Dense CNN with Self-Attention for Time-Domain Speech Enhancement

Ashutosh Pandey,
DeLiang Wang

Outline

- Introduction
- Methodology
- Architecture
- Experiments
- Conclusion

Introduction

When the voice is polluted by background noise, not only the magnitude will be affected,but also the phase will also change, but the risk of adjusting the phase is extremely high,and it is very likely that the voice quality will become very bad.

Introduction

On the other hand, when processing signals in the time domain, the magnitude and phase can be changed together,and it is safer than processing the phase in the frequency domain.

Therefore, this paper proposes a time-domain speech enhancement model that combines Dense CNN and Self Attention,and uses a new loss function that simultaneously constrains speech and background sounds.

Methodology

U-Net

+

Dense Net

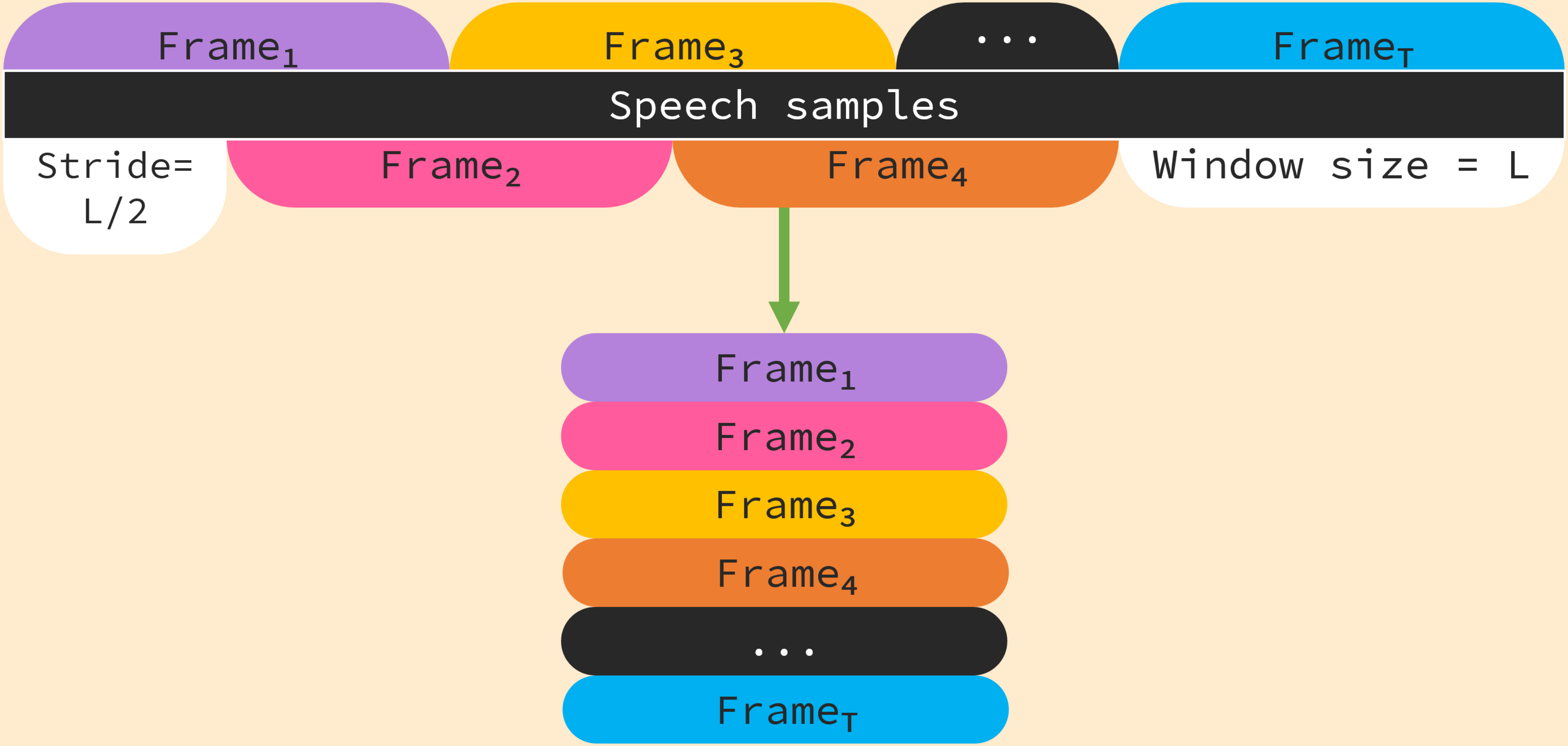
+

Sub-pixel Convolution

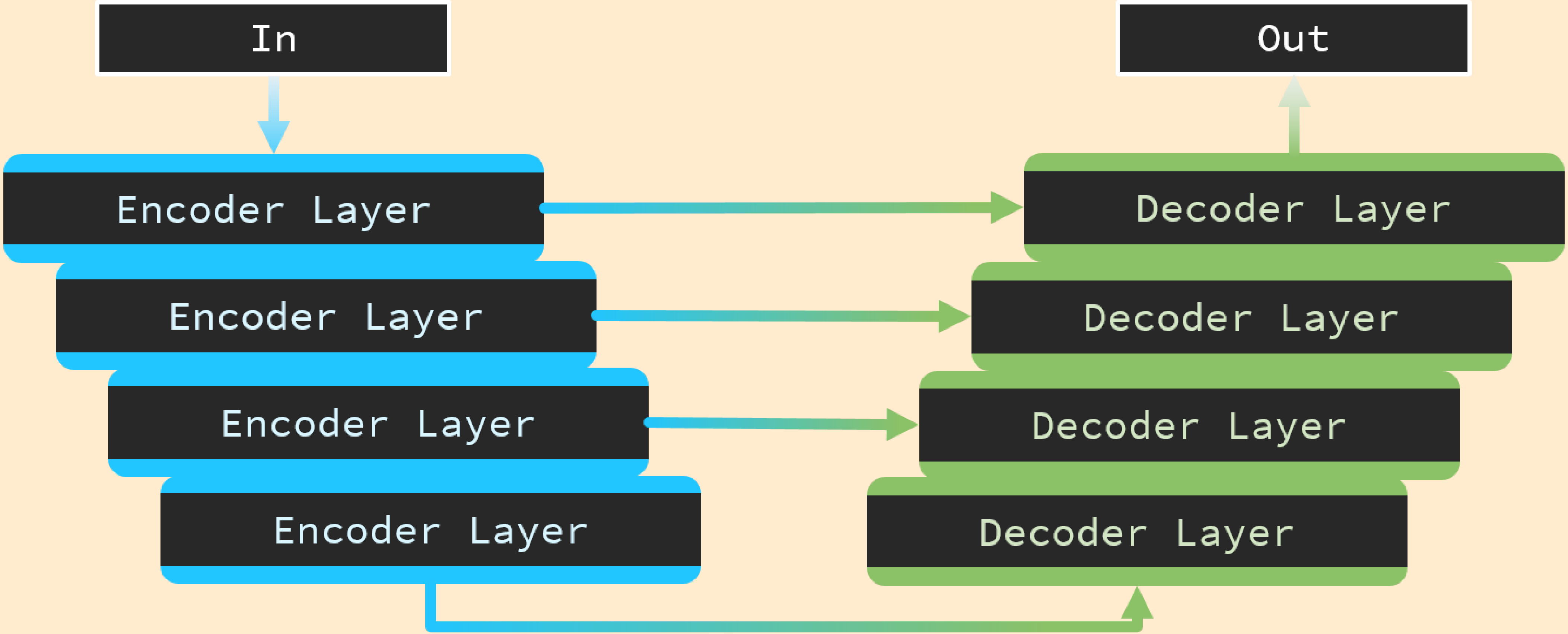
+

Self Attention

Input


Frame_1
Frame_3
...
Frame_T
Speech samples
Stride= L/2
Frame_2
Frame_4
Window size = L
Frame_1
Frame_2
Frame_3
Frame_4
...
Frame_T

U-Net


In

Encoder Layer

Encoder Layer

Encoder Layer

Encoder Layer

Out

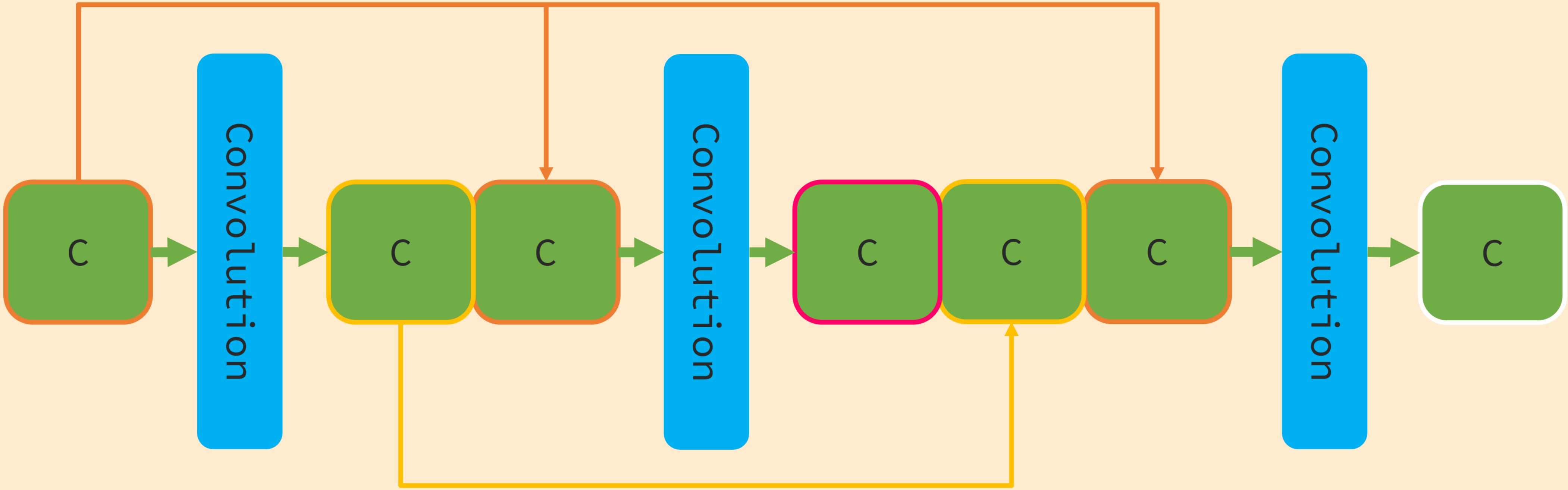
Decoder Layer

Decoder Layer

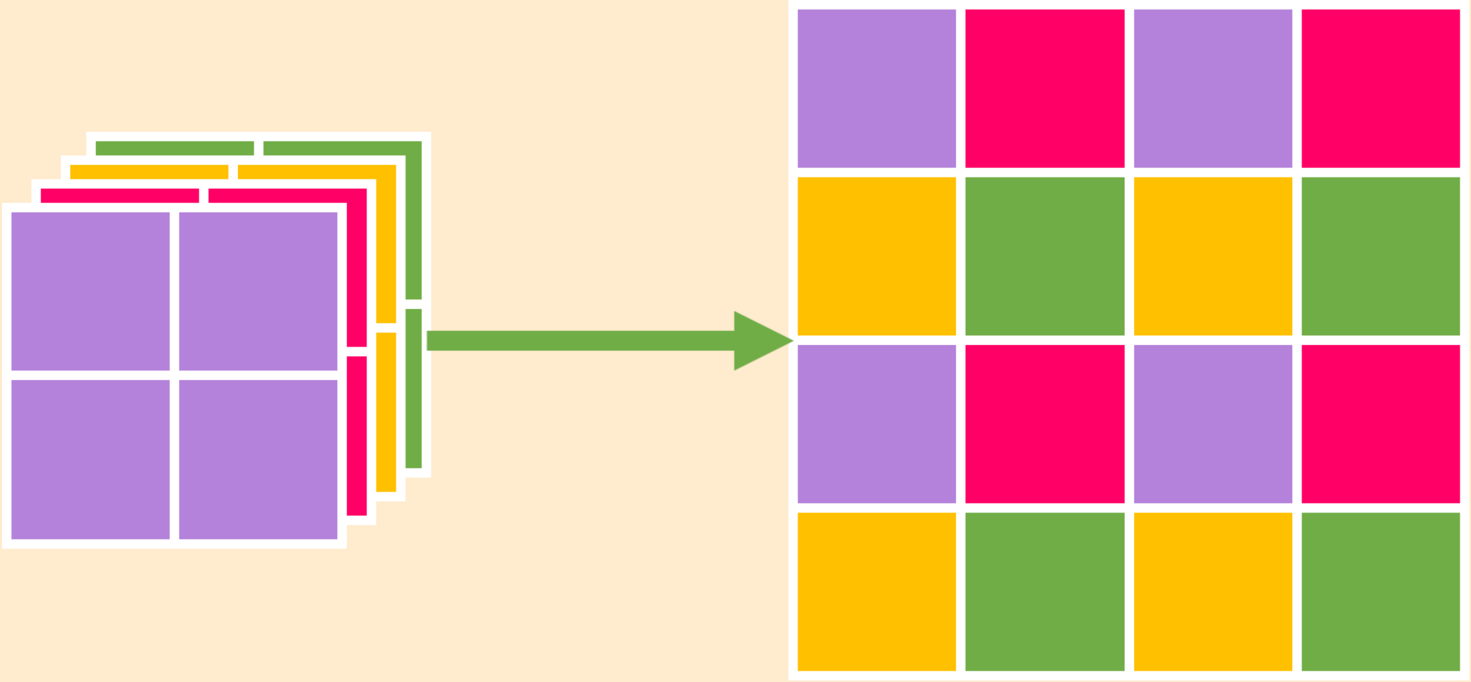
Decoder Layer

Decoder Layer

Dense Net



Sub-pixel Convolution



Self Attention

Causal：$\mathrm{Softmax}(\mathrm{Mask}(QK^{T}))V$

Non Causal：$\mathrm{Softmax}(QK^{T})V$

Loss

- Time-Domain Loss

$$\mathcal{L}_T(s,\hat{s}) = MSE(s,\hat{s})$$

- STFT Magnitude Loss

$$\mathcal{L}_{SM}(s,\hat{s}) = MAE(mag(s), mag(\hat{s}))$$

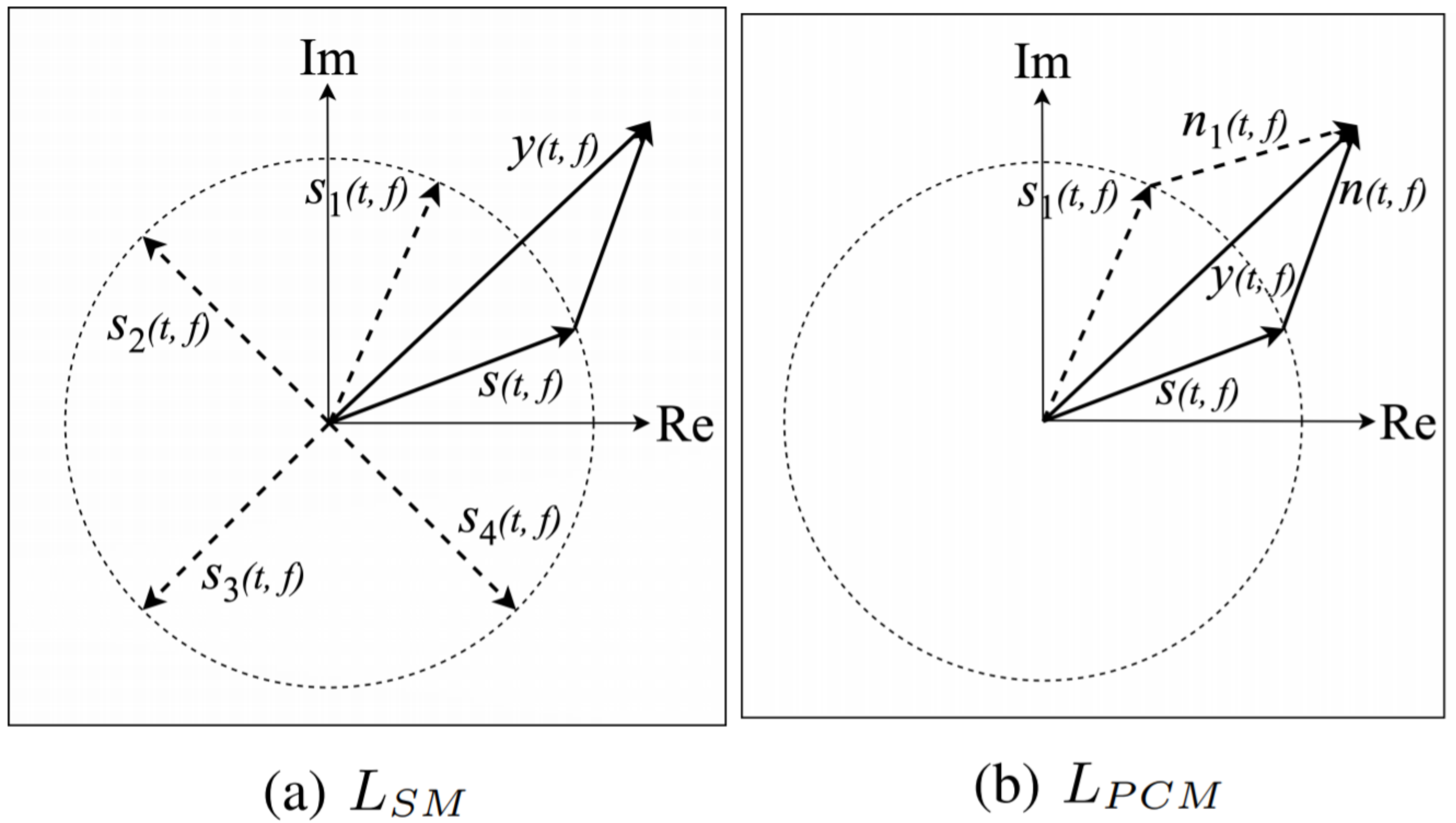
- Time-frequency Loss

$$\mathcal{L}_{TF}(s,\hat{s}) = \alpha\mathcal{L}_T + (1\text{-}\alpha)\,\mathcal{L}_{SM}$$

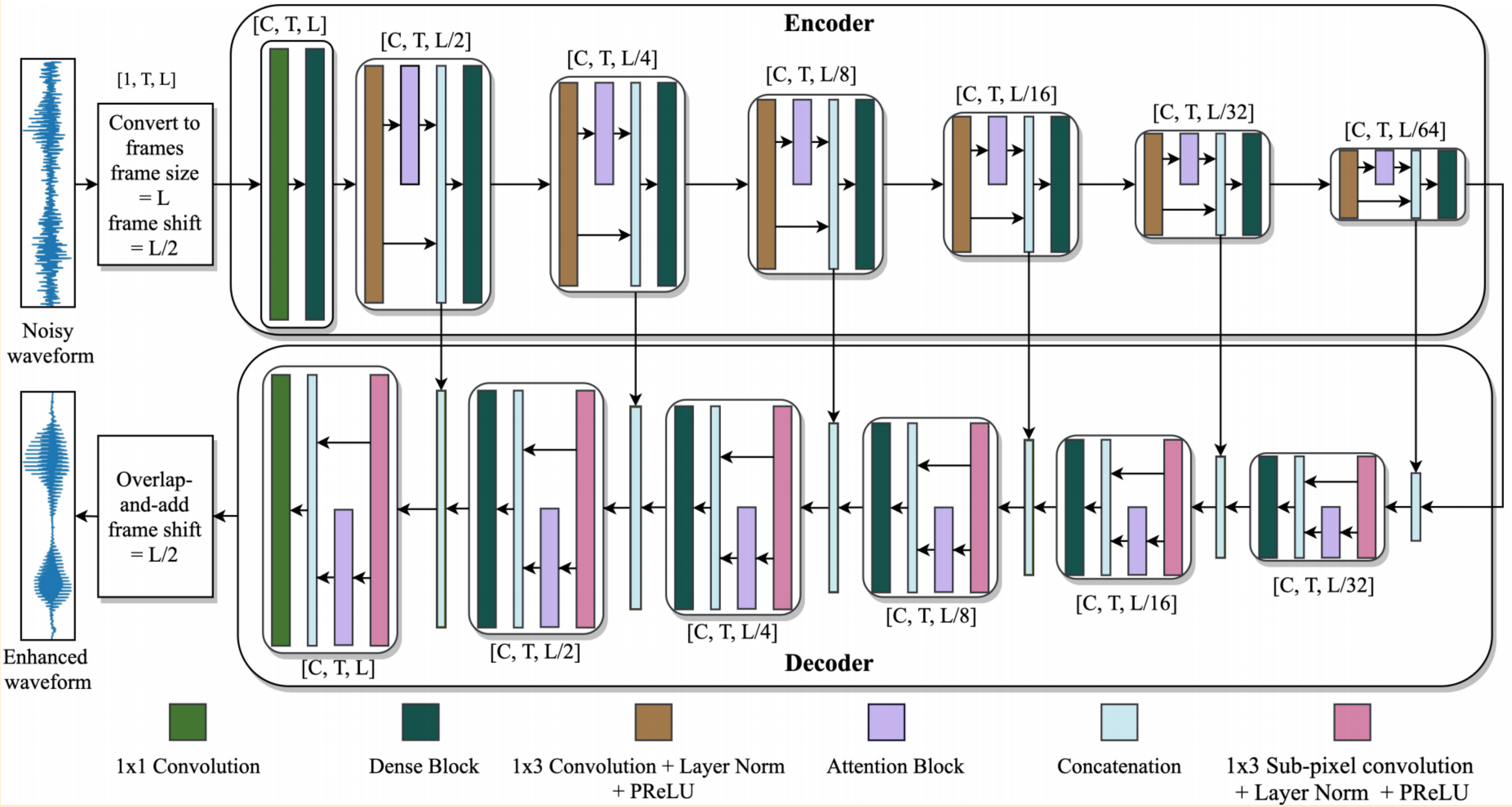
- Phase Constrained Magnitude Loss

$$\mathcal{L}_{PCM}(s,\hat{s}) = 0.5\mathcal{L}_{SM}(s,\hat{s}) + 0.5\mathcal{L}_{SM}(n, x-\hat{s})$$

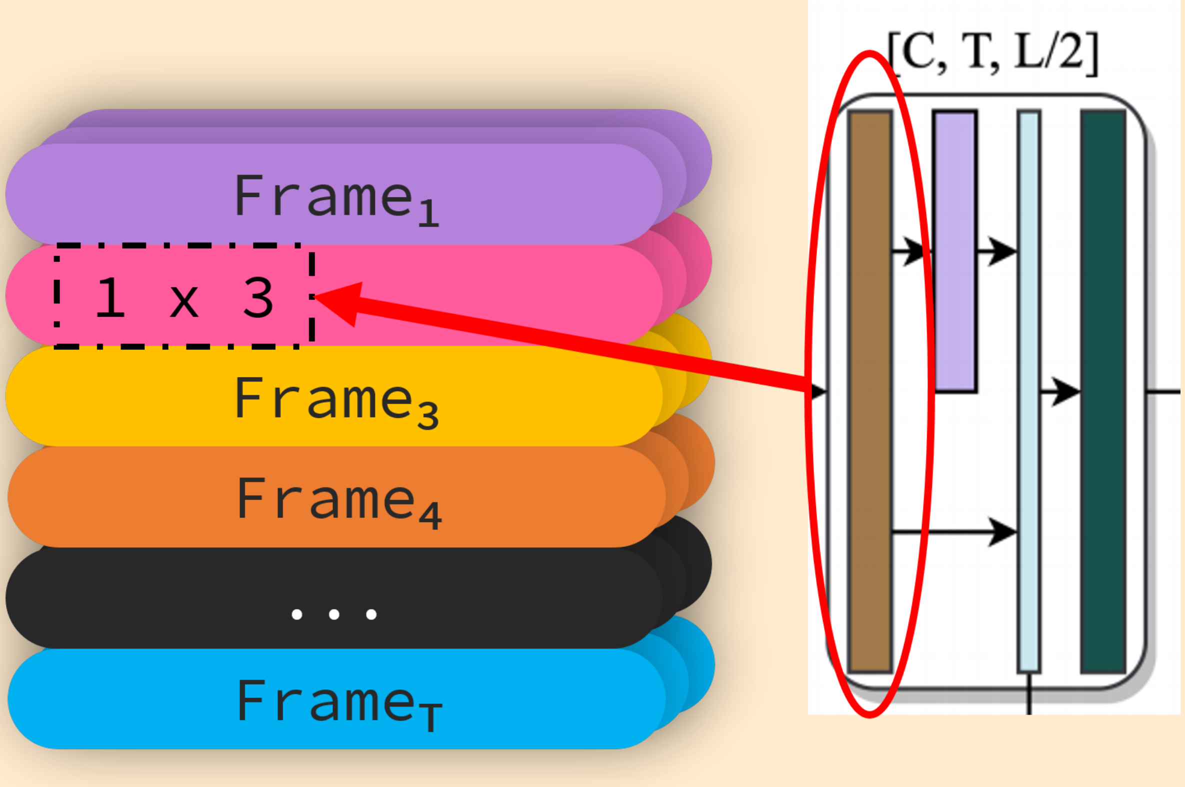
PCM Loss


Im
Re
$y(t,f)$
$s_1(t,f)$
$s_2(t,f)$
$s(t,f)$
$s_3(t,f)$
$s_4(t,f)$
(a) L_{SM}
Im
Re
$n_1(t,f)$
$s_1(t,f)$
$n(t,f)$
$y(t,f)$
$s(t,f)$
(b) L_{PCM}

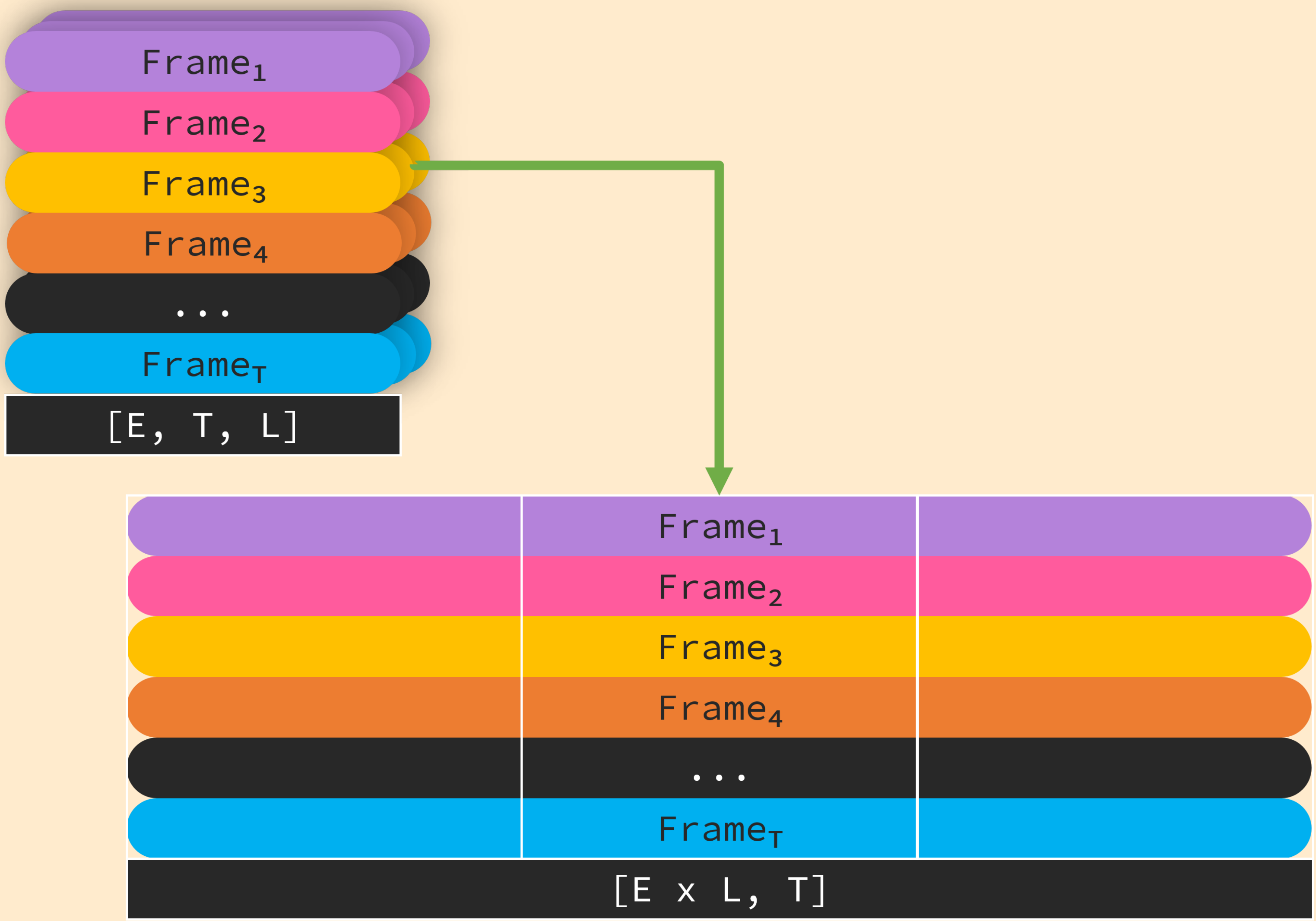
Architecture



1 x 3 Conv

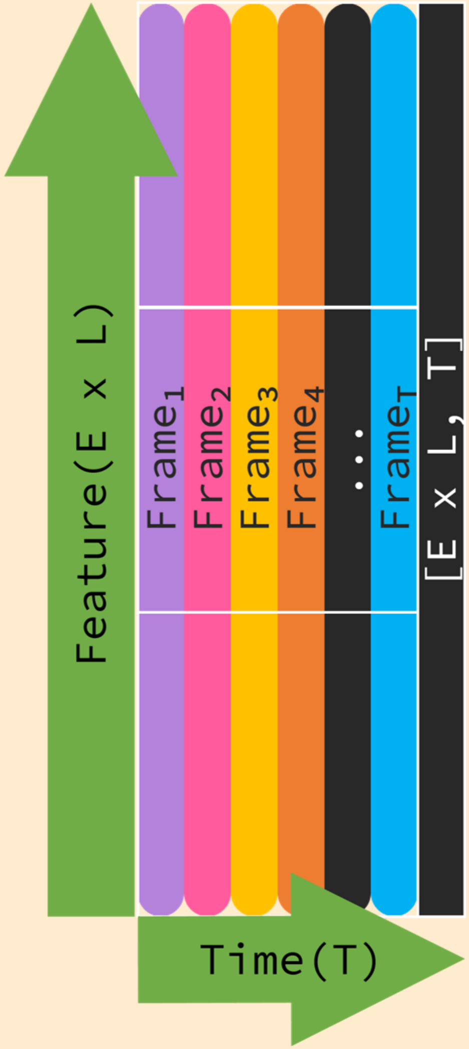


Self Attention Shape

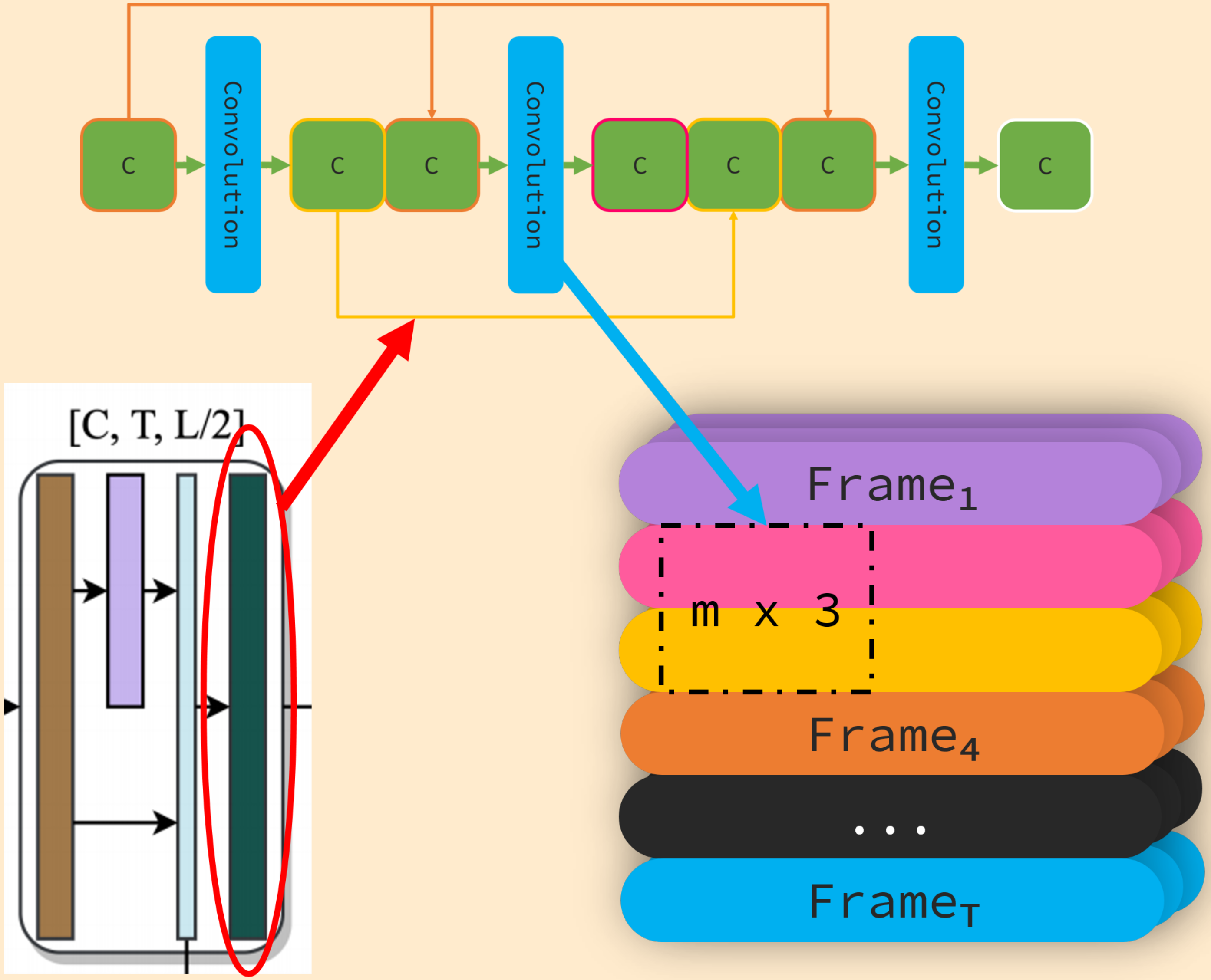


Architecture

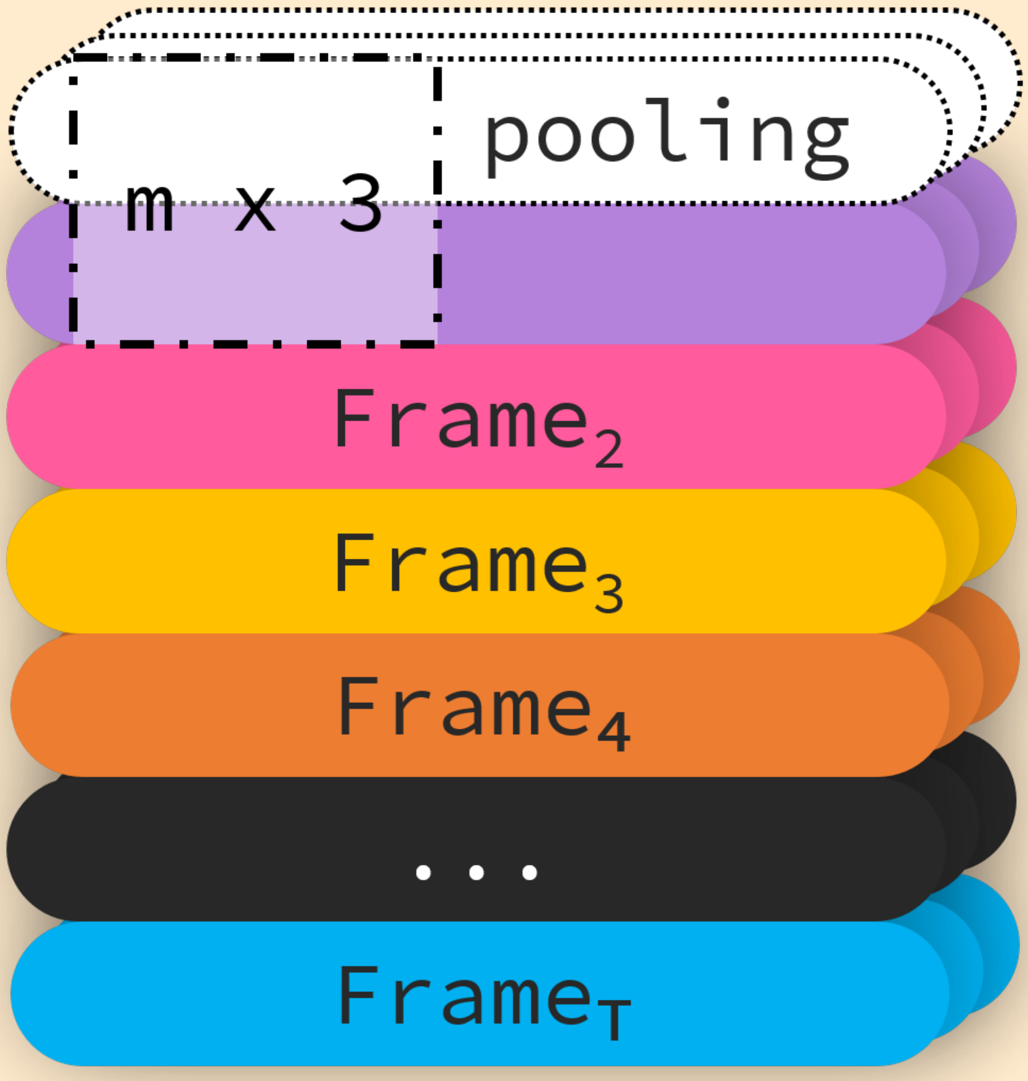
Self Attention Shape



Dense Net Conv



Causal


pooling
m x 3
Frame_2
Frame_3
Frame_4
. . .
Frame_T

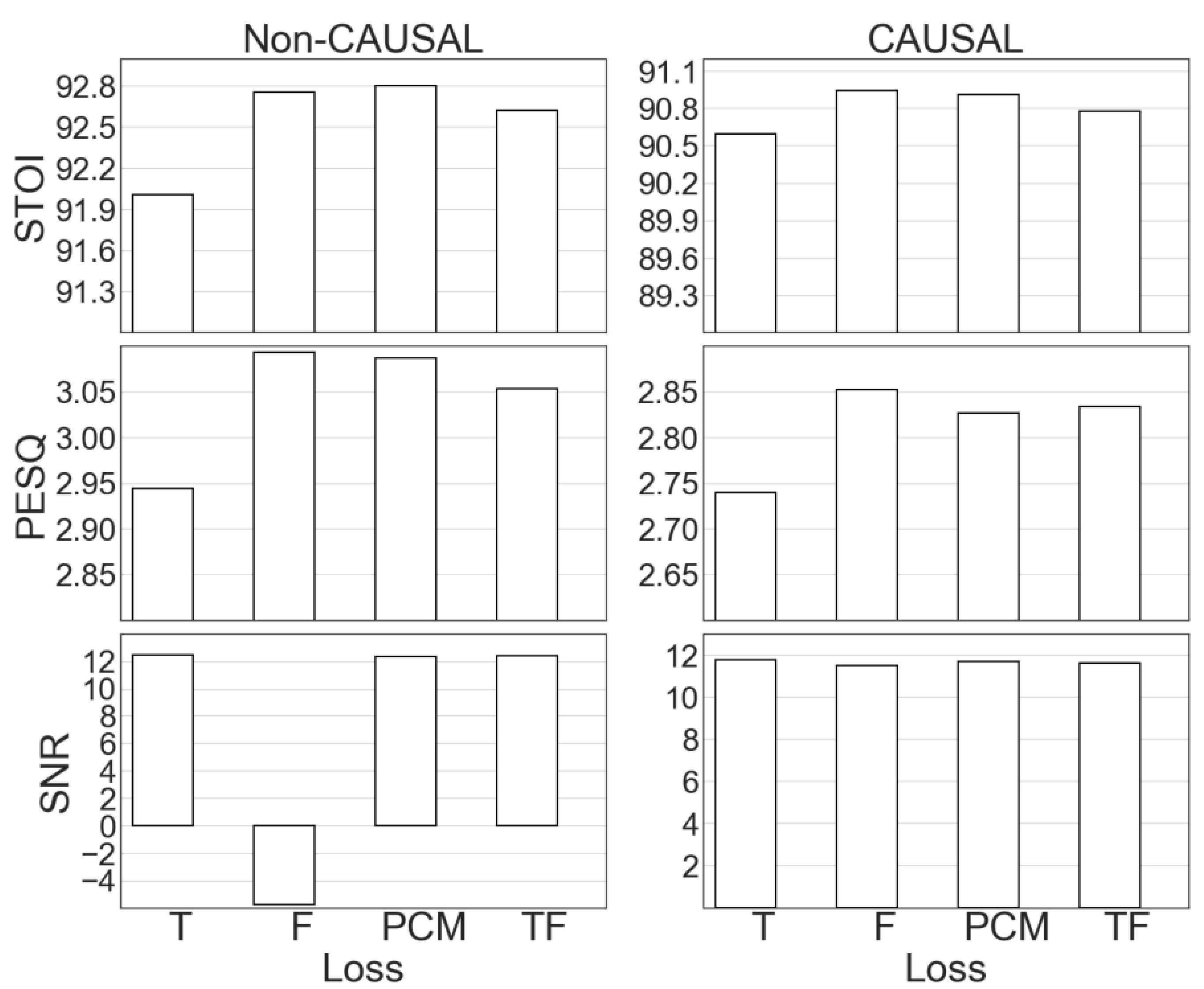
Experiments

- Sample rate：16kHz
- Hamming window
  - size：512
  - stride：256
- Optimizer：Adam

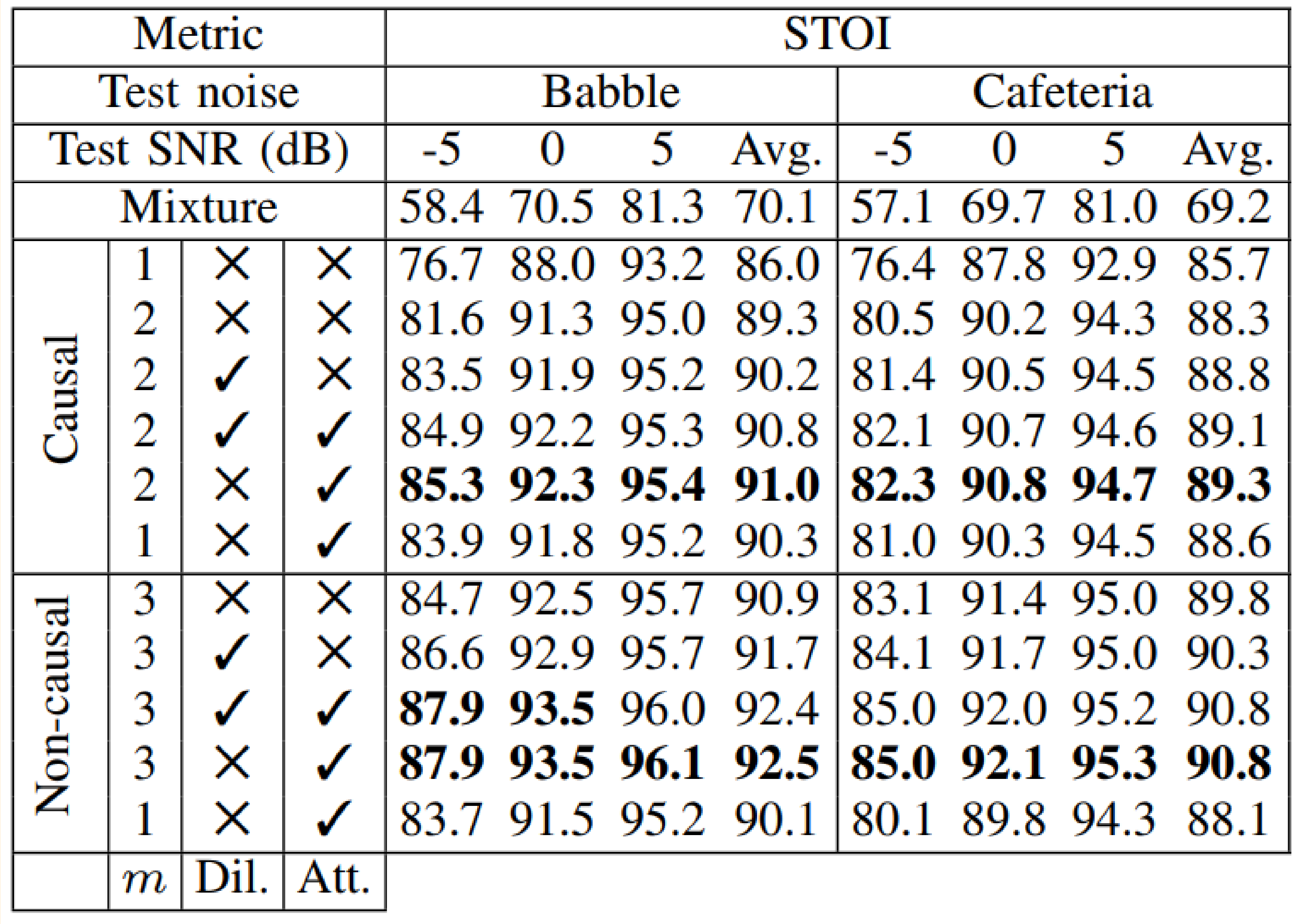
Data Set

- 語音：WSJ0 SI-84 dataset
- 訓練用噪音：10000 non-speech sounds from Sound Ideas
- 測試用噪音：babble and cafeteria noises from an Auditec CD

Experiments

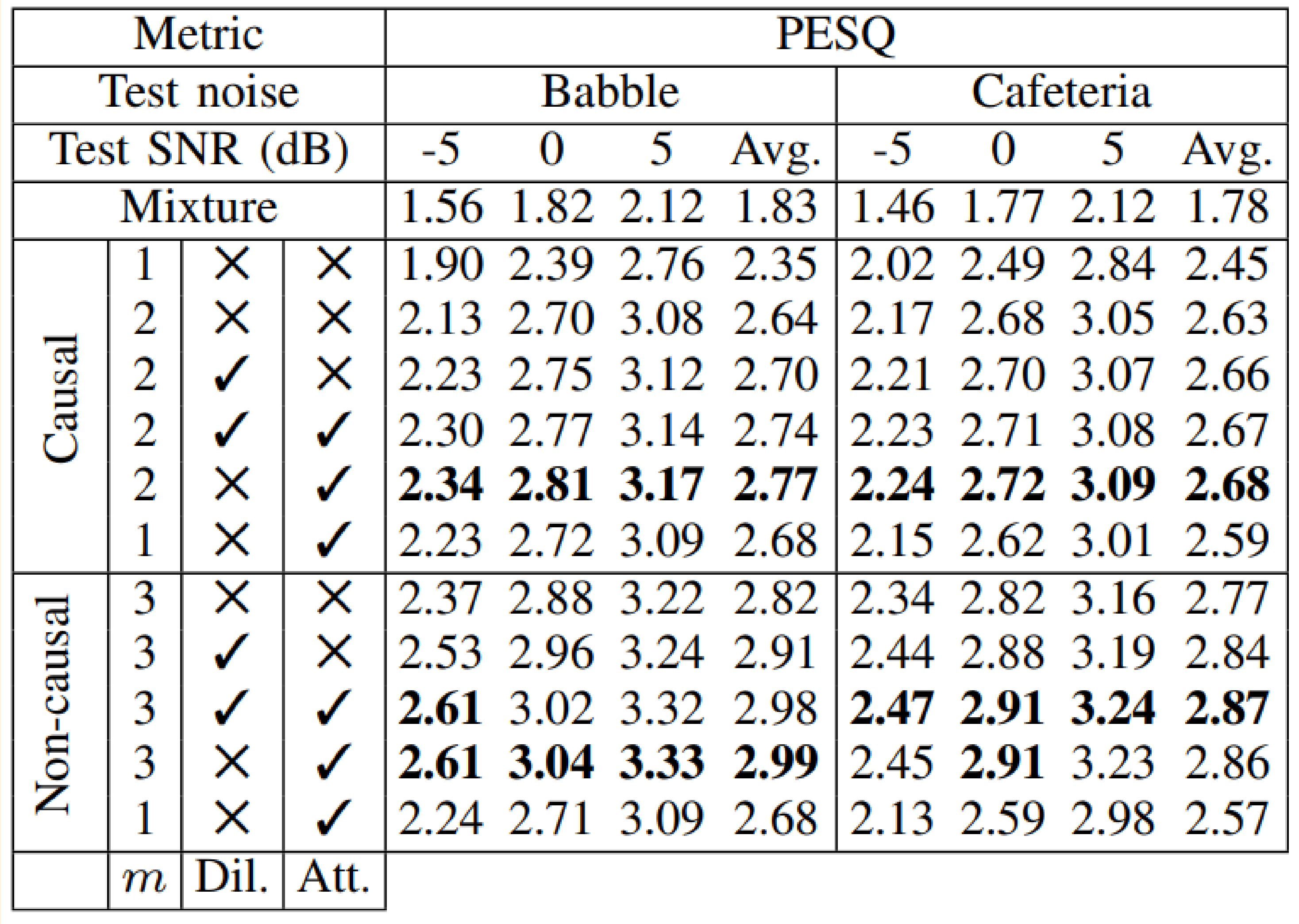


Experiments



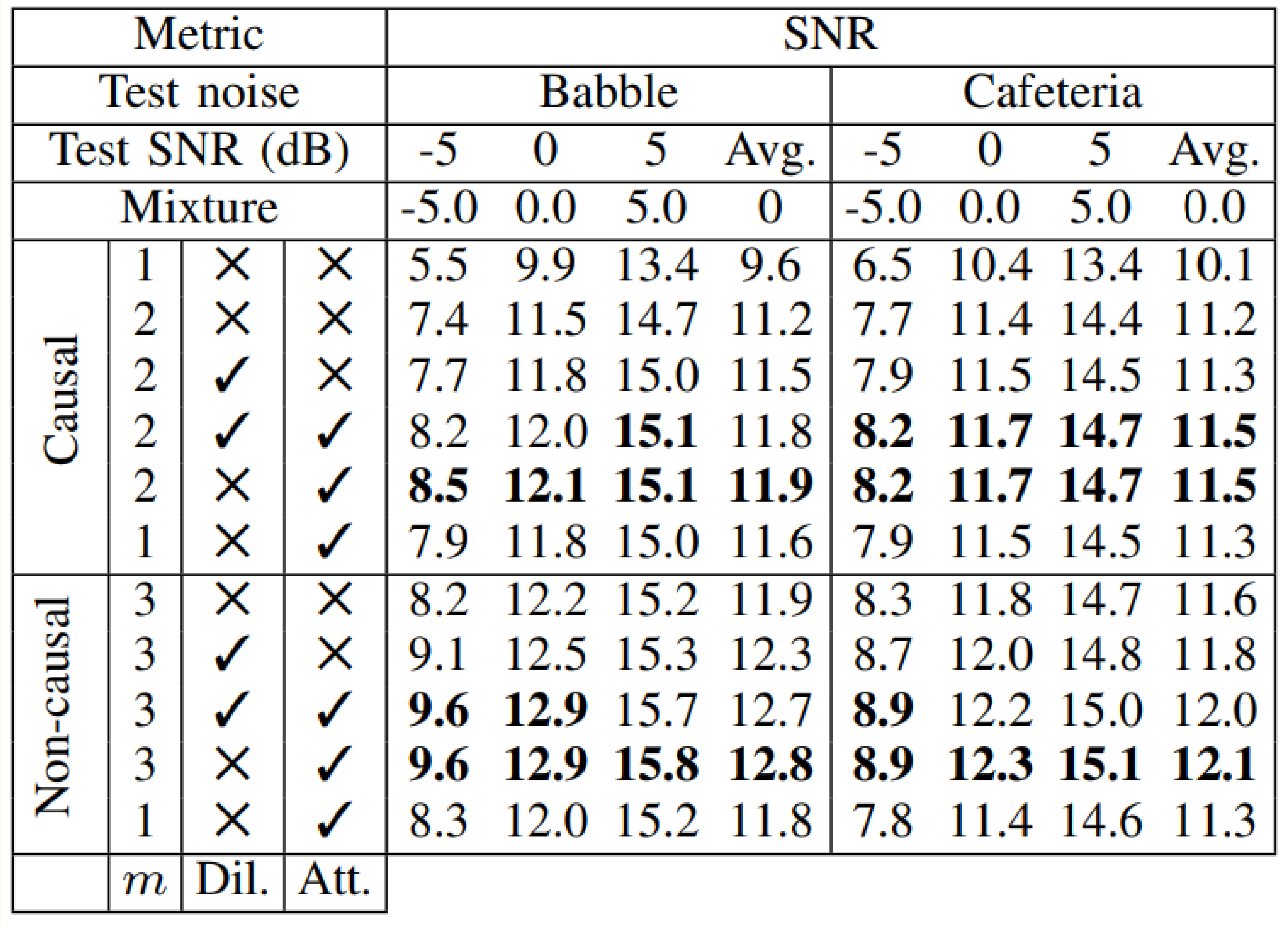
Metric				STOI							
Test noise				Babble				Cafeteria			
Test SNR (dB)				-5	0	5	Avg.	-5	0	5	Avg.
Mixture				58.4	70.5	81.3	70.1	57.1	69.7	81.0	69.2
Causal	1	×	×	76.7	88.0	93.2	86.0	76.4	87.8	92.9	85.7
	2	×	×	81.6	91.3	95.0	89.3	80.5	90.2	94.3	88.3
	2	✓	×	83.5	91.9	95.2	90.2	81.4	90.5	94.5	88.8
	2	✓	✓	84.9	92.2	95.3	90.8	82.1	90.7	94.6	89.1
	2	×	✓	**85.3**	**92.3**	**95.4**	**91.0**	**82.3**	**90.8**	**94.7**	**89.3**
	1	×	✓	83.9	91.8	95.2	90.3	81.0	90.3	94.5	88.6
Non-causal	3	×	×	84.7	92.5	95.7	90.9	83.1	91.4	95.0	89.8
	3	✓	×	86.6	92.9	95.7	91.7	84.1	91.7	95.0	90.3
	3	✓	✓	**87.9**	**93.5**	96.0	92.4	85.0	92.0	95.2	90.8
	3	×	✓	**87.9**	**93.5**	**96.1**	**92.5**	**85.0**	**92.1**	**95.3**	**90.8**
	1	×	✓	83.7	91.5	95.2	90.1	80.1	89.8	94.3	88.1
	m	Dil.	Att.								

Experiments



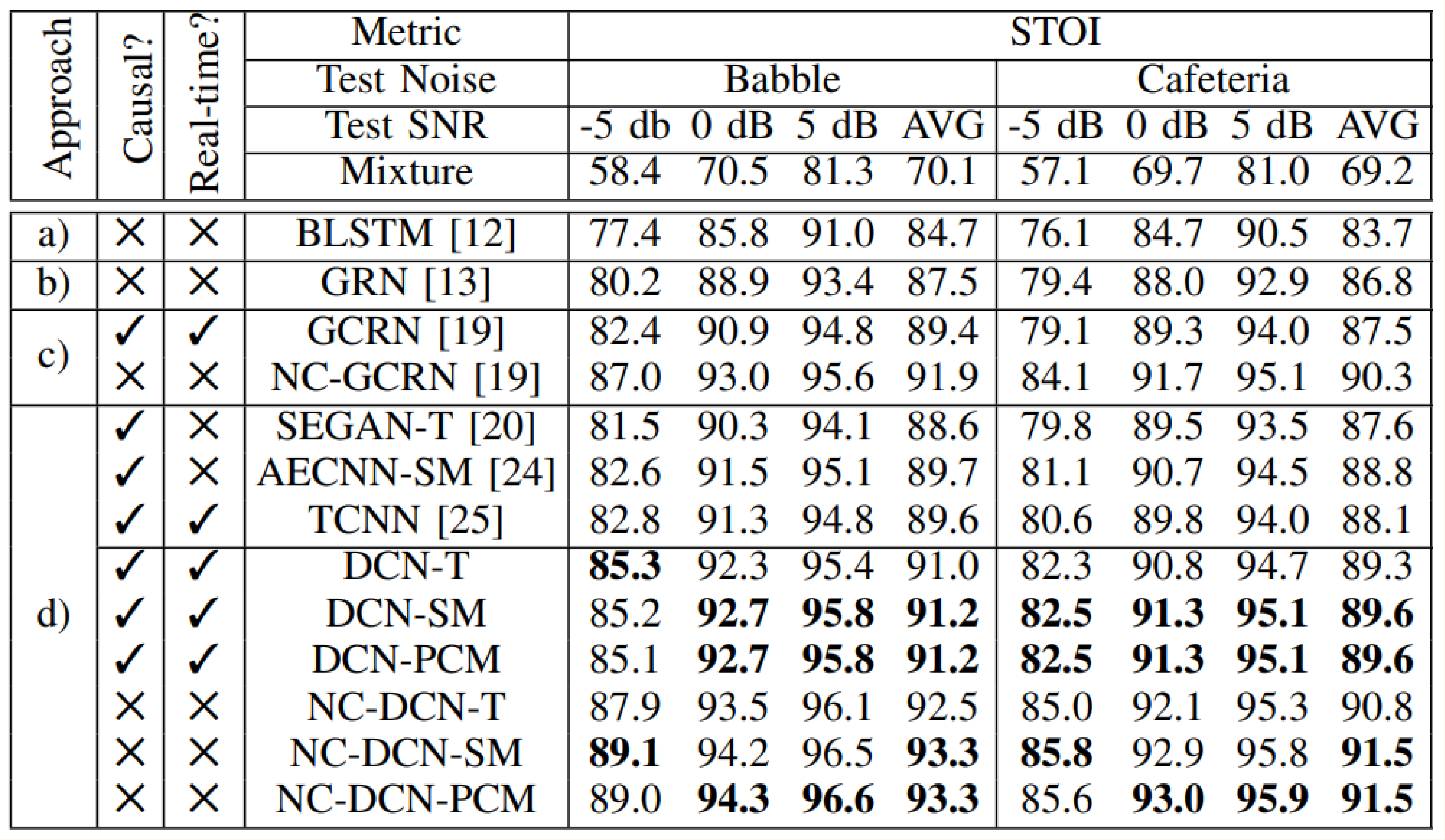
Metric				PESQ							
Test noise				Babble				Cafeteria			
Test SNR (dB)				-5	0	5	Avg.	-5	0	5	Avg.
Mixture				1.56	1.82	2.12	1.83	1.46	1.77	2.12	1.78
Causal	1	✕	✕	1.90	2.39	2.76	2.35	2.02	2.49	2.84	2.45
	2	✕	✕	2.13	2.70	3.08	2.64	2.17	2.68	3.05	2.63
	2	✓	✕	2.23	2.75	3.12	2.70	2.21	2.70	3.07	2.66
	2	✓	✓	2.30	2.77	3.14	2.74	2.23	2.71	3.08	2.67
	2	✕	✓	**2.34**	**2.81**	**3.17**	**2.77**	**2.24**	**2.72**	**3.09**	**2.68**
	1	✕	✓	2.23	2.72	3.09	2.68	2.15	2.62	3.01	2.59
Non-causal	3	✕	✕	2.37	2.88	3.22	2.82	2.34	2.82	3.16	2.77
	3	✓	✕	2.53	2.96	3.24	2.91	2.44	2.88	3.19	2.84
	3	✓	✓	**2.61**	3.02	3.32	2.98	**2.47**	**2.91**	**3.24**	**2.87**
	3	✕	✓	**2.61**	**3.04**	**3.33**	**2.99**	2.45	**2.91**	3.23	2.86
	1	✕	✓	2.24	2.71	3.09	2.68	2.13	2.59	2.98	2.57
	m	Dil.	Att.								

Experiments



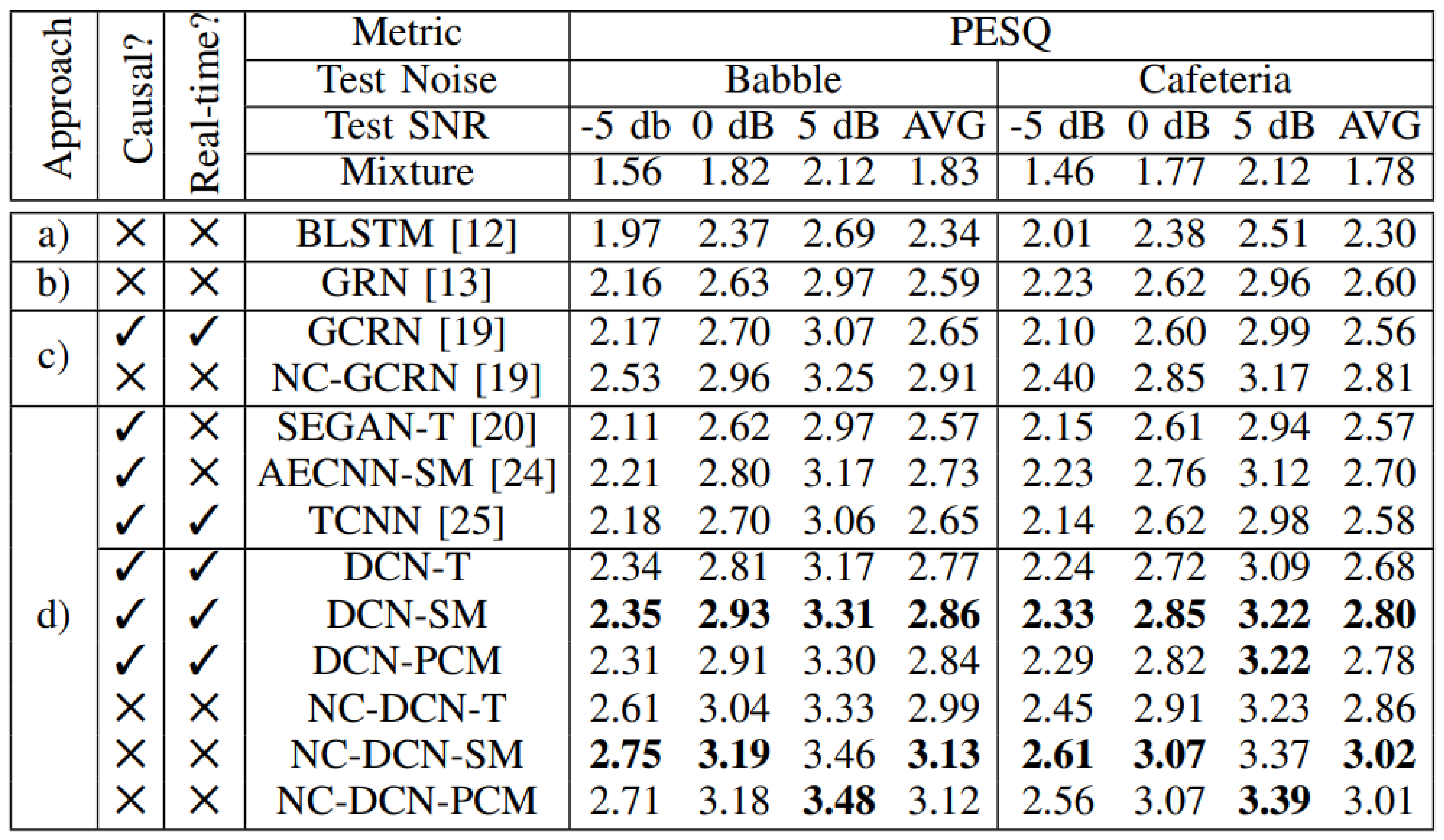
Metric				SNR							
Test noise				Babble				Cafeteria			
Test SNR (dB)				-5	0	5	Avg.	-5	0	5	Avg.
Mixture				-5.0	0.0	5.0	0	-5.0	0.0	5.0	0.0
Causal	1	×	×	5.5	9.9	13.4	9.6	6.5	10.4	13.4	10.1
	2	×	×	7.4	11.5	14.7	11.2	7.7	11.4	14.4	11.2
	2	✓	×	7.7	11.8	15.0	11.5	7.9	11.5	14.5	11.3
	2	✓	✓	8.2	12.0	**15.1**	11.8	**8.2**	**11.7**	**14.7**	**11.5**
	2	×	✓	**8.5**	**12.1**	**15.1**	**11.9**	**8.2**	**11.7**	**14.7**	**11.5**
	1	×	✓	7.9	11.8	15.0	11.6	7.9	11.5	14.5	11.3
Non-causal	3	×	×	8.2	12.2	15.2	11.9	8.3	11.8	14.7	11.6
	3	✓	×	9.1	12.5	15.3	12.3	8.7	12.0	14.8	11.8
	3	✓	✓	**9.6**	**12.9**	15.7	12.7	**8.9**	12.2	15.0	12.0
	3	×	✓	**9.6**	**12.9**	**15.8**	**12.8**	**8.9**	**12.3**	**15.1**	**12.1**
	1	×	✓	8.3	12.0	15.2	11.8	7.8	11.4	14.6	11.3
	m	Dil.	Att.								

Experiments



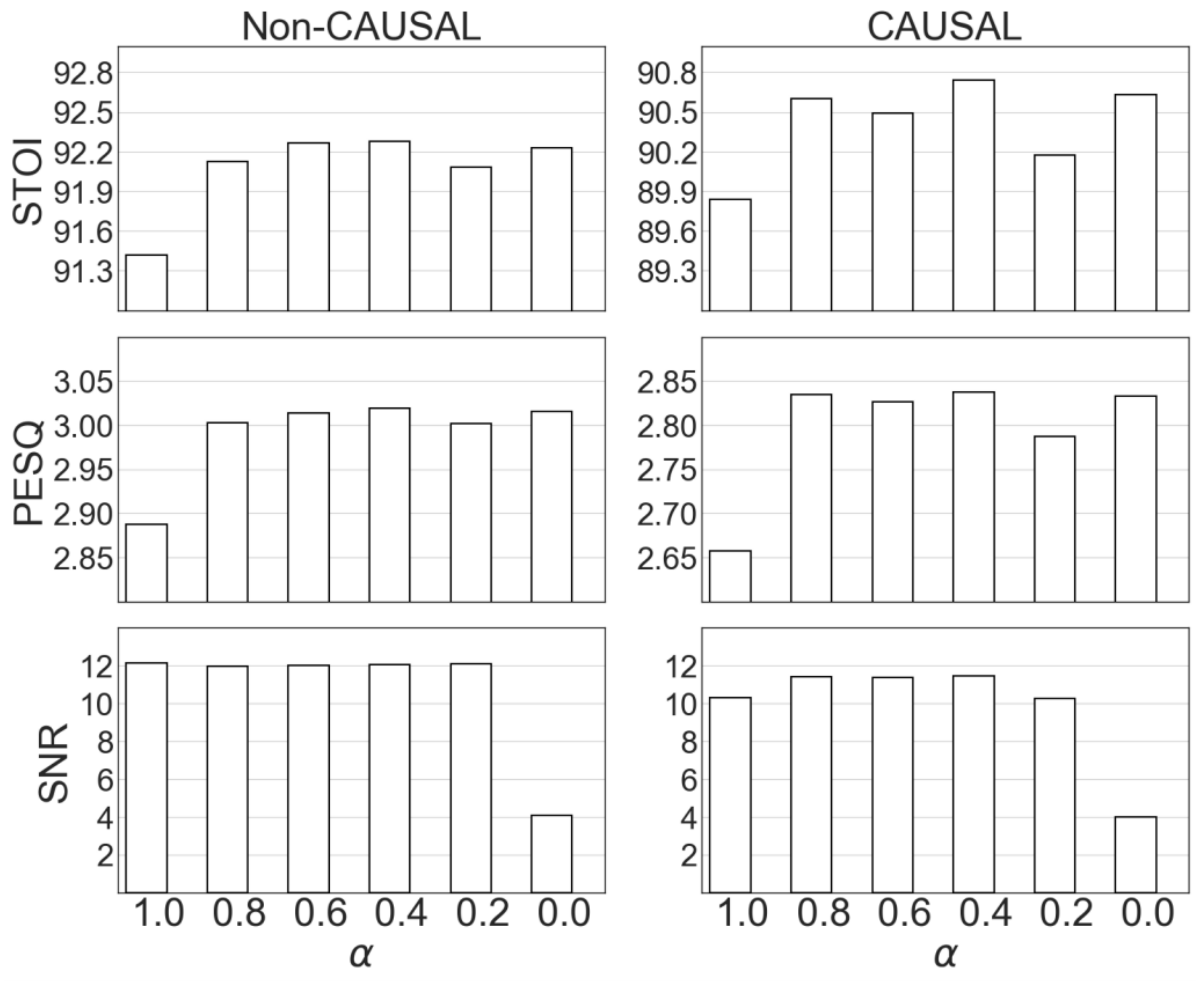
Approach	Causal?	Real-time?	Metric	STOI							
			Test Noise	Babble				Cafeteria			
			Test SNR	-5 db	0 dB	5 dB	AVG	-5 dB	0 dB	5 dB	AVG
			Mixture	58.4	70.5	81.3	70.1	57.1	69.7	81.0	69.2
a)	✗	✗	BLSTM [12]	77.4	85.8	91.0	84.7	76.1	84.7	90.5	83.7
b)	✗	✗	GRN [13]	80.2	88.9	93.4	87.5	79.4	88.0	92.9	86.8
c)	✓	✓	GCRN [19]	82.4	90.9	94.8	89.4	79.1	89.3	94.0	87.5
	✗	✗	NC-GCRN [19]	87.0	93.0	95.6	91.9	84.1	91.7	95.1	90.3
d)	✓	✗	SEGAN-T [20]	81.5	90.3	94.1	88.6	79.8	89.5	93.5	87.6
	✓	✗	AECNN-SM [24]	82.6	91.5	95.1	89.7	81.1	90.7	94.5	88.8
	✓	✓	TCNN [25]	82.8	91.3	94.8	89.6	80.6	89.8	94.0	88.1
	✓	✓	DCN-T	**85.3**	92.3	95.4	91.0	82.3	90.8	94.7	89.3
	✓	✓	DCN-SM	85.2	**92.7**	**95.8**	**91.2**	**82.5**	**91.3**	**95.1**	**89.6**
	✓	✓	DCN-PCM	85.1	**92.7**	**95.8**	**91.2**	**82.5**	**91.3**	**95.1**	**89.6**
	✗	✗	NC-DCN-T	87.9	93.5	96.1	92.5	85.0	92.1	95.3	90.8
	✗	✗	NC-DCN-SM	**89.1**	94.2	96.5	**93.3**	**85.8**	92.9	95.8	**91.5**
	✗	✗	NC-DCN-PCM	89.0	**94.3**	**96.6**	**93.3**	85.6	**93.0**	**95.9**	**91.5**

Experiments



Approach	Causal?	Real-time?	Metric	PESQ							
			Test Noise	Babble				Cafeteria			
			Test SNR	-5 db	0 dB	5 dB	AVG	-5 dB	0 dB	5 dB	AVG
			Mixture	1.56	1.82	2.12	1.83	1.46	1.77	2.12	1.78
a)	✗	✗	BLSTM [12]	1.97	2.37	2.69	2.34	2.01	2.38	2.51	2.30
b)	✗	✗	GRN [13]	2.16	2.63	2.97	2.59	2.23	2.62	2.96	2.60
c)	✓	✓	GCRN [19]	2.17	2.70	3.07	2.65	2.10	2.60	2.99	2.56
	✗	✗	NC-GCRN [19]	2.53	2.96	3.25	2.91	2.40	2.85	3.17	2.81
d)	✓	✗	SEGAN-T [20]	2.11	2.62	2.97	2.57	2.15	2.61	2.94	2.57
	✓	✗	AECNN-SM [24]	2.21	2.80	3.17	2.73	2.23	2.76	3.12	2.70
	✓	✓	TCNN [25]	2.18	2.70	3.06	2.65	2.14	2.62	2.98	2.58
	✓	✓	DCN-T	2.34	2.81	3.17	2.77	2.24	2.72	3.09	2.68
	✓	✓	DCN-SM	**2.35**	**2.93**	**3.31**	**2.86**	**2.33**	**2.85**	**3.22**	**2.80**
	✓	✓	DCN-PCM	2.31	2.91	3.30	2.84	2.29	2.82	**3.22**	2.78
	✗	✗	NC-DCN-T	2.61	3.04	3.33	2.99	2.45	2.91	3.23	2.86
	✗	✗	NC-DCN-SM	**2.75**	**3.19**	3.46	**3.13**	**2.61**	**3.07**	3.37	**3.02**
	✗	✗	NC-DCN-PCM	2.71	3.18	**3.48**	3.12	2.56	3.07	**3.39**	3.01

Experiments



Demo

https://web.cse.ohio-state.edu/~wang.77/pnl/demo/PandeyDCN.html

Conclusion

- This paper proposes a time-domain-based DCN model with a time-frequency loss function to obtain good results in the task of speech enhancement.
- Although SM loss has good results in the evaluation indicators of STOI and PESQ, when judged by human ears, the effect of PCM loss is closer to clean speech.

Conclusion

- The author mentioned that DNN-based speech enhancement methods are not easy to generalize to speech that has not been learned.

- Time domain loss can help improve SNR, and frequency domain loss can improve scores on STOI and PESQ.