



# Coarse-Grained Attention Fusion with Joint Training Framework for Complex Speech Enhancement and End-to-End Speech Recognition

Xuyi Zhuang, Lu Zhang, Zehua Zhang, Yukun Qian, Mingjiang Wang

Harbin Institute of Technology, Shenzhen, China

{19S052014, 18B952047, 21B95200, 20S052011}@stu.hit.edu.cn, mjwang@hit.edu.cn

## Abstract

Joint training of speech enhancement and automatic speech recognition (ASR) can make the model work robustly in noisy environments. However, most of these models work directly in series, and the information of noisy speech is not reused by the ASR model, leading to a large amount of feature distortion. In order to solve the distortion problem from the root, we propose a complex speech enhancement network which is used to enhance the speech by combining the masking and mapping in the complex domain. Secondly, we propose a coarse-grained attention fusion (CAF) mechanism to fuse the features of noisy speech and enhanced speech. In addition, perceptual loss is further introduced to constrain the output of the CAF module and the multi-layer output of the pre-trained model so that the feature space of the CAF is more consistent with the ASR model. Our experiments are trained and tested on the dataset generated by AISHELL-1 corpus and DNS-3 noise dataset. The experimental results show that the character error rates (CERs) of the model are 13.42% and 20.67% for the noisy cases of 0 dB and -5 dB. And the proposed joint training model exhibits good generalization performance (5.98% relative CER degradation) on the mismatch test dataset generated by AISHELL-2 corpus and MUSAN noise dataset.

**Index Terms:** coarse-grained attention fusion, complex speech enhancement, robust end-to-end speech recognition

## 1. Introduction

In recent years, end-to-end automatic speech recognition (ASR) [1, 2, 3] methods have achieved unprecedented development because of the significant improvement of sequence-to-sequence models [4, 5]. Most of the end-to-end ASR networks are structured and trained for the clean speech. While the ASR models have tolerable performance in everyday situations with mild noise and reverberation, but they perform very poorly in very noisy and extreme environments, such as airports and factories. Therefore, it is still a challenge to construct a high-robustness and stable ASR model to meet the tasks in complex scenes.

In order to boost the noise robustness of ASR, there are two mainstream approaches: data augmentation and joint optimization. Multi-condition training (MCT) [6, 7] and SpecAugment [8, 9, 10] are the two most commonly used data augmentation methods to enhance the robustness of ASR. They all make partial adjustments to the training data. SpecAugment is usually used to help the ASR model reduce overfitting. It does not require additional training conditions, but can only be described as moderately effective when dealing with noisy speech. MCT trains the model with a corpus mixed with noise of different signal-to-noise ratios (SNRs). However, when the input noise type or SNR is not matched, its recognition ability is greatly compromised. The joint training method is to add a speech enhancement component in front of the ASR model to

help enhance the speech. In recent years, deep learning-based speech enhancement methods have achieved remarkable performance, and the main approaches can be divided into two categories: mapping [11, 12] which is to reconstruct the target clean speech from the noisy speech, and masking [13, 14] which is to apply the mask to the noisy speech to obtain the enhanced speech. However, due to the influence of the evaluation metrics of speech enhancement, the enhancement model tends to produce over-smoothed spectrum, which leads to speech distortion and degrades the performance of ASR. These enhancement methods that fail to optimize in the intended direction of ASR are suboptimal solutions. To further improve the problem of optimization direction, joint training has become a popular method. The joint optimization of speech enhancement and ASR can reduce the distortion caused by different optimization directions. It can bring a robustness bonus to ASR. Recently, joint training methods with fusion mechanism have been applied to robust ASR networks [7, 15, 16]. But none of them are structurally designed and evaluated for matching the output of the fusion module with the input of the ASR. And most of existing robust ASR models have not been tested on cross-speech corpora and cross-noise datasets. Therefore, it is still necessary to improve their real robustness, especially in very extreme environments such as 0 dB or lower SNRs than scenes have to be tested.

In this paper, we propose a coarse-grained attention mechanism to fully squeeze the internal information of noisy speech, and also optimize the fusion module with a pre-trained model. To sum up, the main contributions of this paper have three aspects. Firstly, the complex spectrum enhancer (CSE) is proposed to enhance the noisy speech by combining masking and mapping in the complex domain, so as to reduce the speech distortion as much as possible. Secondly, coarse-grained attention fusion (CAF) fuses the noisy features and enhanced features with coarse-grained information. In addition, the input of the loss function for CAF is computed from pre-trained ASR, and the loss can guarantee that the output space of CAF is optimized towards the input space of ASR.

## 2. The proposed joint training method

We propose a non-autoregressive coarse-grained attention fusion (CAF) method with a joint training framework to act as a bridge between speech enhancement networks and recognition networks for robust end-to-end speech recognition. The whole model consists of a complex speech enhancement network, a CAF network and an end-to-end speech recognition network. Firstly, the complex spectrum enhancer (CSE) enhances the noisy complex spectrum in multiple stages. Secondly, the CAF network fuses the features of the enhanced speech with the potential speech features extracted from the noisy speech. Finally, the fused features of the CAF module will be used as

the input of the end-to-end ASR network to complete the speech recognition task. During training, the losses of the three parts together constitute the overall loss, and the model needs to optimize the overall loss. As shown in Fig.1, a joint combination scheme is proposed to optimize the entire model and update the model parameters. To reduce the performance degradation of ASR model caused by feature distortion, the enhanced feature space needs to be closer to the latent feature space of the ASR network. Therefore, we further optimize the output of the CAF module by using the common intermediate features of a fixed pre-trained model  $G$ , such as VGG-19 [17] as the target, so that it can fully learn the latent feature space information of the ASR network.

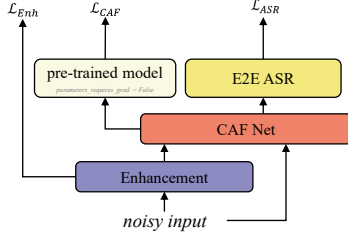


Figure 1: CAF with joint training framework.

## 2.1. Complex spectrum enhancer

The noisy  $y(t)$  in time domain gets  $Y(k, t)$  in time-frequency (TF) domain after short-time Fourier transform (STFT).  $Y(k, t)$  is formed by the superposition of the speech  $X(k, t)$  and the noise  $N(k, t)$  in TF domain. The responsibility of complex spectrum enhancer (CSE) is to eliminate the influence of  $N(k, t)$  on the following networks as much as possible.

Spectral leakage and audio distortion are the main reasons for the high character error rate (CER) in previous noisy speech recognition models. For this reason, we use complex spectrum as our target to this enhancement algorithm. Preserving the imaginary part feature means that the phase information can be preserved while preserving the clear harmonic structure, which is very helpful in suppressing phase-induced distortion.

$$\begin{aligned} Y(k, l) &= \text{Re}(X(k, l)) + i \cdot \text{Im}(X(k, l)) \\ &\quad + \text{Re}(N(k, l)) + i \cdot \text{Im}(N(k, l)) \\ &= X_{RI} + N_{RI} = Y_{RI} \end{aligned} \quad (1)$$

where  $\text{Re}(\cdot)$  and  $\text{Im}(\cdot)$  represents the real and imaginary parts (RI) of STFT spectrum.

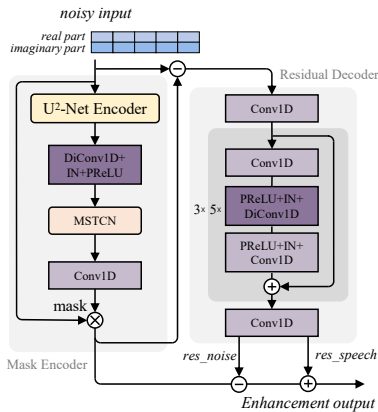


Figure 2: Complex spectrum enhancer.

The overall diagram of proposed CSE network is shown in Fig.2. CSE consists of a mask encoder and a residual decoder. The mask encoder is responsible for estimating the complex mask  $M$  as the initial enhancement stage. The noisy complex spectrum completes point-wise matrix multiplication with the mask  $M$  to obtain the preliminary enhancement result  $X_{pre}$ , as shown in Eq.2:

$$X_{pre} = M \odot Y \quad (2)$$

The signal residuals of  $Y$  and  $X_{pre}$  are fed into the residual decoder as inputs. The residual decoder can further suppress the residual noise components  $X_{res.noise}$  and compensate the missing speech components  $X_{res.speech}$  from the signal residuals.

$$X_{res.noise}, X_{res.speech} = \text{ResDecoder}(Y - X_{pre}) \quad (3)$$

The output of the entire CSE module consists of three sets of complex signals, which are the masked spectrum  $X_{pre}$ , the residual noise spectrum  $X_{res.noise}$  and the residual speech spectrum  $X_{res.speech}$ :

$$\tilde{X} = \text{CSE}(Y) = X_{pre} + X_{res.speech} - X_{res.noise} \quad (4)$$

L2 loss are used as the objective function to optimize the model parameters:

$$\mathcal{L}_{CSE} = \left\| \tilde{X}_{RI} - X_{RI} \right\|_2 \quad (5)$$

In the mask encoder, firstly, the encoder part of the U<sup>2</sup>-Net [18] is connected with a 1D dilated convolution layer (DiConv1D). Instance normalization (IN) and PReLU activation are performed after DiConv1D to produce intermediate RI feature representation. The output dimension of the intermediate RI feature is 322, where each of the real and imaginary parts is 161 dimensions. The multi-scale temporal convolution network (MSTCN) [19] is adopted to perform the multi-scale TF feature analysis on the intermediate RI feature, and finally a 1D convolution (Conv1D) layer is used to estimate the complex mask. As for the U<sup>2</sup>-Net encoder, it contains 4 groups of convolutional structures with decreasing internal layers, which are up-sampling and down-sampling consisting of 2D convolution and 2D deconvolution, to obtain the high-level feature representations. The MSTCN we used consists of four groups of multi-scale residual modules, each of which consists of two 1D convolution layers and a multi-scale dilated convolution layer. The multi-scale dilated convolution layer contains 8 sub-band with kernel size set to 3, and the dilation rates are 1, 3, 5, 7, respectively.

The residual decoder consists of two Conv1D layers and three groups of temporal convolution modules (TCM) in the middle. Each group of TCM consists of five residual blocks with dilation rates of 1, 2, 4, 8, and 16 in sequence. Increasing the receptive field of the residual blocks helps the model to capture long-term embedding features. The final Conv1D layer can estimate the complex spectrum of the residual noise signal and the residual speech signal.

## 2.2. Coarse-grained attention fusion

The traditional joint training method [7] only uses enhanced features for the speech recognition component and still suffers from the problem of audio distortion. As for the CAF network shown in Fig.3, we divide it into two stages including self-attention encoder (SAE) and co-attention decoder (CoAD).  $X$

and  $\tilde{X}$  respectively perform logarithmic Mel matrix multiplication calculation to obtain Fbank features  $O$  and  $\tilde{O}$  as the input of CAF network.

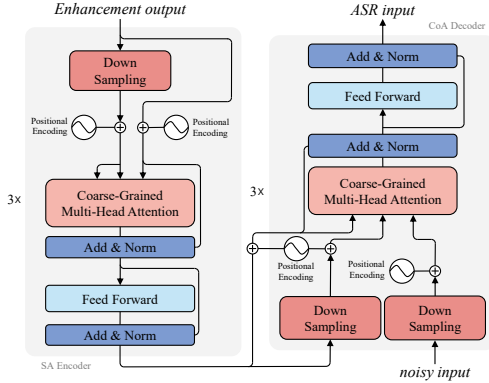


Figure 3: Coarse-grained attention fusion network.

SAE is composed of three repeated blocks with similar structure. Each block contains three sub-layers. The first sub-layer is a downsampling layer implemented with dilated convolutions. The second sub-layer is a coarse-grained multi-head self-attention mechanism (CGMHA), and the third is a feed-forward network. Residual connections [20] are employed around the second and third sub-layer, followed by layer normalization. As the index of the block increases, the DiConv1D with gradually increasing dilated rates of 1, 3, and 5 downsamples  $\tilde{O}$  by a factor of 3 in the time dimension, followed by IN and PReLU. The downsampled  $\tilde{O}$  is added with positional encoding in the time dimension, and then fed into the CGMHA as query and key. The input value is the original  $\tilde{O}$ . To solve the problem of mismatched dimensions of query, key and value, as shown in Fig.4, CGMHA upsamples the transposed product of query and key with deconvolution and extends the attention weights generated by keyframes to all temporal dimensions by position. The matrix of outputs is computed as:

$$\text{CGMHA}(Q, K, V) = \text{softmax}\left(\frac{\text{DeConv2D}(QK^T)}{\sqrt{d_k}}\right)V \quad (6)$$

where  $Q$ ,  $K$  and  $V$  are the results computed by the fully connection layer.  $d_k$  is the dimension of keys.

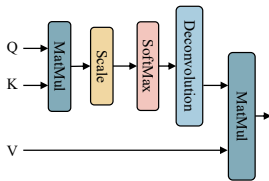


Figure 4: Coarse-grained multi-head attention.

The structure of CoAD is very similar to SAE. It should be noted that in the CGMHA of the CoAD, input key is replaced by the down-sampled  $O$ . This is because the SAE does not introduce the original  $O$ , and the model tends to estimate the distortion part through the existing keyframe information in  $\tilde{O}$ . Using the original  $O$  can effectively help the model recover signal distortion information, such as harmonic information and sibilance. CAF network outputs  $O_{CAF}$ :

$$O_{CAF} = \text{CoAD}\left(\text{SAE}\left(\tilde{O}\right), O\right) \quad (7)$$

To enhance the ability of CAF network to extract ASR features, we extract the hidden features from different layers of the fixed pre-trained model which is the baseline model trained by clean speech datasets. It should be noted that the parameters of the pre-trained model are frozen. Therefore, the CAF network loss function can be formulated as:

$$\mathcal{L}_{CAF} = \sum_{i=1}^n \omega_i \cdot S(G_i(O_{clean}), G_i(O_{CAF})) \quad (8)$$

where  $O_{clean}$  is the clean speech Fbank features.  $G_i$  extracts the  $i$ th hidden features from the fixed pre-trained model.  $S(x, y)$  is the cosine embedding loss between  $x$  and  $y$ ,  $\omega_i$  is a weight coefficient. Perceptual loss [21] measures the difference in cosine similarity between CAF output and clean audio by exploiting multi-layer features from a pre-trained model.

### 2.3. End-to-end automatic speech recognition

In this paper, we employ a state-of-the-art end-to-end ASR method based on low-rank Transformer speech recognition network [22] as our speech recognition component. The end-to-end ASR task can be defined as finding the token sequence  $\tilde{Z}$  to maximize the posterior probability  $Pr(\tilde{Z}|O_{CAF})$  under the given Fbank feature after CAF. Certainly, there are some differences in the performance of the Chinese end-to-end ASR model when outputs different types of tokens. In this work, we choose Mandarin Phonetic Symbols as the form of token output. The number of the heads in multi-head attention is 5. Both the encoder and the decoder block in Transformer are repeated 3 times, and the internal dimension is 512.

The ASR network uses cross entropy as the loss function which can be written as :

$$\mathcal{L}_{ASR} = -\ln P(Z|O_{CAF}) \quad (9)$$

where  $Z$  is the ground truth of the output tokens of the clean speech.

### 2.4. Joint training

The loss functions of the three modules are balanced by hyper-parameters  $\alpha$ ,  $\gamma$  and  $\delta$ .

$$\mathcal{L} = \alpha\mathcal{L}_{CSE} + \gamma\mathcal{L}_{CAF} + \delta\mathcal{L}_{ASR} \quad (10)$$

## 3. Experiment and results

### 3.1. Experimental setups

Our experiments are conducted on the Mandarin speech corpus AISHELL-1 [23] corrupted by the noises from DNS-3 [24]. In addition, to fully test the generalization performance of our model, we also used another open-source clean Mandarin speech corpus AISHELL-2 [25], and MUSAN [26] noise dataset to construct the extra test samples. The sampling rate for all these datasets is 16000 Hz. AISHELL-1 contains above 150-hours speech training set, over 10-hours speech validation set and about 5-hours speech test set. DNS-3 Noise dataset is provided by the Deep Noise Suppression Challenge in 2021. It consists of noise clips containing about 150 different audio categories. We randomly selected 50,000, 4,000, and 2,000 clips from DNS-3 Noise dataset without repetition to construct a noise training set, a noise validation set, and a noise test set, respectively. To construct the cross-corpus test set, we randomly picked up about 5-hours clean speech audio from the test dataset

Table 1: CER results of the different models on different SNR test cases generated by AISHELL-1 and DNS-3 datasets

Model	CER(%)								Paras
	Clean	15dB	10dB	5dB	0dB	-5dB	Ave	Random	
Transformer*	5.81	23.15	30.22	40.34	55.48	77.64	45.37	43.04	49.4M
MCT+TR	6.11	8.63	11.30	16.44	25.18	39.36	20.18	19.47	49.4M
MCT+SpecAugment+TR	4.78	7.12	9.17	13.13	19.59	30.10	15.82	15.36	49.4M
CSE+TR	6.21	14.03	19.95	28.65	41.21	59.15	32.60	31.44	57.0M
(CSE+TR)	5.47	7.72	9.99	13.45	20.64	34.61	17.28	16.51	57.0M
(CSE+CAF)+TR	6.97	14.28	18.22	25.83	38.59	58.33	31.05	30.25	64.4M
(CSE+CAF+TR)	<b>4.10</b>	<b>5.49</b>	<b>6.76</b>	<b>9.12</b>	<b>13.42</b>	<b>20.67</b>	<b>11.09</b>	<b>10.57</b>	64.4M

from AISHELL-2 and 2000 noise clips from MUSAN as the extended test set.

During training, the noisy generator is triggered with a probability of 0.9. The generator synthesizes noisy speech according to SNR which samples from -5dB to 15dB randomly. The final training set is randomly synthesized from the speech training set and the noise training set. Similarly, the validation set and the test set are all generated in the same way. In addition, we sequentially generated the test sets with fixed SNRs of -5, 0, 5, 10, and 15. The model is trained on the training set and evaluated using the validation set to stop the training of the model. Our models are optimized in the same way as [22]. The warmup step is 8000 and the width of beam search in Transformer decoder is 3. Hyperparameters  $\alpha$ ,  $\gamma$  and  $\delta$  are 0.2, 0.3 and 0.5 respectively.

### 3.2. Ablation study

To demonstrate the effectiveness of the proposed CAF joint training framework, we conduct an ablation study to analyze different elements. Different end-to-end ASR models of Transformer (TR) are constructed for comparison. As shown in Table 1, we add the different modules into base network, subsequently. Except for the baseline Transformer\*, which is trained by the clean training dataset of AISHELL-1, the other models are trained on the training set we generated. In the Table 1, '+' represents the combination of different models, and '(.)' represents the model trained by joint training strategies. We measure the performance of models on the clean test set, 5 sets of fixed SNR test sets, and the random SNR test set. The average CER calculated from the results of the 5 sets of fixed SNR cases are also reported. For convenience we call the CER performance of the model on the random SNR test set as random CER.

As we can see that in the same model structure, the performance of the models with joint training is significantly better than that of the models without joint training. If Transformer in those model learns the output features of the front-end models without joint training, the CER performance of those models tested by the clean test set are even worse than the baseline model. On noisy speech datasets, the performance of the front-end model and Transformer can be significantly improved after joint training. Compared to the performance of CSE+TR, the average CER and the random CER performance of (CSE+TR) drop by 15.32% and 14.05%, respectively. More obviously, compared to the performance of (CSE+CAF)+TR, the average CER and the random CER performance of (CSE+CAF+TR) drop by 19.96% and 19.68%, respectively.

As for CAF, if joint training is used, compared to the performance of (CSE+TR), the average CER and the random CER performance of (CSE+CAF+TR) drop by 6.19% and 5.94%, respectively. These show that the loss of CAF calculated by the

pre-trained model during training, which can effectively help the output feature space of CAF to be consistent with the input space expected by ASR. Finally, by testing the CAF with joint training framework we proposed, the CER in the clean speech test set is 4.10%, the average CER in five sets of fixed SNR test sets is 11.09% and the CER in the random SNR dataset is 10.57%.

### 3.3. Generalization performance

As shown in Table 2, we test MCT+SpecAugment+TR and (CSE+TR) and (CSE+CAF+TR) together on the extended test sets. They all have a noticeable performance degradation across datasets. However, compared with the other two models, the model of (CSE+CAF+TR) on all kinds of CER are still the best. By testing the CAF with joint training framework we proposed, the CER in the extended clean speech test set is 9.80%, the average CER in extended five sets of fixed SNR cases is 16.70% and the CER in the extended random SNR test set is 15.64%. Compared to the performance of (CSE+CAF+TR) on the test dataset generated by AISHELL-1 and DNS-3, the performance of (CSE+CAF+TR) on across datasets drops by 5.98%, achieving a stability of 94.02%.

Table 2: CER results of the different models on different SNR test cases generated by AISHELL-2 and MUSAN datasets

Model	CER(%)		
	Clean	Ave	Random
MCT+SpecAugment+TR	11.40	19.37	18.58
(CSE+TR)	14.30	22.71	22.09
(CSE+CAF+TR)	<b>9.80</b>	<b>16.70</b>	<b>15.64</b>

## 4. Conclusion

In this paper, we propose a coarse-grained attention fusion with joint training framework for complex speech enhancement and end-to-end speech recognition. From the perspective of complex domain, complex spectrum enhancer is designed to denoise the noisy speech. In order to solve the problem of speech distortion and make the output space of the front-end model more consistent with the input space expected by ASR network, we apply the coarse-grained attention fusion to further fuse the enhanced speech features with the noisy speech features to extract more robust features for end-to-end ASR. At the same time, we adopt a joint training method to optimize the model towards the overall optimal direction. Experiments on the dataset generated by AISHELL-1 and DNS-3 and the dataset generated by AISHELL-2 and MUSAN show that the method we proposed is effective for robust end-to-end ASR and can solve the speech distortion problem well.

## 5. References

- [1] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [2] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 193–199.
- [3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [4] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [6] F. Li, P. S. Nidadavolu, and H. Hermansky, "A long, deep and wide artificial neural net for robust speech recognition in unknown noise," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 358–362.
- [7] B. Liu, S. Nie, S. Liang, W. Liu, M. Yu, L. Chen, S. Peng, C. Li *et al.*, "Jointly adversarial enhancement training for robust end-to-end speech recognition," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 491–495.
- [8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
- [9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [10] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7829–7833.
- [11] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.
- [12] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.
- [13] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 2472–2476.
- [14] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, 2020, pp. 9458–9465.
- [15] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu, and Z. Wen, "Gated recurrent fusion with joint training framework for robust end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 198–209, 2020.
- [16] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.
- [17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.
- [19] L. Zhang and M. Wang, "Multi-Scale TCN: Exploring better temporal DNN model for causal speech enhancement," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 2672–2676.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 694–711.
- [22] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, and P. Fung, "Lightweight and efficient end-to-end speech recognition using low-rank transformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6144–6148.
- [23] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *IEEE Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.
- [24] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matussevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.
- [25] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.
- [26] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.