ADAPTIVE-FSN: INTEGRATING FULL-BAND EXTRACTION AND ADAPTIVE SUB-BAND ENCODING FOR MONAURAL SPEECH ENHANCEMENT

Yu-sheng Tsao[1], *Kuan-Hsun Ho*[1], *Jeih-weih Hung*[2], *Berlin Chen*[1]

[1]Department of Computer Science and Information Engineering, National Taiwan Normal University
[2]Department of Electrical Engineering, National Chi Nan University
{samtsao, 61047017s, berlin}@ntnu.edu.tw, jwhung@ncnu.edu.tw

2022 IEEE Spoken Language Technology Workshop (SLT) | 979-8-3503-9690-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/SLT54892.2023.10023439

ABSTRACT

An important more recent thread of speech enhancement work is to utilize fine-grinded local spectral patterns with sub-band processing that complement full-band features nicely. To extend the efficacy of sub-band spectral information, we propose Adaptive-FSN, a fully convolutional real-time speech enhancement framework, to dynamically acquire a sub-band embedding within a wide range of sub-band frequencies. We exploit an adaptive subband encoder to portray sub-band processing that encapsulates a wide range of sub-band units. Then we build this effective sub-band embedding with a Conformer-based structure and multi-view attention. As for the full-band features, we make use of the FullSubNet+ architecture with its full-band extractor to get global spectral information. Finally, a Conformer-based fusion model combines the above information sources to predict the complex ideal ratio mask (cIRM). Experimental results on the VoiceBank-DEMAND benchmark task reveal that this novel framework outperforms FullSubNet+ by promoting the quality of processed utterances and reducing the implementation complexity for faster real-time computation.

Index Terms— Speech enhancement, sub-band processing, FullSubNet, complex spectrum, real-time computation

1. INTRODUCTION

Speech enhancement (SE) focuses on removing noise interference from noisy speech signals and improving their quality and intelligibility. SE techniques have been employed in many use cases, such as online meetings, hearing aids, and speech recognition, to name a few. The recent approach to SE research is to employ a deep neural network (DNN) structure, as the DNN-enabled SE methods exhibit superior performance on noise reduction than conventional statistics-based methods, especially for non-stationary noise scenarios. The DNN-fueled SE methods can be broadly categorized as mapping-based [1] and masking-based [2]. The masking-based methods have gradually become the mainstream as they are easier to constrain dynamic range and converge faster [3]. Multiple time-frequency (T-F) masking-based methods such as ideal binary mask [4], ideal ratio mask [5], and complex ideal ratio mask (cIRM) [6] have been proposed. Specifically, cIRM can implicitly deal with the phase information without modeling its unclear structure. On the other hand, the time-domain mask approaches are also emerging to obtain flexible speech representations through learning a filterbank-like structure to process magnitude and phase information simultaneously [7, 8]. Nevertheless, T-F domain methods are usually more favorable than time-domain methods since the speech and noise components tend to be more separately distributed and distinguishable in the T-F domain [9].

A celebrated SE method, FullSubNet [10], adopts a sub-band modeling technique to process the neighborhood frequency components in time-frequency features and has exhibited impressive results. Laying more emphasis on the sub-band features, FullSubNet can deal with local spectral patterns with a subtle view to distinguish clean speech from noise. DPT-FSNet [11] embraces a similar idea to generate sub-band features by stacking dense blocks [12] composed of dilated CNNs, and then feeds these features into a subsequent dual-path network for full-band and sub-band modeling. However, the underlying dense blocks involve complex computation, which is resource-intensive and thus unconducive to real-time SE implementation.

Although FullSubNet is a highly effective real-time SE method, there is still room for improvement. First, FulSubNet limits the range of adjacent sub-bands by predefined hyperparameters. In [13], increasing the number of adjacent sub-bands benefits the SE performance. However, as mentioned in [14], local patterns in the spectrogram often vary in different frequency bands. As such, it might lead to sub-optimal performance by processing a fixed number of sub-band units. In addition, the size of sub-band units determines the input size of the used sub-band model (we refer to it as the fusion model), which possibly causes redundant calculation. Meanwhile, using the LSTM network to combine full-band and sub-band features in FullSubNet significantly increases the computation complexity. Without the parallel computing ability, the LSTM can be a burden for FullSubNet to operate in real-time on CPU.

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

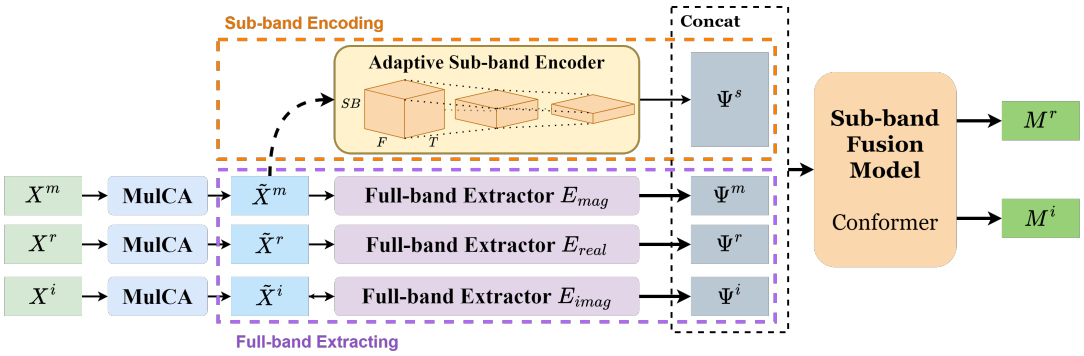


Fig. 1: The overall architecture of Adaptive-FSN.

Partly illuminated by the above studies, this paper presents a fully convolutional sub-band fusion SE framework called Adaptive-FSN, which addresses the potential downsides of FullSubNet. Adaptive-FSN processes full-band and sub-band streams separately and finally integrates the features of the two streams by a fusion model to produce a cIRM. The process pertaining to the full-band stream is similar to the conventional SE model, where a global spectral embedding is generated by modeling all the frequency components. As for the subband stream, we propose an adaptive-subband encoder module, which follows the down-sampling mechanism often used in encoder-decoder networks [8]. The encoder enables the model to learn to adapt and compress effective local patterns in the sub-band frequency range while obtaining more informative and appropriate sub-band features. To fuse the two streams, we use Conformer [15] to reduce the computation time of the model. A set of evaluation experiments conducted on the VoiceBank-DEMAND task show that the proposed Adaptive-FSN outperforms the baseline and FullSubNet in speech quality for the processed utterances, while speeding up the inference time. The model and recipes for use in benchmarking and replication experiments are shared public at https://github.com/samx81/Adaptive-FSN.

2. METHODOLOGY

This study focuses on developing a real-time noise reduction network, Adaptive-FSN, with the short-time Fourier transform (STFT) spectrogram as inputs, which generates a mask to suppress the noise component in the noisy spectrogram and restore the speech signal. Adaptive-FSN employs FullSubNet+ [16] as an archetype, which improves FullSubNet by introducing phase information and the feature weighting mechanism. We based on this to extend the sub-band processing in depth. A schematic diagram of Adaptive-FSN is depicted in Fig. 1, which contains four main components: three parallel multi-scale time-sensitive channel attention (MulCA) modules, three parallel full-band extractors, an adaptive sub-band encoder, and a fusion model.

Adaptive-FSN takes the magnitude, real, and imaginary

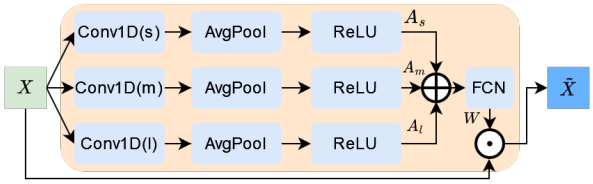


Fig. 2: The diagram of MulCA module.

STFT spectrogram representations $\mathbf{X}^m, \mathbf{X}^r, \mathbf{X}^i \in \mathbb{R}^{F\times T}$ as inputs, where F and T denote the total number of frequency bands and frames, respectively. The three spectrogram representations allow the model to process the magnitude and the ambiguous phase information simultaneously. The MulCAs provide weighting information of frequency bins in different scales to reveal their unequal importance.

Next, the three MulCA-weighted spectrograms $\tilde{\mathbf{X}}^m$, $\tilde{\mathbf{X}}^r$, $\tilde{\mathbf{X}}^i$ are individually passed through a full-band extractor to produce $\tilde{\mathbf{\Psi}}^m, \tilde{\mathbf{\Psi}}^r, \tilde{\mathbf{\Psi}}^i \in \mathbb{R}^{F\times T}$, in which the global context information is highlighted. Simultaneously, the weighted magnitude spectrogram $\tilde{\mathbf{X}}^m$ is first unfolded to be a tensor $\mathbf{S} \in \mathbb{R}^{F\times C_0\times T}$, where $C_0 = 2N+1$ denotes the total number of two-sided adjacent frequency bins around any specific frequency of $\tilde{\mathbf{X}}^m$ with N being the number of each side (circular Fourier frequencies are used for boundary frequencies). Then the tensor $\mathbf{S}$ additionally undergoes an adaptive sub-band encoding to obtain the sub-band units $\tilde{\mathbf{\Psi}}^s \in \mathbb{R}^{F\times C_E\times T}$, where E is the total number of encoder layers. Briefly speaking, the presented adaptive sub-band encoder is learned to extract and compress the neighborhood (sub-band) information around any specific frequency bin of $\tilde{\mathbf{X}}^m$ rather than simply unfolding $\tilde{\mathbf{X}}^m$ as done in FullSubNet+. The details of the sub-band encoding process will be fleshed out in subsection 2.3.

Finally, the three full-band extractor outputs $\tilde{\mathbf{\Psi}}^m, \tilde{\mathbf{\Psi}}^r, \tilde{\mathbf{\Psi}}^i$ and the adapative sub-band encoder output $\tilde{\mathbf{\Psi}}^s$ are concatenated and fed to a sub-band fusion model G_{sub} to predict the target complex ideal ratio mask (cIRM).

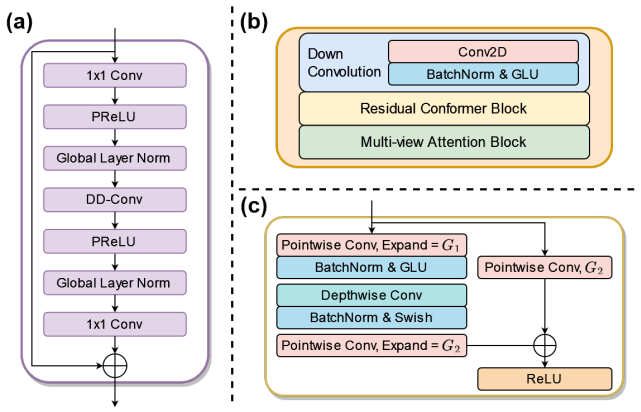


Fig. 3: (a) The diagram of TCN blocks. (b) Adaptive Sub-band Encoder layer. (c) Residual Conformer blocks.

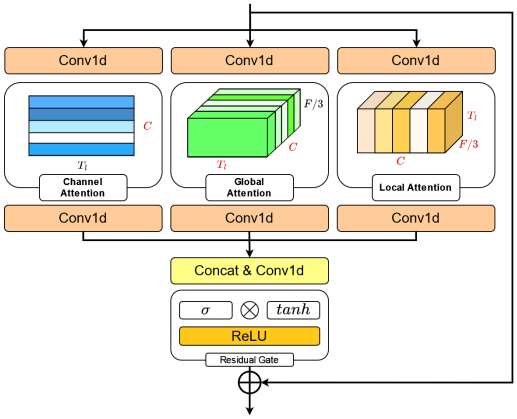


Fig. 4: The diagram of Multi-view Attention blocks.

2.1. MulCA

Previous studies [14, 17] have pointed out that the energy of a speech utterance distributes non-uniformly in frequencies, and different frequency components are unequally crucial to human perception. For example, humans are more sensitive to high frequencies than low frequencies. With this clue, FullSubNet+ exploits MulCA to assign weights to the frequency bins in spectrogram to facilitate the subsequent model to reduce noise at target essential frequency bands.

The operations of an MulCA can be expressed by the following equations:

$$\begin{aligned} \mathbf{w}_i &= \mathrm{ReLU}(\mathrm{AvgPool}(\mathrm{Conv1D}_i(\mathbf{X}))), i = 1, 2, 3, \\ \mathbf{w} &= \mathrm{FCN}([\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3]), \mathbf{w} \in \mathbb{R}^F \\ \mathbf{W} &= \mathrm{Broadcast}(\mathbf{w}), \mathbf{W} \in \mathbb{R}^{F \times T}, \\ \tilde{\mathbf{X}} &= \mathbf{X} \odot \mathbf{W}. \end{aligned} \tag{1}$$

That is, the frame-wise spectra in the spectrogram $\mathbf{X}$ is passed through three convolution layers with different kernel sizes, in parallel, each followed by the average pooling and a ReLU activation to deliver a weight vector $\mathbf{w}_i$, $i = 1, 2, 3$. Then a fully connected (FCN) layer merges three weight vectors to create a frequency-wise weight vector $\mathbf{w} \in \mathbb{R}^F$. Finally, the MulCA output $\tilde{\mathbf{X}}$ is the dot-product of the spectrogram $\mathbf{X}$ and $\mathbf{W}$, the time-broadcasted version of $\mathbf{w}$, where $\odot$ denotes the dot product. As such, all the frames in the spectrogram share an identical frequency-dependent weight vector.

2.2. Full-Band Extractor

To extract the full-band information, we use the stacked temporal convolutional network (TCN) [7] blocks to serve as the full-band extractor. As shown in Fig. 3a, each block contains three main components: an 1×1 convolution layer 1x1Conv, a depth-wise dilated convolution layer (DD-Conv), and another 1x1Conv. A parametric ReLU and a normalization layer are inserted between these convolution layers, and the residual connection is applied to stabilize training. In particular, several TCN blocks that possess different dilation factors are stacked. Here, the used full-band extractor consists of two groups of TCN blocks connected in series to process the input $\tilde{\mathbf{X}}$. Each group contains four TCN blocks with a kernel size of 3 and dilation factors of {1,2,5,9} to capture long-distance speech signal features. Finally, these TCN blocks are followed by a FCN to produce the full-band embedding.

These full-band streams $\mathbf{\Psi}^m$, $\mathbf{\Psi}^r$, $\mathbf{\Psi}^i$ can provide complementary information to the sub-band feature stream, and they are jointly fed into the subsequent fusion model.

2.3. Adaptive Sub-band Encoder

As mentioned earlier, the characteristics in different frequency regions are often distinct, and thus processing a fixed sub-band range around different frequencies might not be the optimal solution. Inspired by the idea of down-sampling in the time-domain SE model that adjusts the time segment length for processing, we design an adaptive sub-band encoder module to process the unfolded magnitude spectrogram. A schematic diagram of the encoder layer is shown in Fig. 3b, and it refers to the recently released MANNER [18] encoder, which has yielded excellent results with a well-designed structure consisting of a down convolution layer, Residual Conformer (ResCon) block, and Multi-view Attention (MA) block. We adjust the encoder to make it work along the sub-band axis without breaking causality. With several encoder layers connected in series to form the sub-band encoder module, fewer but more effective sub-band features can be captured through down-sampling and feature encoding.

Referring to Fig. 3b as the e^{th} encoder layer of the presented sub-band encoder module ($e = 1, 2, 3, ..., E$), the down convolution layer consists of a 2-D convolution (Conv2D) with kernel size $(k_e, 1)$ and stride $(s_e, 1)$ to down-sample the input tensor $\mathbf{S}_{e-1} \in \mathbb{R}^{F \times C_{e-1} \times T}$ ($\mathbf{S}_0 = \mathbf{S}$) along the sub-band axis for each frame, followed by batch nor-

malization and ReLU activation to produce down-sampled features, where C_e is calculated as follows:

$$C_e = \lfloor \frac{C_{e-1} - k_e}{s_e} \rfloor + 1 \tag{2}$$

The ResCon Block is shown in Fig. 3c, which is essentially Conformer in [15] except that all the linear layers are replaced with convolution layers to better process the dynamic length of the input sequence, and the attention layer is moved out as a standalone block. We then change the 1-D convolution (Conv1D) in the MANNER's encoder into a Conv2D to perform encoding.

For the MA block, the modified process is shown in Fig. 4. After adjusting the frequency size from F to $F/3$ by Conv1D, the compressed input $x' \in \mathbb{R}^{F/3 \times C_e \times T}$ is then transposed according to each attention mechanism so that the sub-band dimension can interact with other feature dimensions.

Global Attention. Based on the self-attention [19], we calculate the transposed features x' as $x_G \in \mathbb{R}^{F/3 \times T \times C_e}$ to emphasize the global sequence information:

$$\begin{gathered} Q = x_G W_q, \quad K = x_G W_k, \quad V = x_G W_v, \\ x'_G = W_o(softmax(\frac{QK^\top}{\sqrt{d}})V). \end{gathered} \tag{3}$$

Local Attention. To learn the effective sub-band information within the local sequence, depthwise convolution layers are first employed to represent the input features $x_L \in \mathbb{R}^{T \times F/3 \times C_e}$ focusing on the sub-bands. We then predict the local attention weights $\alpha_L \in \mathbb{R}^{T \times 1 \times C_e}$ by compressing the channel-wise average-pooling and max-pooling results with a convolution layer. The whole process is expressed by the following equations:

$$\begin{gathered} x'_L = \text{DepthwiseConv}(x_L), \\ x_L^{avg} = \text{AvgPool}(x'_L), \quad x_L^{max} = \text{MaxPool}(x'_L), \\ \alpha_L = \sigma(\text{Conv}([x_L^{avg}; x_L^{max}])) \\ x''_L = x_L \times \alpha_L. \end{gathered} \tag{4}$$

Channel Attention. As for the channel attention, we treat different frequency bins as channels. With the input features $x_C \in \mathbb{R}^{F/3 \times C_e \times T}$, we use adaptive average-pooling and max-pooling followed by FCN along the time axis to obtain the sub-band attention features of the overall frames.

$$\begin{gathered} x_C^{avg} = \text{FCN}(\text{AdaptiveAvgPool}(x_C)), \\ x_C^{max} = \text{FCN}(\text{AdaptiveMaxPool}(x_C)), \\ \alpha_C = \sigma(x_C^{avg} + x_C^{max}) \\ x'_C = x_C \times \alpha_C, \end{gathered} \tag{5}$$

where $\alpha_C \in \mathbb{R}^{F/3 \times C_e \times 1}$ is the channel attention weight.

After finishing the above three attention stages, the corresponding outputs are concatenated and passed through a convolution layer. Then the mask gate is applied to the output as a residual gate process, followed by a residual connection, producing $\mathbf{S}_e \in \mathbb{R}^{F \times C_e \times T}$ as the the e^{th} encoder layer.

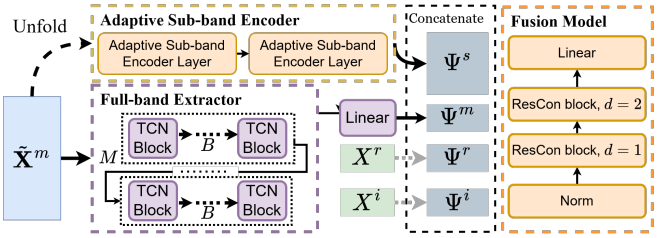


Fig. 5: The flowchart of magnitude spectrogram processing.

2.4. Sub-band Fusion Model

The sub-band fusion model G_{sub} integrates the output of the full-band extractors (Ψ^m, Ψ^r, and Ψ^i) and the adaptive sub-band encoder ($\Psi^s = \mathbf{S}_E$) to predict the cIRM. As shown in the rightmost part of Fig. 5, we construct the sub-band fusion model G_{sub} mainly by stacking the ResCon block like a TCN block with increasing dilation parameters to capture long-term and short-term speech characteristics. It is noteworthy that FullSubNet+ exploits the LSTMs to fulfill the sub-band fusion model, and we replace its LSTMs with ResCon blocks in Adaptive-FSN to expedite the implementation.

3. EXPERIMENTS

3.1. Dataset

We have conducted the evaluation experiments for the Adaptive-FSN with the Voicebank-DEMAND task [20]. The training set containing 11,572 utterances from 28 speakers is mixed with ten types of noise at four different SNRs: 0, 5, 10, and 15 dB. The test set contains 872 utterances from 2 speakers, which are contaminated by five unseen noises at four SNRs of 2.5, 7.5, 12.5, and 17.5 dB. The utterances are at a 16-kHz sampling rate and have no reverberation.

3.2. Training setup

We follow the experimental settings of FullSubNet and FullSubNet+ for a fair comparison. A window length of 512 points with a Hanning window and a hop length of 256 points is used to convert the signal to the STFT spectrogram. The length of input sequence is set to 192 frames. The number of look-ahead frames for enhancing the current frame is set to 2

As for the model training, we use the Adam optimizer with a learning rate of $1e-3$. Meanwhile, for the sub-band encoder, we set $N = 64$, making the number of adjacent frequencies bins $C_0 = 64 \times 2 + 1 = 129$ for creating the initial sub-band features. In addition, the encoder module contains $L = 2$ layers with the kernel size and the stride of down convolutions being $(16, 8)$ and $(4, 2)$, respectively, resulting $C_e = \{129, 16, 7\}$ for $e = 0, 1, 2$. The fusion model is set to have expansion factors $G_1 = 4$ and $G_2 = 1$ for its Residual Conformer Blocks.

Table 1: PESQ, STOI, CSIG and COVL results of various SE methods on the VoiceBank-DEMAND task.

Model	Domain	PESQ	STOI (%)	CSIG	COVL
Noisy	-	1.97	92.1	3.35	2.63
Wiener	Time	2.22	92.0	3.23	2.67
Conv-TasNet [7]	Time	2.53	-	3.95	3.23
DCCRN [21]	Frequency	2.68	93.7	3.88	3.27
DEMUCS [8]	Time	**2.93**	**95.0**	**4.22**	**3.52**
FullSubNet [10]	Frequency	2.77	94.0	3.94	**3.35**
FullSubNet+ [16]	Frequency	2.74	93.9	**3.95**	3.33
Adaptive-FSN	Frequency	**2.81**	**94.1**	3.92	**3.35**

Table 2: Ablation study on the VoiceBank-DEMAND. CF = Conformer Fusion, AS = Adaptive Sub-band

Model	CF	AS	PESQ	STOI	RTF (CPU)	MACs (G)	# of param. (M)
FullSubNet	-	-	2.77	94.0	0.246	66.91	5.64
FullSubNet+	×	×	2.74	93.9	0.260	67.42	8.63
/w Conformer	✓	×	2.72	93.7	**0.064**	**2.51**	**6.86**
/w Ada.Subband	×	✓	**2.81**	93.9	0.341	56.06	11.67
Adaptive-FSN	✓	✓	**2.81**	**94.1**	0.105	4.95	9.88

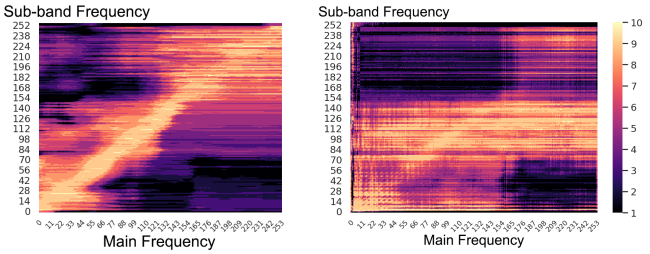
3.3. Results and Discussions

The various quantitative evaluation results of the proposed Adaptive-FSN and several state-of-the-art (SOTA) causal methods are shown in Table 1. From this table, we have several observations as follows:

1. FullSubNet and FullSubNet+ behave quite close to each other. Therefore, the superiority of FullSubNet+ over FullSubNet in Deep Noise Suppression (DNS) Challenge dataset is not apparent for the VoiceBank-DEMAND task.
2. Compared with its archetype FullSubNet+, Adaptive-FSN behaves significantly better in almost all SE metrics (except for the CSIG index). These results confirm that the adaptive sub-band encoder can highlight the clean-speech information in the unfolded magnitude spectrogram.
3. Among the SOTA causal SE models, which usually directly employ time-domain features to preserve phase information, DEMUCS behaves the best and outperforms the presented Adaptive-FSN. Nevertheless, one of the particularities of Adaptive-FSN is that it exploits the magnitude spectrogram on top of the complex-valued spectrogram, as FullSubNet+ does.

3.4. Ablation Study

To explore the influence of every model component in Adaptive-FSN, we conduct an ablation study by altering the components of FullSubNet+, whose results are presented in Table 2. First, replacing the LSTM with the Residual Conformer as the backbone of the sub-band fusion model significantly reduces the CPU processing time in terms of the real-time factor (RTF), whereas it slightly deteriorates both PESQ and STOI. Second, equipping FullSubNet+ with the adaptive sub-band encoder improves PESQ significantly at the cost of increasing RTF. Finally, the presented Adaptive-FSN, which uses the Residual Conformer instead of the LSTM as well as the adaptive sub-band encoder, benefits FullSubNet+ in promoting PESQ and STOI and reducing RTF.



(a) FullSubNet+ (b) Adaptive-FSN

Fig. 6: Normalized correlation between Ψ^m and Ψ^s for (a) FullSubNet+, (b) Adaptive-FSN, which is depicted in deciles for better visualization.

3.5. Visualization of Sub-band dependency

To examine how the sub-band encoding features from Adaptive-FSN highlight the spectral information, we evaluate the normalized correlation between the full-band magnitude spectrogram Ψ^m at frequency bin i and the sub-band spectrogram Ψ^s centered at frequency bin j, which is calculated as follows:

$$R(i,j) = \mathrm{Norm}\left(\sum_k \left|\mathrm{Corr}(\Psi^m[i], \Psi^s[j][k])\right|\right), \quad i,j = 0,1,2,...,256, \tag{6}$$

where k is the sub-band frequency index, Norm and Corr indicate the unity-based normalization and Pearson correlation coefficient calculation, respectively.

The obtained correlation values for FullSubNet+ and Adaptive-FSN for the test set are depicted in Fig. 6. From Fig. 6(a), we see that for FullSubNet+, any arbitrary frequency bin in Ψ^m correlates more with its nearby frequency bins in Ψ^s, which somehow agrees with our intuition that neighboring frequencies are relevant. However, Adaptive-FSN is shown to create a wider range of frequency interdependency for low and middle frequencies in Fig. 6(b), which probably benefits noise reduction. Moreover, Fig. 6(b) reveals that Adaptive-FSN highlights the mid-frequency components in sub-band features by promoting their correlation with the entire frequency band, whereas the high-frequency sub-bands become less relevant to their center frequencies.

4. CONCLUSION

In this study, we have proposed a novel speech enhancement framework, Adaptive-FSN, that captures local spectral patterns dynamically and specifically to different frequencies. In addition, we leverage Conformer layers as an alternative to the LSTM blocks used in FullSubNet+ to enable parallel computation and expedite implementation, making Adaptive-FSN more suitable for CPU implementation. Preliminary experimental results show that Adaptive-FSN outperforms state-of-the-art frequency-domain SE methods on the VoiceBank-DEMAND benchmark task. The ablation study results also indicate that the updates to FullSubNet+ from the presented Adaptive-FSN achieve higher speech quality and lower computation complexity. In the future, we plan to perform adaptive sub-band encoding on complex spectrograms and evaluate Adaptive-FSN on a larger-scale dataset task such as the DNS-Challenge.

5. REFERENCES

[1] Ke Tan and DeLiang 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, 2019.

[2] DeLiang Wang and Guy J Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[3] Shubo Lv, Yanxin Hu, Shimin Zhang, and Lei Xie, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," *arXiv preprint arXiv:2106.08672*, 2021.

[4] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[5] Arun Narayanan and DeLiang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[6] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[7] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[8] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[9] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9458–9465.

[10] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.

[11] Feng Dang, Hangting Chen, and Pengyuan Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6857–6861.

[12] Ashutosh Pandey and DeLiang Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6629–6633.

[13] Xiaofei Li and Radu Horaud, "Narrow-band deep filtering for multichannel speech enhancement," *arXiv preprint arXiv:1911.10791*, 2019.

[14] Naoya Takahashi and Yuki Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 21–25.

[15] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[16] Jun Chen, Zilin Wang, Deyi Tuo, Zhiyong Wu, Shiyin Kang, and Helen Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[17] Rong Chao, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao, "Perceptual contrast stretching on target feature for speech enhancement," *arXiv preprint arXiv:2203.17152*, 2022.

[18] Hyun Joon Park, Byung Ha Kang, Wooseok Shin, Jin Sob Kim, and Sung Won Han, "Manner: Multi-view attention network for noise erasure," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7842–7846.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *SSW*, 2016, pp. 146–152.

[21] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.