Group Multi-Scale convolutional Network for Monaural Speech Enhancement in Time-domain

Juntao Yu* , Ting Jiang* and Jiacheng Yu*
* Beijing University of Posts and Telecommunications, Beijing, China
E-mail: {yujuntao, tjiang, yujiacheng333}@bupt.edu.cn

***Abstract*—Recent researches show that convolutional neural networks (CNN) can effectively enhance speech signal by modeling its long-term dependence in time-domain. However, the unscaled speech sequence length challenges the receptive field of the convolutional speech enhancement system. This paper proposes a plug-and-play bottleneck module named group multi-scale (GMS) module to alleviate the receptive field craving of convolutional neural networks. The GMS module adopts Group-Communication fashion, where each feature group can send messages to both adjacent groups and output features by convolutional encoding. In this way, the series group forms a sub temporal convolutional network (TCN) in a single residual block, bringing several times the receptive field of the standard bottleneck module. Experimental results on TIMIT datasets show that the proposed module achieves 1.2 dB SI-SNR gain in the TasNet framework compared with the baseline Con-TasNet.**

Index Terms: speech enhancement, monaural, multi-scale, group

I. Introduction

Speech enhancement [1][2], aiming to improve the intelligibility and quality of speech in the complex acoustic environment, is always an active research topic in speech processing. These technologies are widely used in computational auditory fields, such as automatic speech recognition [3], telecommunications, online conference, hearing aid.

Recently, speech enhancement based on convolutional neural networks (CNN) [4] has made remarkable progress. There are several advantages of parallel computing and stable training compared to RNN [5][6] and self-attention [7]-[9] based methods. However, the limitation on their receptive field introduces an upper bound to the modeling ability [10], corresponding to models' topology. To alleviate this limitation, most works adopt short-term Fourier transform (STFT) to transform the short-term speech sequences into frequency components and model inter-frame dependencies. Although these means are sufficient to process long sequences, they still introduce the phase reconstruction problem [11][12], which leads to distortion on estimated waveforms. In this case, some methods are proposed to conquer this problem in time-domain, pointing out the direction for the follow-up work.

There are two mainstream frameworks for waveform backbone. One is to use a temporal mask enabled front-end which can be trained with the post-network jointly. The other is to fit the regression function to reconstruct the pure waveform directly. Most speech enhancement methods with the time-domain audio separation network (TasNet) [13] framework apply a stride convolution encoder-decoder to replace the traditional STFT to build a high efficient speech separation system. Unfortunately, the receptive field lacking problem [14] is also occurred when they model the especially long-term dependencies. Building models with U-net framework [15] or sub U-net modules [16] can be the typical solution to achieve substantial modeling length, which is benefited by multi-scale modeling. However, the chessboard artifacts [17] introduced by standard inverse convolution up-sampling would depress sequential inference performance of these neural network. Alternatively, temporal convolutional network (TCN) [18]-[20] designs a stacked convolution layer with increasing dilation rates to obtain the receptive field close to the U-net series models instead of scaling. Since long sequence speech modeling requires a sufficient receptive field, many dilated modules[21] are stacked in such models, which can lead to insufficient feature processing and redundant features.

This paper propose a novel group multi-scale (GMS) module to extend the receptive field and encourage feature reuse for TasNet framework. GMS module can introduce the multi-scale and the features reuse attributes into the TCN backbone. In GMS, the splitting operation along the channel dimension divides the input features into several groups. Each group uses the dilated convolution module to process the corresponding scale features and identity path to avoid gradient vanish. The series group forms a sub-TCN, making the GMS module have a receptive field much larger than the traditional bottleneck module. We equipped the proposed module in the TCN module with an adaptive front-end like Conv-TasNet. As a subsidiary research, the adaptability and performance gain of the proposed module to deep separable convolution and dense links is also explored. Experimental results show that the proposed module has higher application priority than the bottleneck module used by Conv-TasNet.

The rest of the paper is organized as follows.We introduce the proposed Group Multi-Scale network in Section 2, describe the experiment procedures in Section 3, analyze the experiment results in Section 4, conclude this paper in Section 5.

II. Speech enhancement with Group Multi-scale Network

As shown in Fig. 1, the TasNet with GMS modules consists of the following blocks: an encoder, a mask estimation system, and a decoder. Firstly, the encoder block embeds the noisy speech into corresponding feature maps into the hidden space

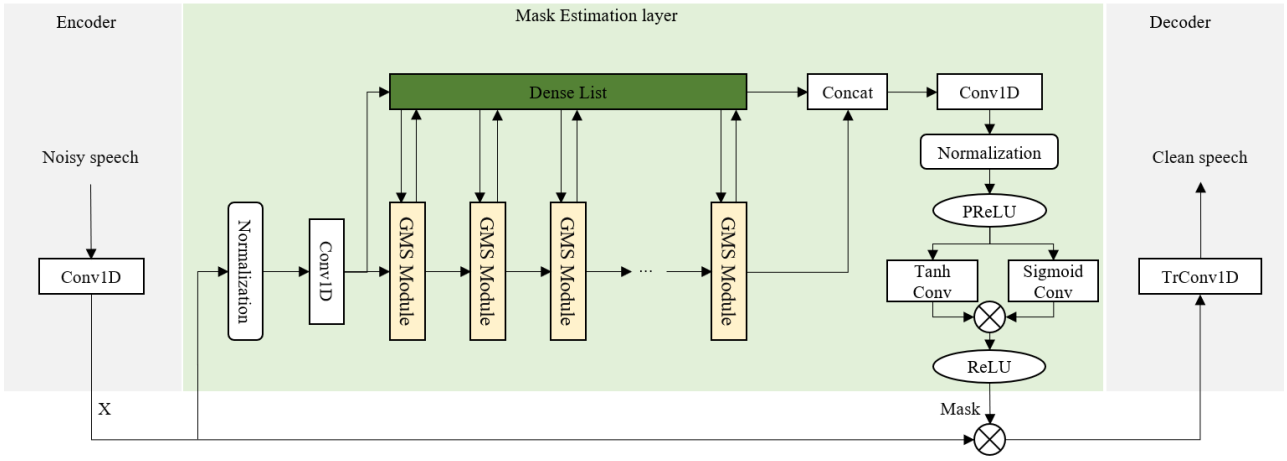

Fig. 1. The architecture of TasNet with GMS modules.

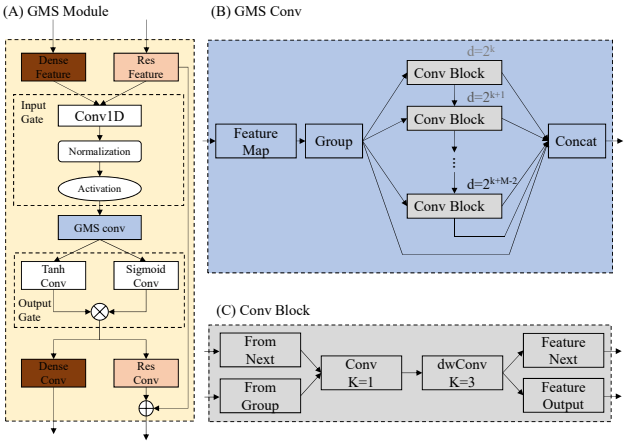

Fig. 2. Detail of GMS modules.

frame-by-frame. After that, these maps are feed into the mask estimation system to generate a time-domain mask to enhance waveform in hidden space. Finally, the decoder block estimates clean waveform using the masked feature maps. In this section, we describe the details of each block.

A. Encoder

We denote the noisy speech $s \in \mathbb{R}^{1\times T}$. Then, the encoder transform s into hidden representation $\mathbf{X} \in \mathbb{R}^{N\times L}$ with frame length L and frame count N. The encoder, a 1-D convolutional layer, can be formulated as follow:

$$\mathbf{X} = \text{ReLU}(s * \mathbf{W}) \tag{1}$$

where $*$ is the convolution operation, $\mathbf{W}$ is the weights with P kernel size and stride S_{enc}. After weighted layer, Rectified Linear Unit (ReLU) is a nonlinear function applied to filter out negative coefficient.

B. Mask estimation system

In mask estimation system, the non-negative features $\mathbf{X}$ are firstly linear compressed by a bottleneck module, which is composed of a global normalization layer (GLN) and a point-wise 1D convolutional layer. Similarly, an extra bottleneck layer is also applied to initialize the data flow in dense path. After such processing, features are feed into K stacked GMS modules.

$$\mathbf{f}_{des} = \text{Conv1D}([\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_{k-1}]), \quad \mathbf{f}_{res} = \mathbf{X}_{k-1} \tag{2}$$

where $\mathbf{f}_{des} \in \mathbb{R}^{N\times H/4}$ is the dense input, $\mathbf{f}_{res} \in \mathbb{R}^{N\times H}$ is the residual input, $\mathbf{X}_{k-1}$ is the residual feature of $k-1$-th GMS module and $[\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_{k-1}]$ is the concatenation of those former dense features. The GMS module integrate both dense and residual features by a 1D convolution:

$$\mathbf{f} = \text{h-Swish}((\text{GLN}(\text{Conv1D}([\mathbf{f}_{des}, \mathbf{f}_{res}])))) \tag{3}$$

where $\mathbf{f} \in \mathbb{R}^{N\times M\times H}$ is the input of GMS conv with channels $M \times H$. A nonlinear activation function and a global layer normalization operation are applied after 1-dimension convolution. The nonlinear activation function is hard-swish:

$$\text{h-Swish}(x) = x\frac{\text{ReLU6}(x+3)}{6} \tag{4}$$

After fusion, GMS module applies a group multi-scale operation to process the features. The operation is shown in Fig. 2(B). First, the features are split into M groups along the channel dimension:

$$\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_M = Group(\mathbf{f}) \tag{5}$$

where the group feature map $\mathbf{g}_p \in \mathbb{R}^{N\times H}, \text{p} = 1, 2, ..., M$ is the index of conv block and $\mathbf{g}_M$ as the identity mapping. $M-1$ serial conv blocks with dilation rates $2^{k+0}, 2^{k+1}, 2^{k+2}..., 2^{k+M-2}$ are applied to model the features and specifically shown in Fig. 2(C).It is worth noting that we set the dilation rate of p-th conv block as 2^{k+p-1} when $p \leq 8$ and 2^{k+p-9} when $p > 8$. In p-th conv block, a

dilated standard 1-D convolution block and a dilated depth-wise separable convolution is used to handle the feature $\mathbf{n}_p$ from the $p-1$-th conv block and the feature $\mathbf{g}_p$ after a group operation:

$$\mathbf{i}_p = \text{dw-Conv1D}(\text{Conv1D}([\mathbf{g}_p, \mathbf{n}_{p-1}])) \quad (6)$$

where $\mathbf{i}_p \in \mathbb{R}^{N \times C_p}$ is the output of conv block, and $C_p = H + C_{p-1}/2, \quad C_1 = H$. Then, we divide $\mathbf{i}_p$ into two feature maps along the channels dimension: $\mathbf{n}_p \in \mathbb{R}^{N \times C_p/2}$ and $\mathbf{o}_p \in \mathbb{R}^{N \times C_p/2}$. The feature maps $\mathbf{o}_p$ of M conv blocks are concatenated as the out of GMS conv :

$$\mathbf{x}_g = [\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_{M-1}, \mathbf{n}_{M-1}, \mathbf{g}_M] \quad (7)$$

where $\mathbf{x}_g \in \mathbb{R}^{N \times C}$ and $C = (\sum_{p=1}^{M-1} C_p + C_p)/2 + H$.

After GMS conv, an output gate is used to process $\mathbf{x}_g$ in Fig. 2(A). In the output gate, there are two parallel 1-D convolutional blocks: one with a tanh activation function and the other with a sigmoid activation function:

$$\mathbf{X}_{gms} = \tanh(\text{Conv1D}(\mathbf{x}_g)) \otimes \text{sigmoid}(\text{Conv1D}(\mathbf{x}_g)) \quad (8)$$

where $\mathbf{x}_g \in \mathbb{R}^{N \times C}$ and $\otimes$ denotes the Hadamard product. The feature $\mathbf{x}_g$ is fed to the dense branch and residual branch. In the dense branch, $\mathbf{x}_g$ is converted to $\mathbf{d}_k$ by a 1-D convolutional operation. And in the residual branch, $\mathbf{x}_g$ is processed by one-dimensional convolution layer and added with the identity mapping $\mathbf{X}_{k-1}$.

$$\mathbf{X}_k = \text{GLN}(\text{Conv1D}(\mathbf{X}_{gms}) + \mathbf{X}_{k-1}), \ \mathbf{d}_k = \text{Conv1D}(\mathbf{X}_{gms}) \quad (9)$$

Finally, the last GMS module's output $\mathbf{X}_K$ and the K dense features are concatenated as the input of a 1-D convolutional layer with GLN and PReLU nonlinear function. And the output is passed to an output gate for estimating mask $\mathbf{M} \in \mathbb{R}^{N \times L}$.

C. Decoder

In decoder, estimated speech $\hat{s} \in \mathbb{R}^{1 \times T}$ is reconstructed by a 1-D transposed convolution block:

$$\hat{s} = (\mathbf{M} \otimes \mathbf{X}) * \mathbf{W}' \quad (10)$$

where $\mathbf{W}'$ is the weight matrix of the 1-D transposed convolution block.

III. Experiments

A. Datasets

We evaluated the GMS-Net on the TIMIT dataset [22] and MUSAN-Noise-Free dataset [23]. The clean speech set includes 6300 utterances from 630 speakers. The noise dataset includes 600 kinds of noises. 80% of the clean speech set are chosen for training, 10% for testing, and 10% for evaluating. 484 kinds of noise are chosen for training, 58 kinds for testing, and 58 kinds for evaluating. The noisy speeches with 16384 sample points are generated by randomly selecting utterances from different speakers in the clean speech dataset and different noise samples in the noise dataset. They are additive at random signal-to-noise ratios(SNR) between -10dB and 0dB. The generated dataset contains 5040, 630 and 630 utterances in the train/test/evaluation set. All the waveforms are resampled at 8kHz.

TABLE I
HYPER PARAMETER SETTINGS.

symbol	description	value
L	number of output channels in encoder	512
P	kernel size of the encoder	17
S_{enc}	stride of the encoder	8
K	numbers of GMS modules	16
H	number of channels in residual feature	128
M	number of groups	5

TABLE II
ABLATION STUDY RESULT.

system	SI-SNR (dB)	PESQ	STOI (%)	Parameters (M)
GMS-Net	15.21	3.06	92.94	8.1
GMS-Net-d	14.41	2.89	91.19	8.1
GMS-Net-dw	15.08	3.06	92.68	10.1

B. Experiment Configurations

In the training stage, the networks with hyperparameters shown in Table I. Each model are trained for 100 epochs on 2s long segments. The learning rate of Adam optimizer is initialized to 1e-3. The learning rate will be halved if the score of the training dataset is not increased in 3 consecutive epochs.

C. Training Objective

The objective of training the proposed network is maximizing the scale-invariant source-to-noise ratio (SI-SNR). SI-SNR can be denoted as:

$$\begin{cases} s_{target} := \frac{\langle \hat{\mathrm{x}}, \mathrm{x} \rangle \mathrm{x}}{\|\mathrm{x}\|^2} \\ e_{noise} := \hat{\mathrm{x}} - s_{target} \\ SI-SNR := 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \end{cases} \quad (11)$$

where $\mathrm{x} \in \mathbb{R}^{1 \times T}$ is and $\hat{\mathrm{x}} \in \mathbb{R}^{1 \times T}$ are the clean speech and reconstructed speech, respectively and $\|\mathrm{x}\|^2 = \langle \mathrm{x}, \mathrm{x} \rangle$ denotes the power of x. Before the calculation, both of them are normalized to zero mean. In the stage of training, utterance-level permutation invariant training(uPIT) [24] is used to solve the problem of source permutation.

D. Ablation Study

In the ablation study, different GMS-Net are trained 100 epochs to reflect the differences between different settings. Table II shows the result of the ablation study on the test noisy set with 0 dB SNR level. Mean SI-SNR, PESQ [25], and STOI [26] on the evaluation dataset are selected as the evaluation metric in this paper. We compare three settings of GMS-Net: the entire model, the model without dilation rates, and the model without depth-wise separable convolution. All of these methods contain 16 GMS modules.

GMS-Net-d: In each GMS module, M blocks are stacked to form a sub-TCN to improve the receptive field of GMS-Net. In this experiment, the dilated depth-wise separable convolutional

TABLE III
SYSTEM ENHANCEMENT PERFORMANCE COMPARISON STUDY ON NOISY SET.

system	SI-SNR (dB)	PESQ	STOI (%)
GMS-Net	15.21	3.06	92.94
Conv-TasNet	14.03	2.93	91.07
DPRNN	15.03	3.02	91.97
Wave-u-net	12.88	2.81	89.17
chimera++	9.62	2.63	84.7
TasNet	13.24	2.79	89.38
IRM	13.45	3.20	92.99
IBM	13.32	2.88	90.64

layer is replaced by a standard depth-wise separable convolutional layer, while other hyperparameters remain unchanged. By comparing GMS-Net to GMS-Net without dilation rates, we find that the dilated convolutional layer can provide 0.8dB, 0.17, and 1.75 gain on SI-SNR, PESQ, and STOI, respectively. The result has shown that dilated convolutional layer can effectively improve the quality of reconstructed speech.

GMS-Net-dw: The proposed model uses the depth-wise separable convolution in each block to decrease the model size. In the ablation study, we replace the dilated depth-wise separable convolutional layer with a standard dilated convolutional layer with the same dilation rates. Without depth-wise separable convolution, GMS-Net has 0.13dB, 0.03, and 0.26 drop on SI-SNR, PESQ, and STOI. The results show that deep-wise separable convolution can improve the performance of GMS-Net slightly and decrease the parameters of the model significantly.

From the results of these experiments, we can find the differences between the three settings of GMS-Net. This undoubtedly proves the effectiveness of our design.

E. Comparison of GMS-Net With Previous Methods

We compared the speech enhancement ability of GMS-Net with previous methods using SI-SNR, PESQ, and STOI. Table III shows the performance of GMS-Net and other advanced methods on the TIMIT dataset. For all speech enhancement systems, the results are based on the evaluation dataset. The hyperparameters of different systems are set according to the original papers.

The results show that the proposed model has a significant gain over Conv-TasNet on all these three metrics. This proves that GMS-Net has a larger receptive field than Conv-TasNet in the TasNet framework. And GMS-Net has better performance than ideal time-frequency magnitude masks, including the ideal binary mask (IBM [27]), ideal ratio mask (IRM [28], [29]). Also, GMS-Net has shown a significant improvement over time-domain approaches like DPRNN, Wave-u-net, indicating the higher application priority of GMS-Net.

IV. CONCLUSIONS

This paper proposes a GMS module to address the problem of the receptive field in CNN architecture. At the same time, this paper further explores the performance gain from the combination of path and deep-wise separable convolution in the GMS block. Our ablation experiments show that GMS improved the TasNet sequence modeling capabilities from both bottleneck module and connection design. In the future, we hope to further explore the potential of GMS modules in mobile speech enhancement systems with lightweight computational costs and causal configuration.

ACKNOWLEDGMENT

This work is supported by the National Nature Science Foundation of China (No.61671075) and (No.61631003).

REFERENCES

[1] Y. Ephraim, H. Lev-Ari, and W. J. Roberts, "A brief survey of speech enhancement 1," in *Microelectronics*, J. C. Whitaker, Ed. CRC press, 2018, pp. 201–220.

[2] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Int. Conf.n Learn. Represent.*, 2018.

[3] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 796–806, 2016.

[4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Real-time speech enhancement using equilibriated rnn," in *2020 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2020, pp. 851–855.

[6] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Commun. Micr. Arrays*. IEEE, 2017, pp. 136–140.

[7] L. Zhou, Y. Gao, Z. Wang, J. Li, and W. Zhang, "Complex spectral mapping with attention based convolution recurrent neural network for speech enhancement," *arXiv preprint arXiv:2104.05267*, 2021.

[8] M. Ge, L. Wang, N. Li, H. Shi, J. Dang, and X. Li, "Environment-dependent attention-driven recurrent convolutional neural network for robust speech enhancement." in *2019 Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3153–3157.

[9] R. Giri, U. Isik, and A. Krishnaswamy, "Attention wave-u-net for speech enhancement," in *2019 IEEE Workshop Applic. Signal Process. Audio Acoust.* IEEE, 2019, pp. 249–253.

[10] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *2020 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2020, pp. 46–50.

[11] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *2019 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2019, pp. 71–75.

[12] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceed. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, 2020, pp. 9458–9465.

[13] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2018, pp. 696–700.

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[15] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[16] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th Int. Workshop Mach. Learn. Signal Process.* IEEE, 2020, pp. 1–6.

[17] C.-H. Pham, A. Ducournau, R. Fablet, and F. Rousseau, "Brain mri super-resolution using deep 3d convolutional networks," in *2017 IEEE 14th Int. Symp. Biomed. Imag.* IEEE, 2017, pp. 197–200.

[18] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conf. Comput. Vision*. Springer, 2016, pp. 47–54.

[19] L. Zhang and M. Wang, "Multi-scale tcn: Exploring better temporal dnn model for causal speech enhancement," *Proc. Interspeech 2020*, pp. 2672–2676, 2020.

[20] V. Kishore, N. Tiwari, and P. Paramasivam, "Improved speech enhancement using tcn with multiple encoder-decoder layers," *Proc. Interspeech 2020*, pp. 4531–4535, 2020.

[21] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Int. Conf. Multimedia Model.* Springer, 2020, pp. 653–665.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus CD-ROM {TIMIT}," National Institute of Standards and Technology, Tech. Rep., 1993.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[25] ITUT Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, Tech. Rep., 2001.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2010, pp. 4214–4217.

[27] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed. Springer, 2005, pp. 181–197.

[28] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Commun.*, vol. 51, no. 3, pp. 230–239, 2009.

[29] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.