TWO HEADS ARE BETTER THAN ONE: A TWO-STAGE APPROACH FOR MONAURAL NOISE REDUCTION IN THE COMPLEX DOMAIN–SUPPLEMENTAL MATERIAL

Andong Li[⋆†]*, Chengshi Zheng*[⋆†]*, Renhua Peng*[⋆†]*, Xiaodong Li*[⋆†]

[⋆] Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[†] University of Chinese Academy of Sciences, Beijing, China

In this supplemental material, we present more details about the implementation and advantages of our approach. In Sec. 1, we present the detailed network configurations of the proposed approach. In Sec. 2, we provide more evaluation results among different systems to reveal the superiority of our approach. In Sec. 3, we visualize the spectra of different systems under multiple noise conditions.

1. NETWORK CONFIGURATIONS

As illustrated in the paper, our system consists of two principal components, namely *Coarse Magnitude Estimation Network* called CME-Net and *Complex Spectrum Refine Network* called CSR-Net, and two stages are designed accordingly. In the first stage, CME-Net is tasked with magnitude estimation, which is then coupled with original noisy phase to obtain a coarse complex spectrum. In the second stage, as some residual noises still remain after the first stage, the second stage aims to further remove the residual noises and modify the phase information by estimating both real and imaginary parts of the spectrum.

The detailed network parameter configurations of CME-Net and CSR-Net are presented in Table 1 and Table 2, respectively. The input size is specified with *(Timestep* $\times$ *FeatureSize)* for 2-dimensional (2-D) format and *(ChannelNum* $\times$ *Timestep* $\times$ *FeatureSize)* for 3-D format. For both encoder and decoder, the hyperparameters are specified with *(KernelSize, Stride, ChannelNum)* format and *(KernelSize, DilationRate, ChannelNum)* for TCMs. From the tables, several observations can be made. Firstly, both two networks have similar structures except one decoder is deployed in the CME-Net and two decoders are in the CSR-Net. This is because in the first stage, only magnitude needs to be estimated while in the second stages, both real and imaginary parts of the spectrum need to be considered. Secondly, for the first network, the input is the magnitude of the noisy spectrum, so the channel number is 1. However, for the second stage, the network input consists of both original noisy and coarse complex spectrum, the channel number is 4. Thirdly, in both encoder and decoder, the conventional (de)convolutional blocks are replaced by its gated linear unit (GLU) version, which have been illustrated in [1] to exhibit better performance over the naive versions, so in this paper, we adopt the same strategy as [1].

2. MORE EVALUATION RESULTS

In this section, we present additional detailed results aobout different systems. As stated in the paper, about 20,000 environmental noises are randomly selected from the Deep Noise Suppression (DNS) Challenge. In this way, we can obtain a noise-robust speech enhancement system. To validate the effectiveness of our approach, we select two challenging unseen noises from NOISEX-92 dataset, namely babble and factory1. Please note that the reasons to choose the two noises are two-fold. Firstly, both noises are highly non-stationary and thus quite challenging for current speech enhancement systems to deal. Secondly, as our model is trained with multi-noise condition strategy, it is robust to various environmental noises and we only select two noises as the representative. To verify this point, we also provide the evaluation results on other noises from MUSAN [2], which are given in Sec. 2.2.

2.1. Performance Comparison on Other Evaluation Metrics

In the paper, three metrics are adopted for model comparison, namely PESQ, ESTOI, and SDR. To further verify the advantages of our approach over previous state-of-the-art systems, we also adopt another three metrics, namely SNR, log spectral distance

Table 1. Detailed parameter setup for CME-Net.

	layer name	input size	hyperparameters	output size
Encoder	convglu_1	$1 \times T \times 161$	$2 \times 5, (1,2), 64$	$64 \times T \times 79$
	convglu_2	$64 \times T \times 79$	$2 \times 3, (1,2), 64$	$64 \times T \times 39$
	convglu_3	$64 \times T \times 39$	$2 \times 3, (1,2), 64$	$64 \times T \times 19$
	convglu_4	$64 \times T \times 19$	$2 \times 3, (1,2), 64$	$64 \times T \times 9$
	convglu_5	$64 \times T \times 9$	$2 \times 3, (1,2), 64$	$64 \times T \times 4$
	reshape_size_2	$64 \times T \times 4$	-	$T \times 256$
	MG-TCMs	$T \times 256$	$\left.\begin{pmatrix} 1,1,64 \\ 5,\mathbf{1},64 \\ 1,1,256 \end{pmatrix} \begin{pmatrix} 1,1,64 \\ 5,\mathbf{2},64 \\ 1,1,256 \end{pmatrix} \begin{pmatrix} 1,1,64 \\ 5,\mathbf{4},64 \\ 1,1,256 \end{pmatrix} \begin{pmatrix} 1,1,64 \\ 5,\mathbf{8},64 \\ 1,1,256 \end{pmatrix} \begin{pmatrix} 1,1,64 \\ 5,\mathbf{16},64 \\ 1,1,256 \end{pmatrix} \begin{pmatrix} 1,1,64 \\ 5,\mathbf{32},64 \\ 1,1,256 \end{pmatrix} \right\} \times 3$	$T \times 256$
	reshape_size_3	$T \times 256$	-	$64 \times T \times 4$
Decoder	skip_1	$64 \times T \times 4$	-	$128 \times T \times 4$
	deconvglu_1	$128 \times T \times 4$	$2 \times 3, (1,2), 64$	$64 \times T \times 9$
	skip_2	$64 \times T \times 9$	-	$128 \times T \times 9$
	deconvglu_2	$128 \times T \times 9$	$2 \times 3, (1,2), 64$	$64 \times T \times 19$
	skip_3	$64 \times T \times 19$	-	$128 \times T \times 19$
	deconvglu_3	$128 \times T \times 19$	$2 \times 3, (1,2), 64$	$64 \times T \times 39$
	skip_4	$64 \times T \times 39$	-	$128 \times T \times 39$
	deconvglu_4	$128 \times T \times 39$	$2 \times 3, (1,2), 64$	$64 \times T \times 79$
	skip_5	$64 \times T \times 79$	-	$128 \times T \times 79$
	deconvglu_5	$128 \times T \times 79$	$2 \times 5, (1,2), 64$	$1 \times T \times 161$
	reshape_size_4	$1 \times T \times 161$	-	$T \times 161$
	linear_1	$T \times 161$	161	$T \times 161$

(LSD) [3], and phase distance (PD) [4]. LSD is to evaluate the spectral distance between clean and enhanced spectra, whose calculation process is given as:

$$\text{LSD}\left(S, \tilde{S}\right) = \frac{1}{L}\sum_{l=0}^{L-1}\left[\frac{2}{N}\sum_{m=1}^{M/2}\left(20log_{10}\left|S_{m,l}\right| - 20log_{10}\left|\tilde{S}_{m,l}\right|\right)\right]. \tag{1}$$

PD is proposed to evaluate the phase distance between clean and enhanced complex spectra, where each T-F bin is weighted by the clean spectral magnitude to emphasize the importance of different spectrum regions. The calculation formula is given as:

$$\text{PD}\left(S, \tilde{S}\right) = \sum_{m,l}\frac{\left|S_{m,l}\right|}{\sum_{m',l'}\left|S_{m',l'}\right|}\angle\left(S_{m,l}, \tilde{S}_{m,l}\right). \tag{2}$$

As better speech quality will bring the improvement of SNR and the reduction of LSD and PD, we adopt iSNR, rLSD, and rPD to evaluate the speech quality of different systems, which are given as:

$$\text{iSNR} = \text{SNR}\left(S, \tilde{S}\right) - \text{SNR}\left(S, X\right), \tag{3}$$

$$\text{rLSD} = \text{LSD}\left(S, X\right) - \text{LSD}\left(S, \tilde{S}\right), \tag{4}$$

$$\text{rPD} = \text{PD}\left(S, X\right) - \text{PD}\left(S, \tilde{S}\right), \tag{5}$$

Fig. 1 shows the curves of iSNR, rLSD, and rPD for different speech enhancement systems under different input SNRs. One can find that CTS-Net outperforms other systems by a large margin under different SNR and noise conditions, which reveals the

Table 2. Detailed parameter setup for CSR-Net.

	layer name	input size	hyperparameters	output size
Encoder	convglu_1	$4 \times T \times 161$	$2 \times 5, (1,2), 64$	$64 \times T \times 79$
	convglu_2	$64 \times T \times 79$	$2 \times 3, (1,2), 64$	$64 \times T \times 39$
	convglu_3	$64 \times T \times 39$	$2 \times 3, (1,2), 64$	$64 \times T \times 19$
	convglu_4	$64 \times T \times 19$	$2 \times 3, (1,2), 64$	$64 \times T \times 9$
	convglu_5	$64 \times T \times 9$	$2 \times 3, (1,2), 64$	$64 \times T \times 4$
	reshape_size_2	$64 \times T \times 4$	-	$T \times 256$
	DMG-TCMs	$T \times 256$	$\left.\begin{pmatrix}1,1,64\\5,\mathbf{1},64\\1,1,256\end{pmatrix}\begin{pmatrix}1,1,64\\5,\mathbf{2},64\\1,1,256\end{pmatrix}\begin{pmatrix}1,1,64\\5,\mathbf{4},64\\1,1,256\end{pmatrix}\begin{pmatrix}1,1,64\\5,\mathbf{8},64\\1,1,256\end{pmatrix}\begin{pmatrix}1,1,64\\5,\mathbf{16},64\\1,1,256\end{pmatrix}\begin{pmatrix}1,1,64\\5,\mathbf{32},64\\1,1,256\end{pmatrix}\right\} \times 2$	$T \times 256$
	reshape_size_3	$T \times 256$	-	$64 \times T \times 4$
Decoder ×2	skip_1	$64 \times T \times 4$	-	$128 \times T \times 4$
	deconvglu_1	$128 \times T \times 4$	$2 \times 3, (1,2), 64$	$64 \times T \times 9$
	skip_2	$64 \times T \times 9$	-	$128 \times T \times 9$
	deconvglu_2	$128 \times T \times 9$	$2 \times 3, (1,2), 64$	$64 \times T \times 19$
	skip_3	$64 \times T \times 19$	-	$128 \times T \times 19$
	deconvglu_3	$128 \times T \times 19$	$2 \times 3, (1,2), 64$	$64 \times T \times 39$
	skip_4	$64 \times T \times 39$	-	$128 \times T \times 39$
	deconvglu_4	$128 \times T \times 39$	$2 \times 3, (1,2), 64$	$64 \times T \times 79$
	skip_5	$64 \times T \times 79$	-	$128 \times T \times 79$
	deconvglu_5	$128 \times T \times 79$	$2 \times 5, (1,2), 64$	$1 \times T \times 161$
	reshape_size_4	$1 \times T \times 161$	-	$T \times 161$
	linear_1	$T \times 161$	161	$T \times 161$

superiority of our two-stage paradigm in noise suppression, spectrum recovery, and phase modification. In addition, going from CME-Net to CTS-Net, one can observe notable improvements in all the three metrics, which means that when the magnitude is pre-estimated in the first stage, it serves as the prior information and facilitates the better recovery of speech in the second stage. It also reveals the necessity and significance of introducing a two-stage paradigm in the speech enhancement task. Note that CRN and CME-Net only aim to recover the magnitude spectrum, and the phase information remains unchanged. So rPD is always zero for different cases.

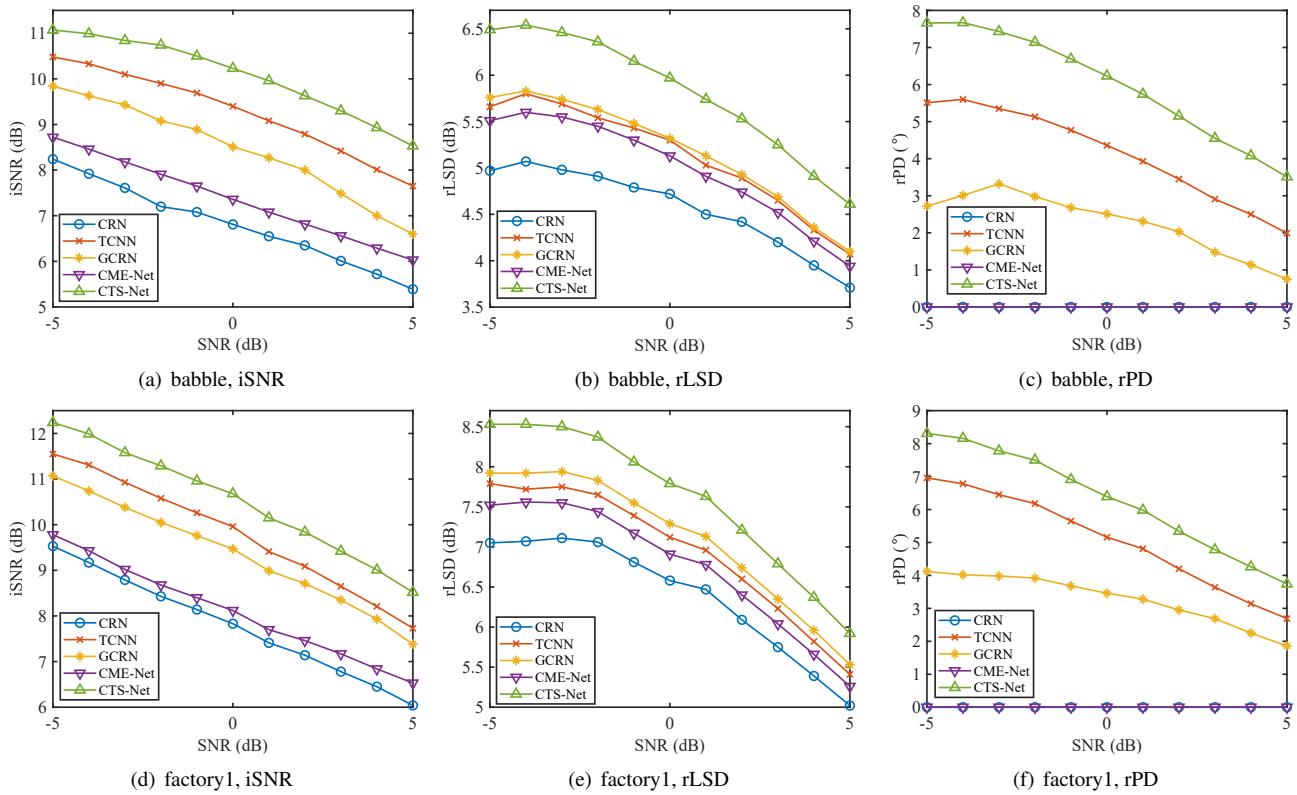
2.2. Performance Comparison on Other Noise Set

Apart from the two challenging noises in the paper, we also provide quantitative results in other noise set. Here 50 noises are selected from MUSAN noise set[1]. Similar noise generation strategy as the paper is adopted to obtain 150 noisy-clean utterance pairs for each SNR case. Three SNRs are considered, namely -5dB, 0dB, and 5dB, which is accordant with the paper. Table 3 presents the objective results of different systems under MUSAN noise set. **BOLD** denotes the best result in each case. From the result, one can observe that for various noises, our proposed CTS-Net still yields notable performance improvements over other state-of-the-art models in different cases. For example, going from GCRN to CTS-Net, around 0.18 and 0.19 PESQ improvements are achieved for seen and unseen speaker cases. When it terms to ESTOI, the improvements are 3.56% and 4.38%, respectively.

2.3. Performance of The Estimated Phase

In the paper, both magnitude and phase are refined in the second stage. To analyze the effect of the estimated phase, after the magnitude is estimated, we synthesize the speech with original noisy, estimated, and clean phase information, respectively. The results are shown in Tabel 4. One can find that when the phase information is obtained by the network, both PESQ and ESTOI

[1]As there are a large number of noises in the MUSAN, only part of them are selected, whose file names are from *noise-free-sound-000.wav* to *noise-free-sound-049.wav*



(a) babble, iSNR (b) babble, rLSD (c) babble, rPD

(d) factory1, iSNR (e) factory1, rLSD (f) factory1, rPD

Fig. 1. iSNR, rLSD, rPD for different input SNRs. Input SNR ranges from -5dB to 5dB. (a): iSNR for different systems under babble noise. (b): rLSD for different systems under babble noise. (c): rPD for different systems under babble noise. (d): iSNR for different systems under factory1 noise. (e): rLSD for different systems under factory1 noise. (f): rPD for different systems under factory1 noise.

can be notably improved over that of the noisy phase. It shows that in the second stage of the network, phase information can be effectively recovered. In addition, we also notice that when compared with the results of oracle phase, there still exist some performance gap. Therefore, a better approach to recover the phase information needs to be explored in the future.

3. SPECTRAL VISUALIZATION

We also provide the spectral visualization of different systems, as shown in Fig. 2-Fig. 7. Three SNRs are selected, namely -5dB, 0dB, and 5dB. For each SNR case, two noises are adopted, namely babble and factory1. To better analyze the effect of the two-stage paradigm in our approach, we also present the learned residual components in the second stage[2]. From the figures, one can find that compared with previous systems, our approach can better suppress the noise components in both non-speech and speech regions. Meanwhile, the formants also become more clear. From the visualization of learned residual, we also find that some spectral components are estimated, indicating that the network indeed captures the spectral details in the second stage, which is in consistent with our expectation.

[2]As stated in the paper, instead of directly estimating the complex spectrum in the second stage, we introduce a global residual connection, *i.e.*, only complex spectral residual is estimated, which is subsequently added to the previous coarse spectrum to obtain the final result.

Table 3. Objective result comparison among different models under MUSAN noise set in terms of PESQ, ESTOI and SDR for both seen and unseen speaker cases. **BOLD** indicats the best score in each case.

Metrics	**Causality**	**PESQ**								**ESTOI(%)**								**SDR(dB)**							
Speaker		**Seen**				**Unseen**				**Seen**				**Unseen**				**Seen**				**Unseen**			
SNR(dB)		-5	0	5	Avg.	-5	0	5	Avg.	-5	0	5	Avg.	-5	0	5	Avg.	-5	0	5	Avg.	-5	0	5	Avg.
Noisy	-	1.78	2.16	2.45	2.13	1.77	2.06	2.45	2.09	47.78	61.16	73.37	60.77	46.03	57.57	71.15	58.25	-4.90	0.05	5.02	0.06	-4.91	0.04	5.03	0.05
CRN	✓	2.51	2.91	3.20	2.88	2.50	2.83	3.19	2.84	67.93	78.88	86.05	77.62	65.68	76.68	85.50	75.96	7.45	10.74	14.02	10.74	7.80	10.90	14.35	11.02
TCNN	✓	2.56	2.90	3.16	2.88	2.52	2.84	3.14	2.84	71.52	81.07	86.68	79.76	69.65	79.70	86.43	78.59	9.71	12.38	15.46	12.52	9.51	13.66	15.59	12.92
GCRN	✓	2.67	3.04	3.30	3.00	2.66	2.98	3.30	2.98	71.62	81.44	87.18	80.08	70.63	80.42	87.38	79.48	9.33	12.44	15.03	12.27	9.85	12.55	15.52	12.64
CME-Net(Pro.)	✓	2.60	2.99	3.27	2.95	2.59	2.91	3.25	2.92	70.02	80.06	86.53	78.87	68.27	78.27	86.24	77.59	7.70	11.04	14.48	11.07	8.20	11.29	14.85	11.44
CTS-Net(Pro.)	✓	**2.88**	**3.22**	**3.45**	**3.18**	**2.86**	**3.17**	**3.47**	**3.17**	**76.78**	**84.78**	**89.37**	**83.64**	**75.60**	**83.69**	**89.60**	**82.97**	**11.02**	**14.02**	**16.80**	**13.95**	**11.40**	**14.11**	**17.15**	**14.22**

Table 4. Objective results of CTS-Net when the estimated phases are replaced by noisy, estimated, and clean phases.

Metrics	**PESQ**									**ESTOI(%)**								
SNR(dB)	**-5**			**0**			**5**			**-5**			**0**			**5**		
noises	babble	factory1	Avg.	babble	factory1	Avg.	babble	factory1	Avg.	babble	factory1	Avg.	babble	factory1	Avg.	babble	factory1	Avg.
noisy_phase	2.05	2.18	2.12	2.57	2.65	2.61	2.97	3.02	2.99	55.39	54.95	55.17	70.84	70.83	70.83	81.10	80.88	80.99
esti_phase	2.18	2.31	2.25	2.75	2.81	2.78	3.13	3.14	3.14	60.10	59.82	59.96	75.37	75.05	75.21	84.12	83.37	83.74
clean_phase	2.42	2.53	2.47	2.97	3.03	3.00	3.35	3.36	3.36	66.51	66.11	66.31	79.82	79.33	79.57	87.38	86.50	86.94

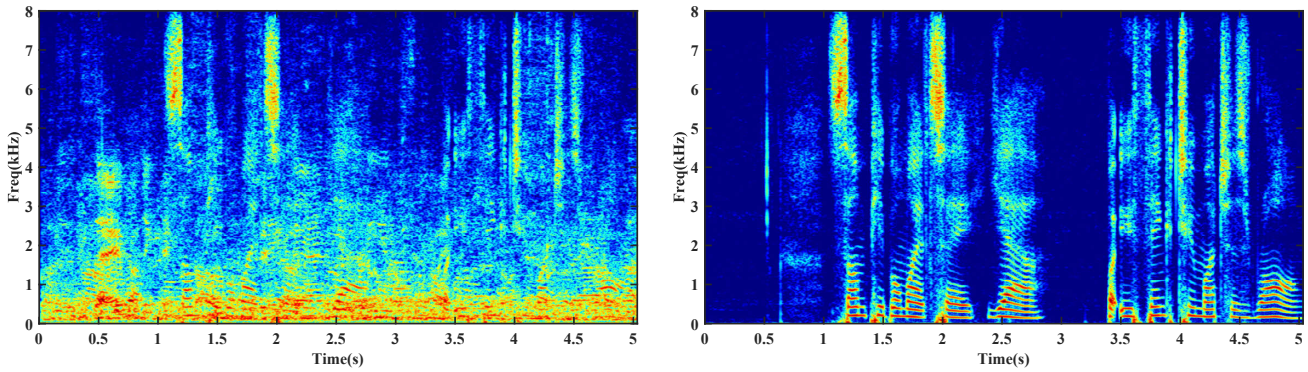
4. REFERENCES

[1] K Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[2] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, *arXiv: 1510.08484v1*.

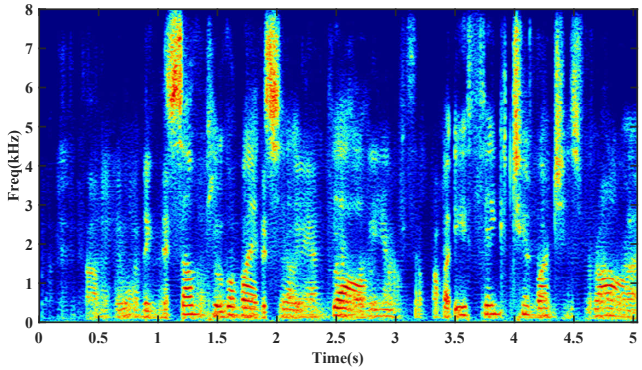
[3] J. Benesty, M. Sondhi, and Y. Huang, *Springer handbook of speech processing*, Spring, 2007.

[4] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," 2019, *arxiv: 1903.03107*.

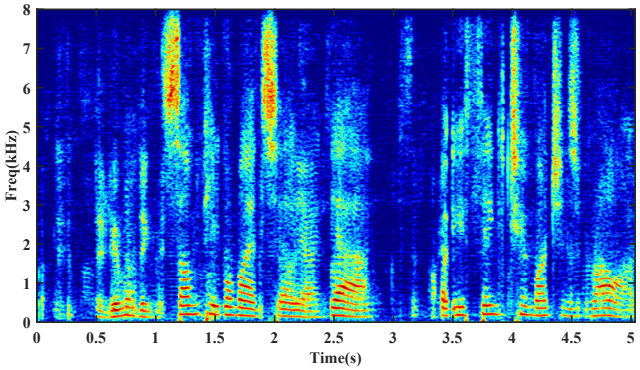


(a) -5dB babble noisy

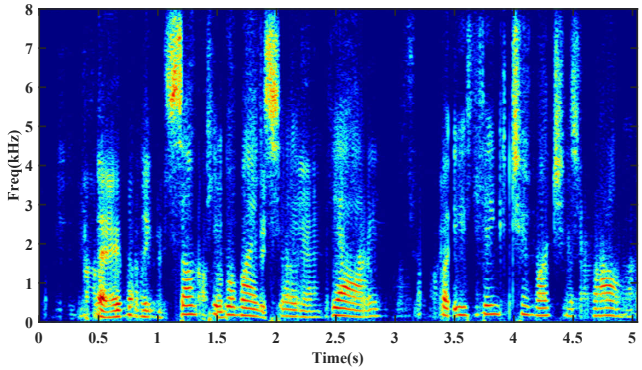
(b) clean



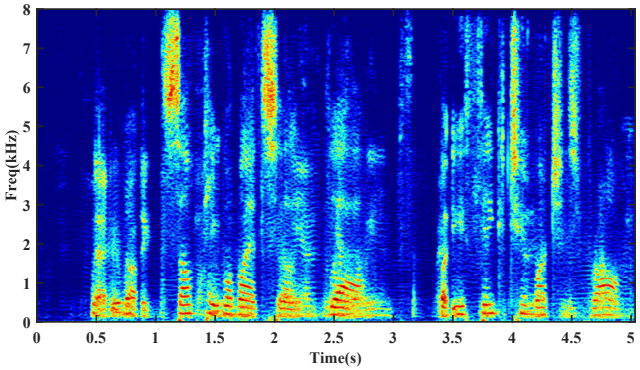
(c) CRN



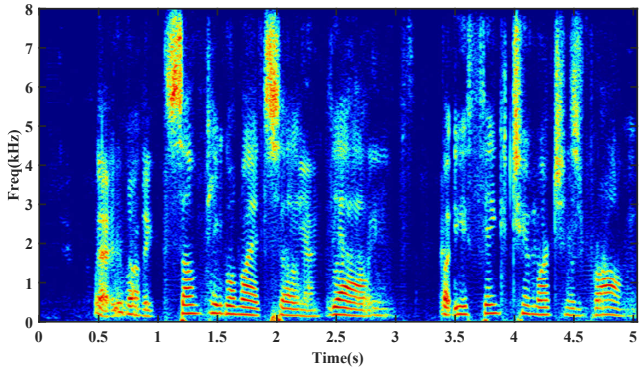
(d) GCRN



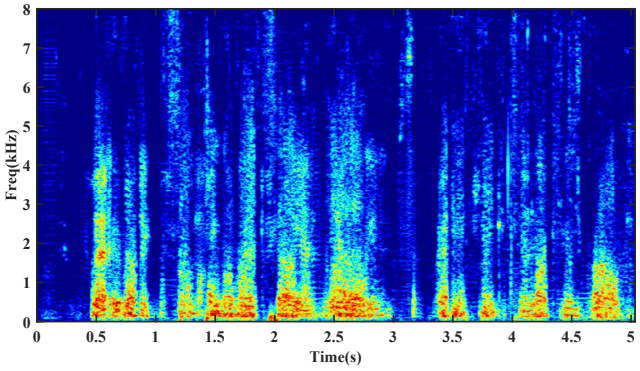
(e) TCNN



(f) CME-Net

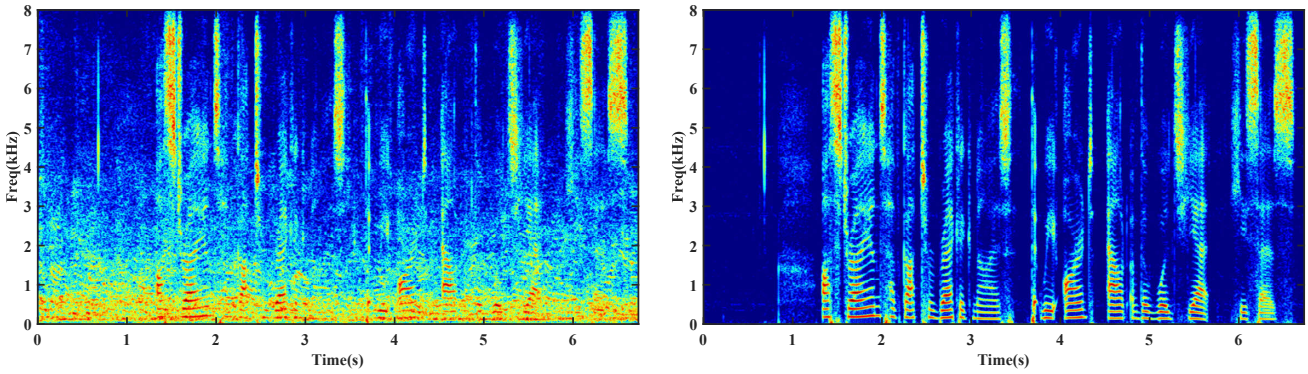


(g) CTS-Net



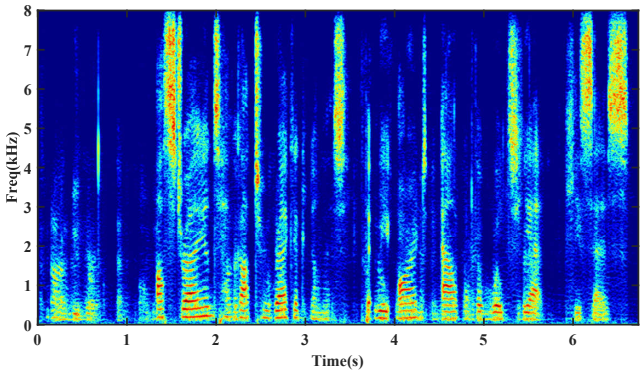
(h) Residual in the second stage

Fig. 2. spectrum visualization for different systems. Utterance file name is *01bc020b.wav*. (a): noisy speech under -5dB babble noise, PESQ=1.62, ESTOI=29.34%. (b): clean. (c): estimated by CRN, PESQ=1.88, ESTOI=44.30%. (d): estimated by GCRN, PESQ=2.16, ESTOI=57.89%. (e): estimated by TCNN, PESQ=2.00, ESTOI=53.43%. (f): estimated by CME-Net, PESQ=1.90, ESTOI=52.20%. (g): estimated by CTS-Net, PESQ=2.28, ESTOI=60.19%. (h): learned residual in the proposed second stage.

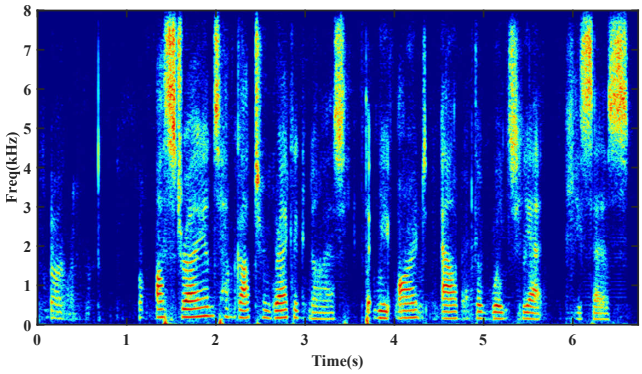


(a) 0dB babble noisy

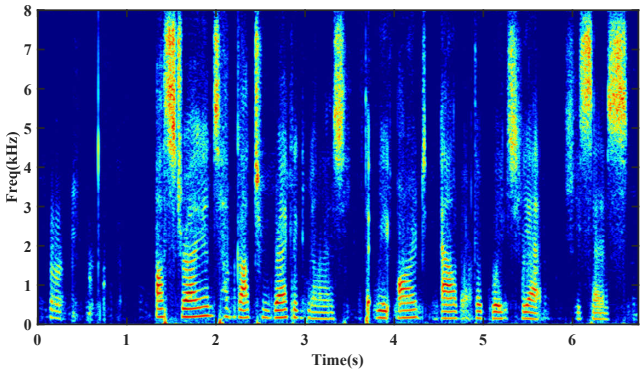
(b) clean



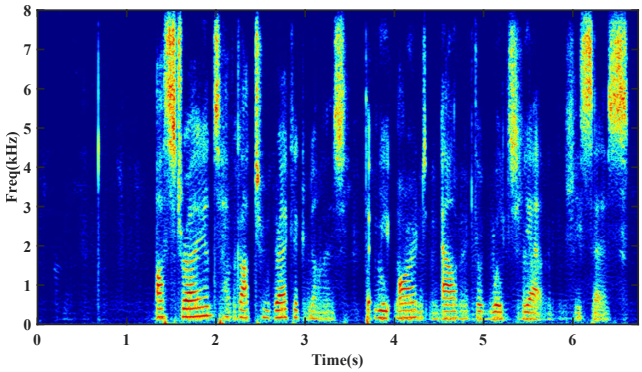
(c) CRN



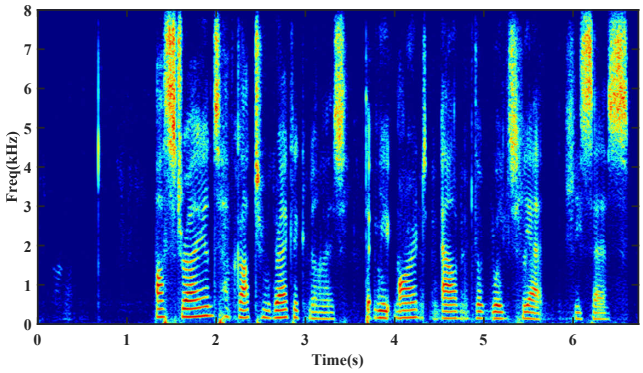
(d) GCRN



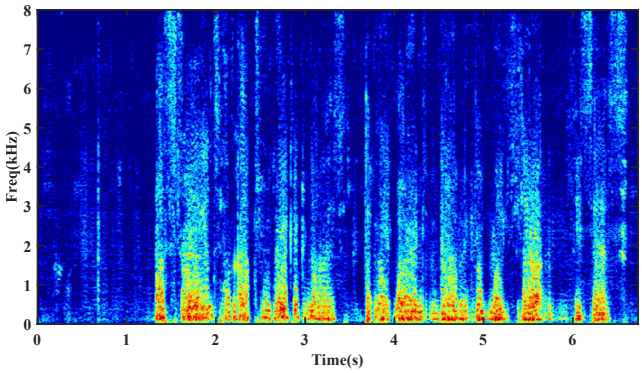
(e) TCNN



(f) CME-Net

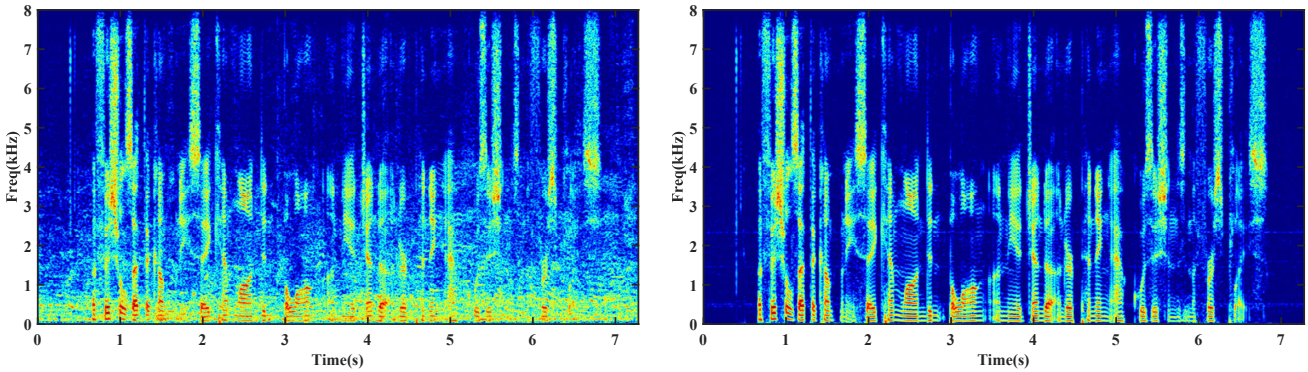


(g) CTS-Net



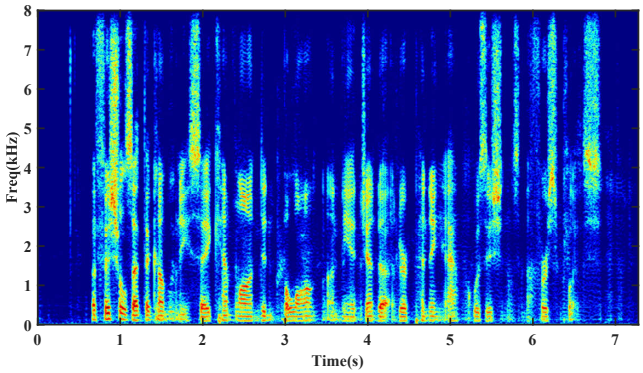
(h) Residual in the second stage

Fig. 3. spectrum visualization for different systems. Utterance file name is *01bc020k.wav*. (a): noisy speech under 0dB babble noise, PESQ=1.88, ESTOI=39.03%. (b): clean. (c): estimated by CRN, PESQ=2.33, ESTOI=65.78%. (d): estimated by GCRN, PESQ=2.42, ESTOI=71.83%. (e): estimated by TCNN, PESQ=2.42, ESTOI=71.53%. (f): estimated by CME-Net, PESQ=2.40, ESTOI=65.41%. (g): estimated by CTS-Net, PESQ=2.66, ESTOI=73.60%. (h): learned residual in the proposed second stage.

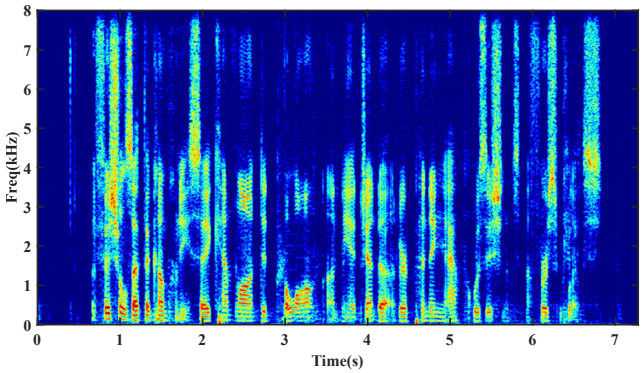


(a) 5dB babble noisy

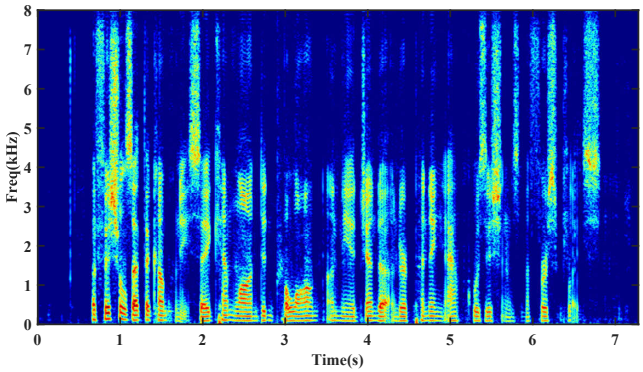
(b) clean



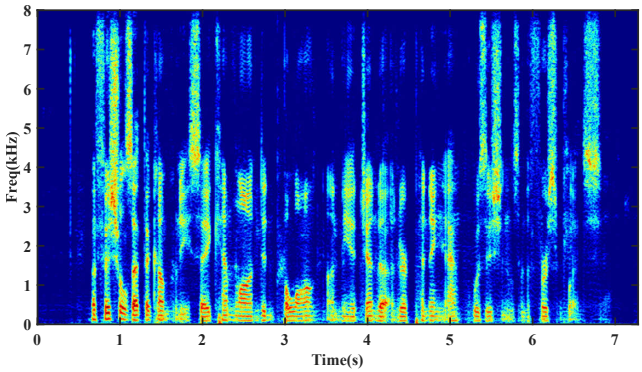
(c) CRN



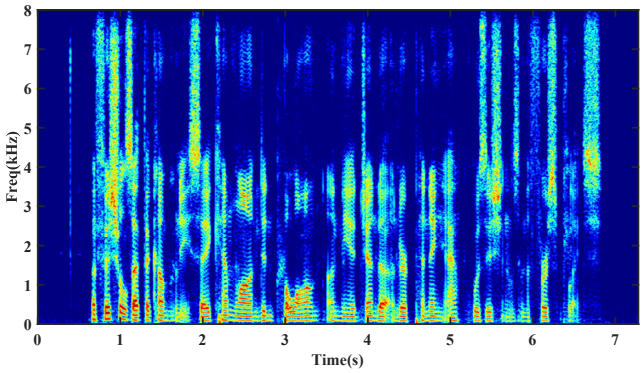
(d) GCRN



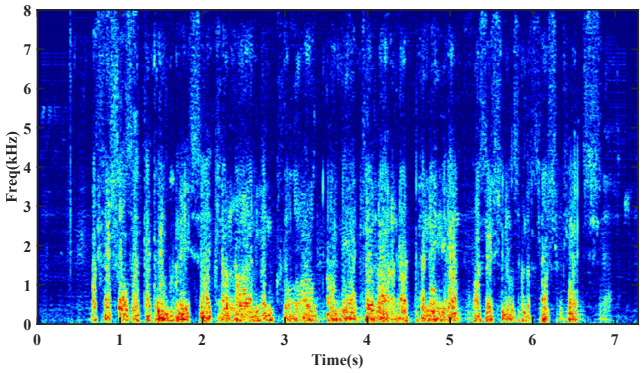
(e) TCNN



(f) CME-Net

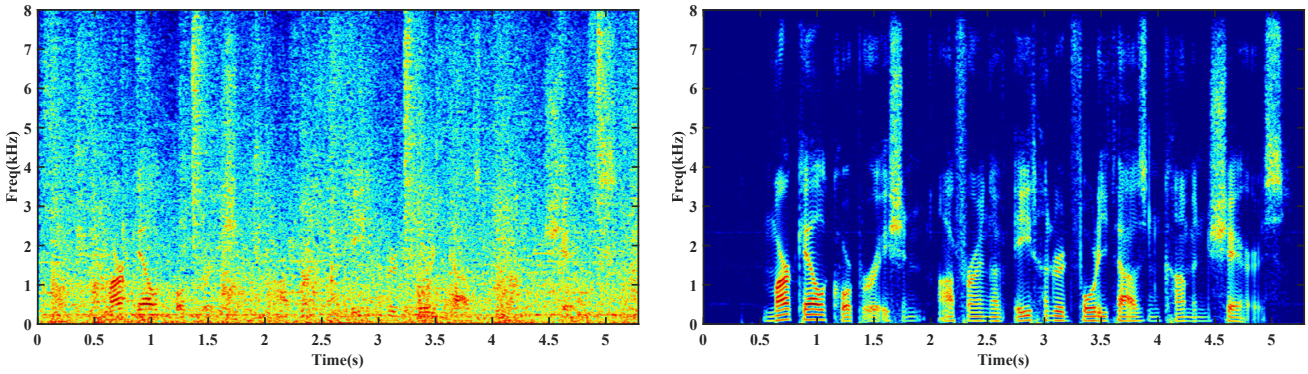


(g) CTS-Net



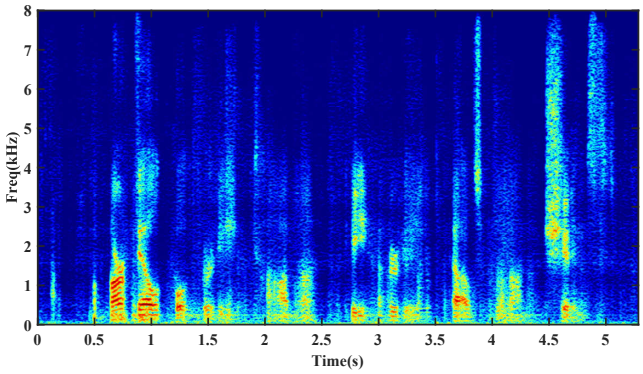
(h) Residual in the second stage

Fig. 4. spectrum visualization for different systems. Utterance file name is *01ec020g.wav*. (a): noisy speech under 5dB babble noise, PESQ=2.17, ESTOI=53.96%. (b): clean. (c): estimated by CRN, PESQ=2.84, ESTOI=70.02%. (d): estimated by GCRN, PESQ=2.80, ESTOI=70.75%. (e): estimated by TCNN, PESQ=2.93, ESTOI=78.03%. (f): estimated by CME-Net, PESQ=2.93, ESTOI=73.98%. (g): estimated by CTS-Net, PESQ=3.09, ESTOI=82.01%. (h): learned residual in the proposed second stage.

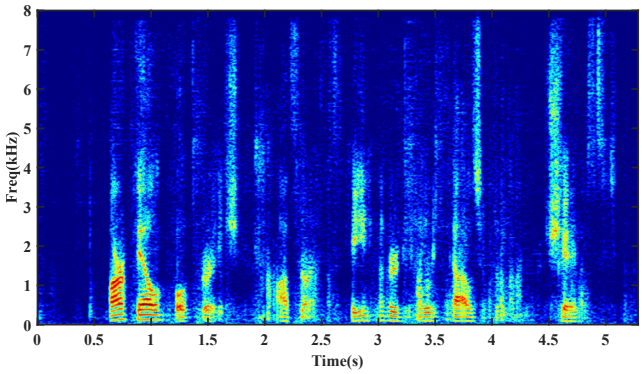


(a) -5dB factory1 noisy

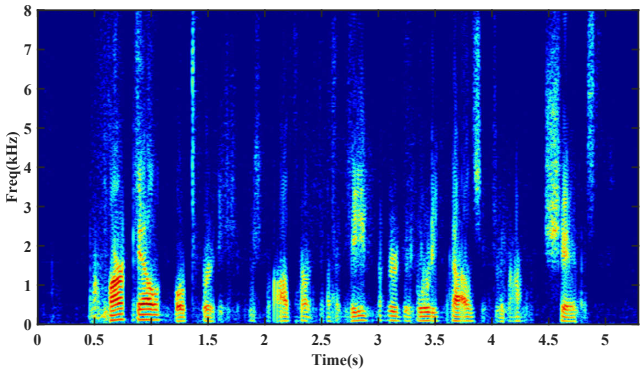
(b) clean



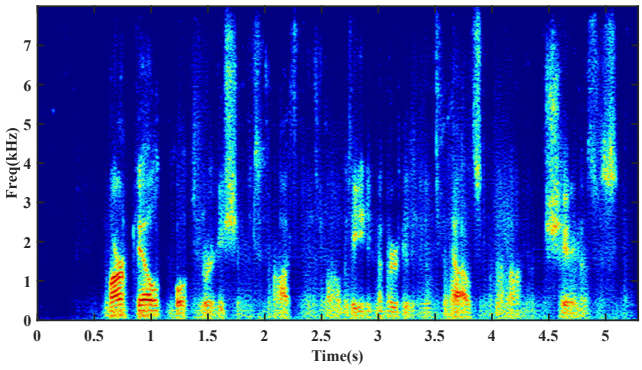
(c) CRN



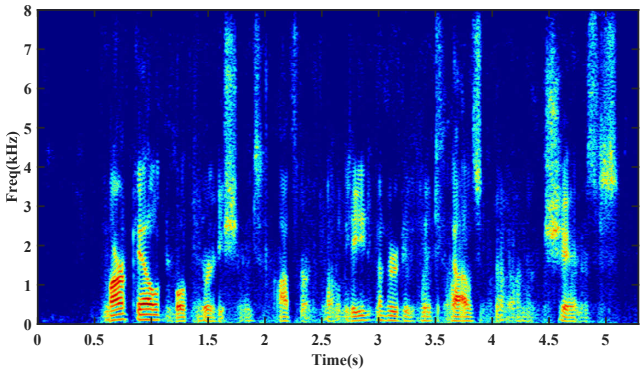
(d) GCRN



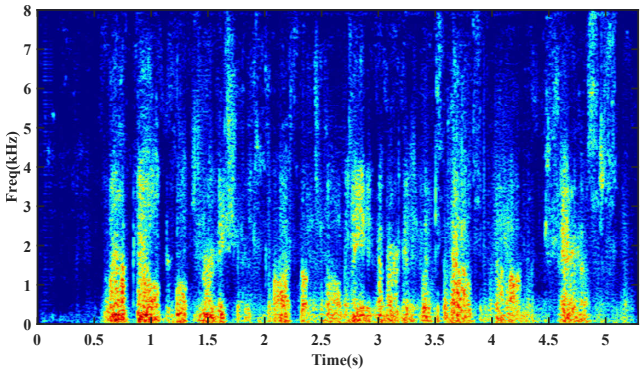
(e) TCNN



(f) CME-Net

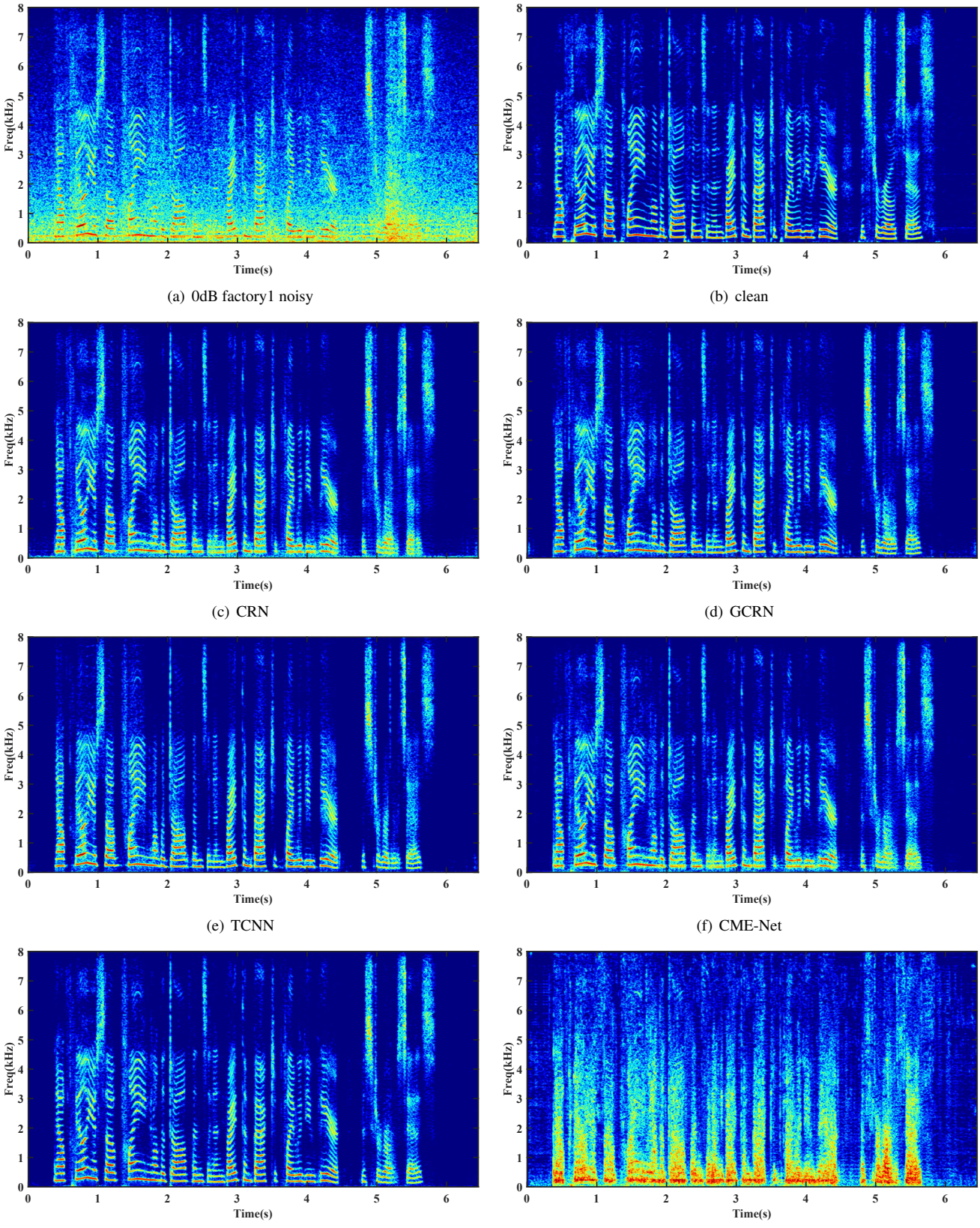


(g) CTS-Net



(h) Residual in the second stage

Fig. 5. spectrum visualization for different systems. Utterance file name is *01ec020p.wav*. (a): noisy speech under -5dB factory1 noise, PESQ=1.24, ESTOI=27.16%. (b): clean. (c): estimated by CRN, PESQ=1.78, ESTOI=35.01%. (d): estimated by GCRN, PESQ=2.06, ESTOI=46.40%. (e): estimated by TCNN, PESQ=2.00, ESTOI=48.55%. (f): estimated by CME-Net, PESQ=2.00, ESTOI=41.51%. (g): estimated by CTS-Net, PESQ=2.04, ESTOI=48.65%. (h): learned residual in the proposed second stage.



(a) 0dB factory1 noisy

(b) clean

(c) CRN

(d) GCRN

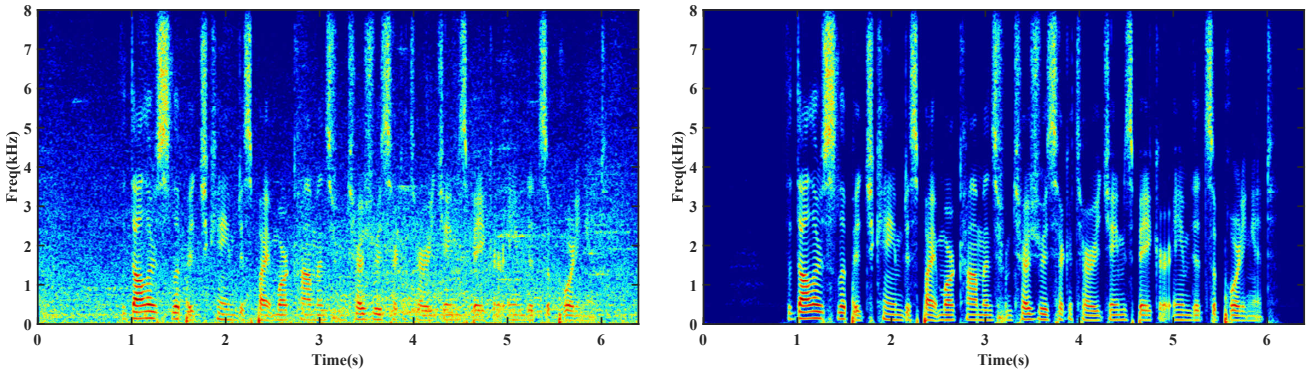
(e) TCNN

(f) CME-Net

(g) CTS-Net

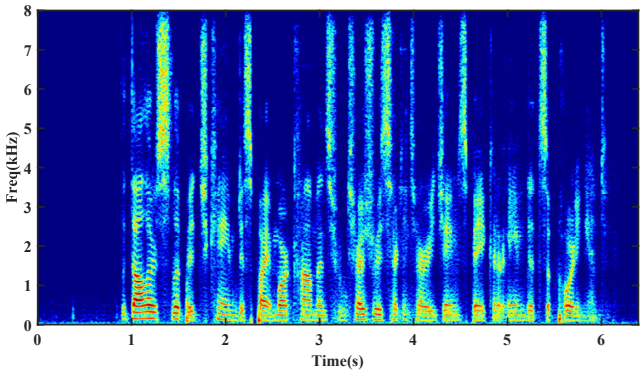
(h) Residual in the second stage

Fig. 6. spectrum visualization for different systems. Utterance file name is *01fc020m.wav*. (a): noisy speech under 0dB factory1 noise, PESQ=1.72, ESTOI=48.10%. (b): clean. (c): estimated by CRN, PESQ=2.48, ESTOI=72.98%. (d): estimated by GCRN, PESQ=2.74, ESTOI=80.25%. (e): estimated by TCNN, PESQ=2.46, ESTOI=76.97%. (f): estimated by CME-Net, PESQ=2.50, ESTOI=73.63%. (g): estimated by CTS-Net, PESQ=2.94, ESTOI=83.12%. (h): learned residual in the proposed second stage.

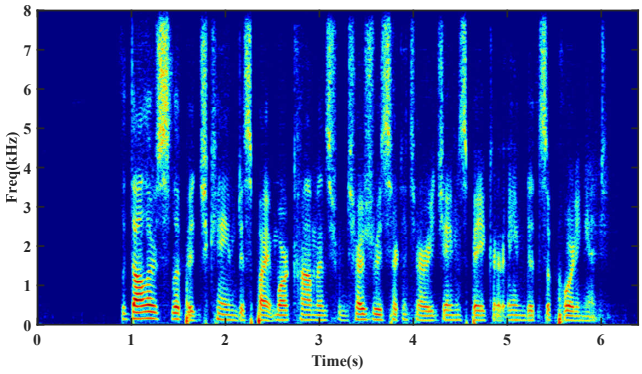


(a) 5dB factory1 noisy

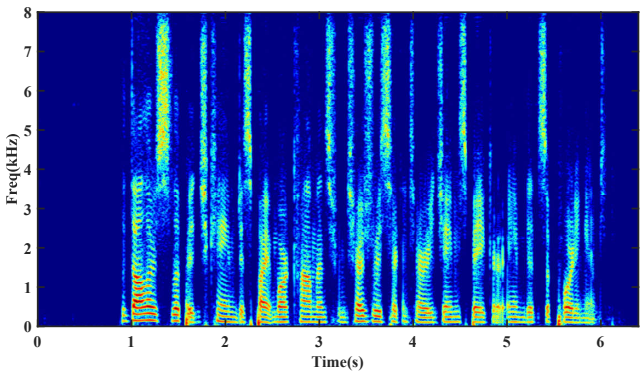
(b) clean



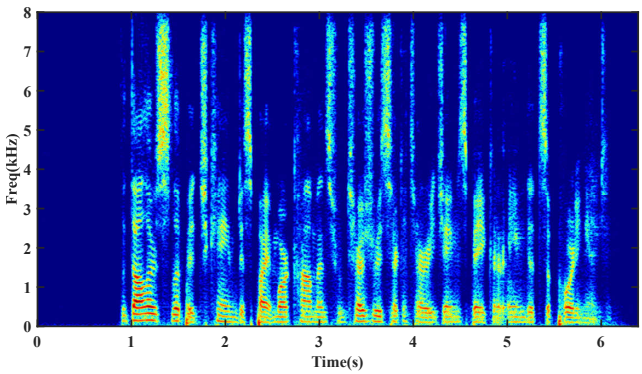
(c) CRN



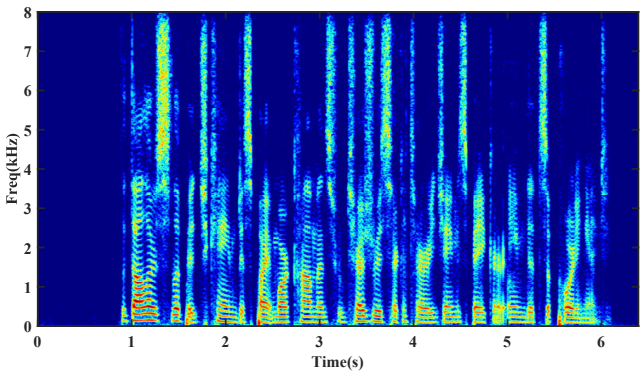
(d) GCRN



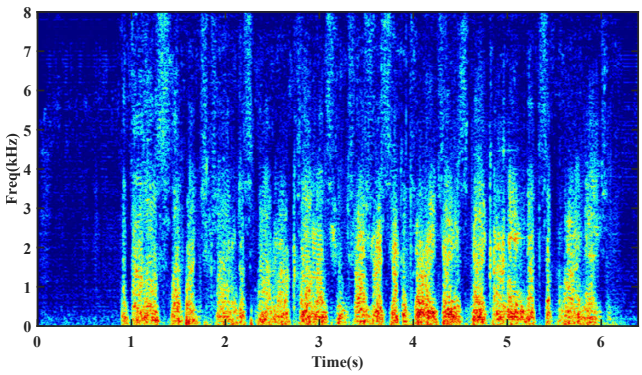
(e) TCNN



(f) CME-Net



(g) CTS-Net



(h) Residual in the second stage

Fig. 7. spectrum visualization for different systems. Utterance file name is *01gc020c.wav*. (a): noisy speech under 5dB factory1 noise, PESQ=2.11, ESTOI=67.44%. (b): clean. (c): estimated by CRN, PESQ=2.88, ESTOI=83.82%. (d): estimated by GCRN, PESQ=3.05, ESTOI=86.75%. (e): estimated by TCNN, PESQ=3.00, ESTOI=87.66%. (f): estimated by CME-Net, PESQ=2.89, ESTOI=84.30%. (g): estimated by CTS-Net, PESQ=3.16, ESTOI=88.42%. (h): learned residual in the proposed second stage.