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

Ashutosh Pandey,
DeLiang Wang

Outline

- Introduction
- Methodology
- Architecture
- Experiments
- Conclusion

Introduction

當語音受到背景噪音污染時，不只是頻率的大小會受到影響，連同相位也會跟著改變，但是調整相位的風險極大，很有可能會使語音品質變得非常糟。

而從在時域處理訊號時，可以將頻率的大小與相位一同改變，而且比從頻域處理相位更加安全。

因此本篇論文提出了一種結合了 Dense CNN 與 Self Attention 的時域語音增強模型，並使用了對語音及背景音同時約束的新損失函數。

Methodology

U-Net

+

Dense Net

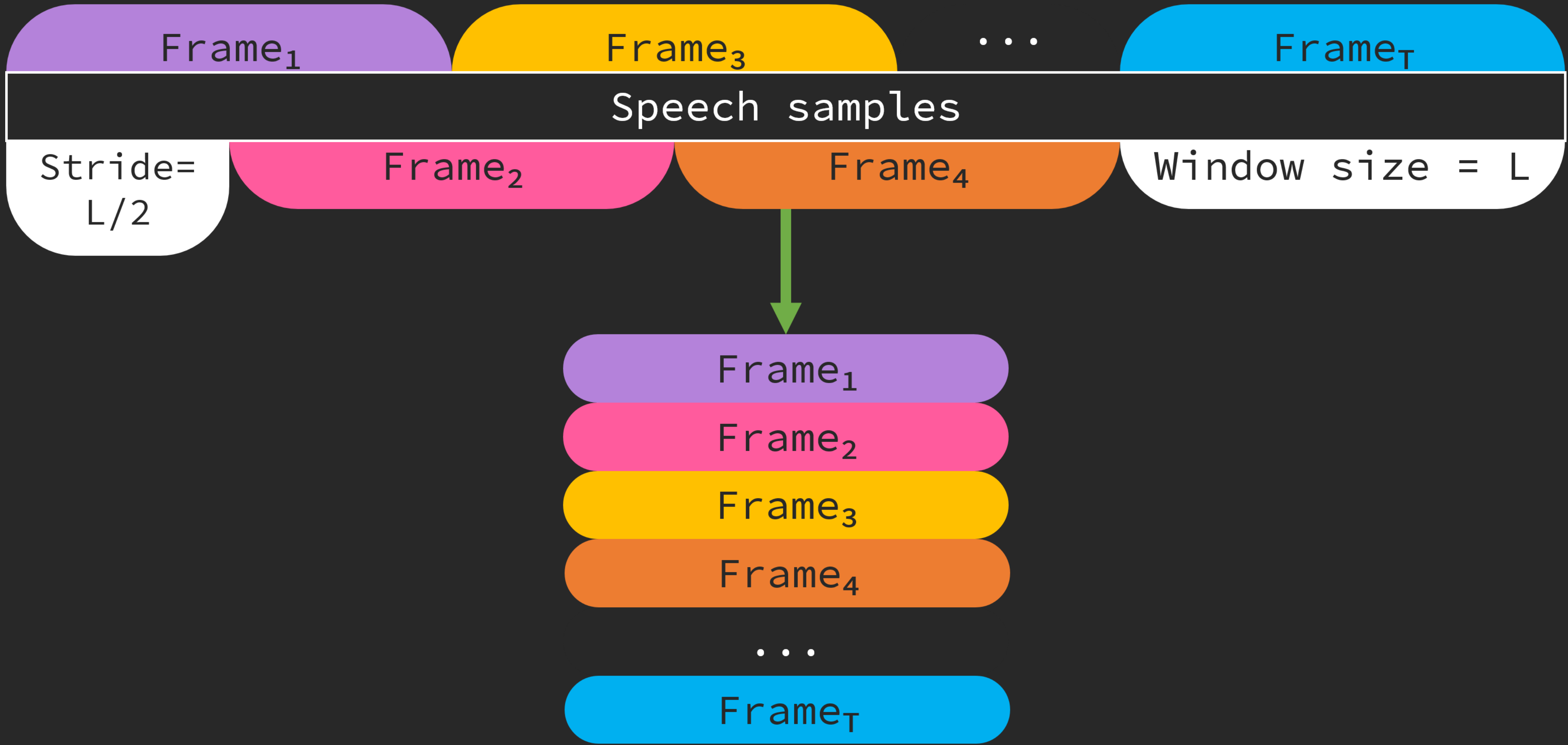
+

Sub-pixel Convolution

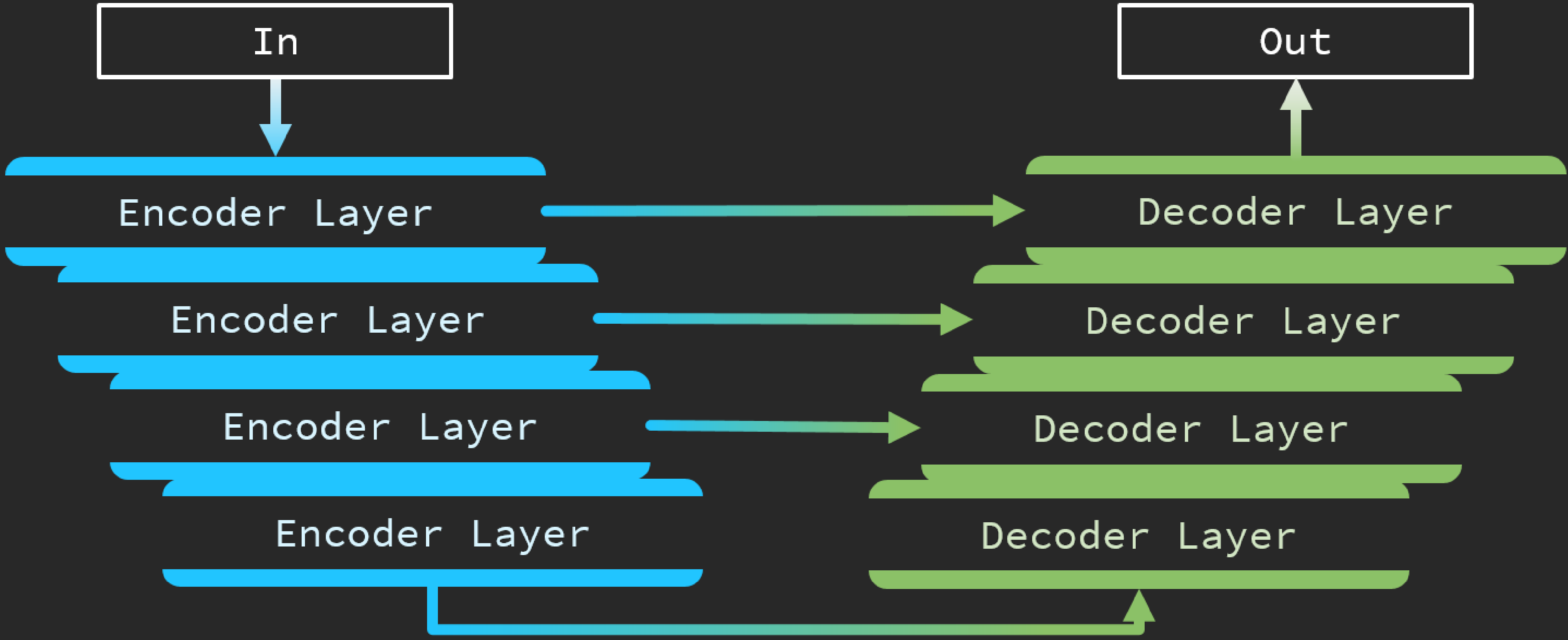
+

Self Attention

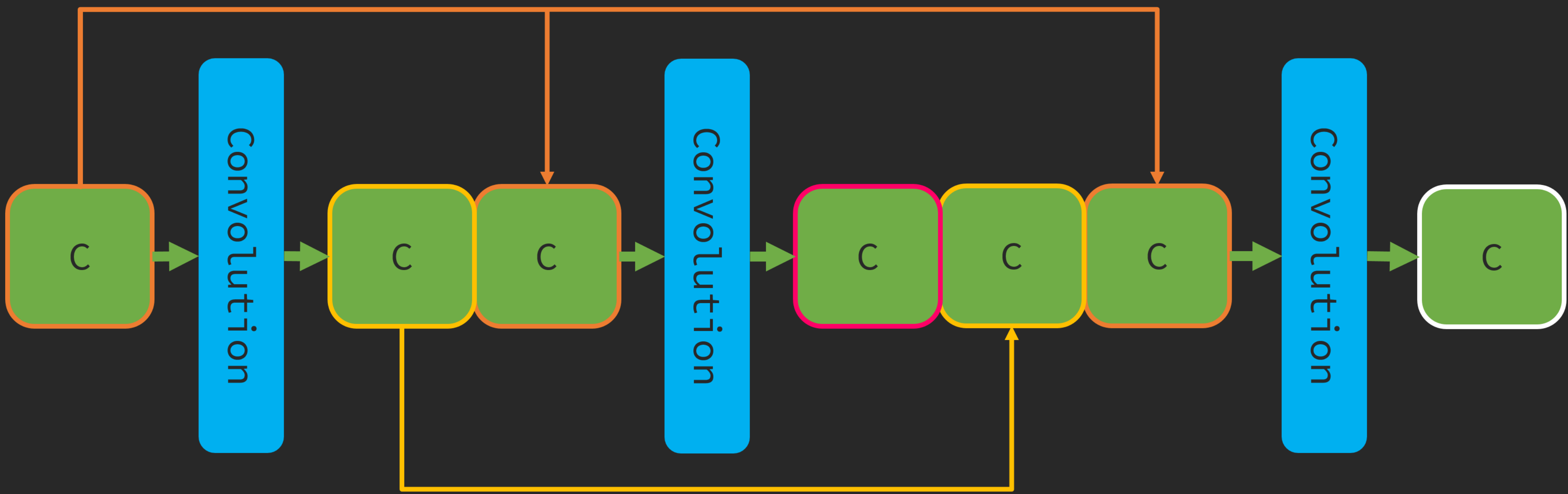
Input


Frame1
Frame3
...
FrameT
Speech samples
Stride=
L/2
Frame2
Frame4
Window size = L
Frame1
Frame2
Frame3
Frame4
...
FrameT

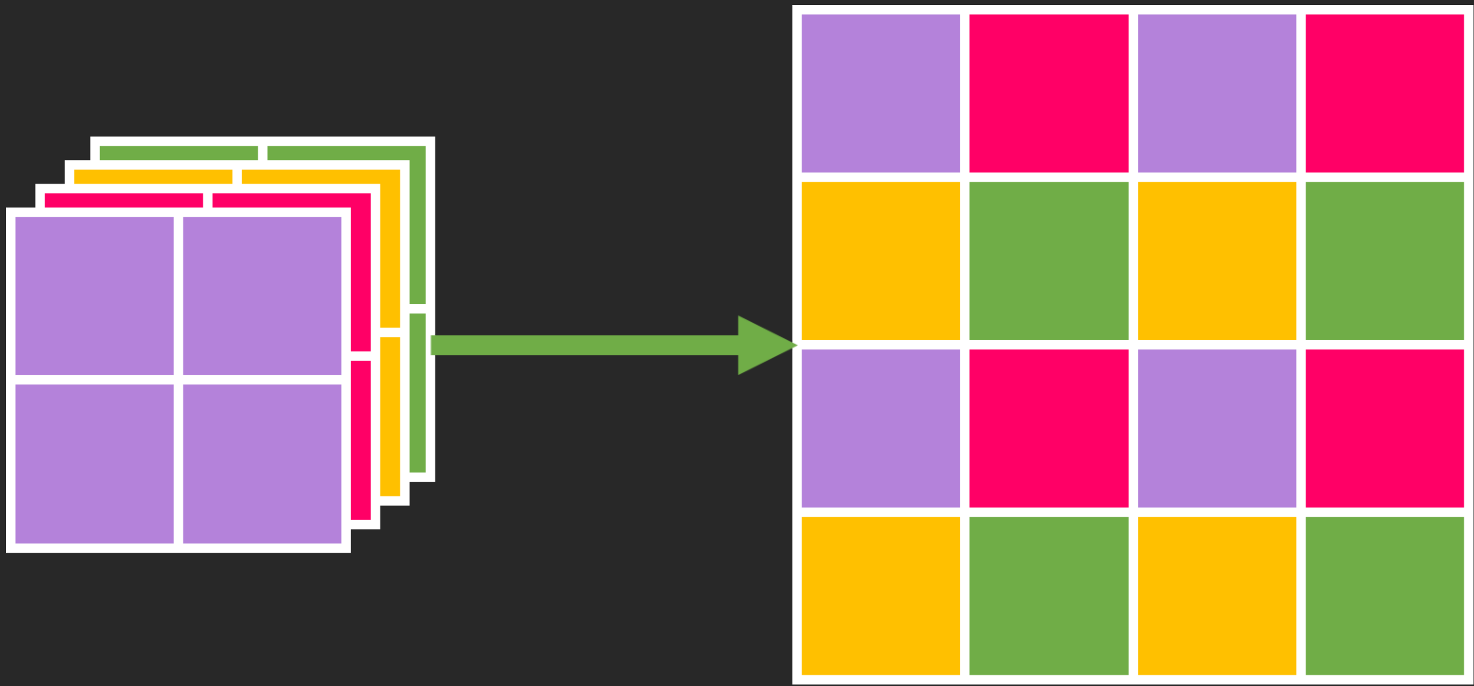
U-Net


In
Out
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Layer
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

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

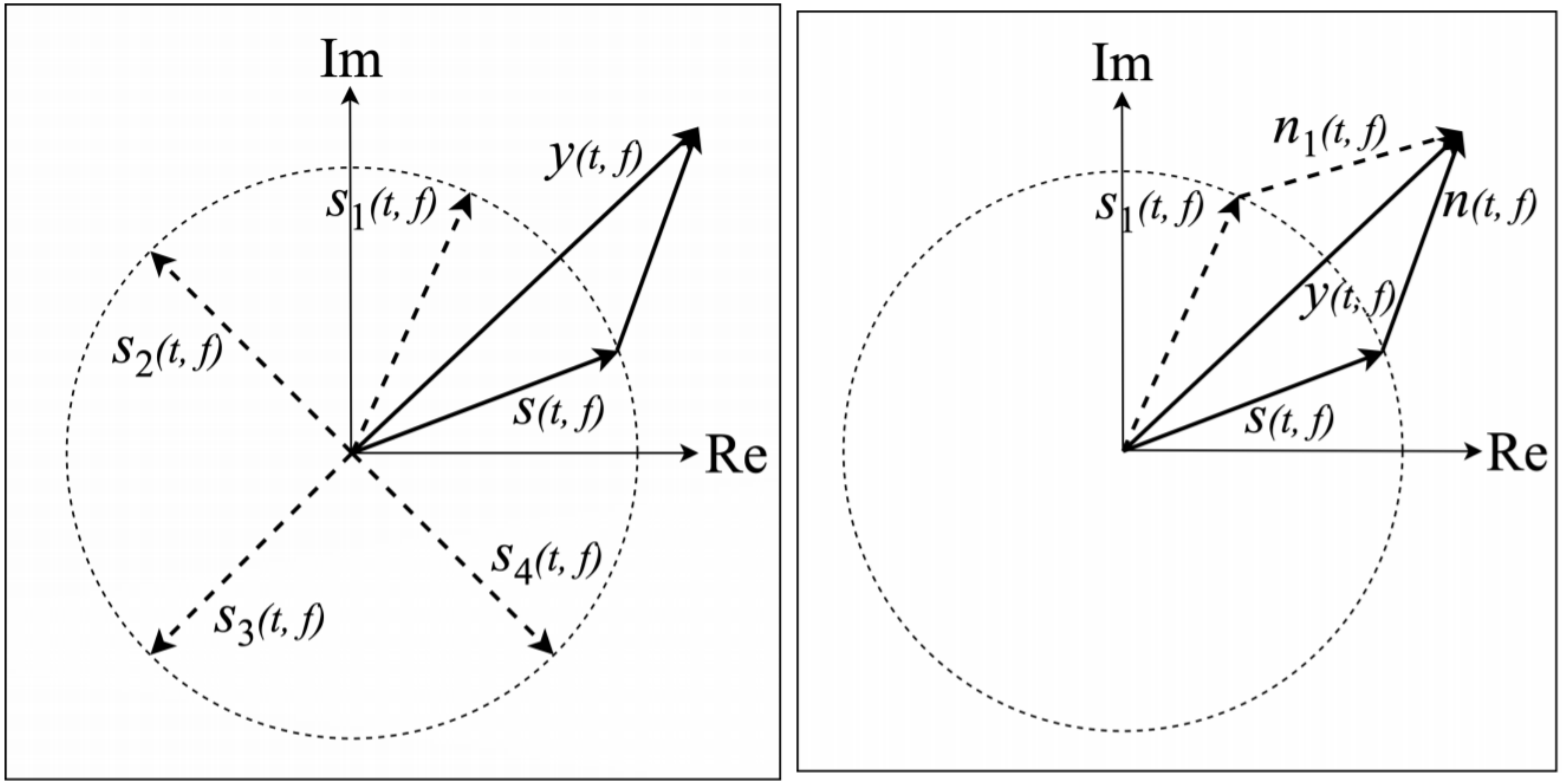
- Time-frequency Loss

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

- Phase Constrained Magnitude Loss

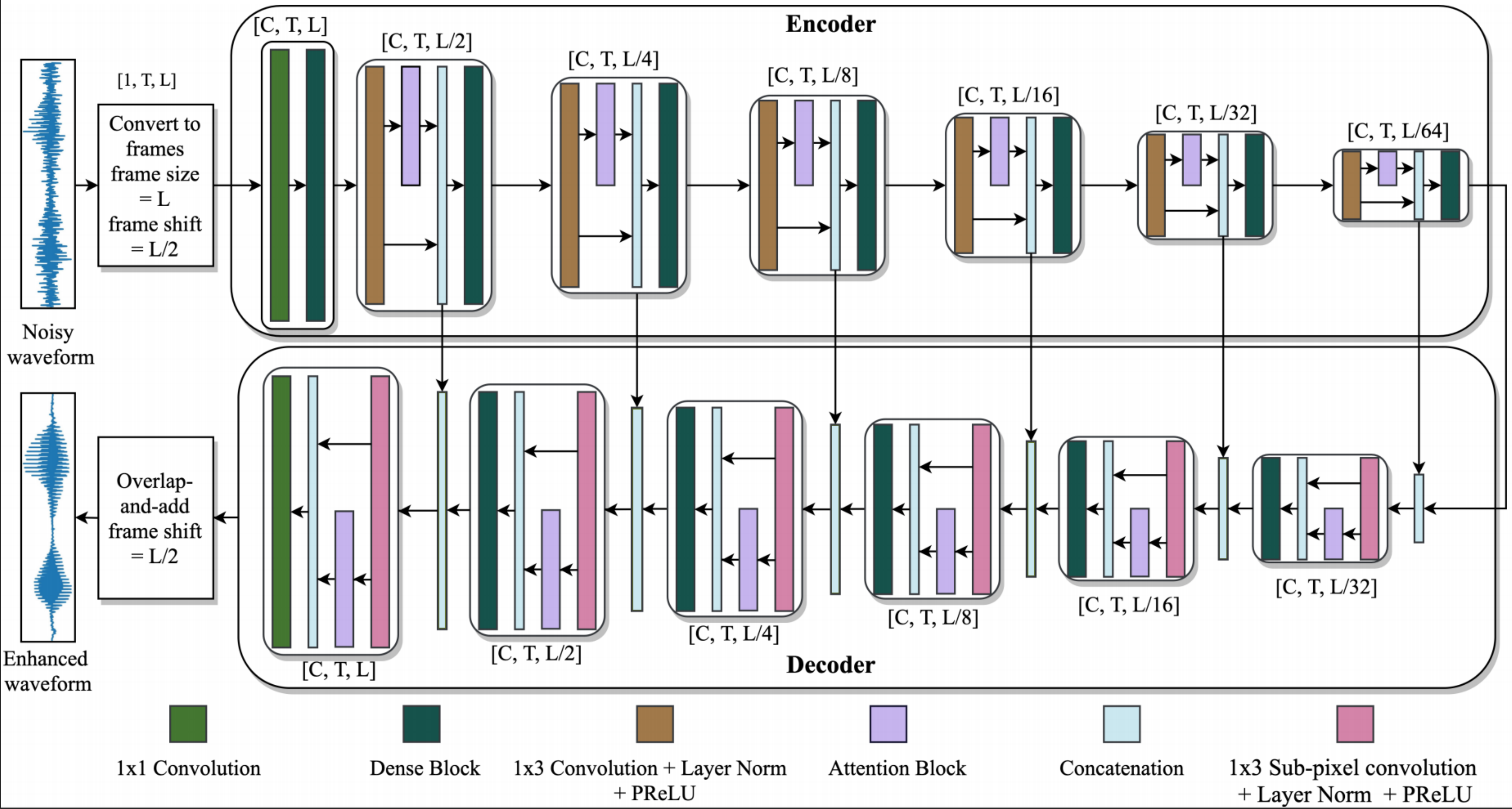
$$\mathcal{L}_{PCM}(s,\hat{s}) = 0.5\mathcal{L}_{SM}(s,\hat{s}) + 0.5\mathcal{L}_{SM}(n, x-\hat{s})$$

PCM Loss

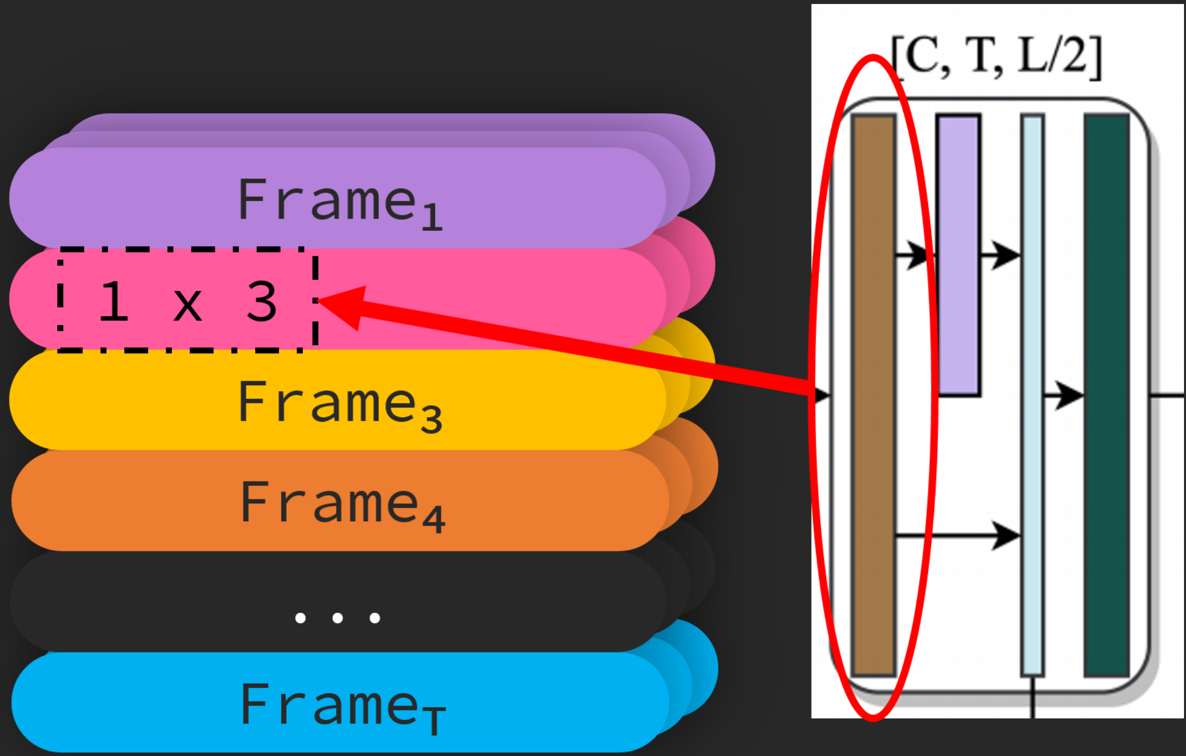


(a) L_{SM} (b) L_{PCM}

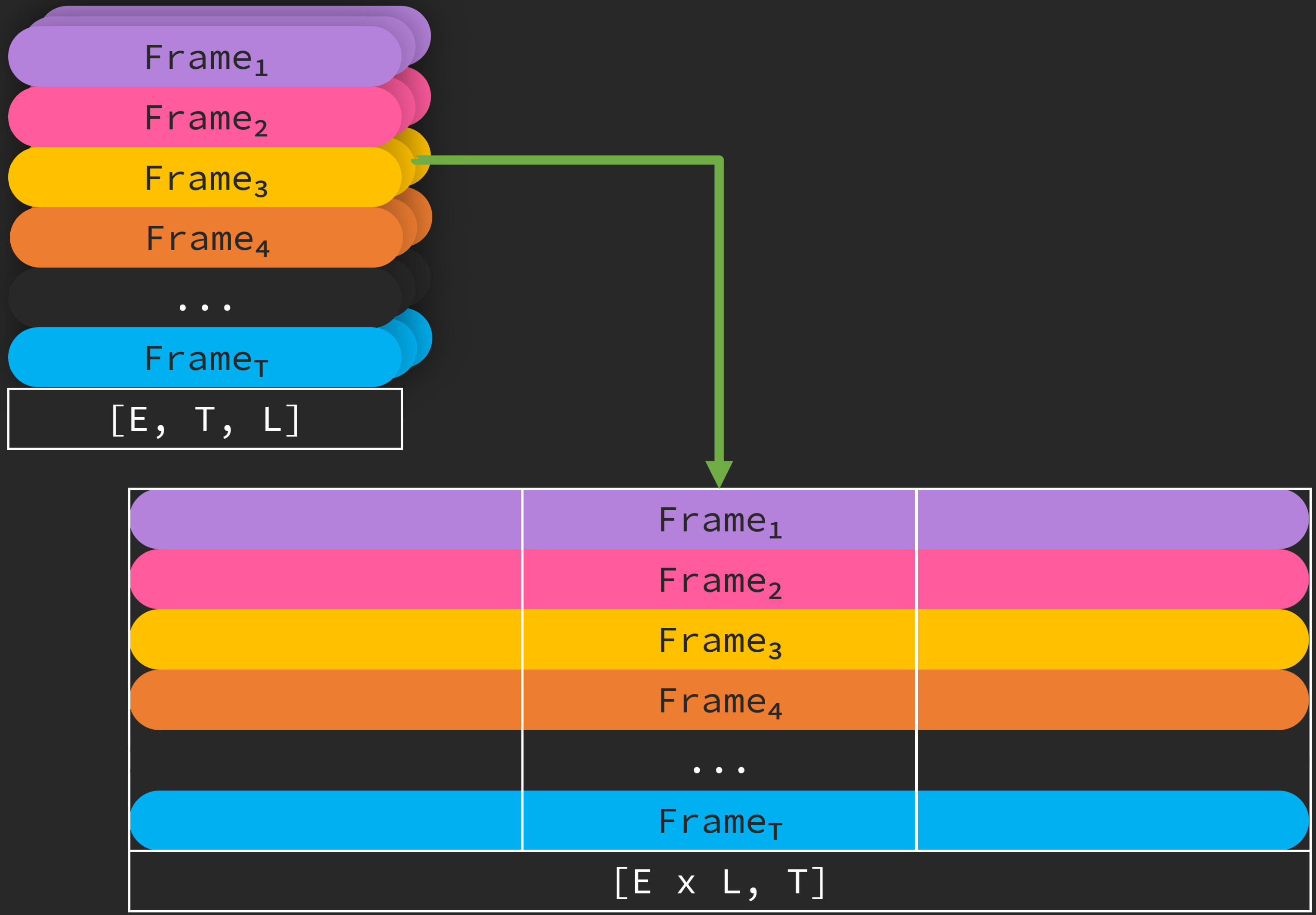
Architecture



1 x 3 Conv

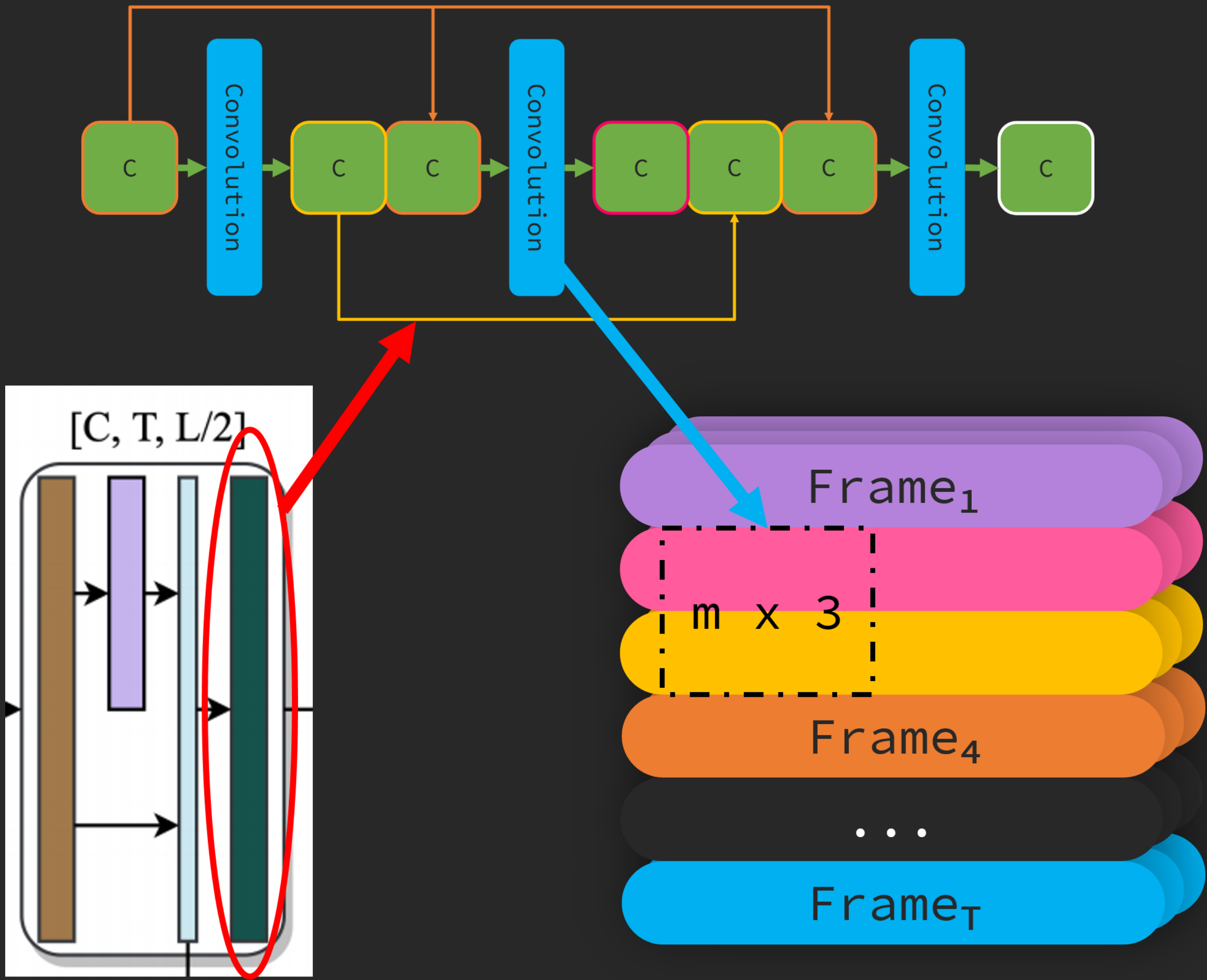


Self Attention Shape

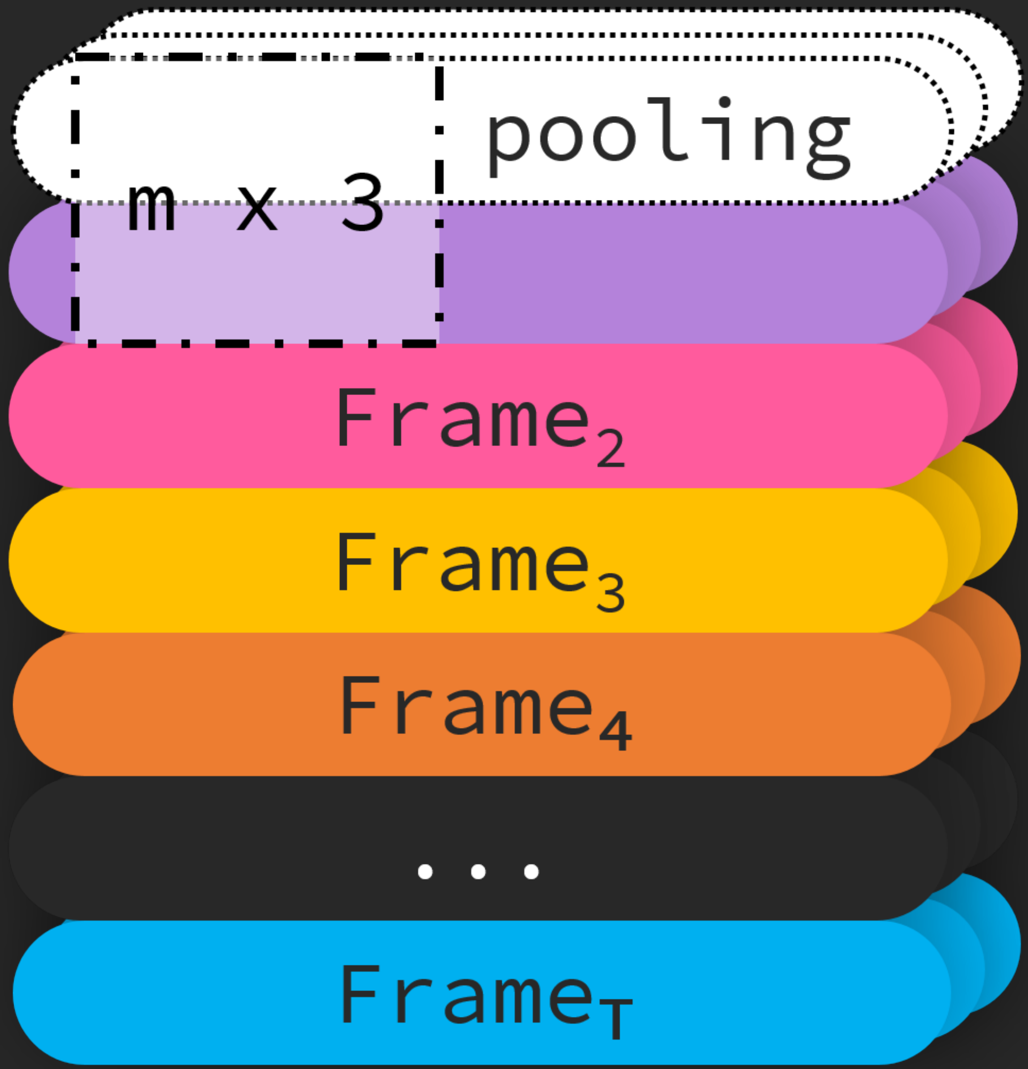


Self Attention Shape

Dense Net Conv



Causal



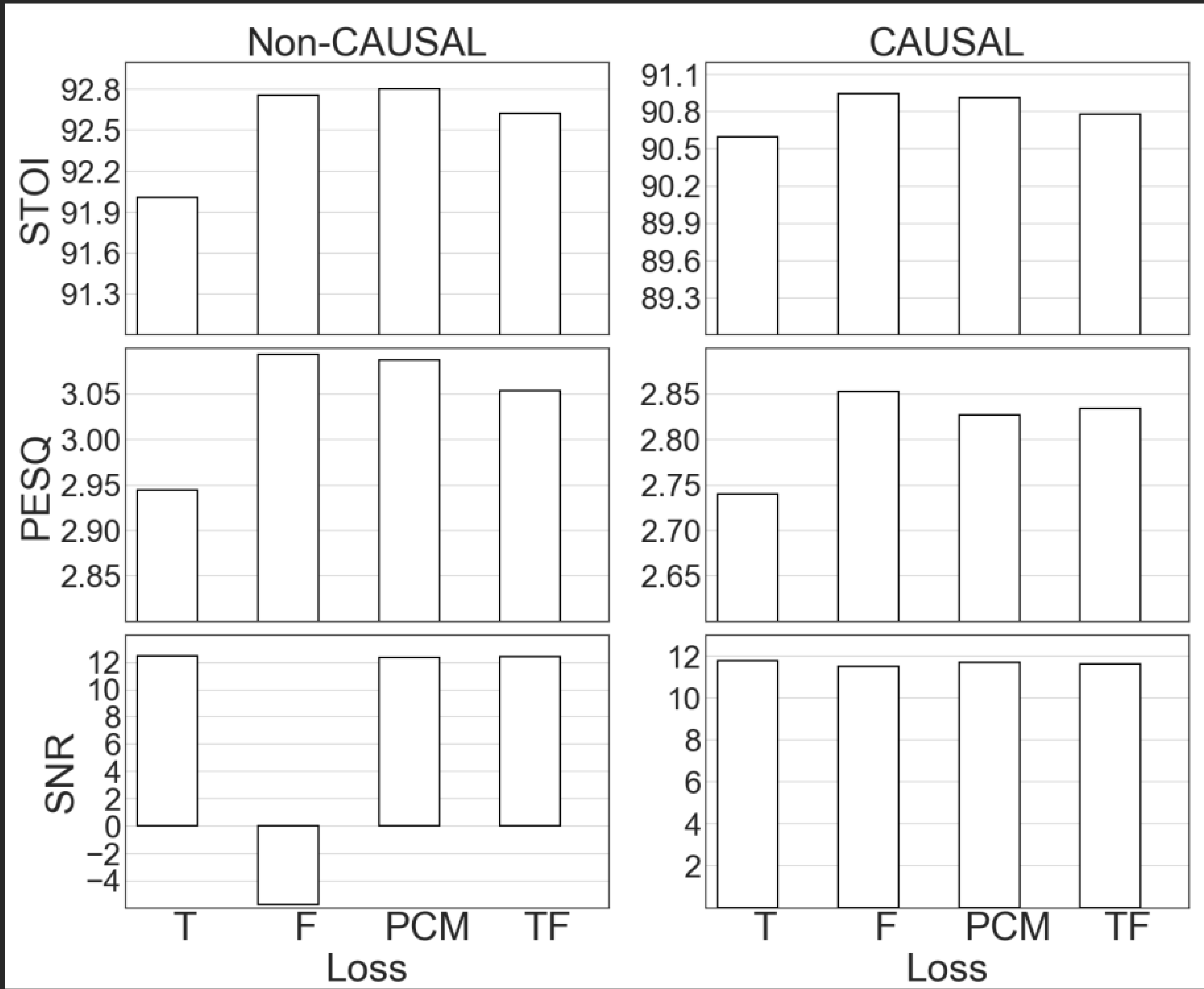
Experiments

- 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



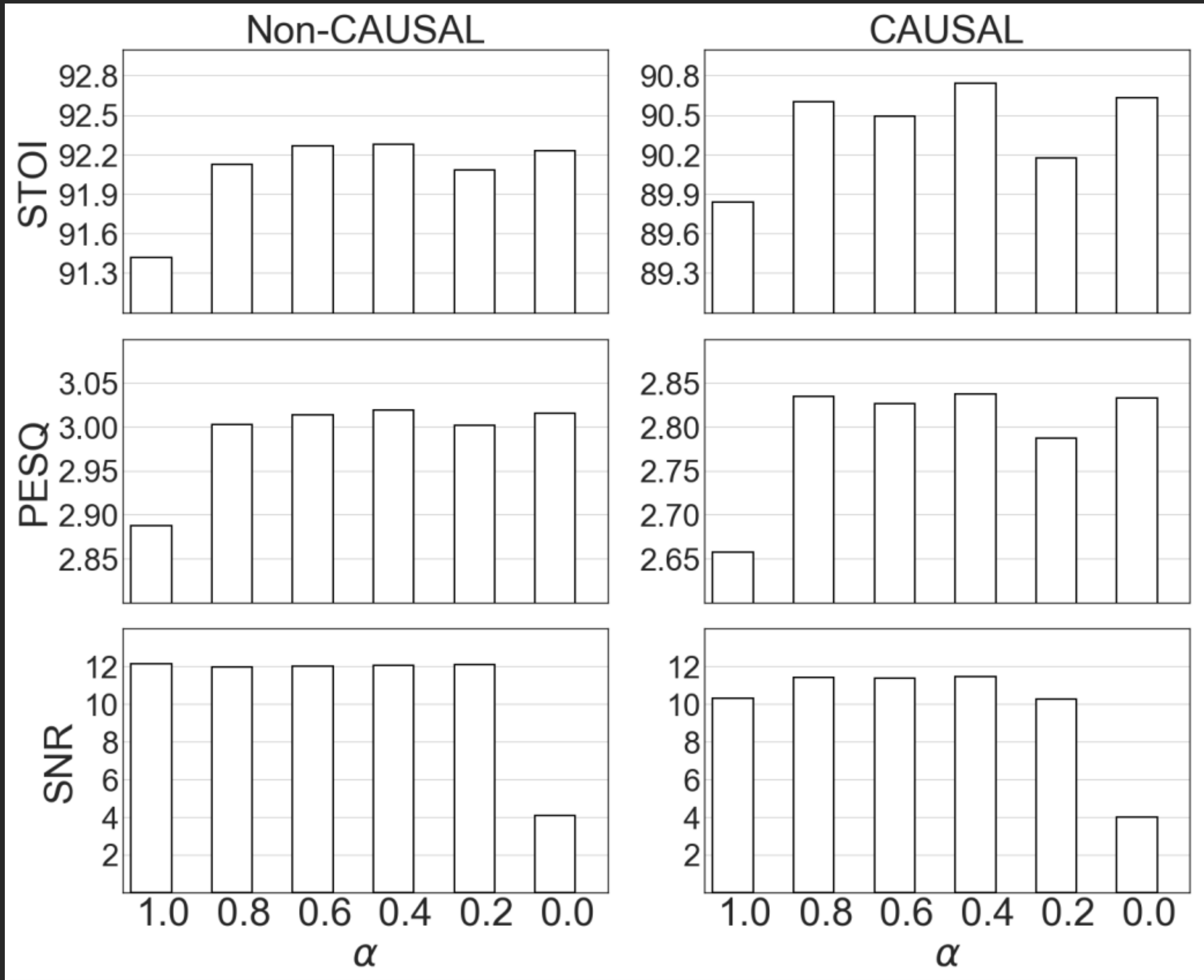
Experiments

Metric				STOI								PESQ								SNR							
Test noise				Babble				Cafeteria				Babble				Cafeteria				Babble				Cafeteria			
Test 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.
Mixture				58.4	70.5	81.3	70.1	57.1	69.7	81.0	69.2	1.56	1.82	2.12	1.83	1.46	1.77	2.12	1.78	-5.0	0.0	5.0	0	-5.0	0.0	5.0	0.0
Causal	1	×	×	76.7	88.0	93.2	86.0	76.4	87.8	92.9	85.7	1.90	2.39	2.76	2.35	2.02	2.49	2.84	2.45	5.5	9.9	13.4	9.6	6.5	10.4	13.4	10.1
	2	×	×	81.6	91.3	95.0	89.3	80.5	90.2	94.3	88.3	2.13	2.70	3.08	2.64	2.17	2.68	3.05	2.63	7.4	11.5	14.7	11.2	7.7	11.4	14.4	11.2
	2	✓	×	83.5	91.9	95.2	90.2	81.4	90.5	94.5	88.8	2.23	2.75	3.12	2.70	2.21	2.70	3.07	2.66	7.7	11.8	15.0	11.5	7.9	11.5	14.5	11.3
	2	✓	✓	84.9	92.2	95.3	90.8	82.1	90.7	94.6	89.1	2.30	2.77	3.14	2.74	2.23	2.71	3.08	2.67	8.2	12.0	**15.1**	11.8	**8.2**	**11.7**	**14.7**	**11.5**
	2	×	✓	**85.3**	**92.3**	**95.4**	**91.0**	**82.3**	**90.8**	**94.7**	**89.3**	**2.34**	**2.81**	**3.17**	**2.77**	**2.24**	**2.72**	**3.09**	**2.68**	**8.5**	**12.1**	**15.1**	**11.9**	**8.2**	**11.7**	**14.7**	**11.5**
	1	×	✓	83.9	91.8	95.2	90.3	81.0	90.3	94.5	88.6	2.23	2.72	3.09	2.68	2.15	2.62	3.01	2.59	7.9	11.8	15.0	11.6	7.9	11.5	14.5	11.3
Non-causal	3	×	×	84.7	92.5	95.7	90.9	83.1	91.4	95.0	89.8	2.37	2.88	3.22	2.82	2.34	2.82	3.16	2.77	8.2	12.2	15.2	11.9	8.3	11.8	14.7	11.6
	3	✓	×	86.6	92.9	95.7	91.7	84.1	91.7	95.0	90.3	2.53	2.96	3.24	2.91	2.44	2.88	3.19	2.84	9.1	12.5	15.3	12.3	8.7	12.0	14.8	11.8
	3	✓	✓	**87.9**	**93.5**	96.0	92.4	85.0	92.0	95.2	90.8	**2.61**	3.02	3.32	2.98	**2.47**	**2.91**	**3.24**	**2.87**	**9.6**	**12.9**	15.7	12.7	**8.9**	12.2	15.0	12.0
	3	×	✓	**87.9**	**93.5**	**96.1**	**92.5**	**85.0**	**92.1**	**95.3**	**90.8**	**2.61**	**3.04**	**3.33**	**2.99**	2.45	**2.91**	3.23	2.86	**9.6**	**12.9**	**15.8**	**12.8**	**8.9**	**12.3**	**15.1**	**12.1**
	1	×	✓	83.7	91.5	95.2	90.1	80.1	89.8	94.3	88.1	2.24	2.71	3.09	2.68	2.13	2.59	2.98	2.57	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								PESQ							
			Test Noise	Babble				Cafeteria				Babble				Cafeteria			
			Test SNR	-5 db	0 dB	5 dB	AVG	-5 dB	0 dB	5 dB	AVG	-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	1.56	1.82	2.12	1.83	1.46	1.77	2.12	1.78
a)	✗	✗	BLSTM [12]	77.4	85.8	91.0	84.7	76.1	84.7	90.5	83.7	1.97	2.37	2.69	2.34	2.01	2.38	2.51	2.30
b)	✗	✗	GRN [13]	80.2	88.9	93.4	87.5	79.4	88.0	92.9	86.8	2.16	2.63	2.97	2.59	2.23	2.62	2.96	2.60
c)	✓	✓	GCRN [19]	82.4	90.9	94.8	89.4	79.1	89.3	94.0	87.5	2.17	2.70	3.07	2.65	2.10	2.60	2.99	2.56
	✗	✗	NC-GCRN [19]	87.0	93.0	95.6	91.9	84.1	91.7	95.1	90.3	2.53	2.96	3.25	2.91	2.40	2.85	3.17	2.81
d)	✓	✗	SEGAN-T [20]	81.5	90.3	94.1	88.6	79.8	89.5	93.5	87.6	2.11	2.62	2.97	2.57	2.15	2.61	2.94	2.57
	✓	✗	AECNN-SM [24]	82.6	91.5	95.1	89.7	81.1	90.7	94.5	88.8	2.21	2.80	3.17	2.73	2.23	2.76	3.12	2.70
	✓	✓	TCNN [25]	82.8	91.3	94.8	89.6	80.6	89.8	94.0	88.1	2.18	2.70	3.06	2.65	2.14	2.62	2.98	2.58
	✓	✓	DCN-T	**85.3**	92.3	95.4	91.0	82.3	90.8	94.7	89.3	2.34	2.81	3.17	2.77	2.24	2.72	3.09	2.68
	✓	✓	DCN-SM	85.2	**92.7**	**95.8**	**91.2**	**82.5**	**91.3**	**95.1**	**89.6**	**2.35**	**2.93**	**3.31**	**2.86**	**2.33**	**2.85**	**3.22**	**2.80**
	✓	✓	DCN-PCM	85.1	**92.7**	**95.8**	**91.2**	**82.5**	**91.3**	**95.1**	**89.6**	2.31	2.91	3.30	2.84	2.29	2.82	**3.22**	2.78
	✗	✗	NC-DCN-T	87.9	93.5	96.1	92.5	85.0	92.1	95.3	90.8	2.61	3.04	3.33	2.99	2.45	2.91	3.23	2.86
	✗	✗	NC-DCN-SM	**89.1**	94.2	96.5	**93.3**	**85.8**	92.9	95.8	**91.5**	**2.75**	**3.19**	3.46	**3.13**	**2.61**	**3.07**	3.37	**3.02**
	✗	✗	NC-DCN-PCM	89.0	**94.3**	**96.6**	**93.3**	85.6	**93.0**	**95.9**	**91.5**	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

- 本篇論文提出基於時域的 DCN 模型並搭配時頻的損失函數在語音增強的任務中獲得了良好的成果。
- 雖然在 STOI 與 PESQ 的評估指標上，SM loss 具有較好的結果，但在實際由人耳評斷時 PCM loss 更接近乾淨的語音。
- 作者提到，基於 DNN 的語音增強方法不易泛化到未曾學習過的資料上面。
- 時域的 loss 有助於提升 SNR、頻域 loss 則能使 STOI 與 PESQ 上的分數提升。