Subband-based Spectrogram Fusion for Speech Enhancement by Combining Mapping and Masking Approaches

Hao Shi*, Longbiao Wang†, Sheng Li‡, Jianwu Dang†, Tatsuya Kawahara*
* Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
† Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
‡ National Institute of Information and Communications Technology (NICT), Kyoto, Japan
§ Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: shi@sap.ist.i.kyoto-u.ac.jp

Abstract—**Deep learning brings effective optimization and significant improvements to speech enhancement (SE). Mapping and masking are currently major approaches in single-channel frequency-domain SE with supervised learning. In this work, we first show that these two approaches are complementary in that mapping is more effective in low-frequency bands, while masking is more suitable in high-frequency bands. This is because the high-frequency bands typically have low energy, so estimating the enhanced spectrogram directly does not make sense. Moreover, learning on the low-energy parts is often annihilated by learning on the high-energy parts during the entire loss calculation. To exploit this complementarity, we propose subband-based spectrogram fusion (SBSF), which combines the spectrogram of low-frequency and high-frequency estimated by different SE models. Experimental evaluations show that the SBSF significantly improved the SE performance.**

Index Terms: Speech enhancement, deep learning, spectrogram fusion, subband fusion

I. Introduction

Speech enhancement (SE) aims to extract clean speech signals from noisy speech signals [1]. The rise in popularity of speech applications has led to a wide variety of use scenarios. Front-end speech signal processing has become more and more important [2]. In particular, the quality of speech is degraded sharply in far-field conditions [3] or when substantial noise occurs. For example, the performance of automatic speech recognition [4] or speaker identification [5] significantly deteriorates in the presence of noise. To solve the problem, many systems now include an SE module at the front-end to perform noise reduction [4], [6]. Moreover, the SE system is of great help for human hearing aids [7].

Deep-learning-based SE has attracted attention because it demonstrates good performance and does not require any mathematical modeling assumptions. Researchers have put significant effort into improving frequency-domain SE systems [8] since 2013, and their performance has been greatly enhanced. The deep autoencoder [9] and deep neural network (DNN) [10], convolutional neural network [11], and recurrent neural network [12] are examples of early network structures for SE. Moreover, the combination of different types of networks [13] and some complex structures [14], [15]—for example, the U-NET structure [16] and the generative adversarial network [17]—have powerful performance.

Although frequency-domain SE systems can be improved in many ways, two types of learning targets are widely used: masking and mapping [18]. Masking targets [19], [20], [21] describe the time—frequency relationships of clean speech to background interference, whereas mapping targets [22], [9], [10] correspond to the spectral representations of clean speech [18]. The motivation for mapping targets is that the features can be estimated directly by the strong nonlinear capability of neural networks [10]. The masking targets are proposed in accordance with computational auditory scene analysis [23]. The earliest ideal binary mask [19] was designed to classify T–F bins of speech signals and non-speech signals, and the ideal ratio mask [20] indicates which T-F bins are dominated by speech. Researchers found that the two types of learning targets have some complementarities [24], [12]. However, few relevant studies analyzed their characteristics, and explored using their complementarity to get better SE performance.

In this paper, we address the aforementioned issues. We use direct mapping (DM) [10] and signal approximation (SA) [25], [26] as mapping and masking targets, respectively. First, we investigate the complementarity between these two learning targets based on their performance at each frequency bins. We find that the mapping-based and masking-based SE systems tend to perform well in the low-frequency and high-frequency parts, respectively. In addition, the recovery of the mapping-based SE system at high and low frequencies is very different, while the recovery of the masking-based SE system at each frequency is more stable.

On the basis of their complementarity, we investigate methods that combine them. Specifically, we propose subband-based spectrogram fusion (SBSF). First, we combine the spectrogram of low-frequency and high-frequency bands, which are estimated by different methods. Next, we combine the full-

Longbiao Wang and Tatsuya Kawahara are corresponding authors.

band SE and subband SE models. The subband enhanced spectrogram is used to replace corresponding subset information in the full-band enhanced spectrogram. The major difference between the proposed method and previous works [27], [28], [29] is that our method divides the full-band spectrogram into subbands from the perspective of the complementarity between the mapping-based and masking-based SE systems. In this study, we divide the full-band spectrogram into low and high subbands considering the loss and endeavor to apply different SE models. The reason for the poor performance of mapping at high frequencies is that the loss is mainly concentrated in the low-frequency part during network training. Thus, the subband optimization is used to optimize the poorly predicted part of the full-band spectrogram. Furthermore, we investigate the effective combination of different learning targets.

The rest of the paper is organized as follows. In Section II, we describe the mapping and masking SE systems. Section III introduces the proposed approach. In Section IV and V, we report the experimental settings and analysis. Finally, we present the conclusion and future work in Section VI.

II. Mapping and masking approaches for speech enhancement

A. Mapping and Masking Approaches

Learning targets are vital for supervised SE. Common learning targets can be divided into two categories: mapping and masking. The DM approach [10] uses a neural network to obtain the enhanced spectrogram directly. The loss function for DM is as follows:

$$\mathcal{L}_{DM}(|X|,|\widehat{X_{\mathrm{DM}}}|) = \frac{1}{TF}\sum_{t,f=1}^{T,F} |||\widehat{X_{\mathrm{DM}}}(t,f)| - |X(t,f)|||_F^2, \tag{1}$$

where t and f represent time frame and frequency bin, respectively. T represents the total number of frames in a speech sample. F represents the total frequency bins. $|\widehat{X_{\mathrm{DM}}}|$ is the output speech magnitude spectrogram, and $|X|$ is the target clean magnitude spectrogram.

Masking-based SE uses deep neural networks to obtain a mask between the speech and noisy speech. This mask is applied to the observed noisy signal to extract speech signal. SA [30], [26] utilizes a masking target. The loss function of SA is as follows:

$$\begin{aligned}\mathcal{L}_{SA}(|X|,|\widehat{X_{\mathrm{SA}}}|) &= \frac{1}{TF}\sum_{t,f=1}^{T,F} |||\widehat{X_{\mathrm{SA}}}(t,f)| - |X(t,f)|||_F^2 \\ &= \frac{1}{TF}\sum_{t,f=1}^{T,F} ||\widehat{M}(t,f)\odot|Y(t,f)| - |X(t,f)|||_F^2,\end{aligned} \tag{2}$$

where $|\widehat{M}|$ is the estimated mask, $|Y|$ is the noisy input speech magnitude spectrogram, $|\widehat{X_{\mathrm{SA}}}|$ is the masking-based speech magnitude spectrogram, and $\odot$ denotes point-wise matrix multiplication. Both DM and SA networks were based on two-layer bidirectional long short-term memory [31] (Bi-LSTM) neural networks.

B. Analysis on Complementarity

We investigate the complementarity of mapping and masking approaches by measuring the SE performance on different frequency bins. To measure the recovery performance of different frequencies between the enhancement signal and unprocessed noisy signal, we compare the loss at different frequencies between the enhanced and unprocessed features. We define the square loss ratio as follows:

$$\mathrm{Ratio}_F = \frac{\mathcal{L}_{square}}{\mathcal{L}_{original}} = \frac{\sum_{t=1}^{T} |||\widehat{X_{\mathrm{enh}}}(t)| - |X(t)|||_F^2}{\sum_{t=1}^{T} |||Y(t)| - |X(t)|||_F^2}, \tag{3}$$

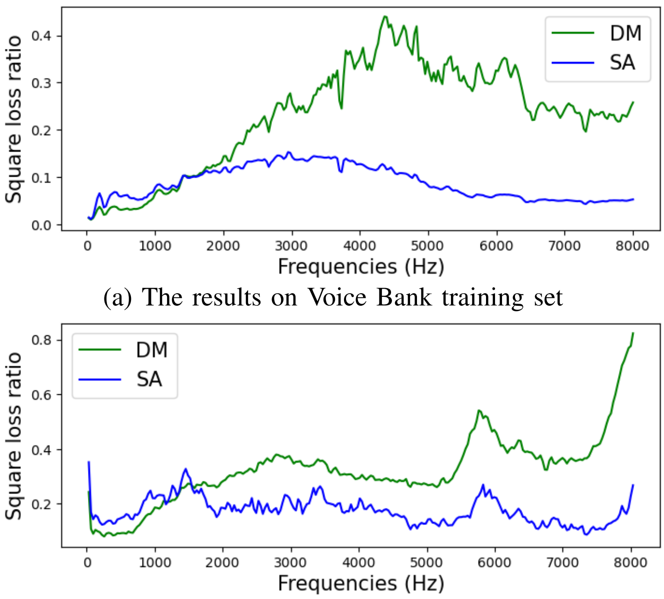
where $|\widehat{X_{\mathrm{enh}}}|$ and $|Y|$ are the enhanced and noisy input speech magnitude spectrogram, respectively. We only sum $\mathcal{L}_{square}$ and $\mathcal{L}_{original}$ along the time axis, so a $(1, F)$ dimensional vector can be obtained according to Eq. (3). The square loss ratio shows the recovery of the enhanced spectrogram at different frequencies compared to the input (noisy) spectrogram.

We use all training set (Voice Bank of 10k utterances and REVERB Challenge of 8k utterances) to compute the square loss ratio. The DM system was trained with Eq. (1), and the SA system was trained with Eq. (2). We used a 257-dimensional spectrum as input and output.

Fig. 1 shows the square loss ratio compared to the input noisy spectrogram. We can see different trends between the two models. The curves of the “DM” and “SA” are clearly demarcated around 1,400 Hz. The mapping-based spectrogram had better recovery in the low-frequency part but worse recovery in the middle and high-frequency parts compared with those of the masking-based spectrogram. The cut-off point of 1,400 Hz is consistent for the two datasets. With the masking-based spectrogram, the recovery of each frequency was uniform and stable.

C. Analysis on Dynamic Ranges of DM-based System

Fig. 2 shows the square loss of the mapping and masking approaches: $\mathcal{L}_{\mathcal{DM}}$ and $\mathcal{L}_{\mathcal{SA}}$. The loss of the low-frequency part was significantly larger than that of the high-frequency part. The 40th point (about 1,400 Hz) is marked with a red dashed line in accordance with the analysis in the previous section. When the frequency was lower than 1,400 Hz, the loss increased significantly, while it was stable for the frequency higher than 1,400 Hz. Thus, the dynamic range of loss differs for low-frequency and high-frequency regions in the DM system. This suggests that the main loss comes from the low-frequency part of the spectrogram, which may affect the recovery of the high-frequency part. Although the output of the masking-based network is not strictly distributed between 0 and 1, it still limits the output of the network and makes the difference between high and low frequencies smaller, which can alleviate the aforementioned problems. In the linear spectrogram, the energy difference between high and low frequencies is 50–90dB. Though the energy distributed at high and low frequencies is very different, there is some correlation between the high and low frequencies of the spectrogram.



(a) The results on Voice Bank training set

(b) The results on REVERB challenge training set

Fig. 1: The square loss ratio of mapping (Eq.(1)) and masking (Eq.(2)) in different frequencies (0 – 8,000 Hz), which are calculated with Eq. (3): the lower, the better.

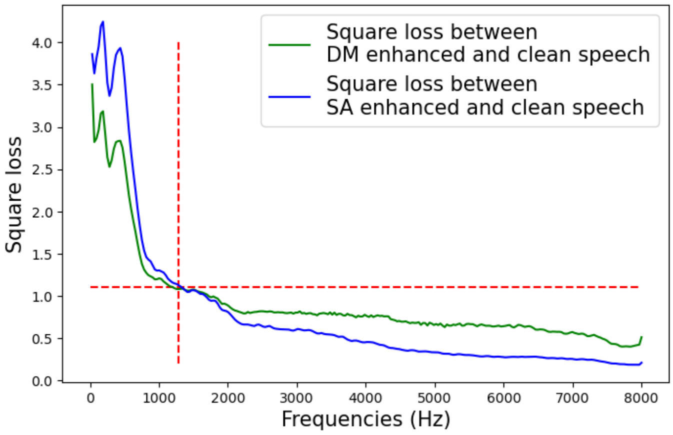


Fig. 2: The square loss of DM & SA in different frequencies (0 – 8,000 Hz) on Voice Bank training set.

Therefore, the mapping-based SE should use the full-band information to help the recovery of the low-frequency signal, while the high-frequency SE can be separately designed. This is not the case in the masking-based SE.

III. Subband-based Spectrogram Fusion

In Section II, we observed that the mapping-based SE system had better recovery in low frequencies, while the masking-based SE system had better recovery in high frequencies. Moreover, the mapping-based SE system can be divided into two dynamic ranges according to the square loss. It suggests the complementary between the mapping-based and masking-based SE systems. Combining these two analyses, we divide the whole spectrogram into two parts around 1,400 Hz and investigate the effective combination of these two learning targets. We call the high-frequency enhancement and low-frequency enhancement HEnh and LEnh, respectively. Both HEnh and LEnh use the full-band spectrogram as input feature to predict the sub-band output feature. We use the 257-dimensional linear spectrogram as an input feature. LEnh predicts 1-th to 40-th low-frequency bins of the spectrogram, and HEnh predicts 41-th to 257-th high-frequency bins. We investigate the following issues in the SBSF:

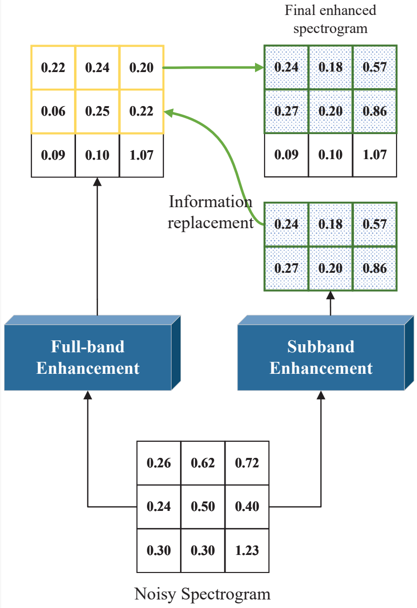


Fig. 3: The flowchart of subband-based optimization: the subband enhanced spectrogram will be used to replace the corresponding information of full-band enhanced spectrogram.

1) Is it effective to enhance different subbands with different learning target?

- **DM_L + SA_H**: directly concatenates DM-based LEnh and SA-based HEnh subband spectrograms. Considering the loss ratio, we select the mapping-based SE to enhance the low frequencies of the spectrogram, and the masking-based SE for the high-frequency spectrogram.

2) Is it effective to process different subbands by mapping separately?

- **DM_L + DM_H**: directly concatenates DM-based LEnh and DM-based HEnh subband spectrograms. In Section II–C, we reason that the poor high-frequency recovery of the mapping SE was because high energy in the low frequencies prevents effective training in the high-frequency regions due to the different dynamic ranges of the mapping targets. Thus, we design a separate DM-based method for the high-frequency region.

3) Is it more effective to combine the full-band and subband SE than combing two subband-based SE?

For DM_L + DM_H, we use subband SE to deal with different dynamic ranges of mapping. However, it ignores the global information, and may cause incoherence in the spectrogram. Thus, we also design a full-band and subband hybrid enhancement methods. Three steps are used: (1) full-band SE, which computes a full-band enhanced spectrogram; (2)

subband SE, which obtains a subband enhanced spectrogram; and (3) information replacement, which uses the subband enhanced information to replace the corresponding information in the full-band spectrogram. Fig. 3 shows the flowchart of the proposed method. Compared with DM_L + DM_H, we propose the following method:

- **DM_F → DM_H**: DM-based full-band enhancement with the DM-based HEnh replacement.

Furthermore, we inverstigate other full-band and subband hybrid combinations. Specifically, we design the following methods:

- **DM_F → DM_L**: DM-based full-band enhancement with the DM-based LEnh replacement.
- **DM_F → SA_H**: DM-based full-band enhancement with the SA-based HEnh replacement.
- **SA_F → DM_L**: SA-based full-band enhancement with the DM-based LEnh replacement.
- **SA_F → SA_H**: SA-based full-band enhancement with the SA-based HEnh replacement.
- **SA_F → DM_H**: SA-based full-band enhancement with the DM-based HEnh replacement.

IV. Experimental Settings

All networks were implemented using TensorFlow. The model parameters were randomly initialized. The implementation of all networks was based on two-layer bidirectional long short-term memory neural networks (Bi-LSTM). The number of nodes in each hidden layer was 1024. For SA and DM, the input was a 257-dimensional spectrogram, and the enhanced spectrogram output also had 257 dimensions. The activation of hidden layers for SA and DM was ReLU. For the activation function of the output layer, ReLU was chosen for SA and a linear function was used for DM. In addition, we estimated the 217-dimensional high-frequency and 40-dimensional low-frequency spectrograms for the HEnh and LEnh, respectively.

A. Datasets

We adopted the Voice Bank and REVERB Challenge datasets to evaluate SBSF under additive noise and reverberation conditions, respectively.

1) Voice Bank: For the training set, we selected 26 speakers from the Voice Bank corpus [32]—13 male and 13 female—from the same accent region (England). Approximately 400 sentences are available from each speaker. The training set contains $10,340$ sentences. For validation set, we selected another 2 speakers from the Voice Bank corpus [32]—1 male and 1 female—from the same accent region (England). The the validation set contains $1,232$ sentences. Two artificially generated (speech-shaped noise and babble) and eight real noise recordings from the Demand database [33] were used to synthesize the training and validation sets. The signal-to-noise ratio (SNR) values used for training were 15, 10, 5, and 0 dB. Two other speakers from England in the same corpus, a male and a female, and five other noises from the Demand database were used to create the test set. The SNR values used for testing were 17.5, 12.5, 7.5, and 2.5 dB. All data were sampled at 16 kHz.

2) REVERB Challenge [34]: The challenge uses utterances spoken by a single stationary distant-talking speaker with 1-channel (1ch), 2-channel (2ch) or 8-channel (8ch) microphone arrays in reverberant meeting rooms. In this paper, we use only single-channel data of channel-1 to train the model. The training set contains 7,861 utterances. We used the development set for model selection.

B. Evaluation Metrics

We used several composite measures to evaluate different SE systems. Composite objective measures are obtained by linearly combining existing objective measures: C_{sig} for a five-point scale of signal distortion (SIG) [35]; C_{bak} for a five-point scale of background intrusiveness (BAK) [35]; C_{ovl} for the overall quality (OVRL, [1=bad, 2=poor, 3=fair, 4=good, 5=excellent]) [35]. The three composite measures obtained from log likelihood ratio (LLR) [35], the perceptual evaluation of speech quality (PESQ) [36], segmental SNR (segSNR) [35], and weighted-slope spectral (WSS) [37] distance:

$$\begin{aligned} C_{sig} = 3.093 - 1.029 * LLR + 0.603 * PESQ \\ -0.009 * WSS \end{aligned} \tag{4}$$

$$\begin{aligned} C_{bak} = 1.634 + 0.478 * PESQ - 0.007 * WSS \\ +0.063 * segSNR \end{aligned} \tag{5}$$

$$\begin{aligned} C_{ovl} = 1.594 + 0.805 * PESQ - 0.512 * LLR \\ -0.007 * WSS \end{aligned} \tag{6}$$

We also adopted the Short-Time Objective Intelligibility (STOI) [38] as evaluation metrics. For all metrics, higher values indicate better performance.

TABLE I: Performance of Different SE Systems on Voice-bank Test Set.

Systems	C_{SIG}	C_{BAK}	C_{OVL}	PESQ
Noisy	3.35	2.44	2.63	1.97
DM	3.85	2.55	3.23	2.60
SA	3.65	2.49	3.07	2.51
DM → DM	3.89	2.55	3.25	2.60
SA → DM	3.89	2.54	3.23	2.56
DM_L + SA_H	3.76	3.04	3.18	2.61
DM_L + DM_H	4.06	3.11	3.38	2.70
DM_F → DM_H	**4.09**	**3.12**	**3.42**	**2.74**
DM_F → DM_L	4.02	2.59	3.35	2.69
DM_F → SA_H	3.94	3.05	3.27	2.63
SA_F → DM_H	3.76	3.05	3.18	2.60
SA_F → DM_L	3.87	2.50	3.21	2.58
SA_F → SA_H	3.71	3.00	3.11	2.52

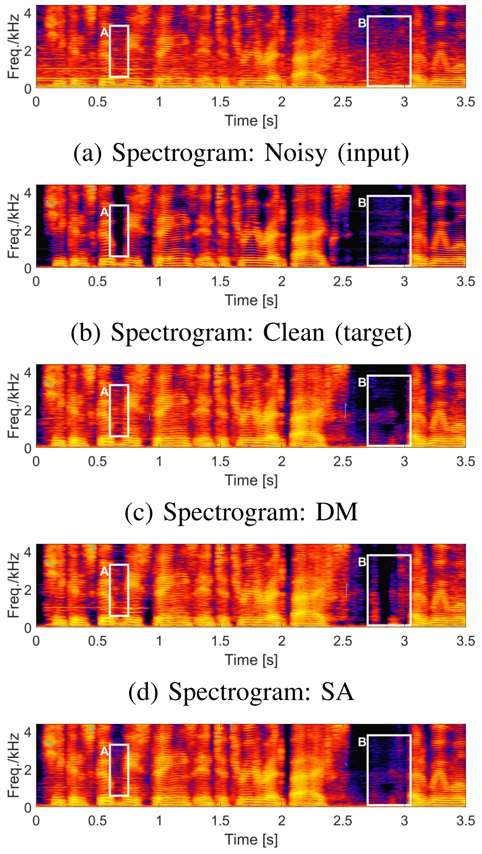
C. Baseline Models

For the baseline methods, we tested the following methods:

- **DM**: SE system trained with Eq. (1).
- **SA**: SE system trained with Eq. (2).
- **DM→DM**: two-stage method, which first uses the DM-based SE system trained with Eq. (1) and then uses another DM-based SE system trained with Eq. (1) for re-enhancement.

TABLE II: Performance of Different SE Systems on Reverb Chanllenge 2014 test set.

Systems	Far room 1		Far room 2		Far room 3		Near room 1		Near room 2		Near room 3	
	PESQ	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**	**PESQ**	**STOI**
Noisy	2.59	84.69%	1.99	78.20%	1.87	71.31%	3.11	95.18%	2.39	92.32%	2.27	89.38%
SA	2.91	89.81%	2.33	84.74%	2.26	83.22%	**3.40**	95.95%	2.78	93.87%	2.59	90.99%
DM	2.74	88.74%	2.45	85.17%	2.36	84.21%	2.94	93.35%	2.78	92.50%	2.70	91.05%
DM_F → DM_H	2.91	89.73%	**2.56**	**86.99%**	**2.48**	**86.10%**	3.18	95.05%	**2.95**	**94.19%**	**2.83**	**92.90%**



(a) Spectrogram: Noisy (input)

(b) Spectrogram: Clean (target)

(c) Spectrogram: DM

(d) Spectrogram: SA

(e) Spectrogram: DM_F → DM_H

Fig. 4: One sample (from Voice-Bank test set) of spectrograms and their corresponding minimum difference masks: (a) is the spectrogram of noisy input; (b) is the spectrogram of the clean signal; (c) is the spectrogram of DM; (d) is the spectrogram of SA; ; (e) is the spectrogram of DM_F → DM_H.

- **SA→DM**: two-stage method, which first uses SA-based SE system trained with Eq. (2) and then uses another DM-based SE system trained with Eq. (1) for re-enhancement.

V. Experimental Results and Analysis

Table I shows the performance of the different SE systems. Generally, "DM" performed better than "SA" in this study. The simple two-stage methods (DM→DM and SA→DM) did not get much improvement from the baseline DM.

We find that the mapping-based SE performed better below 1,400 Hz, while the masking-based SE system was better above 1,400 Hz. However, we find that a simple combination of the two, "DM_L + SA_H," did not improve so much. This is because even in some high-frequency regions, mapping often produced better performance. However, "DM_L + DM_H" had a relatively large improvement. This shows that subband enhancement considering dynamic ranges is more beneficial for the mapping method.

Furthermore, the full-band and subband hybrid approaches showed better performance than directly concatenating subband spectrograms. Nevertheless, the experimental results also show that the masking-based subband optimization was worse than the mapping-based subband optimization. Table II shows the results of different methods on the REVERB Challenge test set. "DM_F → DM_H" further improved the performance of the mapping-based system. Fig. 4 shows that SBSF has better recovery both in the speech and silent segment part.

Fig. 5 shows the square loss ratio of "DM_F → DM_H," "DM_L + DM_H," and "DM_L + SA_H." We divided the full-band frequency into three parts. Part 1 had no significant differences among these methods for low-frequency recovery. We call Part 2 middle-frequency and Part 3 high-frequency. "DM_L + SA_H" had poor recovery in middle frequencies. Although the recovery of "DM_L + SA_H" in other regions was not much different from other methods, a large degradation was observed in PESQ, which illustrates the importance of Part 2 recovery.

We investigated the performance of different models at different frequencies according to the square loss ratio. The curves of the "DM_F → DM_H" and "DM_L + DM_H" are clearly demarcated around 3,200 Hz in the middle and high frequencies. "DM_F → DM_H" worked well for middle frequencies (Part 2) but not for high frequencies (Part 3). This suggests that about 3,200 Hz would be another cut-off point for dividing the frequencies into two dynamic ranges.

We synthesized another test set to evaluate the SE systems on different signal-to-noise ratios (SNRs) and noise conditions. We used all clean speech in the test set from the Voice Bank dataset and chose four noisy conditions: crowd, machine, alarm, siren, and wind. These noise samples were selected from a dataset with 100 non-speech audio clips[1]. The SNRs we chose were -5, 0, 5, 10, and 15 db. Table III shows the performance of enhancement systems under multiple noisy conditions. Except for the machine noise, the improvement of our method was consistent. Fig. 6 shows the performance of different SE systems for multiple SNRs. "SA" did not perform well under low SNRs and was even lower than noisy speech on SIG and OVRL. The performance of other

[1] http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html

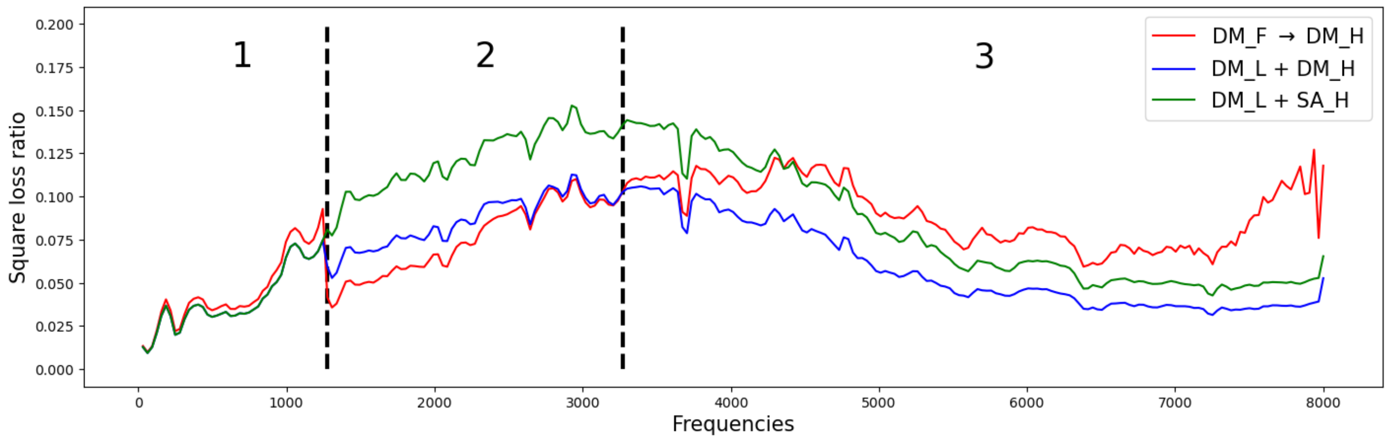
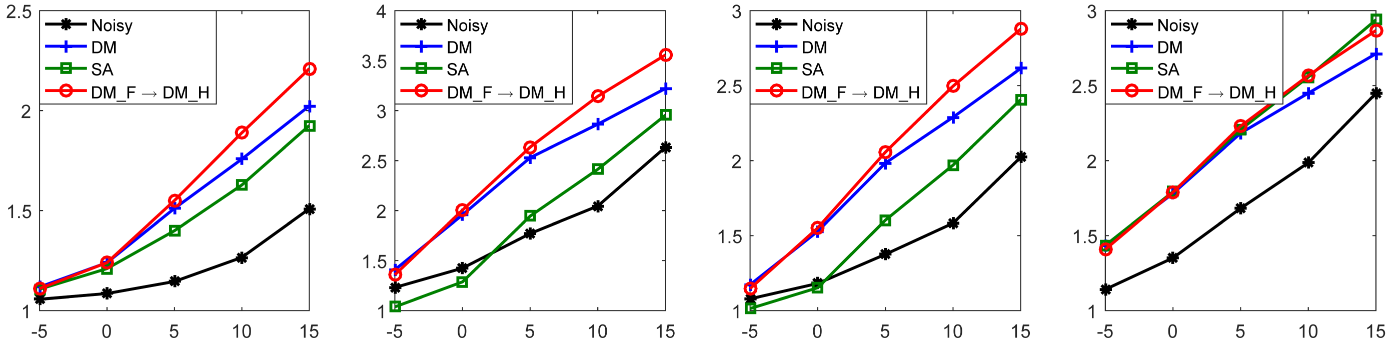


Fig. 5: Square loss ratio (the lower, the better) of DM_F → DM_H, DM_L + DM_H, and DM_L + SA_H at different frequencies (257-dimensional linear spectrogram) on Voice Bank training set, which were calculated with Eq. (3).

TABLE III: Performance of Different SE Systems on Different Noisy Conditions (Unseen, synthesized, clean speech from Voice Bank dataset, noisy from non-speech 100).

Systems	Crowd Noise				Machine Noise				Alarm and Siren				Wind			
	Csig	Cbak	Covl	PESQ	Csig	Cbak	Covl	PESQ	Csig	Cbak	Covl	PESQ	Csig	Cbak	Covl	PESQ
Noisy	1.45	1.74	1.27	1.18	1.85	1.87	1.47	1.19	1.31	1.50	1.15	1.13	2.65	1.77	1.90	1.35
DM	1.97	2.03	1.62	1.35	2.19	2.02	1.75	**1.38**	2.38	2.15	1.90	1.51	3.00	2.22	2.36	1.85
SA	1.66	2.14	1.42	1.30	1.85	**2.14**	1.53	1.33	1.63	2.17	1.44	1.39	2.53	2.27	2.10	1.77
DM_F → DM_H	**2.12**	2.12	**1.74**	**1.44**	2.15	2.02	1.72	1.37	**2.62**	**2.24**	**2.08**	**1.64**	**3.21**	2.29	**2.51**	1.91



(a) PESQ in different SNRs (b) SIG in different SNRs (c) OVRL in different SNRs (d) BAK in different SNRs

Fig. 6: Performance of evaluation measures (PESQ, SIG, OVRL, BAK) of different enhancement systems in SNRs (-5, 0, 5, 10, and 15 db) conditions: -5 db was an unseen condition, and the noisy conditions were unseen. The horizontal axis represents SNRs, and the vertical axis represents the value of the evaluation metric.

enhancement methods at low SNRs was very similar. All methods showed better performance with the increase in SNR. Compared with other methods, the performance of "DM_H → DM_F" increased significantly as the SNR improved.

VI. Conclusion And Future Work

In this paper, we first explored the complementarity between mapping-based and masking-based speech enhancement (SE) systems, which perform well in low and high frequencies, respectively, in accordance with the square loss ratio. The cut-off point was about 1,400 Hz. Meanwhile, the mapping-based method had obvious differences between low and high frequencies, while the performance of the masking-based method was uniform and stable. Therefore, we designed subband-based spectrogram fusion (SBSF), considering the dynamic ranges of the mapping-based SE system. We find that the mapping is more suitable for subband processing. In addition to the low-frequency information below 1,400 Hz, the middle-frequency between 1,400 and 3,200 Hz also has a larger impact on speech quality. Furthermore, we find that using the subband to process the worse part of the full-band can bring greater improvement compared with the simple multi-stage enhancement. The combination of full-band and subband processing was even better. In the future, we will endeavor to design the subband more elaborately and to add propelled phase information for the SE system.

VII. Acknowledge

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123. This work was supported by the National Natural Science Foundation of China under Grant 62176182.

References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, T. Hayashi, C. Boeddeker, Z. Chen, and S. Watanabe, "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. SLT*, 2021, pp. 785–792.

[3] A. S. Subramanian, C. Weng, M. Yu, S.-X. Zhang, Y. Xu, S. Watanabe, and D. Yu, "Far-field location guided target speech extraction using end-to-end speech recognition objectives," in *Proc. ICASSP*, 2020, pp. 7299–7303.

[4] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. International conference on latent variable analysis and signal separation*. Springer, 2015, pp. 91–99.

[5] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," in *Proc. ICASSP*, 2014, pp. 3997–4001.

[6] X. Zhang, Z.-Q. Wang, and D. Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *Proc. ICASSP*, 2017, pp. 276–280.

[7] Y.-H. Lai, W.-Z. Zheng, S.-T. Tang, S.-H. Fang, W.-H. Liao, and Y. Tsao, "Improving the performance of hearing aids in noisy environments based on deep learning technology," in *Proc. EMBC*, 2018, pp. 404–408.

[8] M. Handa, T. Nagai, and A. Kurematsu, "Frequency domain multi-channel speech separation and its applications," in *Proc. ICASSP*, vol. 5, 2001, pp. 2761–2764 vol.5.

[9] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. Interspeech*, vol. 2013, 2013, pp. 436–440.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[11] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[12] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. HSCMA*, 2017, pp. 136–140.

[13] M. Ge, L. Wang, N. Li, H. Shi, J. Dang, and X. Li, "Environment-dependent attention-driven recurrent convolutional neural network for robust speech enhancement." in *Proc. Interspeech*, 2019, pp. 3153–3157.

[14] S. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement." in *Proc. Interspeech*, 2016, pp. 3768–3772.

[15] T. Gao, J. Du, L. Dai, and C. Lee, "SNR-based progressive learning of deep neural network for speech enhancement." in *Proc. Interspeech*, 2016, pp. 3713–3717.

[16] H. Choi, J. Kim, J. Huh, A. Kim, J. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. ICLR*, 2018.

[17] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[18] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[19] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis ," in *Proc. Speech Separation by Humans and Machines*. Springer, 2005, pp. 181–197.

[20] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[21] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *Proc. ICASSP*, 2016, pp. 5220–5224.

[22] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. ICASSP*, 2014, pp. 4628–4632.

[23] D. Wang and G. J. Brown, *Contributors*, 2006, pp. xix–xx.

[24] Y. Xu, J. Du, Z. Huang, L. Dai, and C. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," pp. 1508–1512, 2015.

[25] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[26] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, 2014, pp. 577–581.

[27] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *Proc. WASPAA*, 2019, pp. 298–302.

[28] Q. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "Joint noise and mask aware training for DNN-based speech enhancement with sub-band features," in *Proc. HSCMA*, 2017, pp. 101–105.

[29] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. ICASSP*, 2021, pp. 6633–6637.

[30] H. Shi, L. Wang, M. Ge, S. Li, and J. Dang, "Spectrograms fusion with minimum difference masks estimation for monaural speech dereverberation," in *Proc. ICASSP*, 2020, pp. 7544–7548.

[31] A. Graves, *Long Short-Term Memory*. Springer Berlin Heidelberg, 2012, pp. 37–45.

[32] 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*, 2013, pp. 1–4.

[33] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am*, vol. 133, no. 5, pp. 3591–3591, 2013.

[34] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[37] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. ICASSP*, vol. 7, 1982, pp. 1278–1281.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.