icassp 2022 Singapore China Virtual

Paper #4816

A Training Framework for Stereo-aware Speech Enhancement using Deep Neural Networks

Bahareh Tolooshams and Kazuhito Koishida

Harvard John A. Paulson School of Engineering and Applied Sciences

Microsoft

Summary

Prior Work:

- Mainly focus on speech enhancement when using spatial information.
- Preservation of spatial images such as sensations of depth is barely studied.
- Lack of subjective tests for perceptual evaluations.

Goal:

- Preserve stereo image while performing speech enhancement.
- Provide both objective and subjective evaluation of perceptual improvement.

Propose:

- Quantify stereo aspects of the speech.
- Regularize during training to preserve the stereo image.

Stereo-aware Training

$$\mathcal{L}(\mathbf{s},\hat{\mathbf{s}}) = \mathcal{L}_{\text{speech-rec}}(\mathbf{s},\hat{\mathbf{s}}) + \mathcal{L}_{\text{image-pres}}(\mathbf{s},\hat{\mathbf{s}})$$

$$\mathrm{LSD}(\mathbf{s},\hat{\mathbf{s}}) + \alpha_{\mathrm{TL}}\,\mathrm{TL}(\mathbf{s},\hat{\mathbf{s}}) \qquad \sum_{\mathrm{M}\in\{\mathrm{IID},\mathrm{IPD},\mathrm{IC},\mathrm{OPD}\}} \alpha_{\mathrm{M}}\mathcal{L}_{\mathrm{M}}(\mathbf{S},\hat{\mathbf{S}})$$

Intensity

$$\mathrm{IID}_b(\mathbf{S}) = 10\log_{10}\frac{\sum_{f=f_b}^{f_{b+1}-1}\mathbf{S}_1[f]\mathbf{S}_1^*[f]}{\sum_{f=f_b}^{f_{b+1}-1}\mathbf{S}_2[f]\mathbf{S}_2^*[f]}$$

Phase

$$\mathrm{IPD}_b(\mathbf{S}) = \angle\left(\sum_{f=f_b}^{f_{b+1}-1}\mathbf{S}_1[f]\mathbf{S}_2^*[f]\right)$$

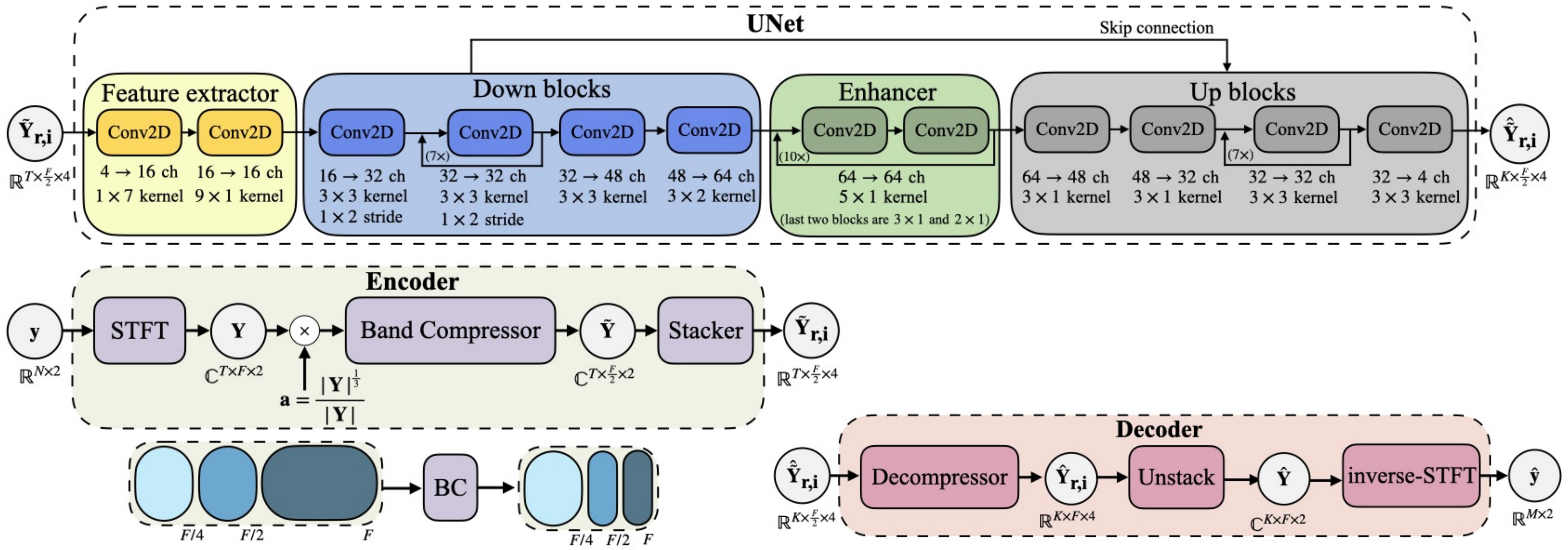
Coherence

$$\mathrm{IC}_b(\mathbf{S}) = \frac{\left|\sum_{f=f_b}^{f_{b+1}-1}\mathbf{S}_1[f]\mathbf{S}_2^*[f]\right|}{\sqrt{\left(\sum_{f=f_b}^{f_{b+1}-1}\mathbf{S}_1[f]\mathbf{S}_1^*[f]\right)\left(\sum_{f=f_b}^{f_{b+1}-1}\mathbf{S}_2[f]\mathbf{S}_2^*[f]\right)}}$$

Overall phase

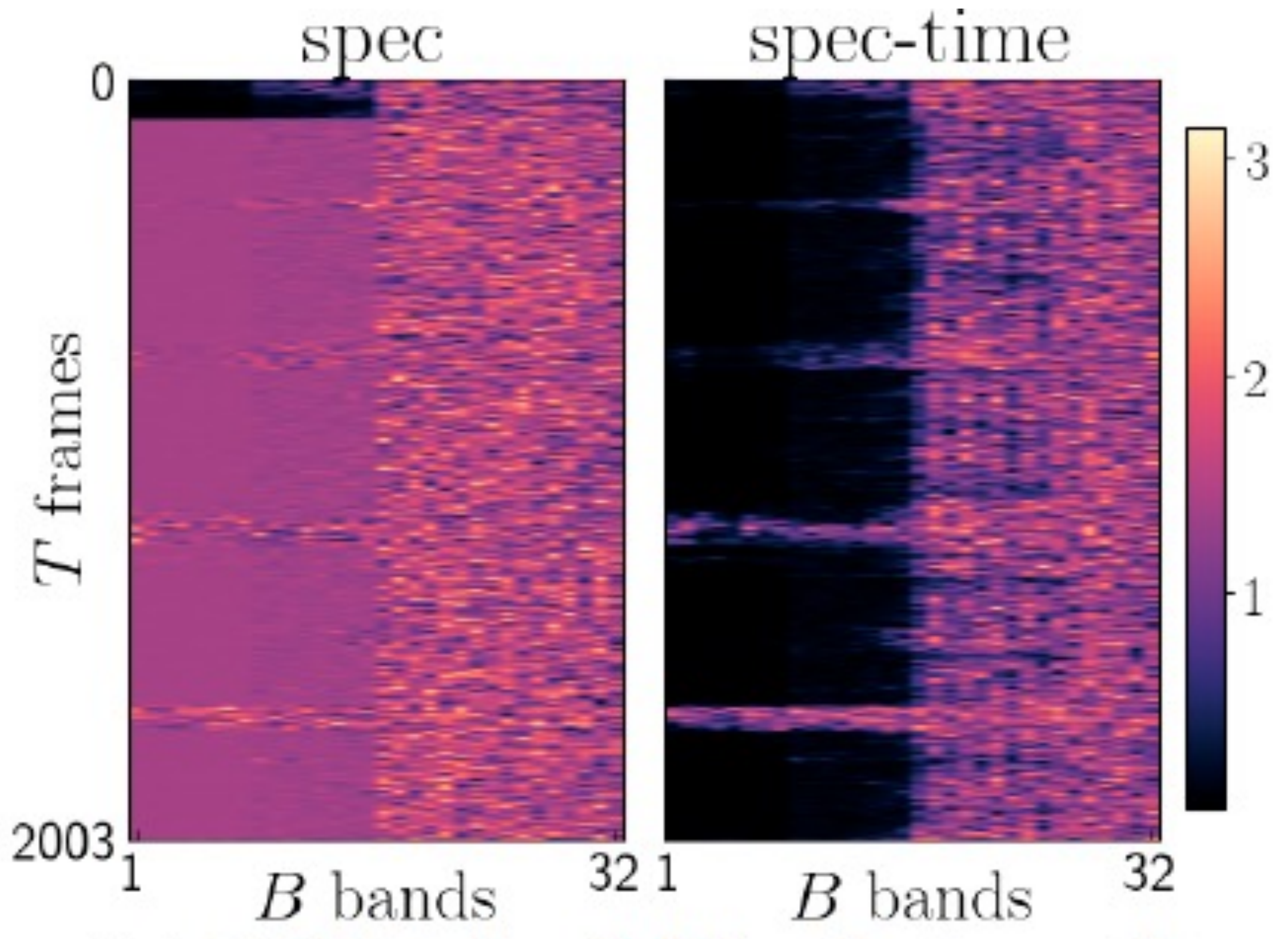
$$\mathrm{OPD}_b(\mathbf{S},\hat{\mathbf{S}}) = \angle\left(\sum_{f=f_b}^{f_{b+1}-1}\mathbf{S}[f]\hat{\mathbf{S}}^*[f]\right)$$

Network Architecture

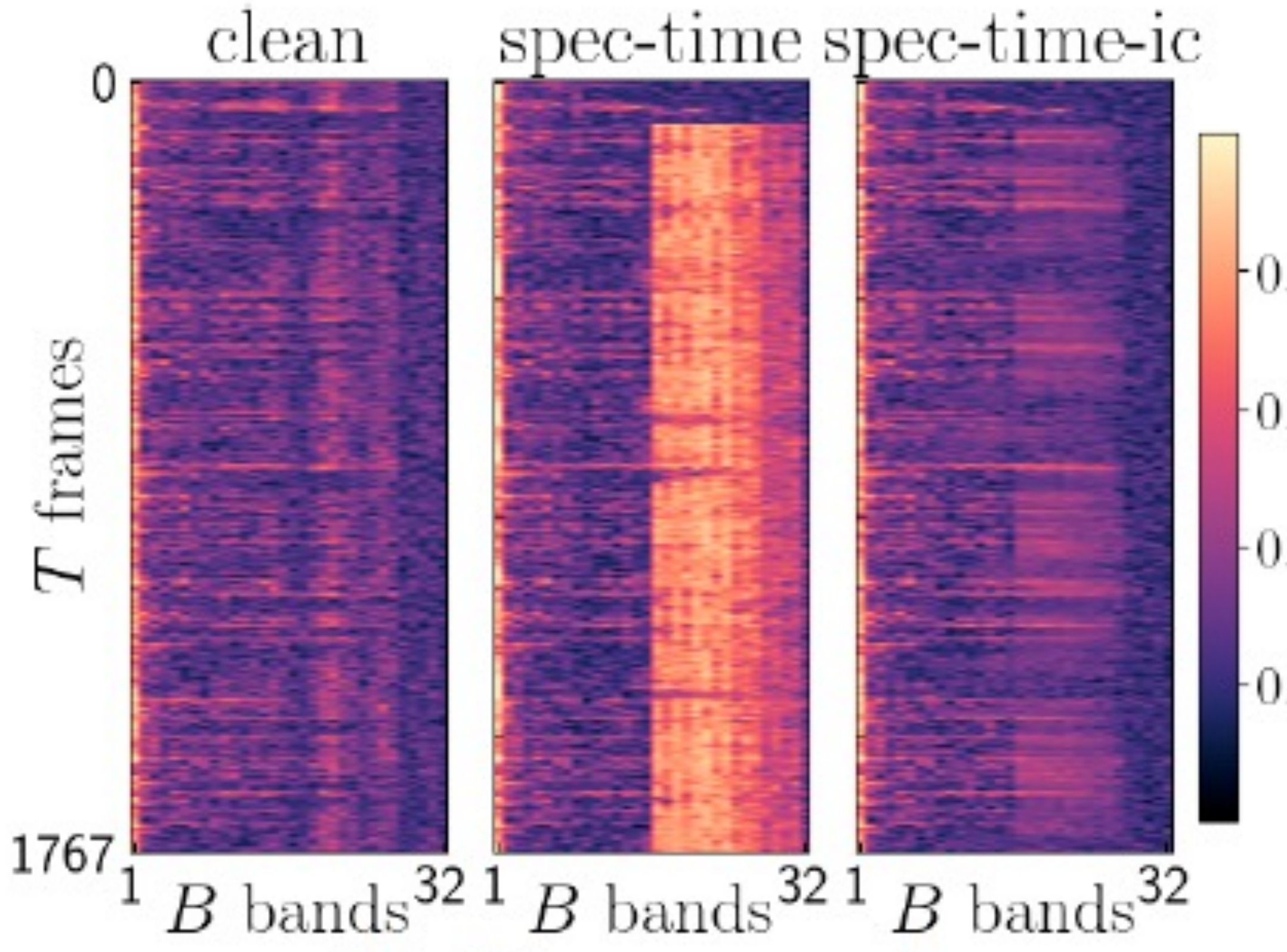


Results

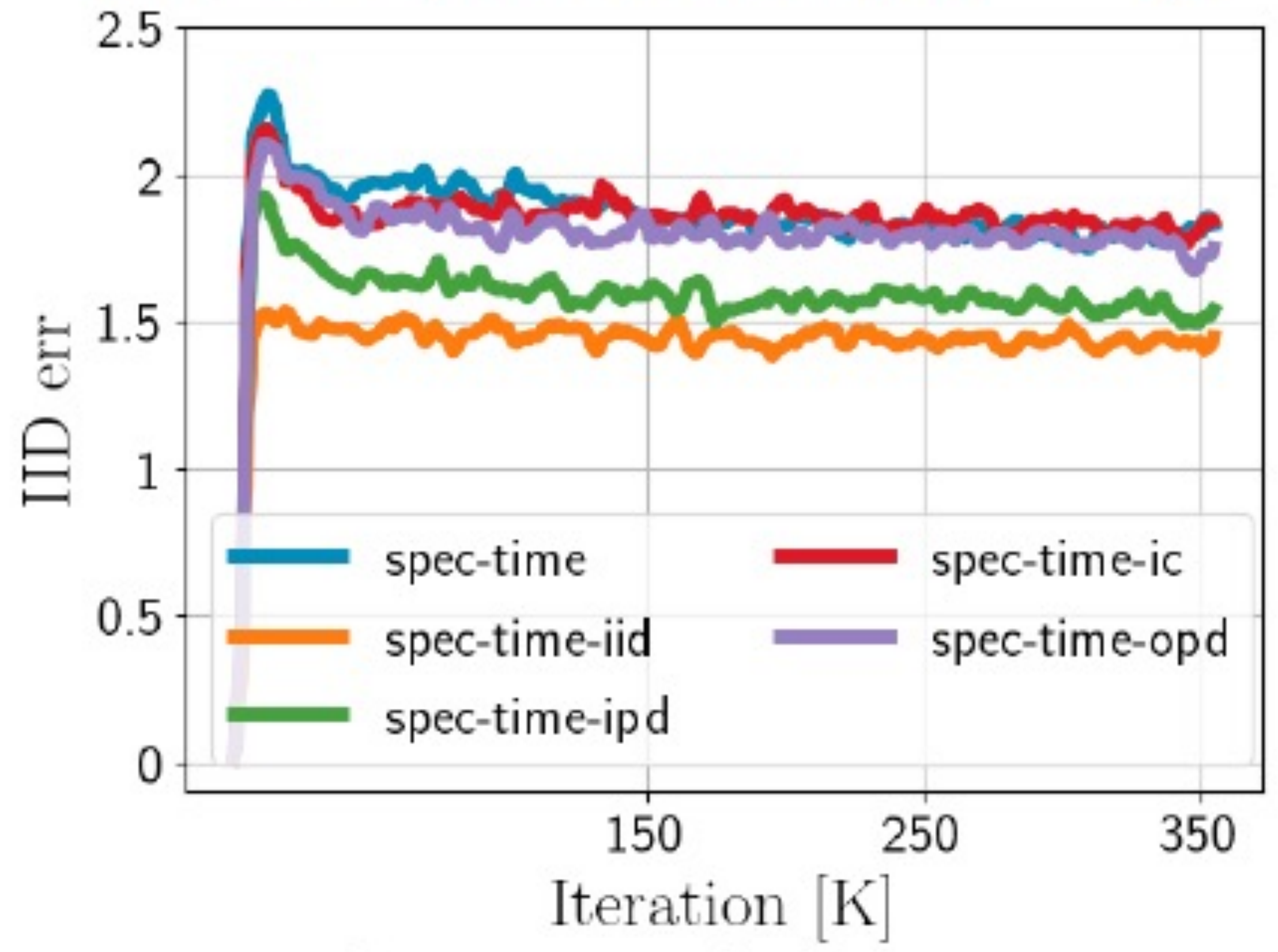
Network	Method	Test set I								Test set II					
		Objective						Subjective		Objective					
		SDR	POLQA	IID	IPD	IC	OPD	OVRL	IMG	SDR	POLQA	IID	IPD	IC	OPD
	noisy	11.61	2.51	1.56	1.92	0.20	0.78	0	0	11.13	2.50	1.60	1.96	0.18	0.79
U-Net	*downmix - spec*	6.46	2.98	2.68	2.79	0.30	1.61	x	x	6.16	2.95	2.70	2.83	0.31	1.62
	LRindp - spec	6.82	3.26	2.36	1.99	0.28	1.62	x	x	6.67	3.19	2.48	2.02	0.27	1.63
	downmix - spec - time	10.10	2.95	2.39	2.78	0.29	1.40	0.34	0.30	9.65	2.92	2.42	2.82	0.29	1.40
	LRindp - spec - time	12.89	3.31	2.42	1.92	0.27	1.27	0.42	0.35	12.27	3.24	2.55	1.95	0.26	1.27
	stereo - spec - time	12.56	3.01	1.85	1.91	0.26	1.25	0.38	0.37	11.97	2.96	1.90	1.93	0.28	1.23
	stereo - spec - time - IID	**14.17**	3.33	**1.55**	1.76	0.35	1.42	0.45	0.41	**13.64**	3.26	**1.59**	1.79	0.39	1.43
	stereo - spec - time - IPD	13.88	**3.36**	1.67	**1.71**	0.32	1.27	**0.63**	0.46	13.24	**3.30**	1.71	**1.73**	0.36	1.28
	stereo - spec - time - IC	12.09	3.04	1.80	2.08	**0.21**	1.43	0.31	0.37	11.47	2.98	1.85	2.12	**0.20**	1.40
	stereo - spec - time - OPD	14.05	3.33	1.86	2.10	0.23	**0.99**	0.42	**0.49**	13.35	3.28	1.90	2.15	0.22	**1.00**
	stereo - spec - time - all	13.78	3.32	1.64	1.81	**0.21**	1.10	0.45	0.43	13.16	3.25	1.69	1.85	**0.19**	1.11
U-NetCM	*stereo - spec*	6.28	**3.34**	2.24	2.14	0.25	2.48	x	x	6.10	**3.27**	2.29	2.18	**0.23**	2.46
	stereo - spec - time - all	**15.02**	3.28	**1.96**	**1.93**	**0.24**	**1.05**	x	x	**14.30**	3.22	**2.01**	**1.97**	**0.23**	**1.06**



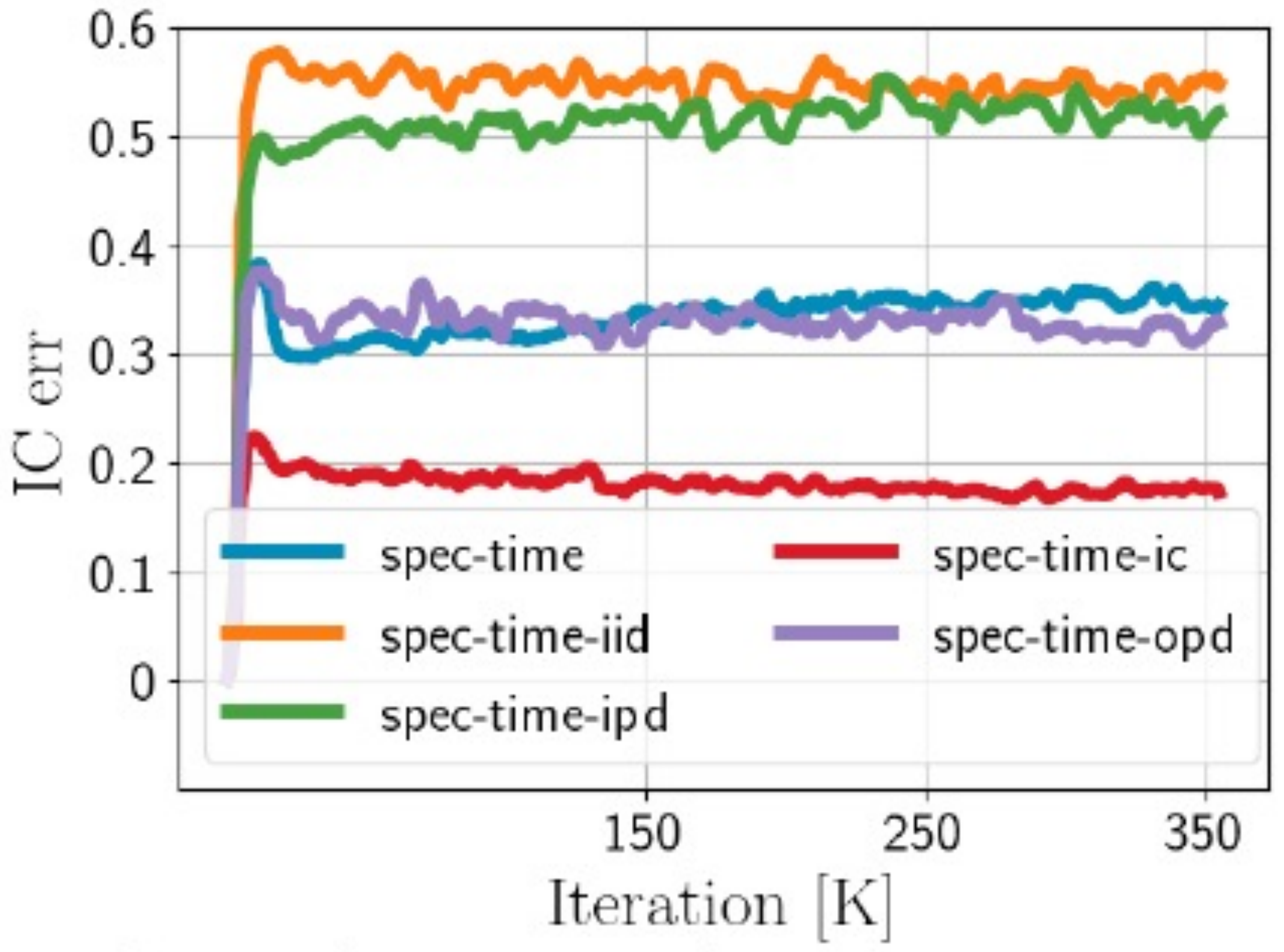
(a) |OPD| for *LRindp* (c = 1).



(b) IC preservation.



(c) IID dynamics during training.



(d) IC dynamics during training.