), y))^2 + (D_{PESQ}(x, y) - Q_{PESQ}(x, y))^2 \quad (8)$$

$$L_{STOI} = (D_{STOI}(y, y) - Q_{STOI}(y, y))^2 + (D_{STOI}(G(x), y) - Q_{STOI}(G(x), y))^2 + (D_{STOI}(x, y) - Q_{STOI}(x, y))^2 \quad (9)$$

where G denotes the generator, speech enhancement network, while $G(x)$ denotes the speech enhanced by G .

The training of the generator network (G) can completely rely on the adversarial loss

$$L_{G(PESQ)} = \mathbb{E}_{x,y} [(D_{PESQ}(G(x), y) - T_{PESQ})^2] \quad (10)$$

$$L_{G(STOI)} = \mathbb{E}_{x,y} [(D_{STOI}(G(x), y) - T_{STOI})^2] \quad (11)$$

$$L_G = \gamma_1 \cdot L_{G(PESQ)} + \gamma_2 \cdot L_{G(STOI)} \quad (12)$$

where T is the desired assigned score. For example, to generate clean speech, we can simply assign T to be 1 (PESQ scores have been normalized). On the contrary, we can also generate more noisy speech by assigning a smaller T . Both γ_1 and γ_2 are weights for loss. Here $\gamma_1 = 3$, $\gamma_2 = 1$.

II. EXPERIMENTS

A. Experimental setup

In our experiments, the TIMIT corpus [25] was used to prepare the training, validation, and test sets. 4000 utterances were selected from the train set of the TIMIT database for training. These utterances were further corrupted with 10000 non-speech sounds randomly selected from a sound effect library (available at www.sound-ideas.com) [26], at eleven SNR levels (from -5 dB to 5 dB with steps of 1 dB) to form 220000 training utterances. 200,000 of them were used to pre-train the PSQN, the remaining 20,000 utterances were used to pre-train the speech enhancement network CRN. To monitor the training process and choose the hyperparameters, we randomly selected another clean 100 utterances from the TIMIT test set to generate our validation set. Each utterance was further corrupted with the noise (different from those already used in the training set) from the above library [26] at eleven different SNR levels (from -5 dB to 5 dB with steps of 1 dB randomly select). To evaluate the performance of different speech enhancement network, 100 utterances from the TIMIT test (different from the above 100 utterances) set were randomly selected as our test set. These utterances were corrupted with four unseen noise types (pink, white,

buccaneer1, and f16), at five SNR levels (from -12 dB to 12 dB with steps of 6 dB).

The speech enhancement network in this study is based on CRNN. It uses the 161-dimensional Short Time Fourier Transform (STFT) amplitude spectrum of noisy speech as the input feature and the amplitude spectrum of clean speech as the training target. The input of the network is encoded into a higher-dimensional latent space, and the sequence of latent feature vectors are then modelled by two LSTM. Subsequently, the decoder converts the output sequence of the LSTM layer back to the original input shape.

As for the discriminator, PSQN, we use four convolution kernels to extract high-order features, the kernel sizes are 5×5 , 7×7 , 9×9 and 11×11 , and the output channels are 15, 25, 40, 64. To handle the variable length input (different speech utterance has different length), a 2-D global average pooling layer was added such that the features can be fixed at 64-dimension. Furthermore, we use two MLPs with 128-dimensional output as experts for extracting global features. Two gating networks with 2-D output are also used to weight the outputs of experts. After weighted summation, the two results will be sent to the tower network of corresponding metric to get the final prediction.

B. Training details

During the experiments, it was found that CRNN as a generator converges faster than the discriminator PSQN. This not only leads to a longer training period, but also leads to poorer prediction ability of the discriminator, resulting in making speech enhancement effect unsatisfactory. To solve this problem, we pre-trained CRNN and PSQN separately. Pre-training PSQN is such that PSQN can start from a good convergence point when used by MetricGAN. As for the pre-training CRNN, on the one hand, it is to generate a baseline model as a comparison model for the proposed Multi-MetricGAN. On the other hand, we argue that pretraining PSQN with speech enhanced by CRNN enables PSQN to start adversarial training with better parameters. So, we need to pre-train CRNN before pre-training PSQN. However, after many experiments, it is found that PSQN with too many pre-training epochs is not conducive to training Multi-MetricGAN. Therefore, we pretrain PSQN for only one epoch and then apply it to train Multi-MetricGAN.

When pre-training CRNN, the clean speech amplitude spectrum is the learning target. The loss function is MSE of the amplitude spectrum of clean speech and that of enhanced. So, here the CRNN pre-trained is denoted as $CRNN(MSE)$. Then use equation (4) as the loss function to pre-train PSQN, which is denoted as $PSQN$. By employing equation (7) (10) as the loss function for the generator and discriminator, respectively, we can fine-tune $CRNN(MSE)$ and $PSQN$ in alternating training. It is worth mentioning to compare with the baseline model, we need to control all variables except the model structure to be consistent with the baseline model. For example, we still use the dataset used for pre-training

Table I: PESQ of enhanced speech for models.

MODEL	PESQ (-12db)	PESQ (-6db)	PESQ (0db)	PESQ (6db)	PESQ (12db)	PESQ (Avg.)
NOISY	0.882	1.131	1.483	1.916	2.362	1.555
CRN (Baseline)	1.036	1.760	2.306	2.697	3.013	2.162
CRN/QN	1.040	1.784	2.384	2.791	3.097	2.220
CRN/PSQN	1.527	2.600	2.928	3.247	3.472	2.755

CRNN in the fine-tuning process. The final model is denoted as CRNN/PSQN.

C. Results and discussion

In addition to using the baseline for comparison with proposed model, we also did a set of comparative experiments employing speech enhancement model based on MetricGAN. The generator used in comparative experiment is still CRNN, while the discriminator is replaced by Quality-Net, a CNN-based single-target discriminator, proposed by MetricGAN. Here this model is denoted as CRNN/QN.

The experimental results are shown in Table I. The results show that CRNN/PSQN presents the best speech enhancement effect, and PESQ of enhanced speech generated by it is 0.593 higher than that by baseline.

D. Generalization ability evaluation of PSQN

To test the generalization ability of PSQN, we select several datasets without clean reference to test. We chose DNS, VOX celeb, and LibriSpeech because these datasets contain speech in natural scenes. We regard the enhanced speech generated by our proposed Multi-MetricGAN as the clean reference and compute the "actual" metric scores by comparing the "reference" and raw utterance. Then we input these utterances into PSQN to generate the predicted metric scores. The test process is shown in Fig.5 and the experimental results are shown in Table II.

It can be seen from the table that on VOX celeb, the metric scores predicted by PSQN is strongly correlated with the actual metric scores. PSQN performs slightly worse on DNS

Table II: Generalization ability of PSQN.

DATASET	PESQ-COR	STOI-COR
DNS	0.754	0.798
VOX celeb	0.833	0.850
LibriSpeech	0.688	0.721

Table III: Generalization ability comparison between PSQN and Quality-Net.

DATASET	PESQ-COR(PSQN)	PESQ-COR(Q-N)
DNS	0.754	0.701
VOX celeb	0.833	0.779
LibriSpeech	0.688	0.655

and LibriSpeech, but the predicted score also slightly correlates with the actual score.

Since Quality-Net is designed for the evaluation behavior of the model PESQ function, we can only compare Quality-Net and PSQN in predicting PESQ. Table III shows the result of comparative experiments.

Experimental results show that our proposed PSQN performs better than Quality-Net when assessment PESQ. At present, the industry is gradually looking for an automated test method for speech quality assessment through Artificial Intelligence (AI). PSQN based on MMoE maybe has a bright future.

III. CONCLUSIONS

In speech enhancement networks, the minimum mean square error between enhanced speech and clean speech cannot directly match human perceptual hearing. This paper proposes an architecture that employs PESQ and STOI as training targets for speech enhancement models. It is based on GAN. In the training process of the discriminator, although PESQ and STOI are closely related, there is a "seesaw" problem between them. This is an inherent problem of multi-task learning, essentially the inconsistency of gradient optimization directions caused by different loss functions. This paper improves the MMoE network so that PESQ and STOI converge quickly. The experimental results show that compared with using speech amplitude spectrum as learning target, the proposed Multi-MetricGAN which regards the speech evaluation metric score as that performs better in speech enhancement. In addition, we also test the generalization of PSQN on other datasets, and the

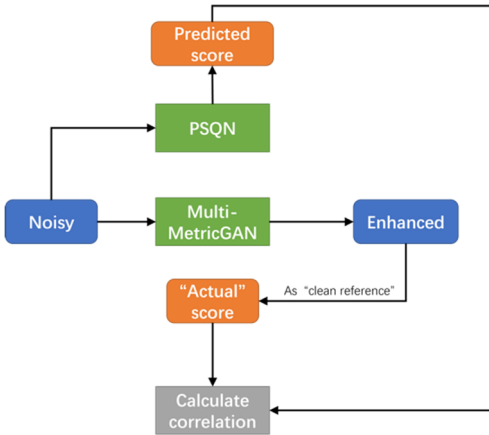


Fig. 5 Test of generalization ability evaluation of PSQN

experimental results show that PSQN performs well on these unseen datasets.

REFERENCES

- [1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
- [2] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 663–678, Apr. 2019.
- [3] Z. Zhao, S. Elshamy, H. Liu, and T. Fingscheidt, "A CNN postprocessor to enhance coded speech," in *Proc. 16th Int. Workshop Acous. Signal Enhancement*, 2018, pp. 406–410.
- [4] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.
- [5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
- [6] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," in *Proc. IEEE Int. Conf. Acous., Speech Signal Processing*, 2018, pp. 5054–5058.
- [7] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
- [8] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, 2018, pp. 1577–1581.
- [9] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 006–012.
- [10] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain," in *Proc. Interspeech*, 2018, pp. 1136–1140.
- [11] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2019.
- [12] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Interspeech*, 2019, pp. 2723–2727.
- [13] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2718.
- [14] Y.-S. Lee, C.-Y. Wang, S.-F. Wang, J.-C. Wang, and C.-H. Wu, "Fully complex deep neural network for phase-incorporating monaural source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 281–285.
- [15] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.
- [16] Y. Koizumi, K. Niwa, Y. Hioka, K. Koabayashi, and Y. Haneda, "DNNbased source enhancement to increase objective sound quality assessment score," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1780–1792, Oct. 2018.
- [17] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech*, 2018, pp. 1873–1877.
- [18] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019.
- [19] Xie R, Liu Y, Zhang S, et al. "Personalized approximate pareto-efficient recommendation," *Proceedings of the Web Conference 2021*. 2021: 3839-3849.
- [20] J. Ma, Z. Zhao, X. Yi, "Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts," *ACM*, 2018.
- [21] Caruana R. "Multitask learning," *Machine learning*, 1997, 28(1): 41-75.
- [22] K. Tan, D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech*, 2018, pp. 3229–1140.
- [23] Caruna R. "Multitask learning: A knowledge-based source of inductive bias," *Machine learning: Proceedings of the tenth international conference*. 1993: 41-48.
- [24] M. Kawanaka, Y. Koizumi, R. Miyazaki, and K. Yatabe, "Stable training of DNN for speech enhancement based on perceptually motivated black-box cost function," in *Proc. ICASSP*, 2020.
- [25] J. Garofolo et al. "Getting started with the darpa timit cdrom: An acoustic phonetic continuous speech database." *National Institute of Standards and Technology (NIST), Gaithersburg, MD*, 107:16, 1988.
- [26] J. Chen, Y. Wang, S. Yoho, D. Wang, and E. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, pp. 2604–2612, 2016.