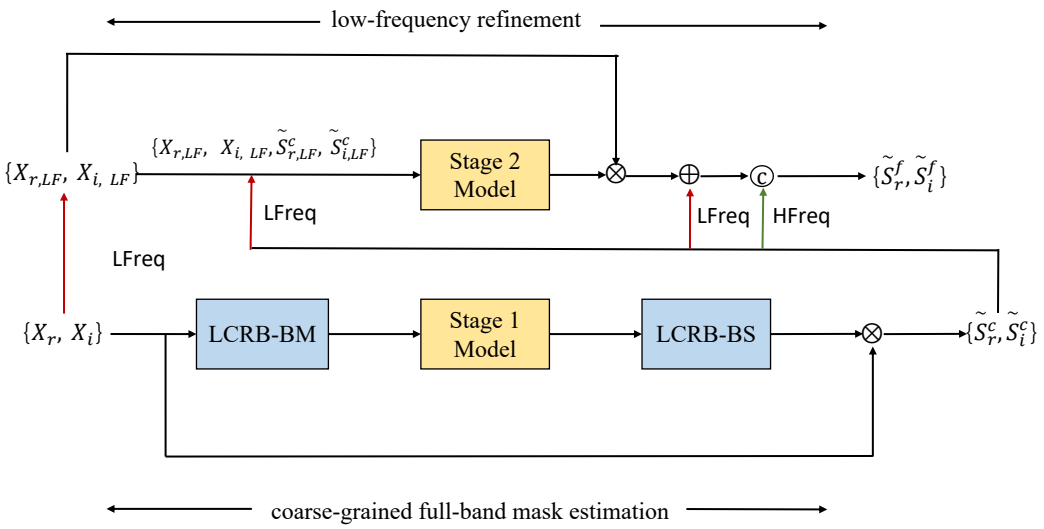

Fig. 1. Overview of proposed system. The red and green solid arrows indicate taking the low-frequency part of the spectrum and the part other than the low frequency, respectively.

1. APPENDIX

1.1. Detailed workflow of the proposed algorithm

The diagram of our proposed system is presented in Fig. 1. It consists of an LCRB filter bank and two sub-networks. The two sub-networks are the coarse-grained full-band mask estimation network (i.e., CoarseNet), and the low-frequency refinement network ($i.e.$, FineNet), where the two stages are jointly trained in an end-to-end fashion. We used a heterogeneous structure, i.e., a U-shaped subnetwork as the backbone of CoarseNet and a single-scale subnetwork as the backbone of FineNet. In SE, the mixed signal in the time domain can be formulated as $x(n) = s(n)+z(n)$, where $x(n), s(n), z(n)$ represent noisy speech, clean speech and noise, respectively. Given the STFT, the signal in the time domain can be converted to the T-F domain, given as:

$$X_{f,t} = S_{f,t} + Z_{f,t}, \tag{1}$$

where $X_{f,t}$, $S_{f,t}$, and $Z_{f,t}$ denote the complex spectrum representations with respect to noisy, clean and noise signals in the frequency index f and time frame index t. For facilitating notations, we omit the index (f, t) henceforth if no conflict arises. In the first stage, the CoarseNet takes LCRB-scaled compact features as input and predicts an LCRB-scaled complex mask M^{LCBR}. Then, the M^{LCBR} is passed through the band splitting module of the LCBR filter bank to obtain the same size mask as the original spectral features. In the second stage, we further refine the low-frequency part of the spectrum using the FineNet, which is fed with the low-frequency part of the complex spectrum of the original speech and the low-frequency part of the enhanced complex spectrum from the first stage. The output of the FineNet, M^f, is a compensation complex mask for the low-frequency part. The final estimate S^f is obtained by summing the predicted compensation with the CoarseNet output S^c. In a nutshell, the whole forward calculation process is formulated as:

$$\begin{aligned}
\left\{X_r^{LCBR}, X_i^{LCBR}\right\} &= LCRB_{BM}\left(X_r, X_i\right), \\
\left\{\widetilde{M}_r^{LCBR}, \widetilde{M}_i^{LCBR}\right\} &= \mathcal{G}_1\left(X_r^{LCBR}, X_i^{LCBR}; \phi_1\right), \\
\left\{\widetilde{M}_r^{c}, \widetilde{M}_i^{c}\right\} &= LCRB_{BS}\left(\widetilde{M}_r^{LCBR}, \widetilde{M}_i^{LCBR}\right), \\
\widetilde{S}_r^c &= \widetilde{M}_r^c \odot X_r - \widetilde{M}_i^c \odot X_i, \\
\widetilde{S}_i^c &= \widetilde{M}_i^c \odot X_r + \widetilde{M}_r^c \odot X_i, \\
\left\{\widetilde{M}_{r,LF}^f, \widetilde{M}_{i,LF}^f\right\} &= \mathcal{G}_2\left(X_{r,LF}, X_{i,LF}, \widetilde{S}_{r,LF}^c, \widetilde{S}_{i,LF}^c; \phi_2\right), \\
\widetilde{S}_{r,LF}^f &= \widetilde{M}_{r,LF}^f \odot X_{r,LF} + \widetilde{S}_{r,LF}^c, \\
\widetilde{S}_{i,LF}^f &= \widetilde{M}_{i,LF}^f \odot X_{i,LF} + \widetilde{S}_{i,LF}^c, \\
\widetilde{S}_{HF}^f &= \widetilde{S}_{HF}^c, \\
\widetilde{S}^f &= Concat\left(\widetilde{S}_{LF}^f, \widetilde{S}_{HF}^f\right),
\end{aligned} \tag{2}$$

where $\mathcal{G}_1\left(\bullet; \phi_1\right)$ and $\mathcal{G}_2\left(\bullet; \phi_2\right)$ denote the mapping functions for the first and the second stages, respectively. $LCRB_{BM}$ and $LCRB_{BS}$ denote the band merging (BM) module and the band splitting (BS) module of the LCRB filter, respectively. The symbol $\odot$ denotes element-wise multiplication. Subscript $(\bullet)_r$ and $(\bullet)_i$ denote the real and imaginary parts of a complex variable, respectively. Subscript $(\bullet)_{LF}$ and $(\bullet)_{HF}$ denote the low-frequency part of the spectrum, and the other parts of the spectrum except for the low frequencies, respec-

tively. X_{LF} takes the first Q frequency bins of X, and X_{HF} consists of the other bins in X except for X_{LF}.

1.2. Motivation of this paper

The aim of this paper is to exploit a novel lightweight SE framework that minimizes the model overhead while maintaining competitive performances. The rationals behind our designation are as follows. Firstly, the vast majority of noise signals have a wide bandwidth with a smooth spectrum. Similarly, both the periodic and the stochastic components of speech have a smooth spectral envelope [1]. That is, the statistical properties of adjacent frequency bins in the spectrum are close. Therefore, we can get the down-sampled spectral feature by a band merging operation such as in the equivalent rectangular bandwidth (ERB) filter bank [2]. Such compact features can retain as much spectral information as possible while effectively reducing the overhead of the model. Secondly, previous studies have shown that harmonics may attend in low-frequency bands; however, rarely in high-frequency bands [3]. Therefore, we only need to take an additional second stage network to enhance the low-frequency part to remove residual noise and recover the harmonic structure. In the two-stage network, the output of the first-stage model can be used as a priori information for the second-stage, which facilitates the second model to better model the low-frequency part of the spectrum. Thirdly, heterogeneous models can play a complementary role, and the use of a specific structure considering the characteristics of each stage helps to improve the performance of the proposed two-stage framework. We adopt the idea of complementary feature processing and consider the respective characteristics of the proposed two-stage task with a U-shaped subnetwork for the CoarseNet and a single-scale subnetwork for the FineNet.

With above discussions, we present a novel SE framework with a Learnable Complex-valued Rectangular Bandwidth (LCRB) filter bank and two stacked heterogeneous subnetworks, one is for coarse-grained full-band mask estimation and the other for low-frequency refinement. Our contributions are detailed below:

- We propose a framework that combines a two-stage task and a lightweight approach, capable of achieving comparable performance to SOTA methods with a low model overhead. Specifically, we design a novel two-stage model that includes a coarse-grained full-band mask estimation stage and a fine-grained low-frequency refinement stage. Instead of using a hand-designed real-valued filter, we use a novel learnable complex-valued rectangular bandwidth (LCRB) filter bank as an extractor of compact features.

- We adopt the idea of complementary feature processing and consider the respective characteristics of the proposed two-stage task, using a U-shaped subnetwork as the backbone of CoarseNet and a single-scale subnetwork as the backbone of FineNet.

- To verify the superiority of the proposed approach, we compare our model with single-stage backbone models and other SOTA systems on two public test sets. The experimental results show that our model achieves comparable results to the single-stage backbone models with greatly reduced parameters and computational effort, and compares favorably with the SOTA models.

2. REFERENCES

[1] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in *Proc. Interspeech 2020*, 2020, pp. 2482–2486.

[2] Brian CJ Moore, *An introduction to the psychology of hearing*, Brill, 2012.

[3] Longbiao Cheng, Junfeng Li, and Yonghong Yan, "Fscnet: Feature-specific convolution neural network for real-time speech enhancement," *IEEE Signal Processing Letters*, vol. 28, pp. 1958–1962, 2021.