DMF-Net: A decoupling-style multi-band fusion model for real-time full-band speech enhancement

Guochen Yu*, Yuansheng Guan†*, Weixin Meng†, Chengshi Zheng†, Hui Wang*
*State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China
†Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

Abstract—Full-band speech enhancement based on deep neural networks is still challenging for the difficulty of modeling more frequency bands and real-time implementation. Previous studies usually adopt compressed full-band speech features in Bark and ERB scale with relatively low frequency resolution, leading to degraded performance, especially in the high-frequency region. In this paper, we propose a decoupling-style multi-band fusion model to perform full-band speech denoising and dereverberation. Instead of optimizing the full-band speech by a single network structure, we decompose the full-band target into multi sub-bands and then employ a multi-stage chain optimization strategy to estimate clean spectrum stage by stage. Specifically, the low- (0-8 kHz), middle- (8-16 kHz), and high-frequency (16-24 kHz) regions are mapped by three separate sub-networks. Comprehensive experiments on two public datasets demonstrate that the proposed method outperforms previous advanced systems and yields promising performance in terms of speech quality and intelligibility in real complex scenarios.

Index Terms—full-band speech enhancement, decoupling-style, multi-stage, multi-band fusion

I. INTRODUCTION

Speech enhancement (SE) aims to extract clean speech target from the noise-corrupted mixture, so as to improve speech quality [1] in real-time speech communication scenarios. Over the past few years, a multitude of SE algorithms based on deep neural networks (DNNs) have demonstrated their superior capability in dealing with non-stationary noise under low signal-to-noise ratio (SNR) conditions [2]. However, due to the high computational complexity caused by a large number of neurons and weights, most of these existing DNN-based SE approaches focus on narrow- or wide-band signals at the sampling rate of 8 or 16 kHz. For the difficulty of modeling the higher-dimensional features and conducting real-time implementation, it is still challenging to implement super-wideband (32 kHz) and full-band (48 kHz) SE.

In this regard, instead of using the Fourier spectrum directly, lower-dimensional compressed features have already been considered as inputs for deep learning-based full-band SE approaches [3–5]. In [3], Bark-scale spectrum with 22-dimensional Bark-frequency cepstral coefficients (BFCC) was adopted as input features and 22 ideal critical band gains were mapped, which can reduce the model size and computational complexity simultaneously. More recently, based on the human perception of speech signals, PercepNet developed a perceptual band representation with 32 triangular spectral bands [4], spaced according to the human hearing equivalent rectangular bandwidth (ERB). Obviously, the frequency resolution of the spectrum in Bark scale and that in ERB scale are much lower than the Fourier spectrum, leading to the leakage of information among frequency bands. In [6], DCCRN [7] was extended for super-wideband SE by simultaneous modeling of sub-band and full-band (SAF) features, in which the input features are the 257-dimensional Fourier spectrum, achieving superior performance when compared with other state-of-the-art (SOTA) full-band SE approaches [6].

Motivated by the curriculum learning, decoupling-style phase-aware methods have thrived in the SE area, where the original complex-spectrum estimation problem is decomposed into two sub-stages, i.e., magnitude and phase [8–11]. Specifically, only the magnitude estimation is involved in the first stage, followed by the complex spectrum refinement in the second stage. This paper proposes a decoupling-style multi-band fusion model, dubbed DMF-Net for full-band SE step by step. To be specific, we adopt a two-step strategy which consists of a pre-trained complex-domain-based SE network (LF-Net) for the low-frequency band (0-8 kHz) and two magnitude-based SE networks (MF-Net and HF-Net) for the middle- and high-frequency bands (8-16 and 16-24 kHz). In LF-Net, inspired by the preliminary study [9], we employ a decoupling-style strategy for wide-band speech enhancement by three sub-networks. In the first two sub-networks, the optimization w.r.t. denoising and dereverberation are serially implemented in the magnitude domain. i.e., we decouple the original complex multi-target optimization into magnitude and phase, and only focus on magnitude estimation. Afterward, the phase information can be effectively recovered based on previous results by introducing a global residual connection. Then, we couple the pretrained LF-Net with two higher-frequency refinement networks, namely MF-Net and HF-Net, to denoise the 8-16 kHz and 16-24 kHz regions. Note that the estimated low-frequency features are also fed into MF-Net and HF-Net to provide extra guidance. Finally, we fuse the low-, middle- and high-frequency regions to obtain the full-band speech, which is then fed into a low-complexity post-processing module to further suppress the residual noise [12].
Experimental results on two public datasets show that DMF-Net achieves remarkable performance and outperforms SOTA super-wideband and full-band baselines.

The remainder of the paper is organized as follows. In Section II, the proposed framework is described in detail. The experimental setup is presented in Section III, while Section IV gives the results. Some conclusions are drawn in Section V.

II. METHODOLOGY

A. Decoupling-style low-frequency band SE

Because human speech contains lots of harmonics and provides more information for the frequency ranging from 0 to 8 kHz, we first employ a low-frequency region estimation sub-network, dubbed LF-Net, to remove the noise and late reverberation in the complex domain. Similar to the previous work [9], LF-Net consists of three sub-networks, namely a denoising network (DN-Net), a dereverberation network (DR-Net), and a spectral refinement network (SR-Net), as illustrated in Figure 2(a). Inspired by the effectiveness of decoupling-style SE methods [8]–[10], we decouple the original spectrum estimation into spectral magnitude and phase, and only the magnitude is processed in DN-Net and DR-Net, without recovering the phase. After denoising and dereverberation, we couple the coarsely processed magnitude and the original noisy phase to obtain the coarse-estimated complex spectrum. For SR-Net, it receives both noisy and coarse-estimated complex spectra, and further refines the complex spectrum details as well as restores the clean phase. Note that different from explicitly estimating the whole complex spectrum, we introduce a global residual connection to only focus on estimating the residual missing details of complex spectra, which alleviates the overall burden of the network. The whole procedure of LF-Net can be given by:

\[
|\tilde{X}^{lf}_{dn}| = |X^{lf}_{dn}| \exp(i \theta_{X^{lf}}),
\]

\[
|\tilde{X}^{lf}_{dr}| = |X^{lf}_{dr}| \exp(i \theta_{X^{lf}}),
\]

\[
|\tilde{X}^{lf}_{sr}| = |X^{lf}_{sr}| \exp(i \theta_{X^{lf}}),
\]  

where \(\cdot\)\_\_ denote the outputs of DN-Net, DR-Net and SR-Net, respectively. \(|X^{lf}_{\cdot}|\) denotes the noisy low-frequency spectral magnitude. \(\tilde{G}_{dn}, \tilde{G}_{dr},\) and \(\tilde{G}_{sr}\) denote the mapping functions of the corresponding three modules with parameter set \(\Phi\_\_\). \(\theta_{X^{lf}}\) denotes the noisy phase of low-frequency bands.

For DN-Net and DR-Net, similar to [9], we employ the multi-frame (MF) filtering to capture the correlations between neighboring frames [13] in the magnitude domain, which can be given by:

\[
|\tilde{X}^{lf}_{i,\tau}| = \sum_{\tau=1}^{k} \tilde{M}^{lf}_{i,\tau} |X^{lf}_{i-\tau,\tau}|,
\]

\[
|\tilde{X}^{lf}_{i,\tau}| = \sum_{\tau=1}^{k} \tilde{M}^{lf}_{i,\tau} |X^{lf}_{i-\tau,\tau}|,
\]

where \(l\) and \(k\) denote the frame index and the filter length, respectively, and we set \(k\) to 5 in this paper.

B. Masking-based middle/high-frequency bands SE

After optimizing the low-frequency band, we employ two sub-networks, dubbed MF-Net and HF-Net, to refine the middle-frequency (8-16 kHz) and high-frequency (16-24 kHz) regions stage by stage, as illustrated in Figure 1. Based on the fact that speech in higher frequency bands tends to contain lower energies and fewer harmonics, we only estimate the spectral gain in the magnitude domain and retain the phase unchanged. This can reduce the complexity of the whole network and evade the implicit compensation effect between magnitude and phase estimation. To introduce explicit information interaction between low- and high-frequency components, the output of LF-Net is also fed into MF-Net and HF-Net. To be specific, we concatenate the estimated spectral magnitude of LF-Net with the original noisy middle-frequency spectral magnitude in the channel axis as the input of MF-Net, while the noisy high-frequency spectral magnitude is concatenated with the output spectral magnitude of LF-Net and MF-Net in the channel axis. In a nutshell, the whole procedure of the middle- and high-frequency bands modeling can be formulated as:

\[
|\tilde{X}^{mf}| = |X^{mf}| \otimes G_{mf} \left(|X^{mf}|; \Phi_{4}\right),
\]

\[
|\tilde{X}^{hf}| = |X^{hf}| \otimes G_{hf} \left(|X^{mf}|; \Phi_{5}\right),
\]

\[
|\tilde{X}^{mf}| = |X^{mf}| \exp(j \theta_{X^{mf}}), \tilde{X}^{hf} = |\tilde{X}^{hf}| \exp(j \theta_{X^{hf}}),
\]

where \(\tilde{X}^{mf}\) and \(\tilde{X}^{hf}\) denote the outputs of MF-Net and HF-Net, respectively. \(G_{mf}\) and \(G_{hf}\) denote the masking-based functions of MF-Net and HF-Net with parameter set \(\Phi_{4}\), \(\theta_{X^{mf}}\) and \(\theta_{X^{hf}}\) denote the noisy phase of middle- and high-frequency bands, respectively. After estimating three sub-band features, we employ a sub-band fusion operation to obtain the full-band spectrum by simply stacking the low-, middle- and high-frequency bands along the frequency axis. Note that we average the overlapped bands in the middle- and high-frequency bands.

C. Network architecture

Inspired by the promising performance of temporal convolutional modules (TCMs) for long-term sequence modeling in the SE area [14], [15], we insert cascaded TCMs in a typical convolutional encoder-decoder topology [16]. For LF-Net, the diagram of DN-Net and DR-Net is shown in Figure 2(b), in which SR-Net utilizes the same encoder as DN-Net and two decoders to recover both real and imaginary (RI) components. Taking DN-Net as an example, the encoder is comprised of five downsampling blocks, each of which is composed of a convolutional layer, InstanceNorm (IN), and PReLU, with kernel size being (2, 3) in the time and frequency axis except (2, 5) in the first block. The number of channels is set to 64, and the stride is set to (1, 2) to halve the frequency dimension. For the decoder, it employs a similar architecture to the encoder except for using the deconvolutional layers.
Specifically, DN-Net and DR-Net adopt MSE loss in the magnitude domain while SR-Net calculates loss functions on the RI components and the magnitude of the estimated spectrum, which can be expressed as:

\[
\mathcal{L}_{lf_{dn}} = \| \tilde{X}^{lf}_{dn} \mathcal{S} - S_{lf}^d \|_F^2, \\
\mathcal{L}_{lf_{dr}} = \| \tilde{X}^{lf}_{dr} \mathcal{S} - S_{lf}^d \|_F^2, \\
\mathcal{L}_{lf_{sr}} = \| \tilde{X}^{lf}_{sr} \mathcal{S} - S_{lf}^d \|_F^2, \\
\mathcal{L}_{lf_{sr}} = \| X^{lf}_{sr} - S_{lf}^f \|_F^2 + \| \tilde{X}^{lf}_{sr} - S_{lf}^f \|_F^2, \\
\mathcal{L}_{lf_{sr}} = \mu \mathcal{L}_{lf_{sr}} + (1 - \mu) \mathcal{L}_{lf_{sr}},
\]

where \(\mathcal{L}_{lf_{dn}}, \mathcal{L}_{lf_{dr}}\) and \(\mathcal{L}_{lf_{sr}}\) denote the loss functions for DN-Net, DR-Net and SR-Net, respectively. Note that \(S_{lf}^d\) denotes the target denoised spectral magnitude for DN-Net with reverberation unaltered and \(S_{lf}^f\) denotes the final target spectral magnitude. \(S_{lf}^f\) and \(S_{lf}^f\) represent the target RI components. With the internal trial, we empirically set \(\mu = 0.5\) in all the following experiments.

In the second stage, we freeze the pretrained LF-Net and optimize the spectral magnitude of the middle- and high-frequency bands ranging from 8 to 24 kHz by training MF-Net and HF-Net jointly, and the overall loss can be given by:

\[
\mathcal{L}_{mf} = \| X^{mf} - S^{mf} \|_F^2, \\
\mathcal{L}_{hf} = \| X^{hf} - S^{hf} \|_F^2, \\
\mathcal{L}_{full} = \alpha \mathcal{L}_{mf} + (1 - \alpha) \mathcal{L}_{hf},
\]

where \(\mathcal{L}_{mf}\) and \(\mathcal{L}_{hf}\) denote the loss function for MF-Net and HF-Net, while \(\mathcal{L}_{full}\) represent the full loss function of the second stage. \(X^{mf}\) and \(X^{hf}\) denote the outputs of MF-Net and HF-Net, respectively, while \(S^{mf}\) and \(S^{hf}\) denote the clean spectral magnitude of middle- and high-frequency bands. We empirically find that \(\alpha = 0.5\) suffices in our evaluation.

### III. Experiments

#### A. Datasets

To evaluate the performance of our framework, we conduct extensive experiments on two public datasets, namely VoiceBank + DEMAND dataset [17] and the 4th DNS-Challenge dataset [15], where all the utterances are sampled at 48 kHz.

**VoiceBank + DEMAND:** The dataset is a selection of the VoiceBank corpus [19] with 28 speakers for training and another 2 unseen speakers for testing. The training set includes 11,572 noisy-clean pairs, while the test set contains 824 pairs.
For the training set, the audio samples are mixed with one of the 10 noise types, including two artificial noise processes, i.e., babble and speech shaped noise, and eight real recording noise processes taken from the Demand database [20]. The test utterances are created with 5 unseen noises taken from the Demand database at SNRs of \{2.5\,dB, 7.5\,dB, 12.5\,dB, 17.5\,dB\}.  

**DNS-Challenge**: It consists of a wide range of clean speech, noise clips, and RIRs, which simulate complicated acoustic scenarios. Clean speech consists of six languages including English, French, German, Italian, Russian and Spanish. For this dataset, we totally generate around 600 hours of noisy-clean pairs. To be specific, we synthesize the noisy sets with around 350 hours for the English language. For the other languages, around 150 hours of training sets are generated (30h for each language). Besides, we generate the rest of 100 hours of training sets with VCTK corpus [19] and emotional utterances. To generate reverberant noisy training data, we use 248 real and 60,000 synthetic room impulse responses (RIR) from openSLR26 and openSLR28 datasets [21] to generate reverberant noisy training data. The direct and early reflected speech components are chosen as the training target when considering that only late reverberation degrades speech quality/intelligibility [22]. During each mixing process, the clean speech is convolved with a randomly selected RIR and is then mixed with the noise with the SNR range of \{-5\,dB, 15\,dB\}. For the testset, we utilize the 4th DNS-Challenge blind set[1] which consists of 859 real clips, each with a duration of 10s.

**B. 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 [23], we conduct the compression toward the spectral magnitude before sending into the network while remaining the phase unaltered, and the compression parameter is $\beta = 0.5$, which is optimal or near-optimal. All the models are optimized using Adam [24] with the learning rate of 5e-4. The batch size is set to 32 at the utterance level.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We use the perceptual evaluation of speech quality (PESQ) [25], short-time objective intelligibility (STOI) [26], CSIG, CBAK, and COVL [27] to evaluate speech enhancement performance. Higher values indicate better performance.

First, we conduct extensive experiments on VoiceBank + DEMAND to compare the proposed methods with several state-of-the-art (SOTA) full-band and super-wideband SE methods, including RNNoise [3], PercepNet [4], DCCRN [7] (super-wideband version), DeepFilterNet [5] and S-DCCRN [6]. We also conduct ablation study to show the importance of phase recovery for the low-frequency band (i.e., w/o SR-Net), which removes the complex spectral refinement in LF-Net. Due to the lack of the full-band PESQ evaluation tool, we resample the outputs of all methods to 16 kHz for a fair comparison when measuring the PESQ. From Table I we can have the following observations. Compared with other SOTA full-band methods, the proposed method achieves relatively better performance in terms of all metrics. On average, DMF-Net yields 0.24 PESQ improvement compared to Percepnet with less parameter burden. This demonstrates that the higher resolution spectral features can leverage the information of frequency bins and improve speech performance. Going from the S-DCCRN to DMF-Net, we also observe consistent improvements in all five objective metrics, e.g., average 0.13, 0.4%, 0.23, 0.55 and 0.19 increase in PESQ, STOI, CSIG, COVL.
CBAK and COVL scores are achieved. This indicates that the multi-stage strategy can obtain better performance over single-stage complex-domain networks.

We further evaluate our model with the 4th DNS-Challenge baseline, dubbed NSNet2 [28], on the DNS Challenge dataset. In Table [1] we present the results with ITU-T P.835 [29] subjective evaluation scores, including speech quality (SIG), background noise quality (SIG), and overall audio quality (OVRL). With less speech distortion, the proposed approach also yields much better performance in background noise and overall MOS than the baseline. It indicates that the proposed approach can effectively suppress the background noise and recover the speech components in real acoustic scenarios. Spectrograms of a noisy utterance and its corresponding enhanced utterances by NSNet2, LF-Net and DMF-Net are presented in Figure [3(a)-(d)]. Focusing on the low-frequency band ranging from 0 to 8 kHz in Figure [3(c)], one can see that LF-Net better suppresses background noise than NSNet2, which indicates the effectiveness of the proposed decoupling-style approach in handling the low-frequency band. As illustrated in Figure [3(d)], by fusing LF-Net, MF-Net and HF-Net, DMF-Net demonstrates the remarkable performance in reducing the background noise and preserving the speech components for the full-band speech.

V. CONCLUSION

In this paper, we propose a novel decoupling-style multi-band fusion speech enhancement model running on 48 kHz speech signals. Instead of optimizing the full-band speech by a single network structure, we decompose the full-band target into single sub-bands, i.e., low- (0-8 kHz), middle- (8-16 kHz) and high- (16-24 kHz) frequency bands, and employ a multi-stage chain optimization strategy to estimate the clean speech target. Specifically, conducting on the STFT domain, we design three sub-networks to respectively focus on each sub-band and adopt a fusion operation to obtain an estimate of the full-band clean spectrum. Experimental results demonstrate that the proposed method achieves state-of-the-art performance in terms of speech quality and intelligibility in both low- and high-frequency regions.

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] S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, “S-discrn: Super wide band discrn with learnable complex feature for speech enhancement,” arXiv preprint arXiv:2111.08387, 2021.