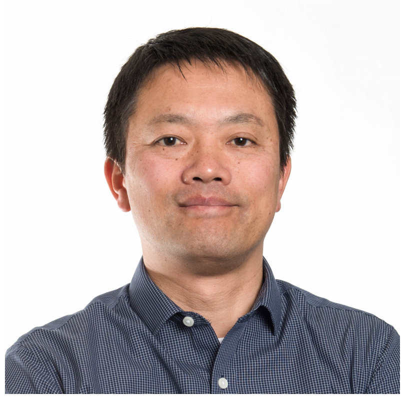
A Training Framework for Stereo-aware Speech Enhancement using Deep Neural Networks



Bahareh Tolooshams[1] and Kazuhito Koishida[2]

[1]Harvard University

Harvard John A. Paulson School of Engineering and Applied Sciences

[2]Microsoft Corporation

Microsoft

IEEE ICASSP 2022

Goal

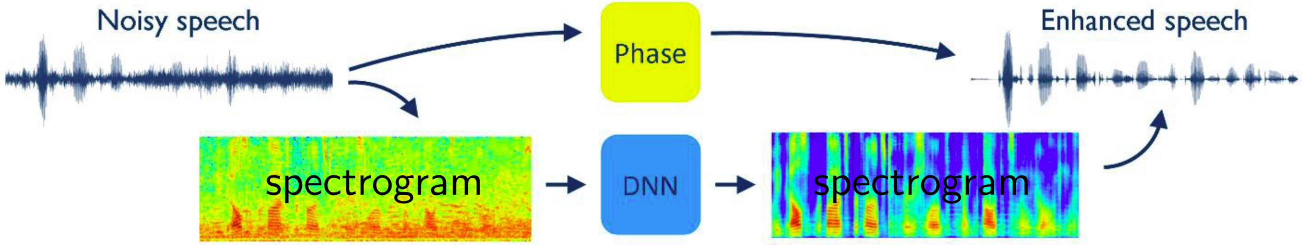
Propose a stereo-aware speech enhancement training framework.

1. Preserve the stereo image while performing speech enhancement.

2. Evaluate perceptual enhancement through subjective tests.

Mono Speech Enhancement

- Enhance spectrogram and add mixture phase at the output.



- Feed stacked (real,imag) and output (real,imag).

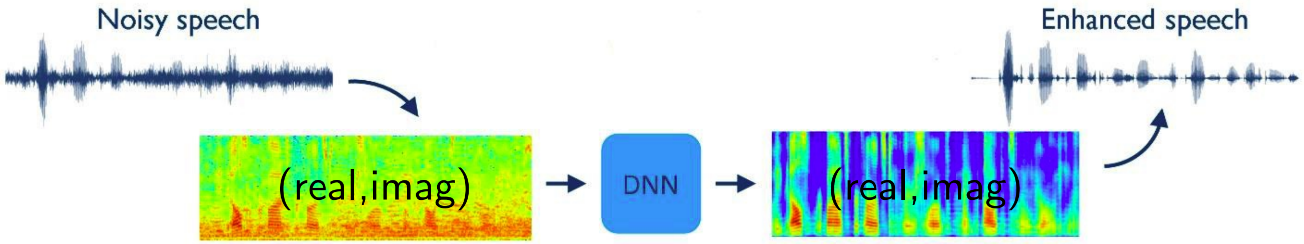
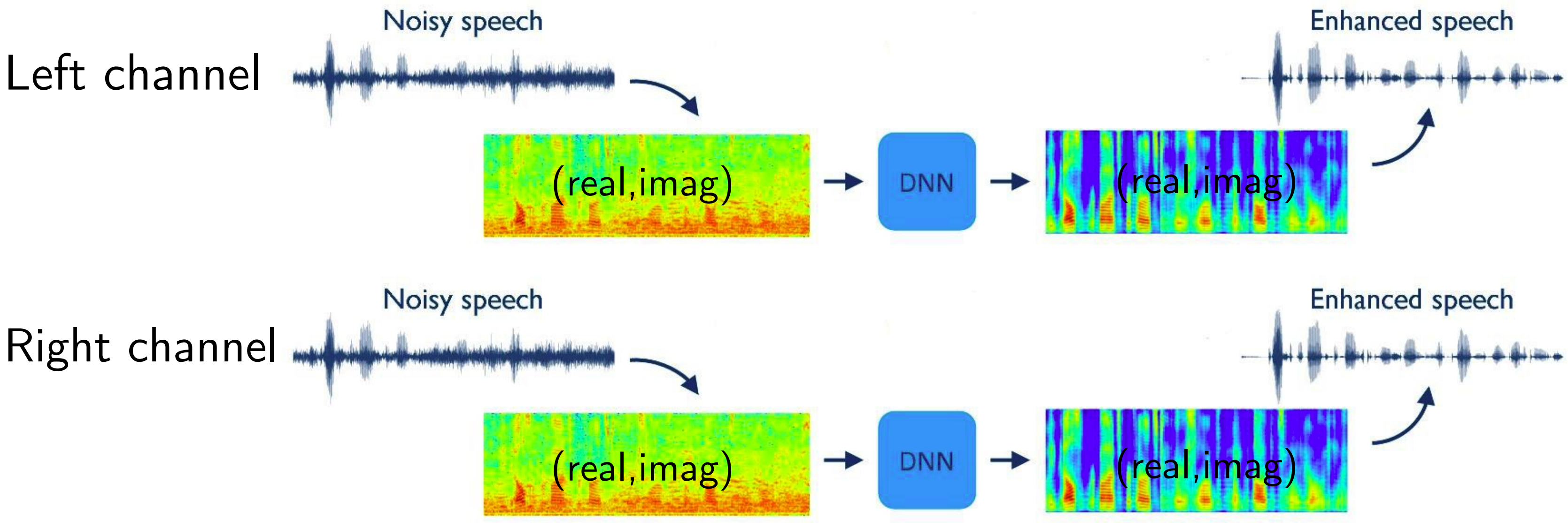


Image from https://acousticsresearchcentre.no/speech-enhancement-with-deep-learning/

Stereo Speech Enhancement (LRindp)

Train one mono network and feed L/R independently.



Drawbacks:

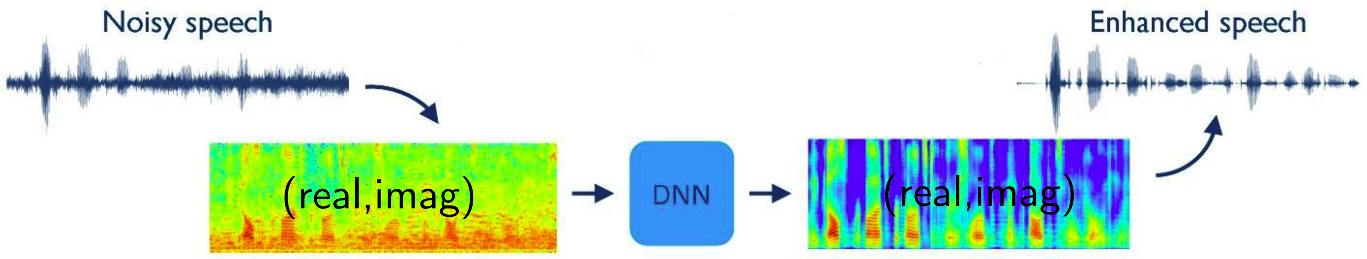
- Channel coherence info is completely ignored.
- Inference time is approximately doubled.

Stereo Speech Enhancement (downmix)

Train using downmix.

Downmix

(L + R)/2



Prediction:

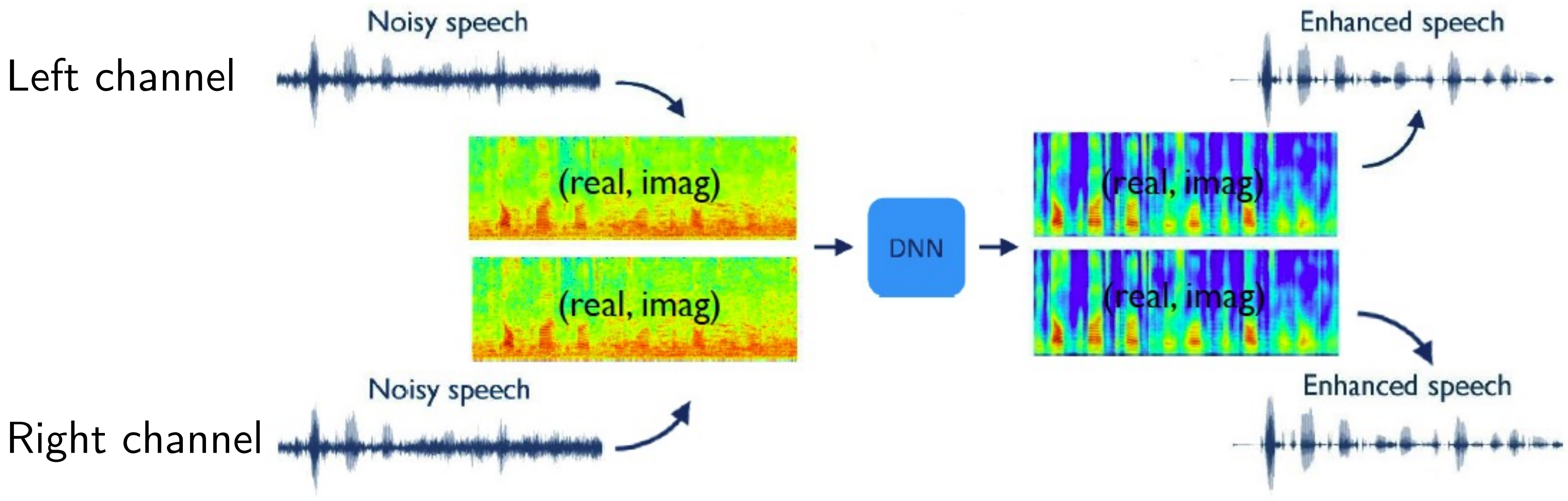
- Enhance downmix.
- Add phase difference between mixture stereo and enhanced downmix.

Drawbacks:

- Added noisy phase at prediction time is not optimal.

Stereo Speech Enhancement (end-to-end)

End-to-end stereo input stereo output.



No guarantee to preserve the stereo image.

Stereo-aware Training

Training loss

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

Speech reconstruction

Stereo image preservation

Speech Reconstruction Loss

$$\mathcal{L}_{\text{speech-rec}}(\mathbf{s},\hat{\mathbf{s}}) = \text{LSD}(\mathbf{s},\hat{\mathbf{s}}) + \alpha_{\text{TL}}\,\text{TL}(\mathbf{s},\hat{\mathbf{s}})$$

Log spectral distortion

$$\text{LSD}(\mathbf{s},\hat{\mathbf{s}}) = \frac{1}{2T}\sum_{c=1}^{2}\sum_{t=1}^{T}\sqrt{\frac{1}{F}\sum_{f=1}^{F}\Big(g(\mathbf{S}_c[t,f]) - g(\hat{\mathbf{S}}_c[t,f])\Big)^2}$$

Time loss

$$\text{TL}(\mathbf{s},\hat{\mathbf{s}}) = \frac{1}{2}\sum_{c=1}^{2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{s}_c[t]-\hat{\mathbf{s}}_c[t])^2}$$

Stereo Image Preservation Loss

$$\mathcal{L}_{\text{image-pres}}(\mathbf{s}, \hat{\mathbf{s}}) = \sum_{\text{M} \in \{\text{IID}, \text{IPD}, \text{IC}, \text{OPD}\}} \alpha_{\text{M}} \mathcal{L}_{\text{M}}(\mathbf{S}, \hat{\mathbf{S}})$$

Intensity

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

Coherence

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

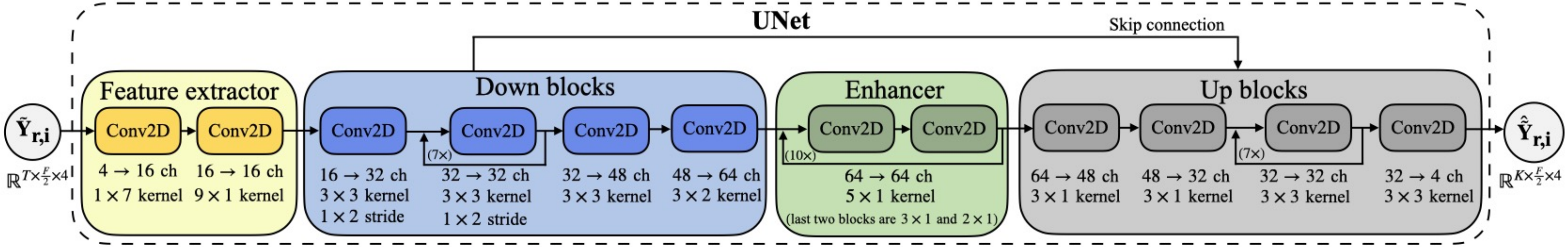
Phase

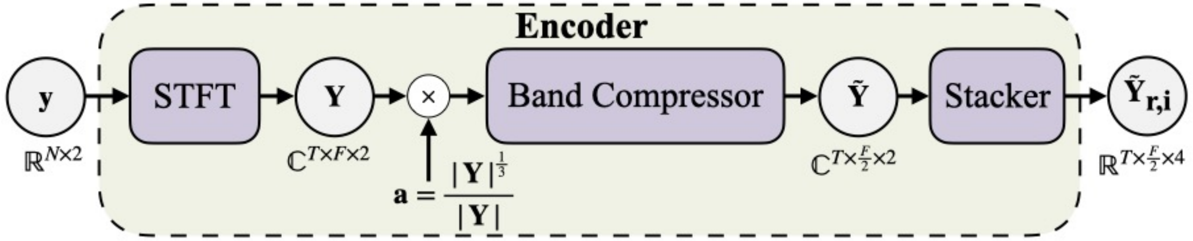
$$\text{IPD}_b(\mathbf{S}) = \angle\left(\sum_{f=f_b}^{f_{b+1}-1} \mathbf{S}_1[f]\mathbf{S}_2^*[f]\right)$$

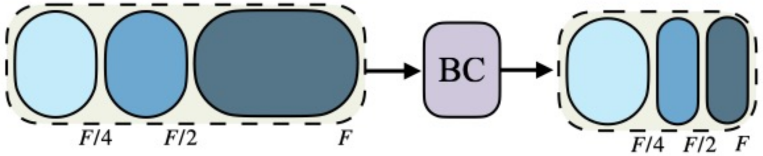
Overall phase

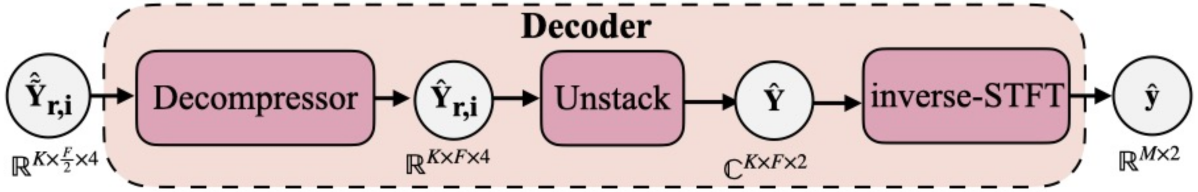
$$\text{OPD}_b(\mathbf{S}, \hat{\mathbf{S}}) = \angle\left(\sum_{f=f_b}^{f_{b+1}-1} \mathbf{S}[f]\hat{\mathbf{S}}^*[f]\right)$$

Network Architecture




Encoder
$\mathbf{y}$
$\mathbb{R}^{N\times2}$
STFT
$\mathbf{Y}$
$\mathbb{C}^{T\times F\times2}$
×
$\mathbf{a} = \frac{|\mathbf{Y}|^{\frac{1}{3}}}{|\mathbf{Y}|}$
Band Compressor
$\tilde{\mathbf{Y}}$
$\mathbb{C}^{T\times\frac{F}{2}\times2}$
Stacker
$\tilde{\mathbf{Y}}_{\mathbf{r,i}}$
$\mathbb{R}^{T\times\frac{F}{2}\times4}$

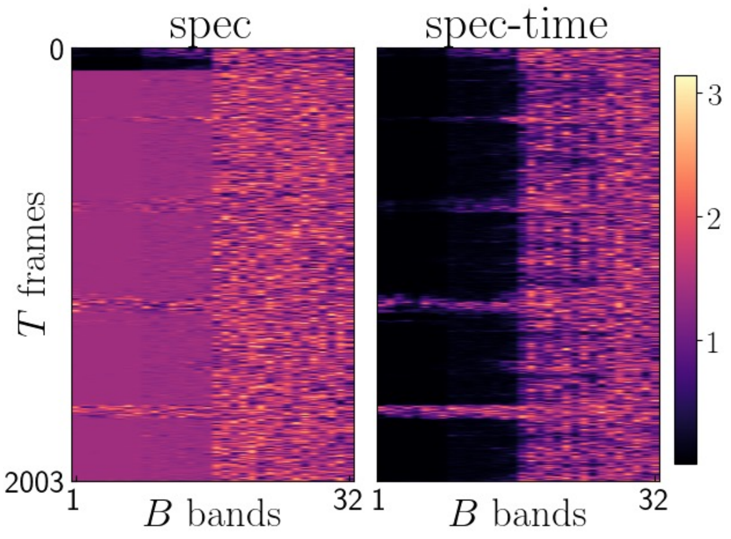


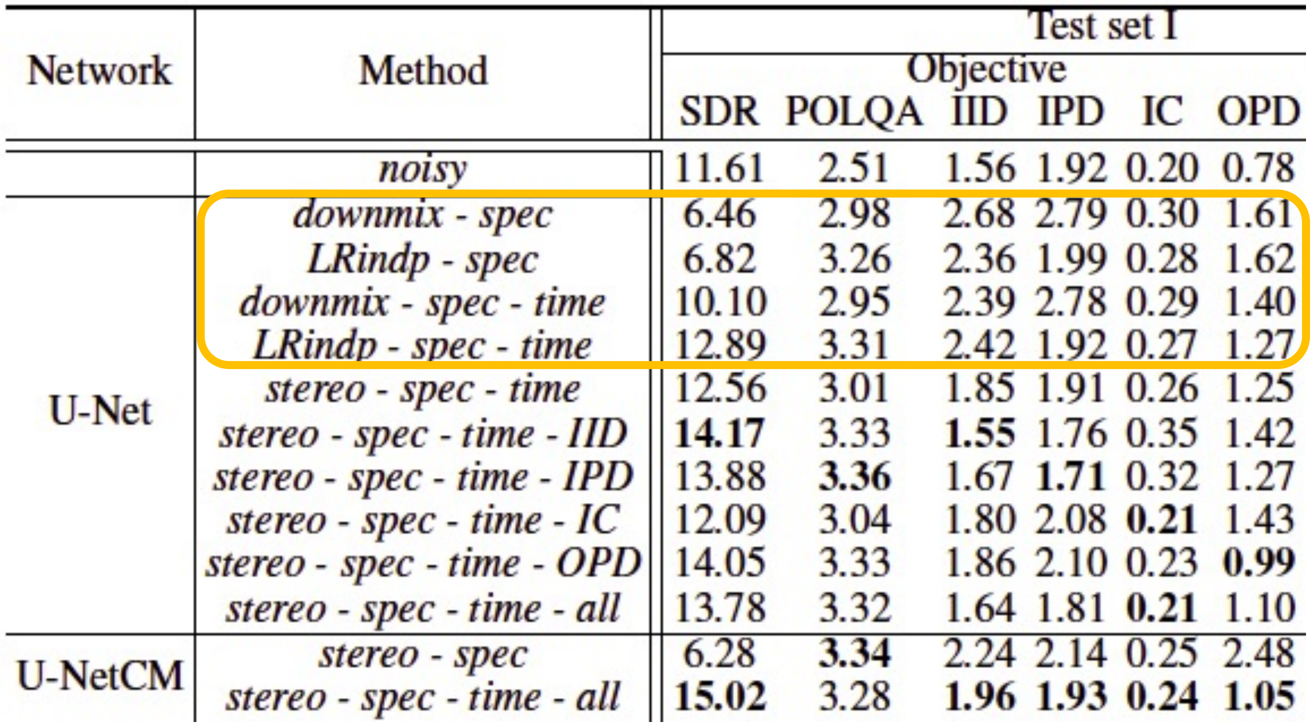

Decoder
$\hat{\tilde{\mathbf{Y}}}_{\mathbf{r,i}}$
$\mathbb{R}^{K\times\frac{F}{2}\times4}$
Decompressor
$\hat{\mathbf{Y}}_{\mathbf{r,i}}$
$\mathbb{R}^{K\times F\times4}$
Unstack
$\hat{\mathbf{Y}}$
$\mathbb{C}^{K\times F\times2}$
inverse-STFT
$\hat{\mathbf{y}}$
$\mathbb{R}^{M\times2}$

Presence of Time Loss

- Higher SDR.

- Overall phase preservation.

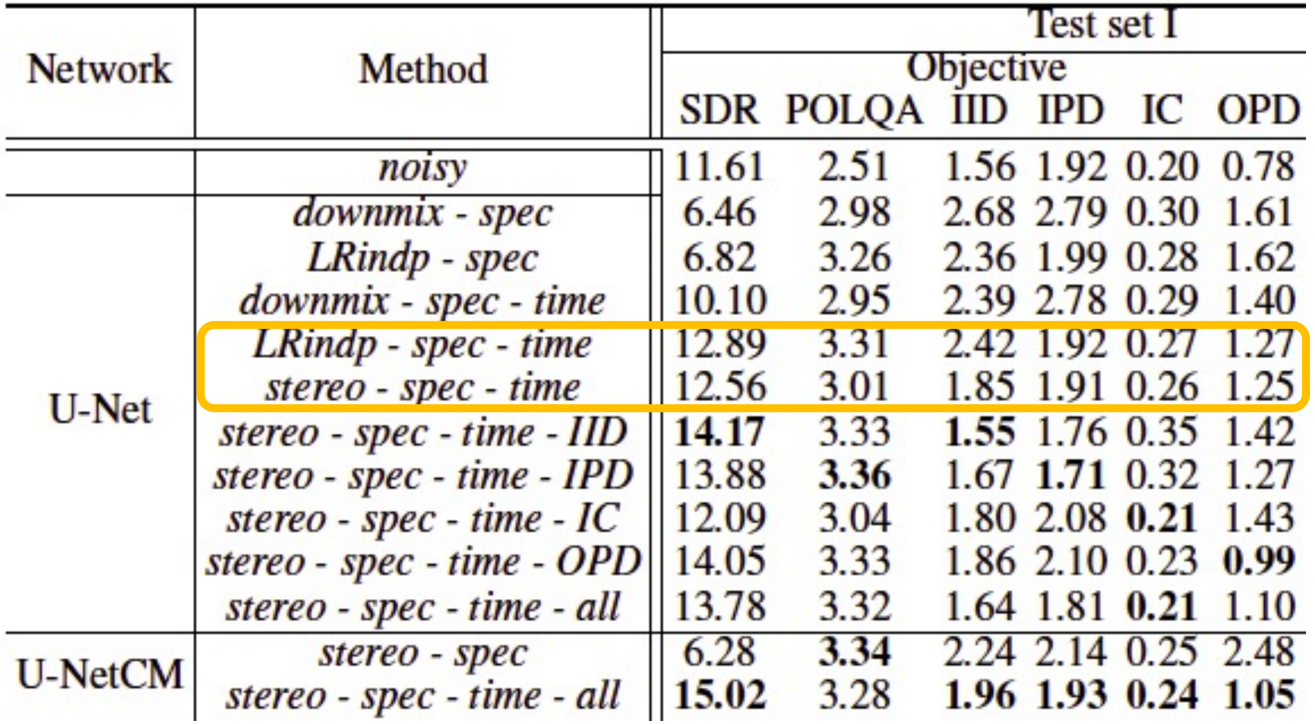




Network	Method	Test set I					
		Objective					
		SDR	POLQA	IID	IPD	IC	OPD
	noisy	11.61	2.51	1.56	1.92	0.20	0.78
U-Net	*downmix - spec*	6.46	2.98	2.68	2.79	0.30	1.61
	LRindp - spec	6.82	3.26	2.36	1.99	0.28	1.62
	downmix - spec - time	10.10	2.95	2.39	2.78	0.29	1.40
	LRindp - spec - time	12.89	3.31	2.42	1.92	0.27	1.27
	stereo - spec - time	12.56	3.01	1.85	1.91	0.26	1.25
	stereo - spec - time - IID	**14.17**	3.33	**1.55**	1.76	0.35	1.42
	stereo - spec - time - IPD	13.88	**3.36**	1.67	**1.71**	0.32	1.27
	stereo - spec - time - IC	12.09	3.04	1.80	2.08	**0.21**	1.43
	stereo - spec - time - OPD	14.05	3.33	1.86	2.10	0.23	**0.99**
	stereo - spec - time - all	13.78	3.32	1.64	1.81	**0.21**	1.10
U-NetCM	*stereo - spec*	6.28	**3.34**	2.24	2.14	0.25	2.48
	stereo - spec - time - all	**15.02**	3.28	**1.96**	**1.93**	**0.24**	**1.05**

Mono to Stereo

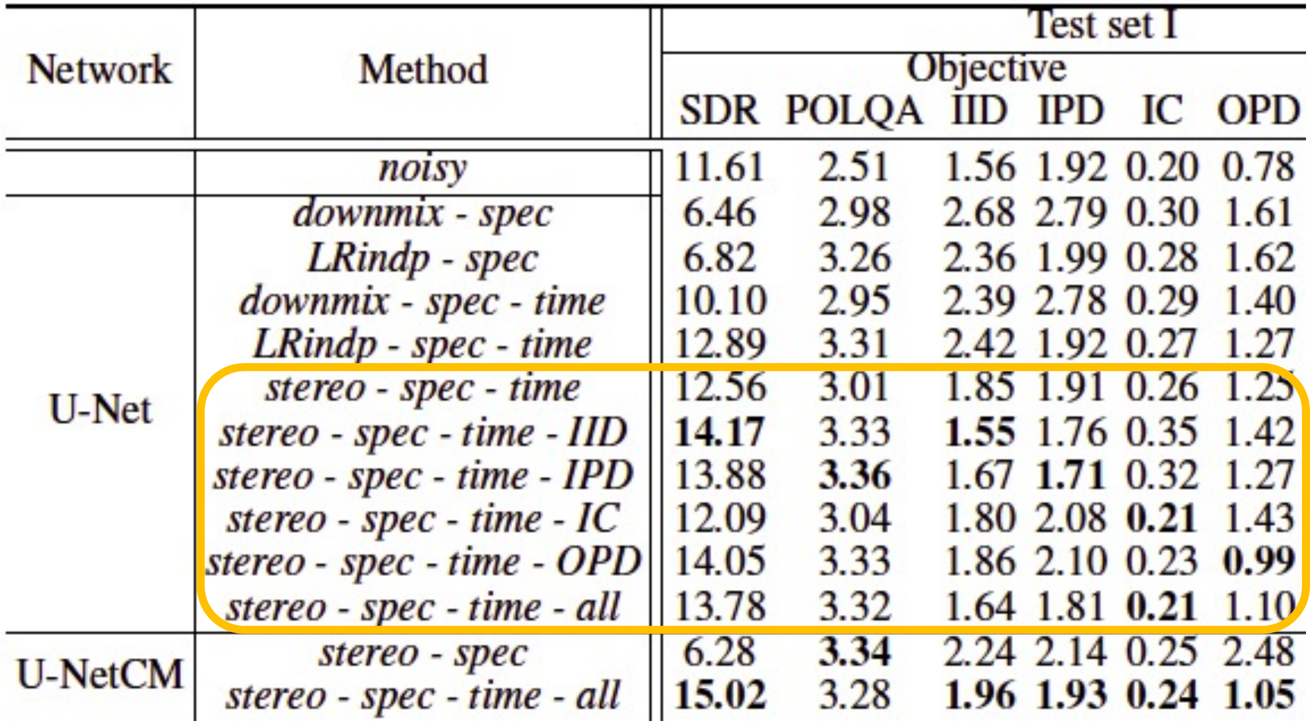
- Stereo preserves IID better than LRindp.
- LRindp has higher SDR and POLQA.



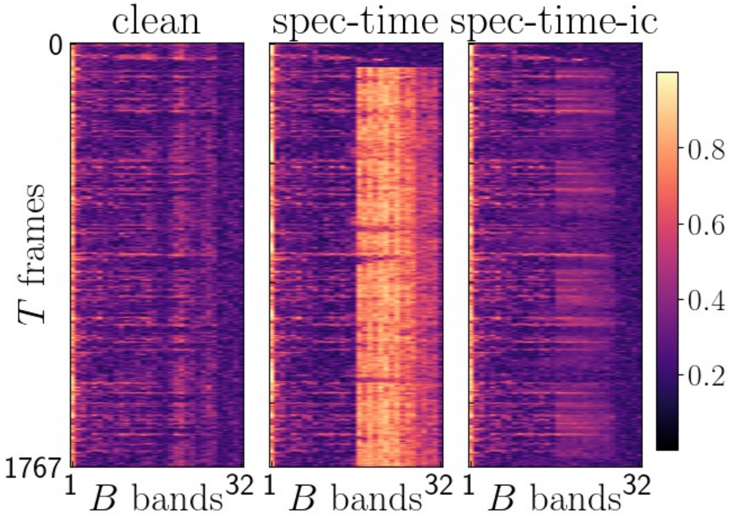
Network	Method	Test set I					
		Objective					
		SDR	POLQA	IID	IPD	IC	OPD
	noisy	11.61	2.51	1.56	1.92	0.20	0.78
U-Net	*downmix - spec*	6.46	2.98	2.68	2.79	0.30	1.61
	LRindp - spec	6.82	3.26	2.36	1.99	0.28	1.62
	downmix - spec - time	10.10	2.95	2.39	2.78	0.29	1.40
	LRindp - spec - time	12.89	3.31	2.42	1.92	0.27	1.27
	stereo - spec - time	12.56	3.01	1.85	1.91	0.26	1.25
	stereo - spec - time - IID	**14.17**	3.33	**1.55**	1.76	0.35	1.42
	stereo - spec - time - IPD	13.88	**3.36**	1.67	**1.71**	0.32	1.27
	stereo - spec - time - IC	12.09	3.04	1.80	2.08	**0.21**	1.43
	stereo - spec - time - OPD	14.05	3.33	1.86	2.10	0.23	**0.99**
	stereo - spec - time - all	13.78	3.32	1.64	1.81	**0.21**	1.10
U-NetCM	*stereo - spec*	6.28	**3.34**	2.24	2.14	0.25	2.48
	stereo - spec - time - all	**15.02**	3.28	**1.96**	**1.93**	**0.24**	**1.05**

Image Preservation Loss

- Higher POLQA and SDR.
- IID to improve SDR.
- IPD to increase POLQA.



Network	Method	Test set I					
		Objective					
		SDR	POLQA	IID	IPD	IC	OPD
	noisy	11.61	2.51	1.56	1.92	0.20	0.78
U-Net	*downmix - spec*	6.46	2.98	2.68	2.79	0.30	1.61
	LRindp - spec	6.82	3.26	2.36	1.99	0.28	1.62
	downmix - spec - time	10.10	2.95	2.39	2.78	0.29	1.40
	LRindp - spec - time	12.89	3.31	2.42	1.92	0.27	1.27
	stereo - spec - time	12.56	3.01	1.85	1.91	0.26	1.25
	stereo - spec - time - IID	**14.17**	3.33	**1.55**	1.76	0.35	1.42
	stereo - spec - time - IPD	13.88	**3.36**	1.67	**1.71**	0.32	1.27
	stereo - spec - time - IC	12.09	3.04	1.80	2.08	**0.21**	1.43
	stereo - spec - time - OPD	14.05	3.33	1.86	2.10	0.23	**0.99**
	stereo - spec - time - all	13.78	3.32	1.64	1.81	**0.21**	1.10
U-NetCM	*stereo - spec*	6.28	**3.34**	2.24	2.14	0.25	2.48
	stereo - spec - time - all	**15.02**	3.28	**1.96**	**1.93**	**0.24**	**1.05**



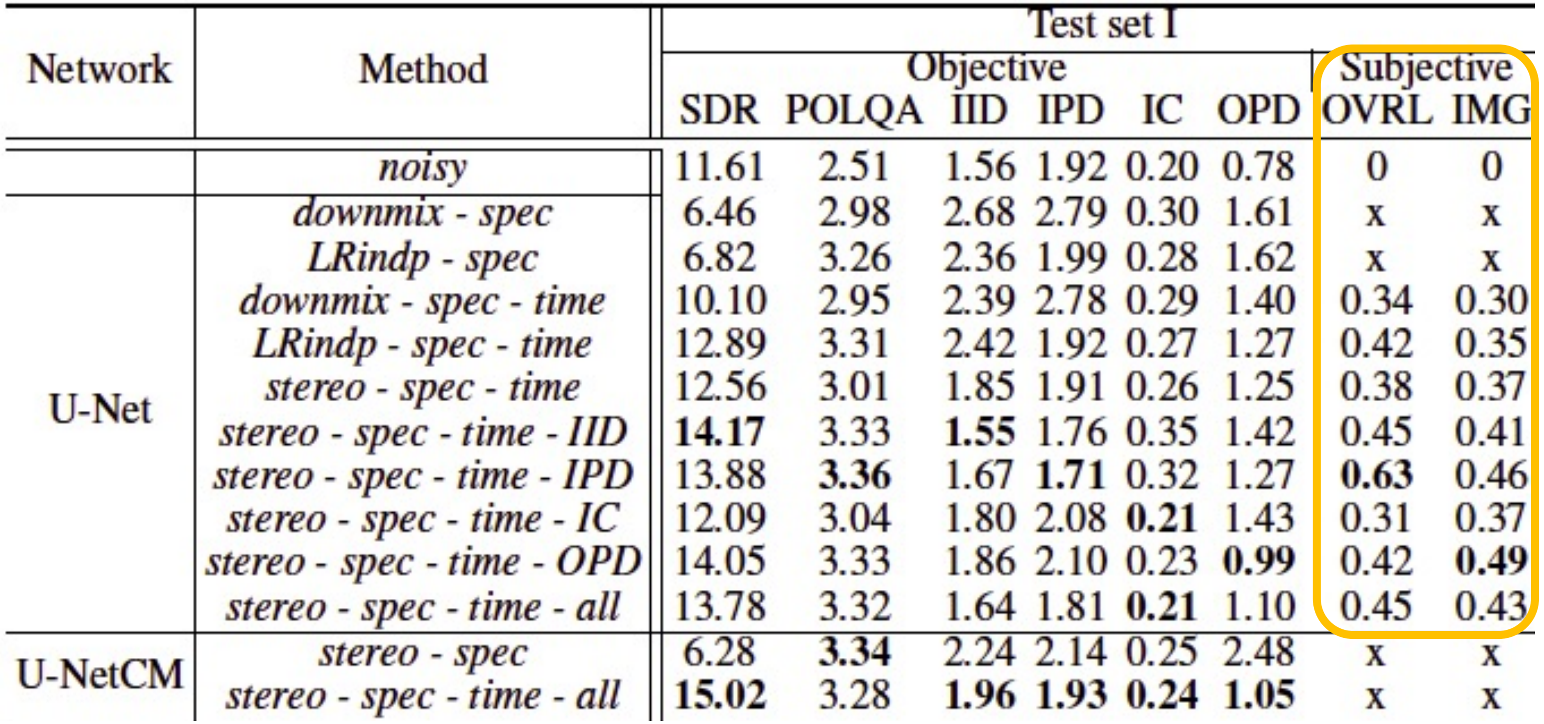
Subjective Evaluation

MUSHRA test.

Approx. 2,750 listeners.

Overall quality (OVRL).

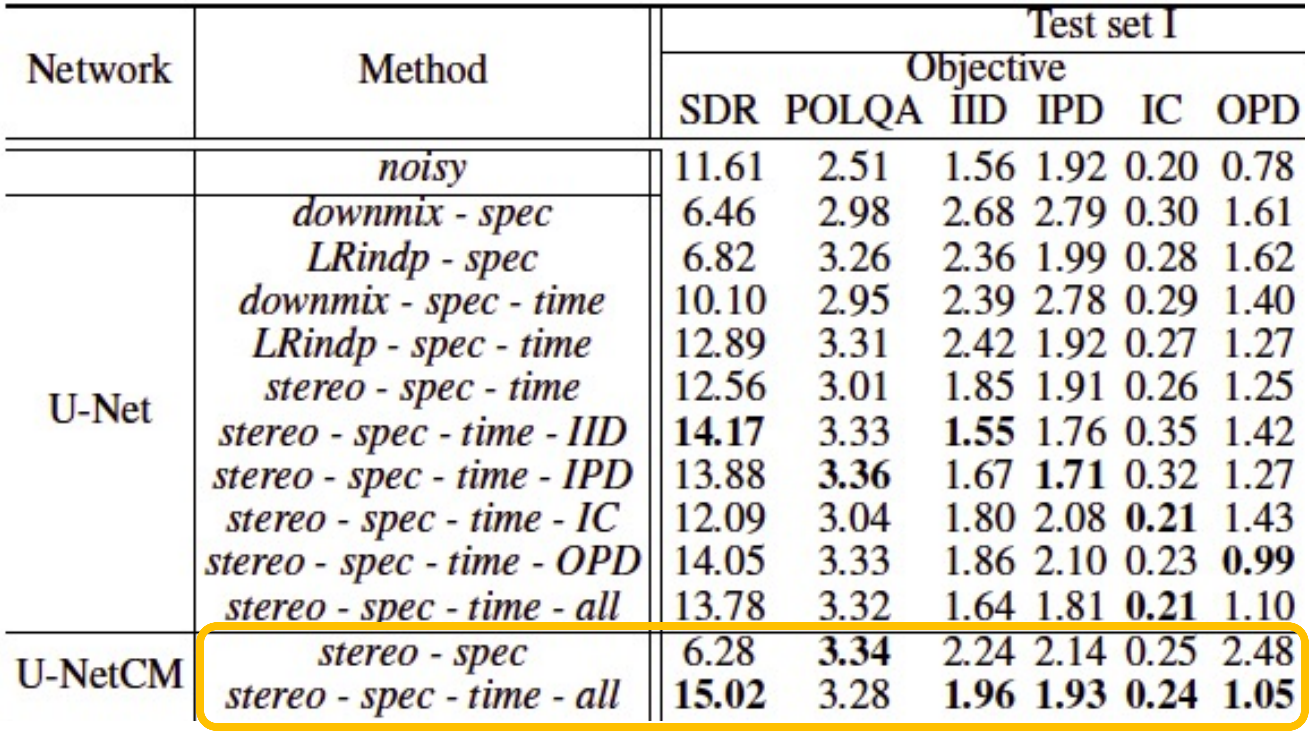
Stereophonic image quality (IMG).



Network	Method	Test set I							
		Objective						Subjective	
		SDR	POLQA	IID	IPD	IC	OPD	OVRL	IMG
	noisy	11.61	2.51	1.56	1.92	0.20	0.78	0	0
U-Net	*downmix - spec*	6.46	2.98	2.68	2.79	0.30	1.61	x	x
	LRindp - spec	6.82	3.26	2.36	1.99	0.28	1.62	x	x
	downmix - spec - time	10.10	2.95	2.39	2.78	0.29	1.40	0.34	0.30
	LRindp - spec - time	12.89	3.31	2.42	1.92	0.27	1.27	0.42	0.35
	stereo - spec - time	12.56	3.01	1.85	1.91	0.26	1.25	0.38	0.37
	stereo - spec - time - IID	**14.17**	3.33	**1.55**	1.76	0.35	1.42	0.45	0.41
	stereo - spec - time - IPD	13.88	**3.36**	1.67	**1.71**	0.32	1.27	**0.63**	0.46
	stereo - spec - time - IC	12.09	3.04	1.80	2.08	**0.21**	1.43	0.31	0.37
	stereo - spec - time - OPD	14.05	3.33	1.86	2.10	0.23	**0.99**	0.42	**0.49**
	stereo - spec - time - all	13.78	3.32	1.64	1.81	**0.21**	1.10	0.45	0.43
U-NetCM	*stereo - spec*	6.28	**3.34**	2.24	2.14	0.25	2.48	x	x
	stereo - spec - time - all	**15.02**	3.28	**1.96**	**1.93**	**0.24**	**1.05**	x	x

Model Independence

Proposed stereo-aware training improves SDR and preserves stereo image (e.g., IID, IPD, IC, and OPD) independent of the network architecture.



Network	Method	Test set I					
		Objective					
		SDR	POLQA	IID	IPD	IC	OPD
	noisy	11.61	2.51	1.56	1.92	0.20	0.78
U-Net	*downmix - spec*	6.46	2.98	2.68	2.79	0.30	1.61
	LRindp - spec	6.82	3.26	2.36	1.99	0.28	1.62
	downmix - spec - time	10.10	2.95	2.39	2.78	0.29	1.40
	LRindp - spec - time	12.89	3.31	2.42	1.92	0.27	1.27
	stereo - spec - time	12.56	3.01	1.85	1.91	0.26	1.25
	stereo - spec - time - IID	**14.17**	3.33	**1.55**	1.76	0.35	1.42
	stereo - spec - time - IPD	13.88	**3.36**	1.67	**1.71**	0.32	1.27
	stereo - spec - time - IC	12.09	3.04	1.80	2.08	**0.21**	1.43
	stereo - spec - time - OPD	14.05	3.33	1.86	2.10	0.23	**0.99**
	stereo - spec - time - all	13.78	3.32	1.64	1.81	**0.21**	1.10
U-NetCM	*stereo - spec*	6.28	**3.34**	2.24	2.14	0.25	2.48
	stereo - spec - time - all	**15.02**	3.28	**1.96**	**1.93**	**0.24**	**1.05**

Check out our poster #4816

Session: SPE-33: Speech Enhancement: Training Schemes and Losses.

Tuesday, 10 May, 20:00 – 20:40 (Singapore Time, UTC +8)

Bahareh Tolooshams

btolooshams@seas.harvard.edu

https://btolooshams.github.io