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]

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

[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

1. Introduction

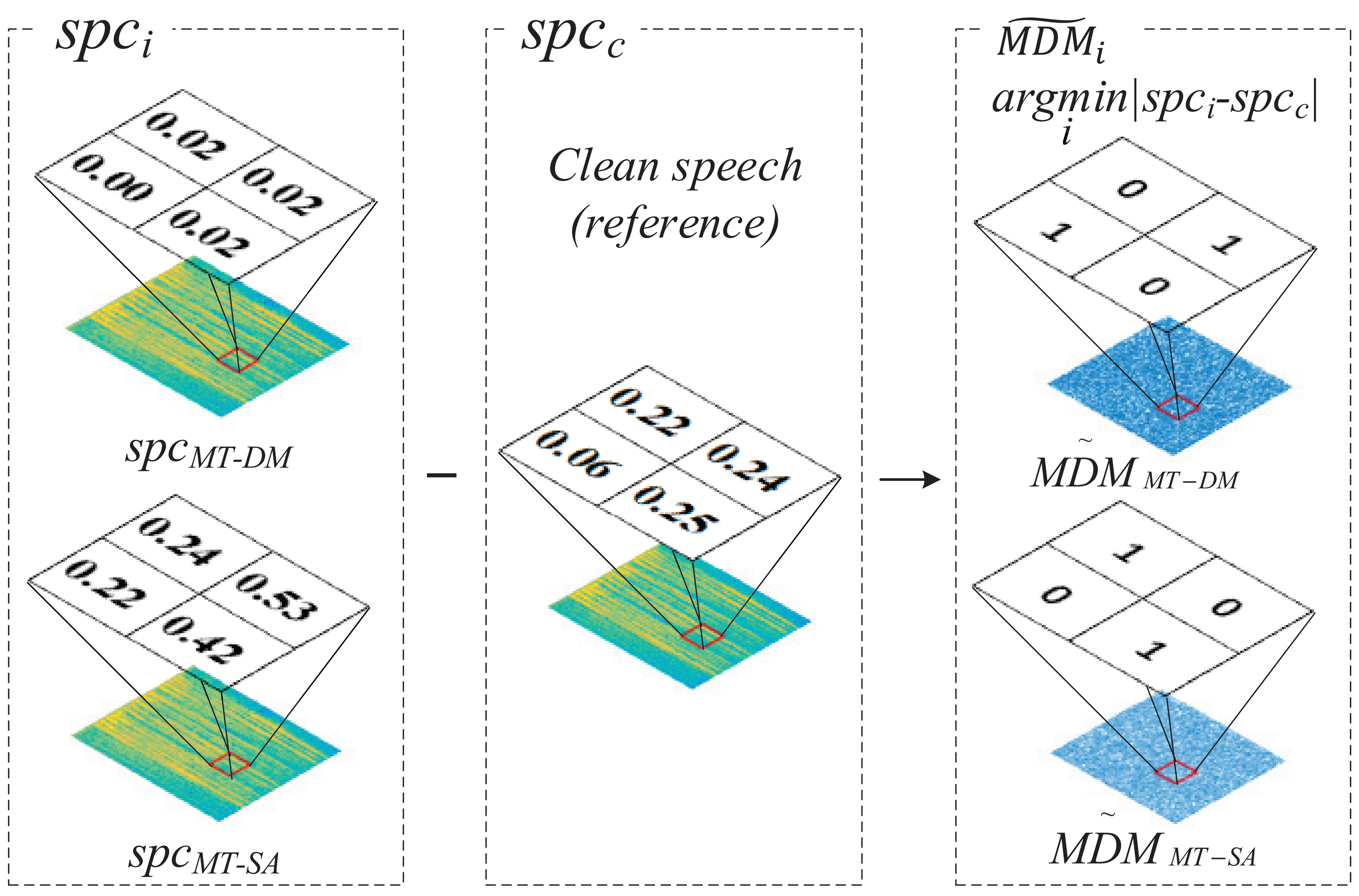
Background and Motivation

- Mapping and masking are two common learning targets used in speech dereverberation, and they have different effects in different scenarios.
- It is not suitable to use linear processing to deal with nonlinear, and the study of correlation between the mapping and masking is still insufficient.
- Many systems are now training according to the mean squared error (MSE) criterion, the MSE of spectrograms in different regions is different.

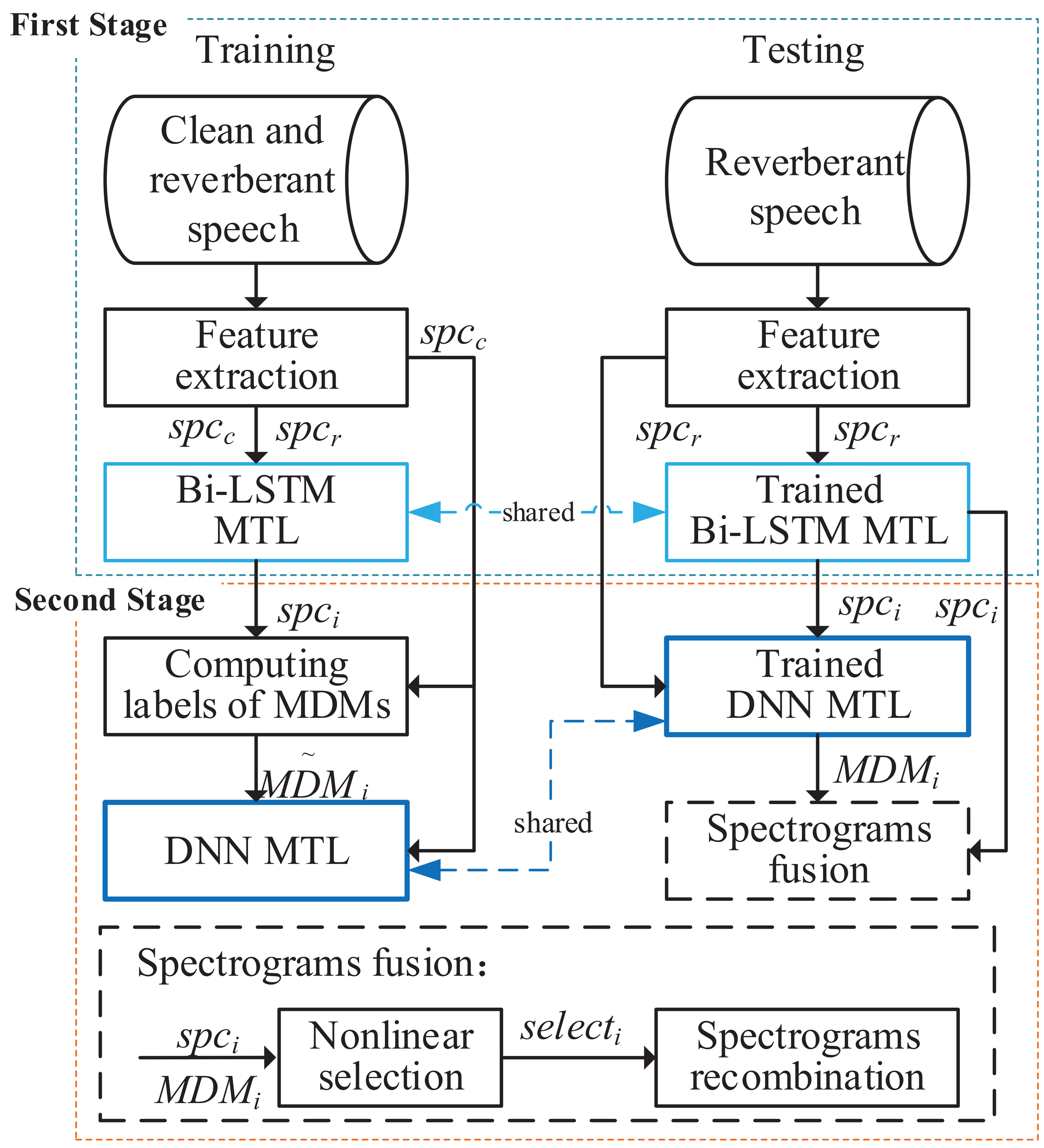
We propose in this paper:

- Design the minimum difference masks (MDMs): to classify T-F bins, which are nearest to the labels in spectrograms.
- Design a nonlinear spectrograms fusion system: to recombine spectrograms into one spectrogram.

2. Minimum Difference Masks Labels



3. Nonlinear Spectrograms Fusion



4. Experiments Results

Table 1. PESQ and SRMR results for simulated data.

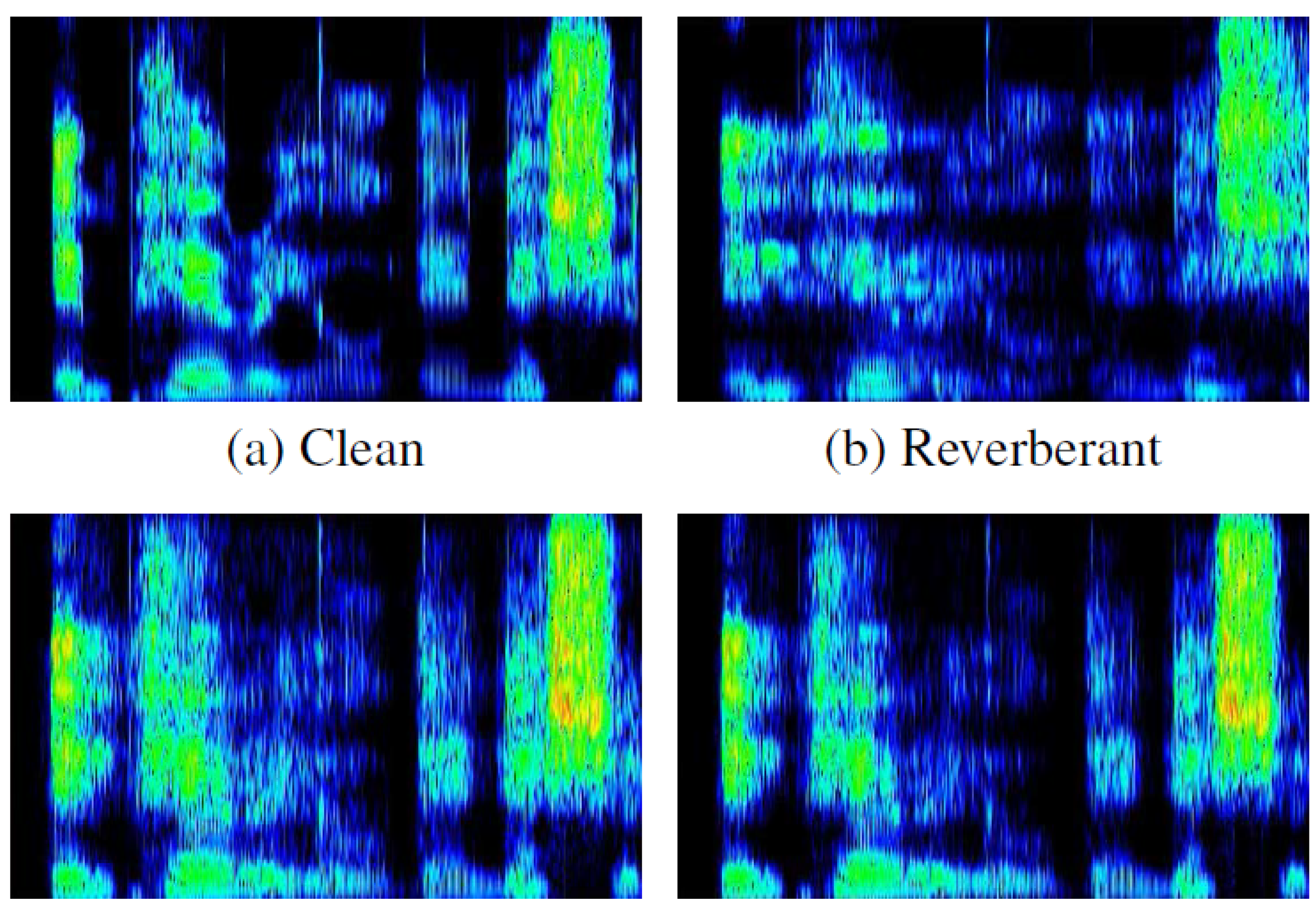
Models	PESQ Far	PESQ Near	PESQ Avg.	SRMR Far	SRMR Near	SRMR Avg.
Reverb	2.15	2.59	2.37	3.43	3.94	3.68
DM	2.58	2.88	2.73	4.39	4.88	4.64
SA	2.54	2.93	2.74	4.48	4.92	4.70
MT-DM	2.56	2.90	2.73	4.42	4.92	4.67
MT-SA	2.60	3.01	2.81	4.64	4.97	4.80
MT-LF	2.64	3.02	2.83	4.58	4.99	4.78
MDM-2O(B)	2.56	2.92	2.74	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	SRMR Near	SRMR 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**

- Real masks worked better than binary masks, indicating that soft masks are more suitable than hard masks.
- An active feature complimentary between spectrograms and MDMs.

5. Enhanced Spectrograms



(a) Clean (b) Reverberant (c) MT-LF (d) MDM-4O

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

6. Conclusions and Future Work

Conclusions We use spectrograms from the first stage and MDMs from the second stage to fuse the best parts of spectrograms. And this mainly improved both the speech quality and speech-to-reverberation modulation energy ratio.

Future Work 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.