INTERACTIVE FEATURE FUSION FOR END-TO-END NOISE-ROBUST SPEECH RECOGNITION

Yuchen Hu, Nana Hou, Chen Chen, Eng Siong Chng
School of Computer Science and Engineering, Nanyang Technological University, Singapore

icassp 2022 Singapore

Poster Number: 2783

Introduction

- Speech enhancement (SE) aims to reduce additive noise from the noisy speech to improve the speech quality for noise-robust automatic speech recognition (ASR). However, recent work observed that the enhanced speech from SE processing could degrade the downstream ASR performance, because some important latent information in the original noisy speech would be reduced by the SE processing together with the additive noise (i.e. over-suppression problem).
- Prior study proposed a joint training approach to optimize SE and ASR modules together via multi-task learning strategy, as shown in Figure 1(a). However, the over-suppression phenomenon still exists since the input information of ASR task only comes from the enhanced speech.
- In this paper, we propose an interactive feature fusion network (IFF-Net) for the end-to-end ASR system to improve its noise-robustness and alleviate the over-suppression problem. We learn a fused representation from the enhanced speech and noisy speech, which acts as the input for the ASR task to complement the missing information in the enhanced speech. Specifically, the IFF-Net consists of two branches to exchange information between enhanced and noisy features. Then, a merge module is proposed to merge these two features, in order to learn clean speech information from the enhanced feature and the complementary information from the noisy feature.

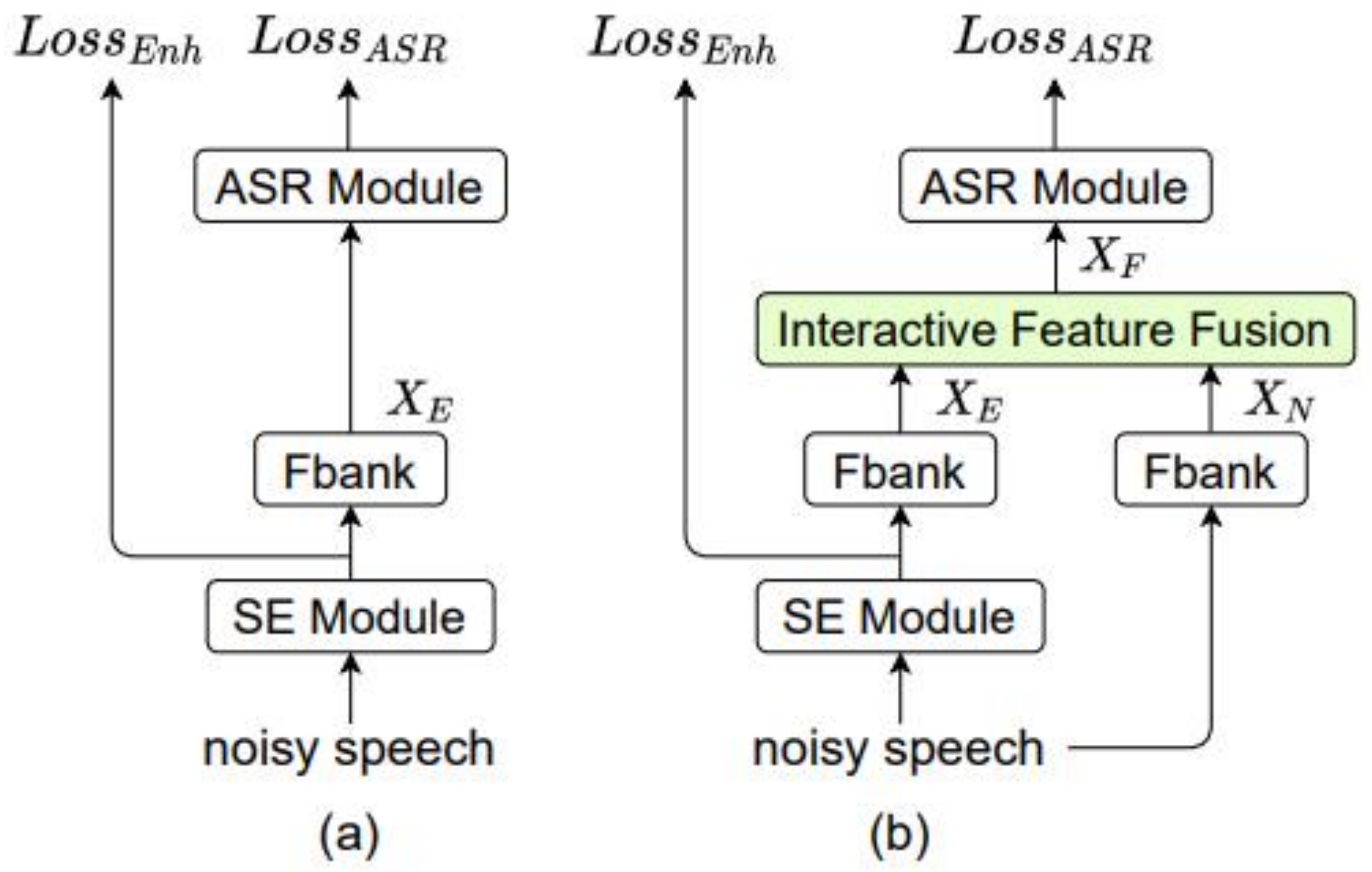

Fig. 1. Block diagrams of (a) joint training approach, (b) joint training approach with our proposed IFF-Net.

Method

IFF-Net (Figure 2) consists of Up/Down Convolutions, Residual Attention (RA) Blocks, Interaction Modules and a Merge Module:

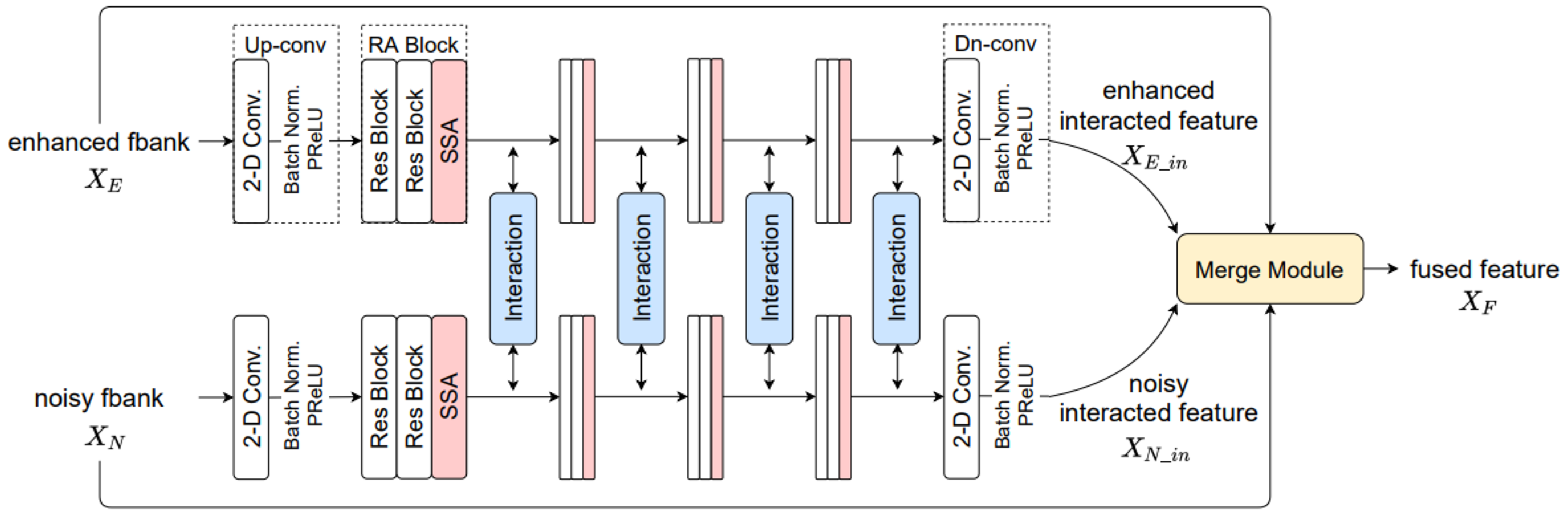

Fig. 2. Block diagram of the interactive feature fusion network (IFF-Net). "Up-conv" is the upsample operation with convolutional layers and "Dn-Conv" is the downsample operation with convolutional layers. "RA" denotes the residual attention block, "SSA" denotes the separable self-attention block, and "Interaction" denotes the interaction module.

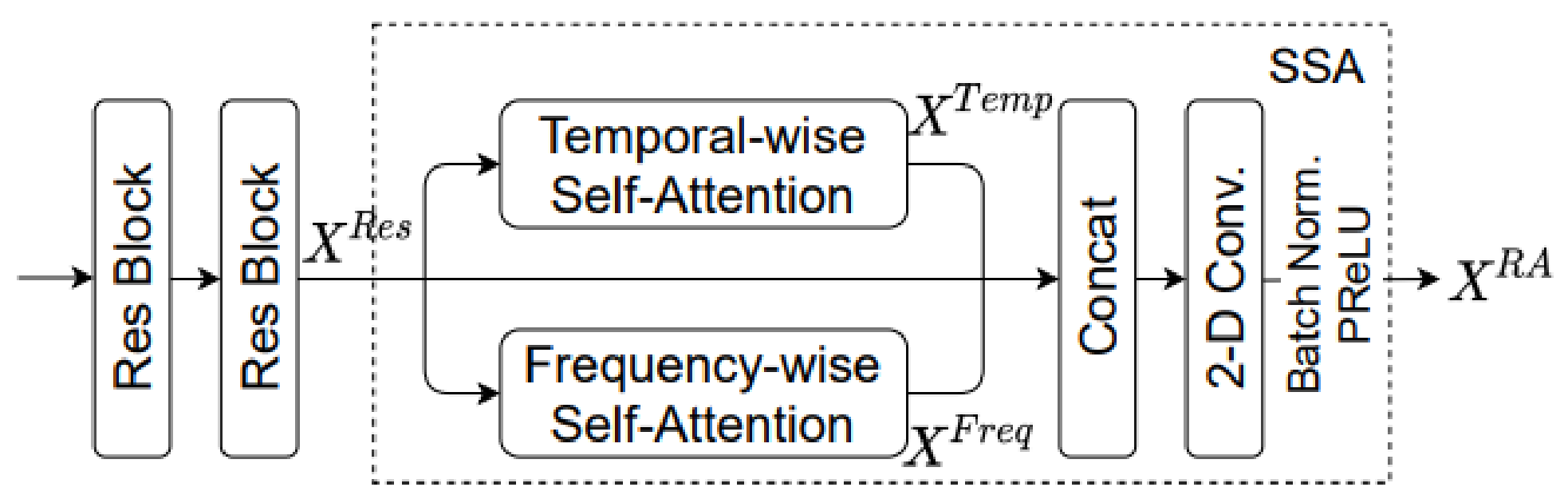

Fig. 3. Block diagram of the residual attention (RA) block. "Res Block" is short for the residual block.

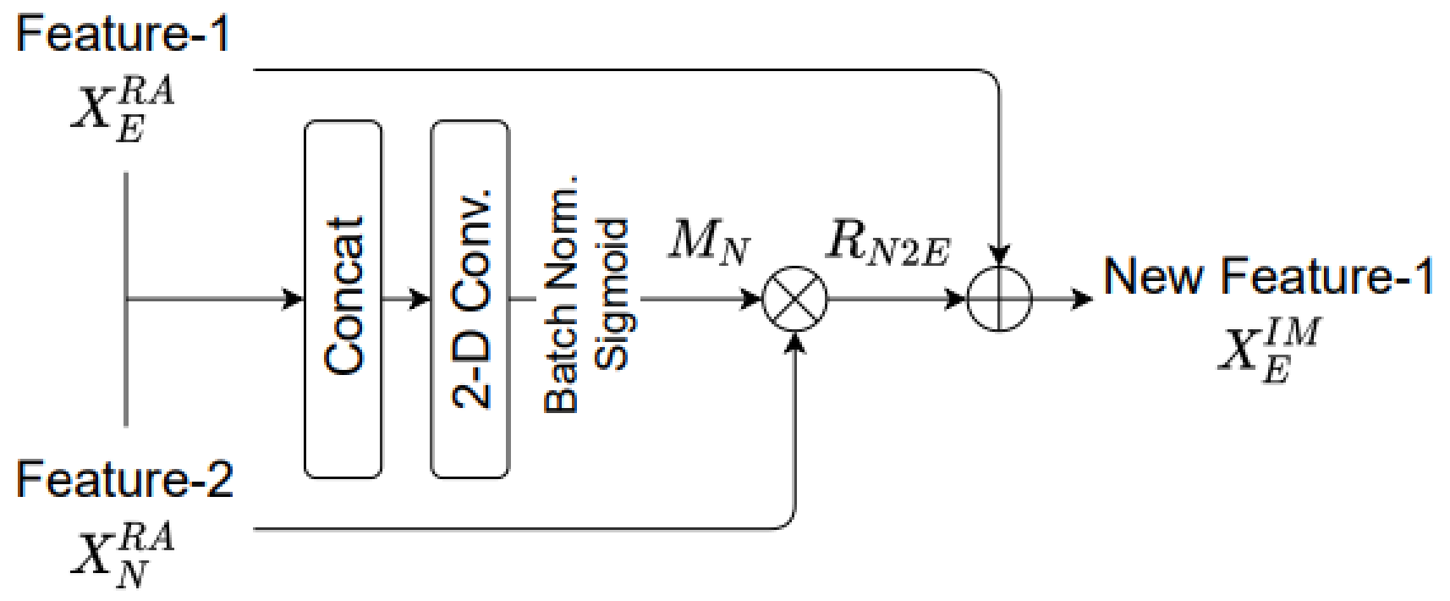

Fig. 4. Block diagram of the interaction module. Here we take the n2e interaction direction as an example for illustration. ⊗ denotes element-wise multiplication, and ⊕ is residual connection.

Table 1. WER% results in an ablation study of the proposed IFF-Net. "# blocks" denotes number of RA blocks in each branch of IFF-Net, "# filters" denotes filter number of the convolutional layer in the RA blocks, and "# params." denotes number of parameters of IFF-Net. Different configurations have been explored to maximize our GPU memory usage.

Method	# blocks	# filters	# params.(M)	WER(%)
IFF-Net	2	32	0.19	49.1
	2	64	0.74	47.9
	4	32	0.37	47.7
	4	64	1.49	**46.2**

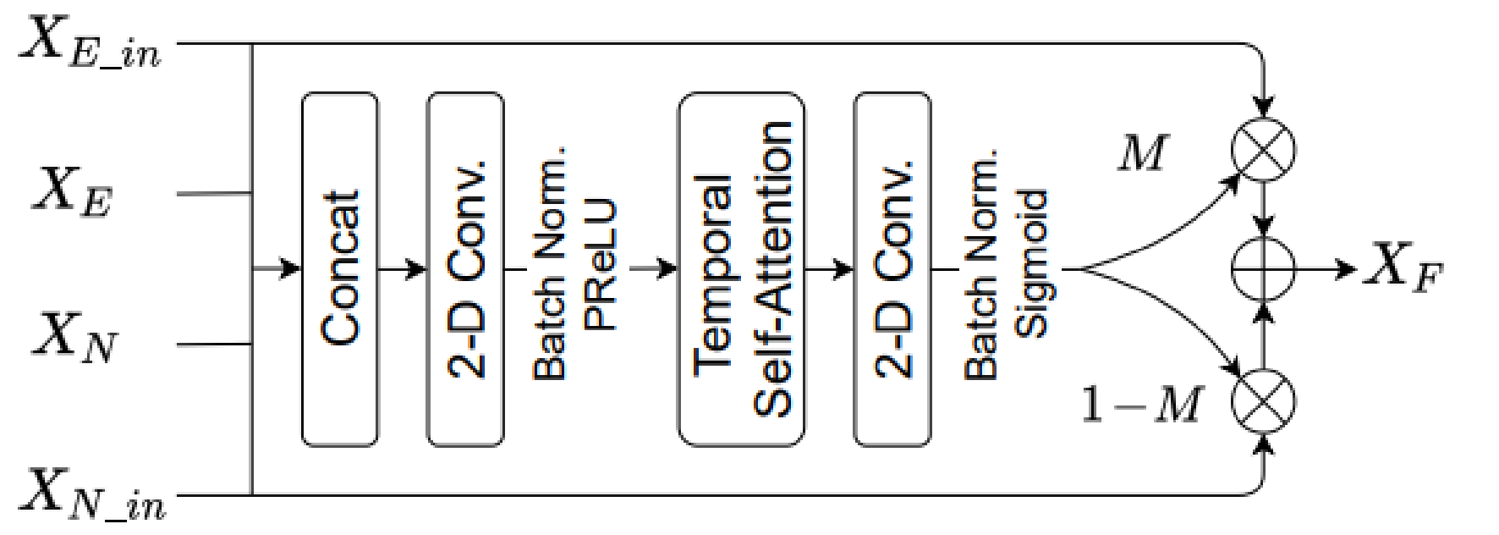

Fig. 5. Block diagram of the merge module. ⊗ denotes element-wise multiplication, and ⊕ is residual connection.

- Up/Down Convolution: 2-D convolution, only to increase/decrease the number of filters, in order to learn deep features along more dimensions;
- Residual Attention (RA) Block: as shown in Figure 3, use residual block and convolution to capture local dependency, and use self-attention block to capture global dependency along both time and frequency axis;
- Interaction Module: as shown in Figure 4, mask based architecture, bi-directionally exchange information between enhanced and noisy features;
- Merge Module: as shown in Figure 5, mask-based architecture, merge the enhanced and noisy features to generate a fused feature for subsequent ASR;

Experiments and Results

We conduct experiments on RATS Channel A corpus (44.3 train data, 0 dB)

Table 2. WER% results in a comparative study of the proposed IFF-Net. "IFF-Net w/o noisy branch" means without the noisy branch, interaction modules and the merge module.

Method	WER(%)
IFF-Net w/o noisy branch	49.5
IFF-Net w/o SSA	49.2
IFF-Net w/o both Interaction Modules	48.0
IFF-Net w/o n2e Interaction Module	47.8
IFF-Net w/o e2n Interaction Module	47.4
IFF-Net	**46.2**

Table 3. WER% results of the proposed IFF-Net and competitive baselines.

Method	WER(%)
E2E ASR System [17]	54.3
Cascaded SE and ASR System [13]	53.1
Joint Training Approach [15]	51.8
GRF Network [16]	50.3
IFF-Net	**46.2**

- More filters and RA blocks lead to better WER results (Table 1);
- Each component in our proposed IFF-Net yield positive effect on the final ASR performance (Table 2);
- IFF-Net has achieved better performance than all competitive baselines, with 4.1% absolute WER reduction over the best baseline (Table 3);

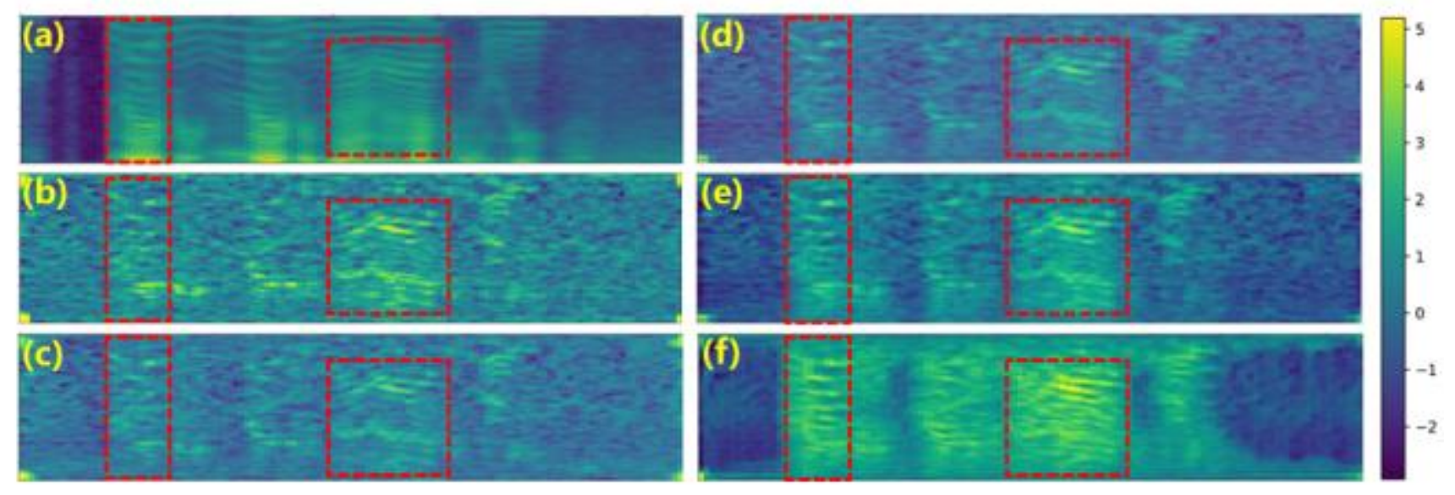

Fig. 6. Spectrums of (a) clean fbank, (b) noisy fbank; and ASR input of (c) Cascaded SE and ASR System, (d) Joint Training Approach, (e) GRF Network, (f) IFF-Net. The colorbar is for all the spectrums.

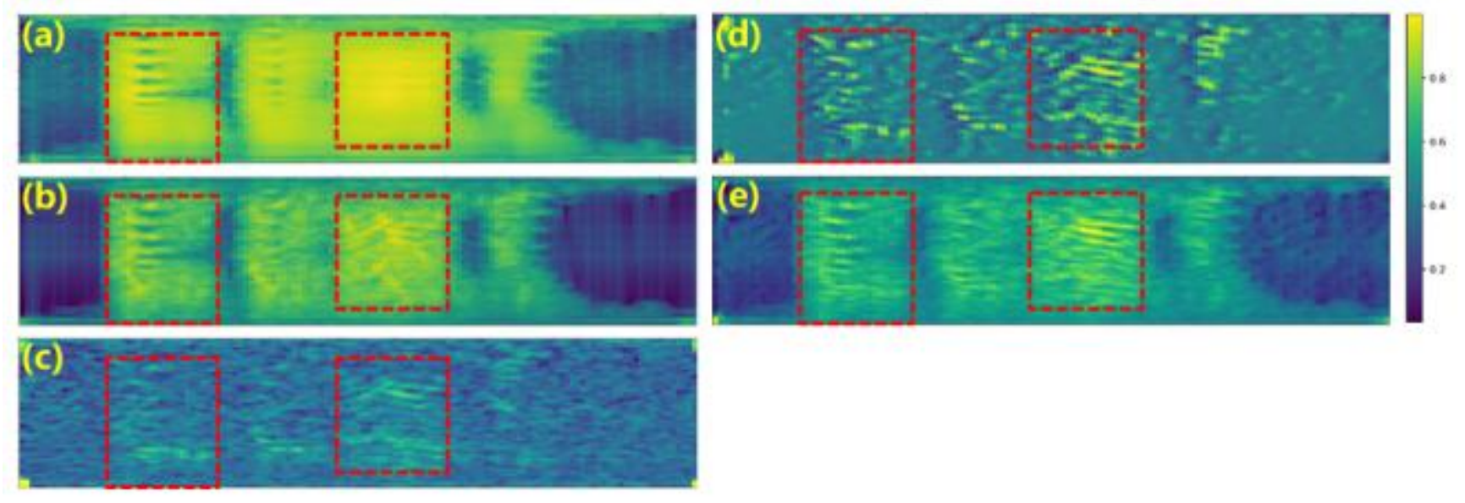

Fig. 7. Spectrums of (a) enhanced fbank, (b) enhanced interacted feature, (c) noisy interacted feature, (d) weight mask, (e) fused feature in IFF-Net. The colorbar is for the weight mask.

- Spectrums of intermediate features (Figure 6) indicate that IFF-Net has generated better speech feature for ASR, with richer speech information and less noise distortions, compared to the baselines;
- We also observe from Figure 7 that RA block and interaction module exchange the enhanced feature and noisy feature, then the merge module learn a mask to generate better feature for ASR;

Conclusion

- In this paper, we propose an IFF-Net for noise-robust speech recognition;
- Specifically, we interactively fuse enhanced speech and original noisy speech to recover lost information in over-suppressed enhanced speech
- The proposed IFF-Net achieves better performance in noise-robust ASR task than the competitive baselines;
- Spectrums indicate that the IFF-Net can indeed alleviate the over-suppression problem