

Optimizing Shoulder to Shoulder: A Coordinated Sub-Band Fusion Model for Real-Time Full-Band Speech Enhancement

Guochen Yu^{1,2}, Andong Li², Wenzhe Liu², Chengshi Zheng², Yutian Wang¹, Hui Wang¹

¹State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

²Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

{yuguochen, wangyutian, hwang}@cuc.edu.cn, {liandong, liuwenzhe, cszheng}@mail.ioa.ac.cn

Abstract

Due to the high computational complexity to model more frequency bands, it is still intractable to conduct real-time full-band speech enhancement based on deep neural networks. Recent studies typically utilize the compressed perceptually motivated features with relatively low frequency resolution to filter the full-band spectrum by one-stage networks, leading to limited speech quality improvements. In this paper, we propose a coordinated sub-band fusion network for full-band speech enhancement, which aims to recover the low- (0-8 kHz), middle- (8-16 kHz), and high-band (16-24 kHz) in a step-wise manner. Specifically, a dual-stream network is first pretrained to recover the low-band complex spectrum, and another two sub-networks are designed as the middle- and high-band noise suppressors in the magnitude-only domain. To fully capitalize on the information intercommunication, we employ a sub-band interaction module to provide external knowledge guidance across different frequency bands. Extensive experiments show that the proposed method yields consistent performance advantages over state-of-the-art full-band baselines.

Index Terms: full-band speech enhancement, sub-bands fusion, dual-stream, decoupling-style concept, multi-stage

1. Introduction

Speech enhancement (SE) aims to rehabilitate the target speech from noise-corrupted mixtures [1]. Recently, deep neural network (DNN) based SE approaches have shown remarkable performance in suppressing highly non-stationary noise over traditional statistical signal processing based methods, especially under low signal-to-noise ratio (SNR) conditions [2]. However, due to the exorbitant computational cost of stepping toward higher frequency bands, most existing DNN-based SE schemes are restricted to the scenarios of narrow- or wide-band speech signals with a sampling rate of 8000 Hz or 16000 Hz. Note that the unit of frequency bands is abbreviated to ‘kHz’ to avoid concept confusion with the unit of the sampling rate in the remainder of this paper.

Instead of performing on the Fourier spectrum directly, previous works usually adopt coarse-grained psycho-acoustically motivated features as the inputs for full-band (sampled at 48000 Hz) SE [3, 4, 5, 6], and thus reduce the overall system complexity. In [3], 22-dimensional Bark-frequency cepstral coefficients (BFCC) defined in the Bark-scale were adopted as input features and 22 ideal critical band gains were mapped as the target. More recently, according to the human hearing equivalent rectangular bandwidth (ERB) scale, PercepNet developed a perceptual band representation with only 34 spectral bands [4]. Although these approaches can lower the frequency-wise feature dimension and the overall calculation complexity can then

be decreased, the frequency resolution of the spectrum in Bark scale and that in ERB scale are much smeared than the original Fourier spectrum, leading to inaccurate spectrum recovery and information loss among frequency bands. Moreover, due to the fact that the wide-band (0-8 kHz) tend to have more energy, tonalities, and harmonics than the higher-frequency bands (8-24 kHz), simultaneously modeling 0-8 kHz and 8-24 kHz frequency bands by a single network may severely degrade SE performance, especially in the high-frequency region [7].

To resolve the aforementioned problems, this paper proposes a coordinated Sub-band Fusion network, dubbed **SF-Net**, for real-time full-band SE using short-time Fourier transform (STFT) features. To be specific, we split the original full-band spectrum into low-band (LB), middle-band (MB) and high-band (HB) spectra, and three sub-networks are elaborately devised to cope with them accordingly. Motivated by the decoupling-mechanism in recent phase-aware wide-band SE methods [8, 9, 10, 11, 12], we first pre-train a dual-stream network, called **DSLB-Net**, to address the LB complex spectrum (0-8 kHz) recovery, which mainly comprises of a magnitude estimation network (ME-Net) and a complex purification network (CP-Net). From the complementary perspective, ME-Net aims to coarsely suppress noise components in the magnitude domain, while CP-Net is established to compensate for the missing spectral details and implicitly recover phase information in the complex domain. Then, we integrate the pretrained DSLB-Net with another two higher-band masking networks, namely **MBM-Net** and **HBM-Net**, to tackle the 8-16 kHz and 16-24 kHz bands. Due to the fact that speech in higher frequency bands contains lower energies and fewer harmonics, we only map the magnitude gain and retain the phase unaltered for the 8-24 kHz bands. Besides, to capitalize on the implicit correlations among different frequency bands, a sub-band interaction module is devised within the MBM-Net and HBM-Net, which aims to extract the knowledge from the estimated LB spectrum as guidance. Finally, the estimated low-, middle- and high-band spectra are fused to obtain the full-band signal. Comprehensive experiments on two public benchmarks well validate the superiority of the proposed method in various evaluation metrics.

The remainder of the paper is organized as follows. In Section 2, the proposed framework is described in detail. The experimental setup is presented in Section 3, while Section 4 gives the results. Some conclusions are drawn in Section 5.

2. Methodology

2.1. Collaborative dual-stream low-band SE

The overall diagram of the proposed approach is shown in Figure 1(a), which is comprised of three sub-networks, namely DSLB-Net, MBM-Net, and HBM-Net. As speech contains more harmonics and semantic information in the frequency range of 0 to 8 kHz, we first employ DSLB-Net to eliminate

Chengshi Zheng is the corresponding author

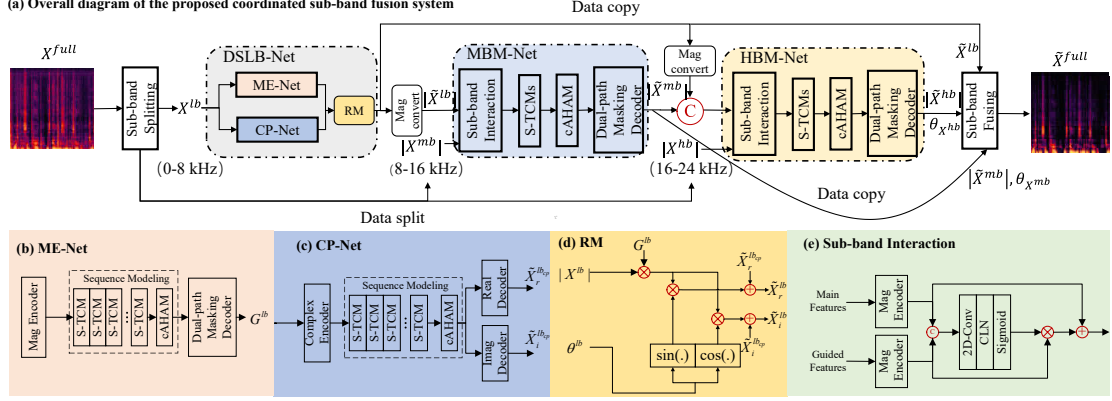


Figure 1: The overall diagram of the proposed system. Different modules are indicated with different colors for better visualization. \odot denotes the concatenation operation in the channel axis.

the noise and recover the clean complex spectrum in the LB regions. Inspired by the efficacy of decoupling-style SE methods [8, 9, 11, 12], we decouple the original complex spectrum estimation into spectral magnitude and phase optimization, and adopt a dual-stream network to collaboratively estimate the magnitude and residual complex components of the low-band (0-8 kHz) spectrum in parallel, which consists of a magnitude estimation network (ME-Net), a complex spectrum purification network (CP-Net), and a reconstruction module (RM), as illustrated in Figure 1(b), (c) and (d), respectively. The input features of ME-Net and CP-Net are denoted as $|X^{lb}| \in \mathbb{R}^{T \times F \times 1}$ and $X_{com}^{lb} = Cat(X_r^{lb}, X_i^{lb}) \in \mathbb{R}^{T \times F \times 2}$, where $\{T, F\}$ denote the number of frames and that of frequency bins, respectively.

In ME-Net, given noisy low-band spectral magnitude $|X^{lb}|$, the network estimates a real-valued gain function G^{lb} , which aims at coarsely filtering out the dominant noise. Then the denoised LB spectral magnitude is coupled with the original noisy phase to obtain the coarse-estimated real and imaginary (RI) spectrum. As the supplementation, we leverage CP-Net to purify the spectral structures and also recover the phase information. Rather than explicitly estimating the complex spectrum from scratch, CP-Net is designed for residual mapping, which alleviates the overall burden of this network. Taking the above estimations as the input, we reconstruct the estimated spectrum via the proposed RM. The whole procedure is thus formulated as:

$$|\tilde{X}^{lb_{me}}| = |X^{lb}| \otimes G^{lb}, \quad (1)$$

$$\tilde{X}_r^{lb_{me}} = |\tilde{X}^{lb_{me}}| \otimes \cos(\theta_{X^{lb}}), \quad (2)$$

$$\tilde{X}_i^{lb_{me}} = |\tilde{X}^{lb_{me}}| \otimes \sin(\theta_{X^{lb}}), \quad (3)$$

$$\tilde{X}_r^{lb} = \tilde{X}_r^{lb_{me}} + \tilde{X}_r^{lb_{cp}}, \quad (4)$$

$$\tilde{X}_i^{lb} = \tilde{X}_i^{lb_{me}} + \tilde{X}_i^{lb_{cp}}, \quad (5)$$

where $\{\tilde{X}_r^{lb_{cp}}, \tilde{X}_i^{lb_{cp}}\}$ denote the output residual RI components of CP-Net and $\{\tilde{X}_r^{lb}, \tilde{X}_i^{lb}\}$ denote the final merged estimation of clean RI components. G^{lb} and $\theta_{X^{lb}}$ denotes the estimated gain of ME-Net and the LB noisy phase, respectively. \otimes is the element-wise multiplication operator.

2.2. Sub-band fusion for full-band SE

Based on the fact that the frequency bands ranging from 8-24 kHz tend to contain less speech information, we further employ two light-weight sub-networks, namely MBM-Net and HBM-Net, as the noise suppressors for the middle-band and high-

band, respectively. To reduce the computational burden of the network and evade the implicit compensation effect between magnitude and phase [13], we only consider the magnitude and retain the phase unaltered in these two bands. Besides, we model the correlations among LB, MB, and HB via the proposed interaction module, where the estimated LB features are employed to guide the spectrum recovery of MB and HB.

To be specific, the estimated LB spectral magnitude by DSLB-Net is fed into MBM-Net and HBM-Net along with the noisy MB and HB spectral magnitude. Taking MBM-Net as an example, as shown in Figure 1(e), we adopt two encoders to extract the magnitude features from noisy MB spectra and the pre-estimated LB spectra. After that, the features are concatenated together and fed into the mask block to derive the gain function, which aims to automatically learn to filter and preserve different areas of the guided LB feature. Then we sum the extracted FB feature and the filtered version from LB to yield the interacted representation and feed for the latter modules. For HB regions, likewise, the denoised LB and MB spectra are concatenated in the channel axis to obtain the guided input features. With the proposed interaction module, external knowledge from DSLB-Net can effectively propagate to MBM-Net and HBM-Net regions and gradually guide the spectrum recovery.

In summary, the operation stream of the middle- and high-band modeling can be formulated as:

$$|\tilde{X}^{mb}| = |X^{mb}| \otimes \mathcal{G}^{mb}(|X^{mb}|, |\tilde{X}^{lb}|; \Phi_1), \quad (6)$$

$$|\tilde{X}^{hb}| = |X^{hb}| \otimes \mathcal{G}^{hb}(|X^{hb}|, |\tilde{X}^{lb}|, |\tilde{X}^{mb}|; \Phi_2), \quad (7)$$

$$\tilde{X}^{mb} = |\tilde{X}^{mb}| \exp(j\theta_{X^{mb}}), \tilde{X}^{hb} = |\tilde{X}^{hb}| \exp(j\theta_{X^{hb}}), \quad (8)$$

where $|\tilde{X}^{mb}|$ and $|\tilde{X}^{hb}|$ denote the denoised MB and HB outputs, respectively. \mathcal{G}^{mb} and \mathcal{G}^{hb} denote the mapping functions of MBM-Net and HBM-Net with parameter set $\Phi(\cdot)$. $\theta_{X^{mb}}$ and $\theta_{X^{hb}}$ denote the noisy MB and HB phase, respectively. Finally, we stack the estimated three sub-band estimation along the frequency axis to obtain the full-band spectrum. Note that we average the overlapped bands in the MB and HB regions.

2.3. Network architecture

The detailed architectures of DSLB-Net, MBM-Net, and HBM-Net are shown in Figure 1. Similar to [8, 9], we employ a classical convolutional encoder-decoder topology [14] for these sub-networks, where multiple temporal convolution modules (TCMs) and a causal adaptive hierarchical attention module (cAHAM) [11] are stacked in the bottleneck for sequence modeling. More specifically, the detailed structures of ME-Net and CP-Net in DSLB-Net are shown in Figure 1(b) and (c), in which

ME-Net utilizes a magnitude encoder and a dual-path masking decoder and CP-Net utilizes a complex encoder and two decoders to recover both RI components.

Taking ME-Net as an example, the encoder consists of five downsampling blocks, each of which consists of a convolutional layer, a normalization layer, and PReLU, with kernel size being (2, 3) in the time and frequency axes except (2, 5) in the first block. The number of channels remains 64 by default, and the stride is set to (1, 2) to gradually halve the frequency size. To enable streaming inference, the features are normalized with cumulative layer norm (cLN) [15], where the statistics are adaptively updated in a frame-wise manner. The dual-path masking decoder consists of five symmetrical deconvolutional layers and a dual-path mask module, which is performed to obtain the magnitude spectral gain by a 2-D convolution and a dual-path tanh/sigmoid nonlinearity operation similar to [11].

Inspired by [16], four groups of squeezed TCMs (S-TCMs) are employed for sequence modeling, each of which stacks six S-TCMs with increasing dilation rates d to obtain a large temporal receptive field, *i.e.*, $d = \{1, 2, 4, 8, 16, 32\}$. To reduce the computational burden, the parameter weights of S-TCMs are shared in ME-Net and CP-Net. Similar to our previous study [11, 17], we utilize a casual AHAM [11] to integrate all intermediate features and global hierarchical contextual information during sequence modeling, where the average pooling layer is adopted along the frequency axis and retain the time resolution unaltered to satisfy the real-time requirement. For MBM-Net and HBM-Net, we leverage a similar structure as ME-Net, which aims at filtering out the dominant noise in MB and HB regions, respectively.

2.4. Loss function

In our SF-Net, we adopt a two-stage training pipeline to recover low-band and full-band spectra progressively. First, we pretrain DSLB-Net with MSE until convergence. The loss function involves both RI and magnitude estimation, given as:

$$\mathcal{L}_{lb}^{RI} = \left\| \tilde{X}_r^{lb} - S_r^{lb} \right\|_F^2 + \left\| \tilde{X}_i^{lb} - S_i^{lb} \right\|_F^2, \quad (9)$$

$$\mathcal{L}_{lb}^{Mag} = \left\| |\tilde{X}^{lb}| - |S^{lb}| \right\|_F^2, \quad (10)$$

$$\mathcal{L}_{lb} = \mu \mathcal{L}_{lb}^{RI} + (1 - \mu) \mathcal{L}_{lb}^{Mag}, \quad (11)$$

where \mathcal{L}_{lb}^{Mag} and \mathcal{L}_{lb}^{RI} denote the loss terms toward magnitude and RI components, respectively. $|S^{lb}|$ denotes the target LB spectral magnitude. S_r^{lb} and S_i^{lb} represent the RI components of target LB regions. With the internal trial, we empirically set $\mu = 0.5$ in the following experiments.

In the second stage, we couple the pretrained DSLB-Net with MBM-Net and HBM-Net, and train them jointly. The overall loss can be given by:

$$\mathcal{L}_{full} = \alpha \mathcal{L}_{lb} + \mathcal{L}_{mb}^{Mag} + \mathcal{L}_{hb}^{Mag} \quad (12)$$

where \mathcal{L}_{mb}^{Mag} and \mathcal{L}_{hb}^{Mag} denote the loss functions for MF-Net and HF-Net in the magnitude domain, while \mathcal{L}_{full} represents the full loss function of the second stage. We empirically find that $\alpha = 0.1$ suffices in our evaluation.

3. Experimental setup

3.1. Datasets

To verify the effectiveness of the proposed model, we conduct extensive experiments on two public full-band benchmarks, namely VoiceBank + DEMAND dataset [18] and the ICASSP 2022 DNS-Challenge (DNS-2022) dataset [19], where all the utterances are sampled at 48000 Hz.

VoiceBank + DEMAND: The dataset is a selection of the VoiceBank corpus [20] with 28 speakers for training and another 2 unseen speakers for testing. The training set consists

Table 1: Ablation study w.r.t. dual-stream structure, sub-band fusion strategy, cAHAM and sub-band interaction module.

Models	Feat.	Para. (M)	MACs (G/s)	wide-band*		full-band	
				PESQ \uparrow	STOI(%) \uparrow	SDR(dB) \uparrow	SSNR(dB) \uparrow
Noisy	—	—	—	1.97	92.1	8.42	1.71
One-stage full-band approaches							
ME-Net (full)	Mag	2.18	3.14	2.72	94.0	16.26	7.23
CP-Net (full)	RI	2.89	5.57	2.67	93.1	15.38	7.34
DS-Net (full)	Mag+RI	3.30	8.71	2.78	94.3	18.86	9.12
Sub-band fusion approaches							
ME-SF-Net	Mag	6.42	3.73	2.83	93.6	16.74	8.13
CP-SF-Net	RI	7.22	6.04	2.90	94.2	17.04	8.28
SF-Net (Pro.)	RI+Mag	6.98	5.62	3.02	94.5	19.90	9.69
- cAHAM	RI+Mag	6.82	5.49	2.94	94.2	19.67	9.13
- Inter.	RI+Mag	6.49	5.11	2.93	94.4	18.99	8.80

*: Calculated on downsampled speech at 16000 Hz.

of 11,572 mono audio samples, while the test set contains 824 utterances by 2 speakers (one male and one female). For the training set, all the audio samples are mixed with one of the 10 noise types, including two artificial noise processes and eight real noise recordings taken from the Demand database [21].

DNS-Challenge: We further train and evaluate our model on the DNS-2022 dataset¹, which consists of various clean speech, noise clips, and room impulse responses (RIRs) to simulate practical acoustic scenarios. For this dataset, we totally generate around 600 hours of noisy-clean pairs. To generate reverberant-noisy training data, we use 248 real and 60,000 synthetic RIRs from openSLR26 and openSLR28 datasets [22]. As only late reverberation degrades the speech quality/intelligibility [23], we preserve both anechoic and early reverberation components as the training target. During each mixing process, the clean speech is convolved with a randomly selected RIR, and is then mixed with the noise in the SNR range of (−5dB, 15dB). We evaluate the model performance upon the DNS-2022 blind test set, which includes 859 real test clips.

3.2. Implementation setup

The 20ms Hanning window is utilized, with 50% overlap between adjacent frames. To extract the features, 960-point FFT is utilized and 481-dimension spectral features are obtained. Due to the efficacy of power compression in both dereverberation and denoising tasks [24], we conduct the power compression toward the spectral magnitude while remaining the phase unaltered, and the compression factor is set to $\beta = 0.5$. All the models are optimized using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [25]. In the first stage, the initialized learning rate (LR) is set to 1e-3 for DSLB-Net. In the second stage, DSLB-Net is fine-tuned with LR of 1e-4, while LR is set to 1e-3 for MBM-Net and HBM-Net. The batch size is set to 16 at the utterance level. **The processed samples are available online, where the source code will be released soon.**²

4. Experimental results and discussion

In this study, we mainly adopt wide-band PESQ [26], STOI [27], CSIG, CBAK, and COVL [28] to evaluate low-band SE performance, while the segmental SNR (SSNR) and SDR [29] are employed for full-band SE evaluation. Higher values indicate better performance.

4.1. Ablation study

We first conduct ablation studies to investigate the effects of the proposed sub-band fusion strategy, dual-stream structure, cAHAM and the sub-band interaction module, including a) three one-stage full-band SE approaches, *i.e.*, the magnitude estimation network (dubbed ME-Net (full)), the complex purification network (dubbed CP-Net (full)), and the dual-stream network (dubbed DS-Net (full)), b) three sub-band fusing-based meth-

¹<https://github.com/microsoft/DNS-Challenge>

²<https://github.com/yuguochencuc/SF-Net>

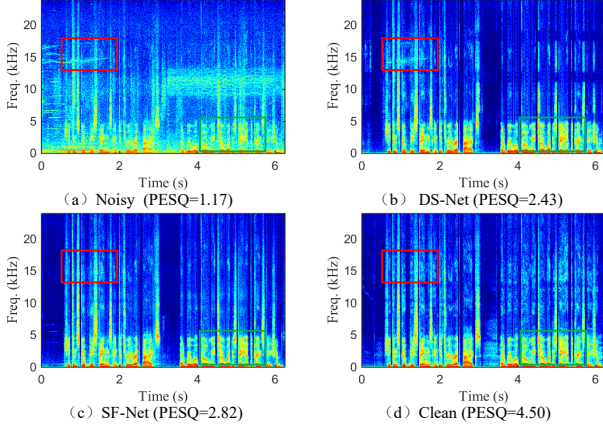


Figure 2: Visualization of spectrograms. (a) Noisy utterance (b) Enhanced utterance by DS-Net. (c) Enhanced utterance by SF-Net. (d) Clean utterance.

ods, namely ME-SF-Net, CP-SF-Net and the proposed coordinated sub-band fusion model (SF-Net), c) SF-Net without the causal adaptive hierarchical attention module (-cAHAM) and sub-band interaction module (-Inter.). Due to the lack of PESQ evaluation for full-band speech, we downsample the outputs of all models at 16000 Hz and measure wide-band PESQ and STOI, while SSNR and SDR are measured on the full-band SE.

Quantitative results are presented in Table 1, one can have the following observations. First, among the one-stage full-band approaches, DS-Net dramatically outperforms other single-stream approaches in terms of wide-band and full-band SE performance. For example, DS-Net provides average 0.21, 1.2%, 3.38dB and 1.78dB improvements than CP-Net in PESQ, STOI, SDR and SSNR, respectively. This reveals the effectiveness of the proposed dual-stream SE topology in improving speech quality. Second, compared with the one-stage full-band SE, when employing the sub-band fusion strategy, considerable performance improvements in terms of both wide-band and full-band cases can be obtained. Besides, we also provide the model size and the number of multiply-accumulate operations (MACs) per second, as shown in Table 1. Note that although SF-Net suffers more parameters than DS-Net (full), it achieves fewer MACs and remarkably better performance. This indicates the effectiveness of the proposed sub-band splitting and fusion strategy. Finally, without using cAHAM and sub-band interaction module, consistent performance degradations in both wide-band and full-band speech are observed, which emphasize the significance of cAHAM and sub-band interaction in improving speech quality.

Moreover, spectrograms of noisy utterance, clean utterance and enhanced utterances by DS-Net and SF-Net are presented in Figure 2 (a)-(d). Focusing on the red and green boxes in Figure 2 (b) and (c), one can see that SF-Net can better suppress background noise than DS-Net in the high-frequency regions, while more spectral details can be restored in the low-frequency regions. This indicates the superiority of SF-Net in improving speech quality in both low- and high-band cases.

4.2. Comparison with full-band SOTA methods

The best configuration of SF-Net in Table 1 is chosen to compare with other SOTA baselines, whose results are presented in Table 2. For VoiceBank + DEMAND dataset, six full-band and one super-wideband SE approaches are selected as the baselines, namely GCRN (full-band version) [30], RNNNoise [3], PercepNet [4], CTS-Net (full-band version) [8], DeepFilterNet [5], DMF-Net [31] and S-DCCRN (super-wide band) [32].

Table 2: Comparison on VoiceBank + DEMAND dataset. “—” denotes that the result is not provided in the original paper.

Models	Year	Para.(M)	PESQ \uparrow	STOI(%) \uparrow	SIG \uparrow	CBAK \uparrow	COVL \uparrow
Noisy	—	—	1.97	92.1	3.35	2.44	2.63
GCRN (full) [30]	2019	10.59	2.71	93.8	4.12	3.23	3.41
RNNNoise [3]	2020	0.06	2.34	92.2	3.40	2.51	2.84
PercepNet [4]	2020	8.00	2.73	—	—	—	—
CTS-Net (full) [8]	2020	7.09	2.92	94.3	4.22	3.43	3.62
DeepFilterNet [5]	2021	1.80	2.81	—	—	—	—
S-DCCRN [32]	2022	2.34	2.84	94.0	4.03	2.97	3.43
DMF-Net [31]	2022	7.84	2.97	94.4	4.26	3.25	3.48
DS-Net (full)	2022	3.30	2.78	94.3	4.20	3.34	3.48
SF-Net (Pro.)	2022	6.98	3.02	94.5	4.36	3.54	3.67

Table 3: DNSMOS P.835 and P.808 results on DNS-2022 blind test set.

Model	DNSMOS P.835*			DNSMOS* P.808
	SIG	BAK	OVRL	
Noisy	4.14	2.94	3.27	3.03
NSNet2 [37]	3.87	4.21	3.58	3.57
DMF-Net [31]	3.92	4.57	3.72	3.61
DS-Net (full)	4.07	4.29	3.77	3.65
SF-Net(Pro.)	4.15	4.49	3.94	3.73

*: Calculated on downsampled speech at 16000 Hz.

Note that we re-implement GCRN and CTS-Net (full) with power compression for a fair comparison, while we directly use the reported results of other baselines. From Table 2, several observations can be made. First, compared with the compressed psycho-acoustically feature-based methods, SF-Net provides considerably better performance in all objective metrics. Second, the proposed system dramatically outperforms the previous one-stage full-band baselines, demonstrating the effectiveness of the proposed sub-band fusion strategy. Third, compared with our preliminary decoupling-style multi bands fusion model (*i.e.*, DMF-Net), SF-Net achieves further improvements with fewer trainable parameters, especially in the background noise suppression (CBAK) and overall quality (COVL). This indicates that the utilization of the proposed dual-stream network can realize better speech recovery and noise suppression.

We further conduct evaluations on the DNS-2022 blind test set, whose results are presented in Table 3. Due to the lack of clean speech as the reference, we utilize the non-intrusive subjective evaluation metrics to evaluate the subjective speech performance, namely DNSMOS P.808 [33] and P.835 [34], which are based on ITU-T P.808 [35] and P.835 [36], respectively. Compared with NSNet2, a standard baseline system for DNS-2022 [37], the proposed approach yields consistently better performance in speech distortion (SIG), background noise (BAK) and overall quality (OVRL). Besides, SF-Net outperforms DMF-Net in terms of speech distortion and overall quality, while a similar BAK DNSMOS score is observed. This further verifies the superiority of our approach in recovering the speech components in practical acoustic scenarios.

5. Conclusions

In this paper, we propose a collaborative sub-band fusion approach, dubbed SF-Net, for real-time speech enhancement running on 48 kHz-sampled speech signals. Motivated by the curriculum learning concept, we split the full-band target into three frequency sub-bands, *i.e.*, low-band (0-8 kHz), middle-band (8-16 kHz), and high-band (16-24 kHz), and three chain sub-networks are elaborately designed to recover the full-band clean spectrum by a stage-wise manner. Specifically, conducting on the STFT domain, a dual-stream decoupling-style network is employed to denoise the low-band complex spectrum, and two magnitude-masking based networks are employed to recover the middle- and high-band spectral magnitude. Experimental results demonstrate that the proposed method yields state-of-the-art performance over previous baselines by a large margin.

6. References

- [1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.
- [2] D. L. 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, 2018.
- [3] J. M. Valin, “A hybrid DSP/deep learning approach to real-time full-band speech enhancement,” in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE, 2018, pp. 1–5.
- [4] J. M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, “A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech,” *arXiv preprint arXiv:2008.04259*, 2020.
- [5] H. Schröter, T. Rosenkranz, A. Maier *et al.*, “DeepFilterNet: A Low Complexity Speech Enhancement Framework for Full-Band Audio based on Deep Filtering,” *arXiv preprint arXiv:2110.05588*, 2021.
- [6] J. Ge, X. Han, Y. Long, and H. Guan, “PercepNet+: A Phase and SNR Aware PercepNet for Real-Time Speech Enhancement,” *arXiv preprint arXiv:2203.02263*, 2022.
- [7] X. Zhang, L. Chen, X. Zheng, X. Ren, C. Zhang, L. Guo, and B. Yu, “A two-step backward compatible fullband speech enhancement system,” in *Proc. ICASSP*. IEEE, 2022.
- [8] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1829–1843, 2021.
- [9] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, “A Simultaneous Denoising and Dereverberation Framework with Target Decoupling,” in *Proc. Interspeech*, 2021.
- [10] A. Li, C. Zheng, L. Zhang, and X. Li, “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” *Applied Acoustics*, vol. 187, p. 108499, 2022.
- [11] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, “Dual-branch Attention-In-Attention Transformer for single-channel speech enhancement,” in *Proc. ICASSP*. IEEE, 2022.
- [12] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, “Dbt-net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement,” *arXiv preprint arXiv:2202.07931*, 2022.
- [13] Z.-Q. Wang, G. Wichern, and J. Le Roux, “On The Compensation Between Magnitude and Phase in Speech Separation,” *arXiv preprint arXiv:2108.05470*, 2021.
- [14] Y. Zhao and D. Wang, “Noisy-Reverberant Speech Enhancement Using DenseUNet with Time-Frequency Attention,” in *INTER-SPEECH*, 2020, pp. 3261–3265.
- [15] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [16] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *Proc. ICASSP*. IEEE, 2019, pp. 6875–6879.
- [17] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, “CycleGAN-based Non-parallel Speech Enhancement with an Adaptive Attention-in-attention Mechanism,” in *Proc. Asia-Pacific Signal Inf. Process. Assoc. (APSIPA)*, 2021, pp. 523–529.
- [18] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust text-to-speech,” in *Proc. SSW*, 2016, pp. 146–152.
- [19] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper *et al.*, “ICASSP 2022 DEEP NOISE SUPPRESSION CHALLENGE,” in *Proc. ICASSP*. IEEE, 2022.
- [20] C. Veaux, J. Yamagishi, and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *Proc. O-COCOSDA/CASLRE*. IEEE, 2013, pp. 1–4.
- [21] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings,” *JASA*, vol. 133, no. 5, pp. 3591–3591, 2013.
- [22] T. Ko, V. Peddinti, M. L. Povey, D. and Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *Proc. ICASSP*. IEEE, 2017, pp. 5220–5224.
- [23] Y. Zhao, D. Wang, B. Xu, and T. Zhang, “Monaural speech dereverberation using temporal convolutional networks with self attention,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1598–1607, 2020.
- [24] A. Li, C. Zheng, R. Peng, and X. Li, “On the importance of power compression and phase estimation in monaural speech dereverberation,” *JASA Express Letters*, vol. 1, no. 1, p. 014802, 2021.
- [25] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [26] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.
- [27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. ICASSP*. IEEE, 2010, pp. 4214–4217.
- [28] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2007.
- [29] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.
- [30] K. Tan and D. L. Wang, “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2019.
- [31] G. Yu, Y. Guan, W. Meng, C. Zheng, and H. Wang, “DMF-Net: A decoupling-style multi-band fusion model for real-time full-band speech enhancement,” *arXiv preprint arXiv:2203.00472*, 2022.
- [32] S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, “S-DCCRN: Super Wide Band DCCRN with learnable complex feature for speech enhancement,” *arXiv preprint arXiv:2111.08387*, 2021.
- [33] C. K. Reddy, V. Gopal, and R. Cutler, “DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proc. ICASSP*. IEEE, 2021, pp. 6493–6497.
- [34] —, “DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proc. ICASSP*. IEEE, 2022.
- [35] B. N. and R. C., “An Open Source Implementation of ITU-T Recommendation P.808 with Validation,” in *Proc. Interspeech 2020*, 2020, pp. 2862–2866.
- [36] —, “Subjective Evaluation of Noise Suppression Algorithms in Crowdsourcing,” in *Proc. Interspeech 2021*, 2021, pp. 2132–2136.
- [37] S. Braun and I. Tashev, “Data augmentation and loss normalization for deep noise suppression,” in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.