SPECTROGRAMS FUSION WITH MINIMUM DIFFERENCE MASKS ESTIMATION FOR MONAURAL SPEECH DEREVERBERATION

Hao Shi[1], *Longbiao Wang*[1*], *Meng Ge*[1], *Sheng Li*[2*], *Jianwu Dang*[1,3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

{hshi_cca, longbiao_wang, gemeng}@tju.edu.cn, sheng.li@nict.go.jp, jdang@jaist.ac.jp

ABSTRACT

Spectrograms fusion is an effective method for incorporating complementary speech dereverberation systems. Previous linear spectrograms fusion by averaging multiple spectrograms shows outstanding performance. However, various systems with different features cannot apply this simple method. In this study, we design the minimum difference masks (MDMs) to classify the time-frequency (T-F) bins in spectrograms according to the nearest distances from labels. Then, we propose a two-stage nonlinear spectrograms fusion system for speech dereverberation. First, we conduct a multi-target learning-based speech dereverberation front-end model to get spectrograms simultaneously. Then, MDMs are estimated to take the best parts of different spectrograms. We are using spectrograms in the first stage and MDMs in the second stage to recombine T-F bins. The experiments on the REVERB challenge show that a strong feature complementarity between spectrograms and MDMs. Moreover, the proposed framework can consistently and significantly improve PESQ and SRMR, both real and simulated data, e.g., an average PESQ gain of 0.1 in all simulated data and an average SRMR gain of 1.22 in all real data.

Index Terms— speech dereverberation, spectrograms fusion, multi-target learning, two-stage, deep learning

1. INTRODUCTION

In real life, speech is always disturbed by various reverberations, especially in confined indoor spaces [1]. The echo can reduce the clarity and intelligibility of speech and seriously affect people's hearing experience. Speech dereverberation provides preprocessing for speech recognition [2, 3, 4], sound source localization, and speaker identification [5].

Recently, supervised deep dereverberation methods [6, 7] have shown powerful capability and achieve better performances than the traditional ways in speech de-reverberation. These supervised deep dereverberation methods can be categorized into two groups according to the learning targets, i.e., masking targets and mapping targets [7]. Masking targets [7, 8, 9] describe the time-frequency relationships of clean speech to background interference, while mapping [6, 7, 10, 11] targets correspond to the spectral representations of clean speech [7, 8].

Based on these two kinds of learning targets, people more explore the improvement of the neural network model [12, 13] or add more information to the network [14], but do not further explore the deeper relationship between the two kinds of learning targets. Spectrograms fusion is an effective method for incorporating complementary information from these two types of speech dereverberation systems. However, there are two challenges to build a spectrograms fusion system. The first is the nonlinear nature of real scenarios. Although previous linear fusion by averaging spectrograms shows good performances [13], it can not fuse various systems with different patterns by simple linear processing. The second is that it is unrealistic to build massive systems for fusion.

To overcome these problems, we design a nonlinear spectrograms fusion system for speech dereverberation. Many systems are now training according to the mean squared error (MSE) criterion, and this motivates us. If the time-frequency (T-F) bins close to the clean spectrogram in the enhanced spectrograms are fused back into a spectrogram, it may be helpful for enhancement:

1. In the first stage, we use multi-target learning (MTL) with both masking and mapping following [15, 13, 16] to obtain different learning targets spectrograms, instead of constructing various systems with a large number of resources.

2. For nonlinear spectrograms fusion, we design the minimum difference masks (MDMs) to classify T-F bins, which are nearest to the labels in spectrograms. In the second stage, the MDMs are estimated using deep neural networks (DNN) to take the best parts of the different spectrograms. We use spectrograms in the first stage and MDMs in the second stage to recombine spectrograms into one spectrogram.

*Corresponding author.

The rest of this paper is organized as follows. Section 2, and 3 describe our proposed method. Section 4 gives the data description and experiment evaluations. Section 5 gives the conclusion and future work.

2. MULTI-TARGET LEARNING AND LINEAR FUSION

The idea of MTL [17, 13] is to learn the different targets in one model. In this study, mapping and masking targets are learned in one single Bi-LSTM (Bi-directional Long Short-Term Memory) model with two outputs:

$$L_{MTL} = L_{DM} + \alpha L_{SA} \tag{1}$$

where α is the weight coefficient of the two mean squared error (MSE) [18] items corresponding to the dual outputs of Bi-LSTM. The **DM** is the proposed direct mapping target [10, 11], uses a linear output layer and MSE loss function:

$$L_{DM} = \sum_{t,f} \left(spc_{DM}(t,f) - spc_c(t,f)\right)^2 \tag{2}$$

where t and f denote time and frequency, respectively. spc_{DM} and spc_c are the estimated spectrogram and the reference clean spectrogram, respectively. The **SA** is the abbreviation of the second learning target, which is called signal approximation [12, 19]. It trains a ratio mask estimator that minimizes the difference between the spectrogram of clean speech and that of estimated speech:

$$L_{SA} = \sum_{t,f} \left(spc_r(t,f) * mask_{SA}(t,f) - spc_c(t,f)\right)^2 \tag{3}$$

where spc_r denotes the reverb spectrogram and $mask_{SA}$ is the estimated mask. We denote the DM outputs and SA outputs of MTL as **MT-DM** and **MT-SA**, respectively. These estimated spectrograms could be combined via a simple average operation to post-processing [13], called linear spectrograms fusion:

$$spc_{MT-LF} = \frac{1}{2}\left(spc_{MT-DM} + spc_{MT-SA}\right) \tag{4}$$

where spc_{MT-LF} denotes the linear fusion spectrogram. spc_{MT-DM} and spc_{MT-SA} are two enhanced spectrograms. We abbreviate the linear fusion approach as **MT-LF**.

3. NONLINEAR SPECTROGRAMS FUSION WITH MINIMUM DIFFERENCE MASKS ESTIMATION

We design a set of masks, called minimum difference masks (MDMs). Each MDM corresponds to an enhanced spectrogram. We use a nonlinear system to fuse T-F bins nearest to the labels in multiple spectrograms into one spectrogram may be better than using linear methods.

3.1. Minimum Difference Masks

We define the distance between each enhanced T-F bin and its corresponding label as d_i:

$$d_i(t,f) = \left|spc_i(t,f) - spc_c(t,f)\right|, i \in \{_{MT-DM}, _{MT-SA}\} \tag{5}$$

where spc_i denotes an enhanced spectrogram from the MTL model. The i in this study is $_{MT-DM}$ or $_{MT-SA}$. The labels of MDMs are defined as:

$$\widetilde{MDM}_i(t,f) = \begin{cases} 1, & i = \arg\min\limits_i d_i(t,f) \\ 0, & otherwise \end{cases} \tag{6}$$

When $d_i(t,f)$ is minimum, set $\widetilde{MDM}_i(t,f)$ to value 1 and 0 otherwise. With labels, MDMs estimation can be treated as a supervised problem. Considering the continuity of the spectrogram, MDMs are real values in (0, 1) in testing. Fig. 1 shows the process of computing labels of MDMs.

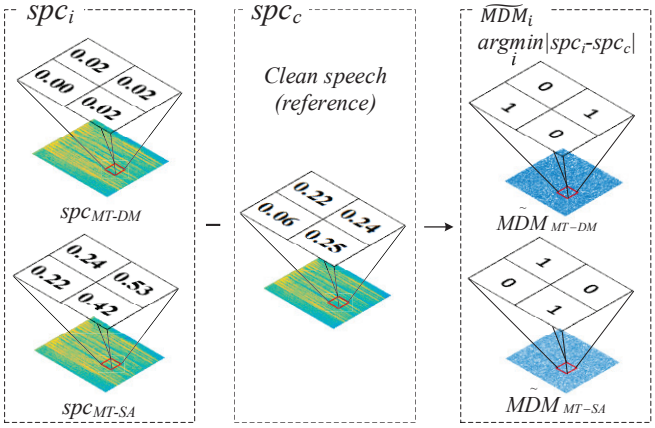


Fig. 1. The process of computing labels of MDMs: spc_c is clean spectrogram, spc_i are enhanced spectrograms from the first stage, $\widetilde{MDM}_i$ are labels of MDMs.

3.2. Nonlinear Spectrograms Fusion

Nonlinear spectrograms fusion consists of two stages. In the first stage, an MTL based speech de-reverberation front-end model is conducted to get spectrograms of different targets, using loss function Eq. (1). Then an MTL for DNN based back-end model is trained to predict MDMs. Consideration of features complementarity [15, 13], two learning targets are conducted. Estimation MDMs only, and estimation MDMs with spectrograms:

$$L_{MDM-2O} = \sum_i \sum_{t,f} \left(MDM_i(t,f) - \widetilde{MDM}_i(t,f)\right)^2 \tag{7}$$

$$L_{MDM-4O} = L_{MDM-2O} + \alpha\left(L_{DM} + L_{SA}\right) \tag{8}$$

where $\widetilde{MDM}_i$ denotes the labels of MDMs while MDM_i denotes estimated MDMs. We abbreviate the models trained using Eq. (7) as **MDM-2O** while **MDM-4O** using Eq. (8). In the testing stage, nonlinear selection processing is conducted:

$$select_i(t,f) = MDM_i(t,f) * spc_i(t,f) \tag{9}$$

where $select_i$ denotes nonlinear selected portion in spc_i. Finally, we recombine each selected portion to gain the final

enhanced spectrograms:

$$spc_{fusion} = \sum_{i} select_i \tag{10}$$

where spc_{fusion} denotes the final nonlinear fusion spectrogram. Fig. 2 shows the training and testing processing in the second stage.

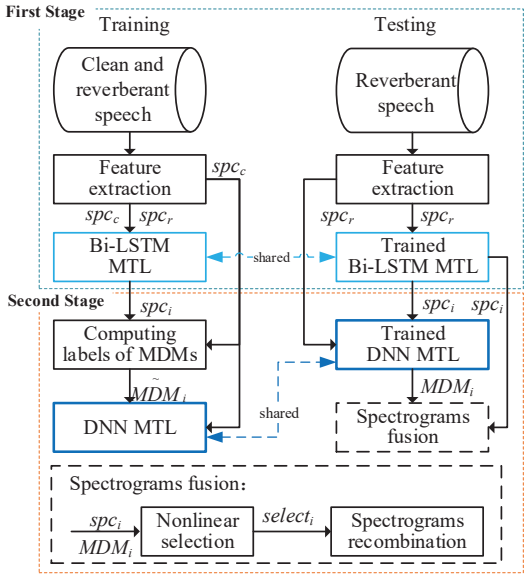


Fig. 2. Nonlinear spectrogram fusion system: spc_r is reverberant spectrogram, spc_i are enhanced spectrograms in the first stage, MDM_i are estimated MDMs and $select_i$ is selected portion in spc_i.

4. EXPERIMENTS

The experiments were conducted on the REVERB challenge task [20]. The REVERB challenge dataset contained simulated and real utterances; the training data only included simulated recordings. The simulated and real recordings on test data were used for evaluation. The speech signals were sampled to 16kHz. The frame length and shift were set to 512 and 256, respectively. The input and output features were both magnitudes of spectrograms for the whole utterance.

All networks were implemented based on TensorFlow. In the first stage, the Bi-LSTM model consisted of the 257-dimensional input layer, two hidden layers with 1024 nodes for each layer, and the 257-dimensional output layer for both mapping and masking target. In the second stage, the DNN model consisted of the 771-dimensional input layer, two hidden layers with 1024 nodes for each layer, and the 257-dimensional output layer for both two (MDM-2O) or (MDM-4O) four outputs. The model's parameters were randomly initialized. A validation set was adopted to control the learning rate (initialized as 0.01), which was decreased by 50% when no improvement between two consecutive epochs. Besides, the performance in the validation set was decided whether to save the trained model in one epoch. Each back-propagation through time (BPTT) or back-propagation (BP) batch contained eight utterances. α for multi-target learning both in the first stage and second stage were set to 1.

Table 1. PESQ and SRMR results for simulated data.

Models	PESQ			SRMR		
	Far	Near	Avg.	Far	Near	Avg.
Reverb	2.15	2.59	2.37	3.43	3.94	3.68
DM	2.58	2.88	2.73	4.39	4.88	4.64
SA	2.54	2.93	2.74	4.48	4.92	4.70
MT-DM	2.56	2.90	2.73	4.42	4.92	4.67
MT-SA	2.60	3.01	2.81	4.64	4.97	4.80
MT-LF	2.64	3.02	2.83	4.58	4.99	4.78
MDM-2O(B)	2.56	2.92	2.74	4.38	4.54	4.46
MDM-2O	2.65	3.06	2.86	4.59	4.96	4.78
MDM-4O(B)	2.66	3.09	2.87	4.61	5.02	4.81
MDM-4O	**2.71**	**3.14**	**2.93**	**5.09**	**5.60**	**5.35**

Table 2. SRMR results in real data.

Models	SRMR		
	Far	Near	Avg.
Reverb	3.187	3.171	3.179
DM	3.291	2.926	3.109
SA	3.657	3.535	3.596
MT-DM	3.707	3.586	3.647
MT-SA	3.852	3.669	3.761
MT-LF	3.842	3.699	3.771
MDM-2O(B)	3.686	3.512	3.599
MDM-2O	3.931	3.767	3.849
MDM-4O(B)	3.956	3.815	3.885
MDM-4O	**5.055**	**4.927**	**4.991**

4.1. Experiments on Multi-target Learning and Linear Fusion

Table 1 shows the perceptual evaluation of speech quality (PESQ) [21] and the speech-to-reverberation modulation energy ratio (SRMR) [22] performance on simulated data sets, and Table 2 shows the SRMR performance on real data sets. "Reverb" denotes the reverb speech. "DM" and "SA" denote the mapping and the masking approach using Eq. (2) and Eq. (3) separately. "MT-DM" and "MT-SA" denote two outputs of the MTL approach using Eq. (1). "MT-LF" denotes linear fusion spectrograms using Eq. (4). All the baseline models consisted of the 257-dimensional input layer, two hidden layers with 1024 nodes for each layer. Several observations could be made from results.

1. First, for the PESQ measure, the DM approach yielded better results than the SA approach in the far-field, while the near field is the opposite.

2. Second, for the SRMR measure, the SA approach consistently outperformed the DM approach. In contrast, the DM approach generated the worst performance.

3. Third, each output of the MTL approach, MT-DM approach, and MT-SA approach, is the most direct way to demonstrate the complementarity of different learning targets. Accordingly, the PESQ and SRMR measures were improved, overusing one single target learning model.

4. Besides, one of the MTL model outputs was consistently superior to another; the MT-SA approach had a better performance than the MT-DM approach.

5. Finally, linear spectrograms fusion was helpful for speech dereverberation, although SRMR had some degradation in the far-field.

4.2. Experiments on Nonlinear Spectrograms Fusion

The fusion approaches, using Eq. (7) and Eq. (8) training models are compared both in the far-field and near-field. The training methods of "MDM-2O(B)" and "MDM-2O" are the same. However, in the fusion, "MDM-2O(B)" restores the predicted result to 0-1 value in binary masks, while "MDM-2O" uses the predicted probability in real value masks. The same as "MDM-4O(B)" and "MDM-4O". The results at the bottom of Table1 and Table2 shown that real masks worked better than binary masks, indicating that soft masks are more suitable than hard masks. Moreover, the MDM-4O approach showed its superiority in all PESQ and SRMR. This result suggests that there is an active feature complimentary between spectrograms and MDMs.

Compared with the MT-LF approach, most nonlinear spectrograms fusion approaches, except the MDM-2O(B) approach, showed superb effects. Using binary masks to fuse spectrograms may cause the loss of time-varying information in the spectrogram, which may be one of the reasons for the poor performance of the MDM-2O(B) approach. MDM-2O and MDM-4O(B) got a more smooth improvement in PESQ and SRMR. In contrast, MDM-4O got a remarkable improvement in PESQ and SRMR, both real and simulated data, e.g., an average PESQ gain of 0.1 in all simulated data and an average SRMR gain of 1.22 in all real data. The success of the MDM-4O approach inspired us to use MDMs as an auxiliary feature to predict a spectrogram or mask in future work.

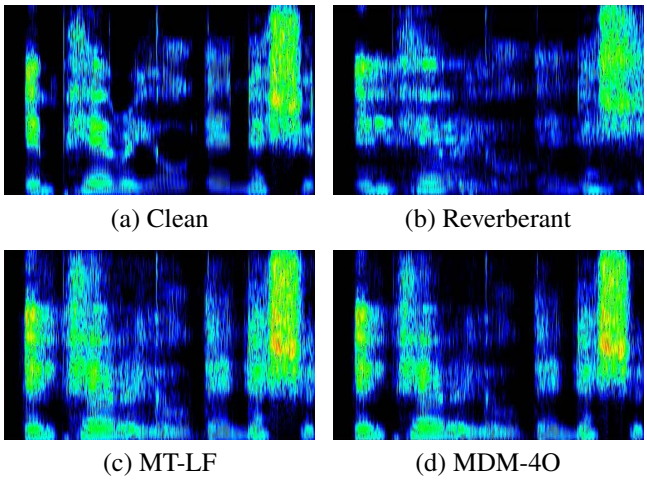
Fig. 3 shows the magnitude spectrograms [1]. Several observations could be made from Fig. 3.

1. First, two enhancement approaches achieved excellent results in reducing the reverberation and restoring the spectrum at low frequencies buried under reverberation.

2. Second, interference usually comes from high frequencies, the MDM-4O approach had an excellent ability to suppress high-frequency interference.

Considering that we used utterance-level features to train the model in this study, as a comparative experiment, the frame-level features with frame expansion will be explored to train the model in future work. Moreover, our method has the potential to be extended to multi-system fusion.

[1]The audio samples are at https://paperdemo.github.io/icassp2020.html



(a) Clean (b) Reverberant

(c) MT-LF (d) MDM-4O

Fig. 3. Magnitude spectrograms from speech (a) clean, (b) reverberant, (c) enhanced with linear fusion and (d) enhanced with nonlinear fusion.

5. CONCLUSION

We proposed a nonlinear spectrograms fusion with minimum difference masks estimation system for speech dereverberation. First, a multi-target learning Bi-LSTM based speech dereverberation front-end was conducted to obtain different learning targets spectrograms. Then, MDMs were estimated to classify T-F bins nearest to the labels. Finally, we used spectrograms from the first stage and MDMs from the second stage to fuse the best parts of spectrograms. We observed an active feature complementarity between spectrograms and minimum difference masks (MDMs) when using multi-target learning. By Nonlinear spectrograms fusion, speech dereverberation mainly improved both the speech quality and speech-to-reverberation modulation energy ratio, e.g., an average PESQ gain of 0.1 in all simulated data and an average SRMR gain of 1.22 in real data. In future studies, we will analyze the spectrogram and use the time-varying information in the spectrogram for fusion. Moreover, feature fusions for other speech tasks will also be explored, such as MFCC, for automatic speech recognition.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61771333, the National Key R&D Program of China under Grant 2018YFB1305200, the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330. Sheng Li is partially supported by JSPS KAKENHI Grant No. 19K24376 and NICT tenure-track startup fund "Research of advanced automatic speech recognition technologies", Japan.

7. REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 745–777, 2014.

[2] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," *In Proc. IEEE ASRU*, pp. 444–451, 2015.

[3] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE TASLP*, vol. 18, no. 7, pp. 1692–1707, 2010.

[4] Y. Ueda, L. Wang, A. Kai, and B. Ren, "Environment-dependent denoising autoencoder for distant-talking speech recognition," *EURASIP J. Adv. Sig. Proc.*, vol. 2015, no. 92, 2015.

[5] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi, "Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification," *EURASIP J. Audio, Speech and Music Processing*, vol. 2015, no. 12, 2015.

[6] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2013.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[9] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM TASLP*, vol. 24, no. 3, pp. 483–492, 2016.

[10] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2015.

[11] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 982–992, 2015.

[12] 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," *LVA/ICA*, pp. 91–99, 2015.

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

[14] 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," *Proc. Interspeech 2019*, pp. 3153–3157, 2019.

[15] 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," *arXiv preprint arXiv:1703.07172*, 2017.

[16] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," *ICML*, pp. 1310–1318, 2013.

[17] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," *Proc. INTERSPEECH*, 2015.

[18] G. Box, "Signal-to-noise ratios, performance criteria, and transformations," *Technometrics*, vol. 30, no. 1, pp. 1–17, 1988.

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

[20] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," *Proc. IEEE WASPAA*, 2013.

[21] 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," *Proc. IEEE-ICASSP*, vol. 2, pp. 749–752, 2001.

[22] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE TASLP*, vol. 16, no. 1, pp. 229–238, 2007.