DLHLP HW3 Report

組長 Githud ID: wubinary
組員: 徐均筑 R08921040
吳彬睿 R08942087

1. (5%)請記錄 evaluate.log 裡面的SiSNR 數值，和當時所用的 hyperparameter(這一題請3-1不用PIT, 3-2用PIT)
   - 3-1 without PIT

     ```
     Average SDR improvement: 17.34
     Average SISNR improvement: 17.09
     ```

   - 3-2 with PIT

     ```
     Average SDR improvement: 10.98
     Average SISNR improvement: 10.46
     ```

   - Hyperparameter

	3-1 without PIT	**3-2 with PIT**
N (# of filters)	128	256
L (lenght of filters)	40	20
B (# of channels in bottlneck)	128	256
H (# of channels in conv)	256	512
P (kernel size)	3	3
X (# of conv block)	7	8
R (# of reapeats)	1	4
norm_type	gLN	gLN
causal	0	0
mask_nonlinear	relu	relu
C	2	2
Epoch	100	14
Si-SNR	**17.09**	**10.46**

2. (5%)嘗試調整不同的hyperparameter，比較其差異，並試著分析結果(至少針對2種不同的hyperparameter進行實驗)

- The experiments below are under closed-condition (CC) (seen speaker) without PIT and open-condition (OC) with PIT.

- Repeat

N	128	128	128
L	40	40	40
B	128	128	128
H	256	256	256
P	3	3	3
X	7	7	7
R	1	2	3
Si-SNR (CC)	**17.71**	**19.32**	**20.41**
Si-SNR (OC)	**7.21**	**8.91**	**9.71**

- The number of filters in autoencoder

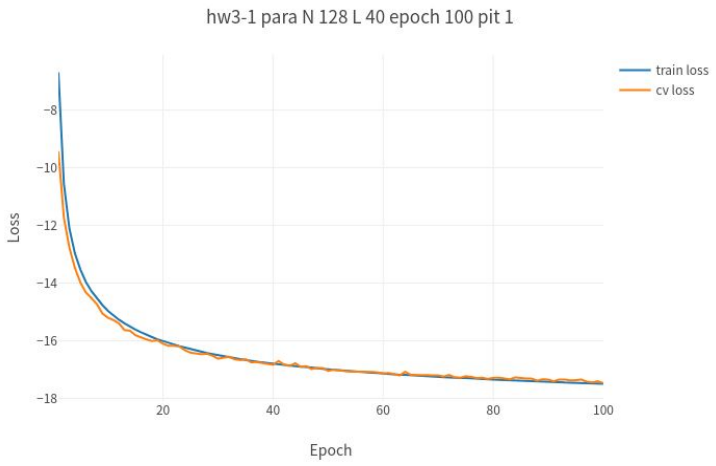
N	64	128	256	512
L	40	40	40	40
B	128	128	128	128
H	256	256	256	256
P	3	3	3	3
X	7	7	7	7
R	1	1	1	1
Si-SNR (CC)	**15**	**17.71**	**18.12**	**18.54**
Si-SNR (OC)	**6.64**	**7.21**	**7.53**	**7.69**

- 分析

  從數據中可以發現 R和 N 越大通常 performance 越佳，但 model size 也隨之增大。另外通常有 saturation 的現象存在，當 model size 到達一定程度時，提升 R 或 N 所帶來的 Si-SNR 的提升會逐漸趨緩。

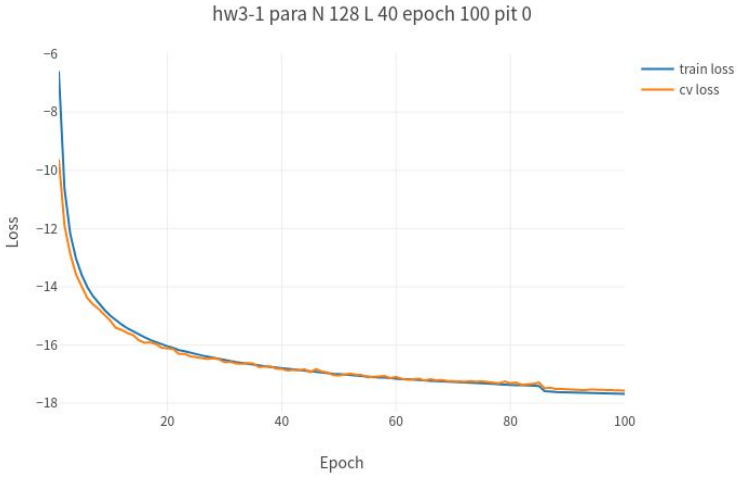
3. (3%)3-1, 3-2請分別試看看有無PIT的差異並記錄結果(loss learning curve, Si-SNR)

- 3-1 with PIT



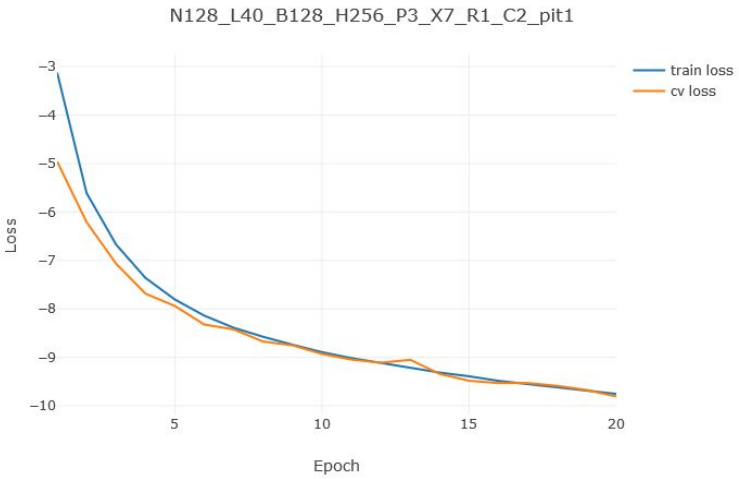
Average Si-SNR: **17.69**

- 3-1 without PIT



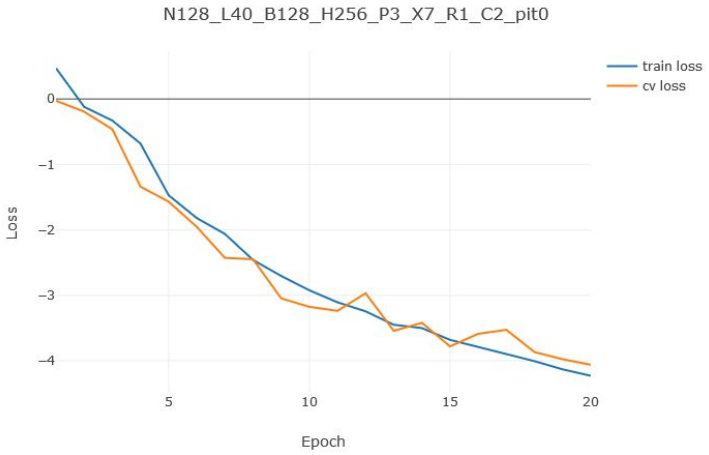
Average Si-SNR: **17.71**

- 3-2 with PIT



Average Si-SNR: **7.21**

- 3-2 without PIT



Average Si-SNR: **1.81**

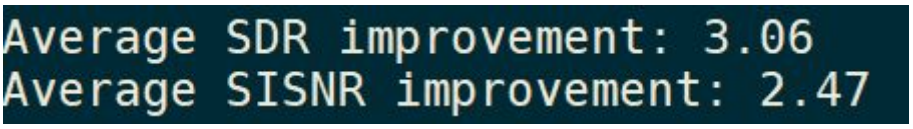
4. (2%)思考一下為何有無PIT會影響3-1, 3-2的結果並寫下你的看法

在不同的 frame 之間最佳的 output-speaker assignment 可能有所改變，因此若沒有使用 PIT 會造成較差的 Si-SNR，特別在 [1] 之中有提到在相同性別以及 output window size 蠻小的時候，有無使用 PIT 的效果會很明顯。另外相較於 open condition (HW3-2)，closed condition (HW3-1) 的效果反而差了些許，原因在於測試時是以訓練所使用的 speaker，在無使用 PIT 的情況下 model 能更好的針對這兩個 speaker 去調整，因此得到的 model 表現在已知的 speaker 較佳，但卻犧牲了 generalization。

- Bonus(2%) :

  請自己找兩段音訊合起來(請不要使用作業給的data)測看看是否能成功分離，上傳音訊(含原音檔、合成後音檔及經過model分離的音檔)，紀錄Si-SNR於report中，並給出至少一種improve Si-SNR的方法(調參數除外)。

  結果:

  Improve Si-SNR 方法 :

  將訓練過的 Conv-Tasnet 視作 pretrained autoencoder，只取 encoder 的部分，利用 Deep Clustering 的方法去進行訓練，得到更好的 embedding，最後再根據此 embedding 做 K-means Clustering，得到最後的 mask。

References

[1] D. Yu, M. Kolbk, Z.-H. Tan, J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation", Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 241-245, 2017

[2] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks", IEEE/ACM Transactions on Audio, Speech,and Language Processing, vol. 25, no. 10, pp. 1901–1913, 2017