# Contrastive Learning for Speech Enhancement

郭品辰 F14066127 NCKU SNAME
黃仁鴻 P76094169 NCKU CSIE

Table 1. Contrastive Learning 方法比較。

| CL Method | Batch Size | Negative Pairs | Momentum Encoder |
|---|---|---|---|
| SimCLR | 4096 | √ | × |
| MoCo v2 | 256 | √ | √ |
| **BYOL** | 256~4096 | × | √ |
| SwAV | 4096 | × | × |
| **SimSiam** | 256 | × | × |

## 1. Introduction

許多與日常生活息息相關的任務都是仰賴語音作爲資訊傳遞的媒介，像是電話通訊、語音辨識與助聽器等等。然而在現實環境中充滿著各種不可預期的噪音干擾，這嚴重影響了語音訊號技術的效能。因此，將這些雜訊去除的語音增強技術就成了對語音任務很重要的前置處理單元。

而語音增強的問題就是不論在何種噪音環境，面對相同的語音，模型都能夠抽取出相同的特徵，進而利用與特徵來重構語音。這部分想法與近年流行自監督方法中的對比學習 (Contrastive Learning) 不謀而合，對比學習希望相似樣本間的特徵編碼能越像越好，而不同樣本的特徵差異則是越大越好。

我們認爲，藉由 CL 的方法來學習語音特徵的潛在編碼，再利用此特徵編碼還原語音，應該會具備比一般深度學習的語音增強方法更高的通用性與泛化能力。因此，本次專題研究的主軸便是將對比學習應用於語音增強任務的可行性。

## 2. Related works

Table.1 對幾種常見的 Contrastive Learning 方法進行比較。包含 SimCLR [1] 在內的多數 CL 方法都需要使用到大量的負樣本輔助進行訓練，否則會發生 collapsing output 的問題，也就是不論輸入任何東西都只會有相同輸出。而大量的負樣本需求導致這些方法必須使用極大的 batch size 才能獲得良好效果，例如在 SimCLR 的論文中，實驗所使用的 batch size 就到達了 4096。此外，Speech Enhancement 問題本身亦並不容易訂定 Frame Level 的負樣本。爲了因應硬體資源與 SE 任務的限制，本次研究使用 BYOL[2] 與 SimSiam[3] 這兩項不需要極大 Batch Size，同時也不需負樣本的 Contrastive Learning 架構，其架構如 Fig.1 與 Fig.2 所示[1]。

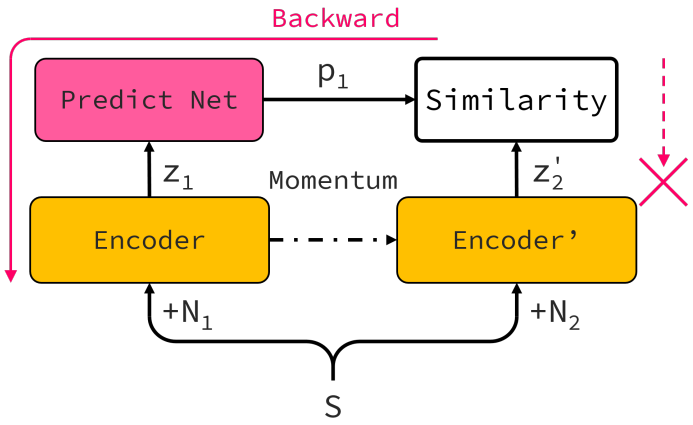


Figure 1. BYOL 的算法架構。Encoder-PredictNet 會將 Encoder' 取出特徵做爲目標， 計算 L2 Loss 後更新 Encoder-PredictNet 的參數， Encoder' 則會把自身目前參數與更新後 Encoder 做加權平均當新參數。

### 2.1. BYOL

BYOL 的架構流程如 Fig.1，模型主要可區分成 Encoder、Encoder′ 與 Predictor 這三大區塊，參數最佳化的目標即爲 Predictor 輸出的特徵向量 p 與 Encoder′ 輸出的特徵向量 z′ 相似度越高越好：

$$Loss_{CL} = -\frac{Sim(p_1, sg(z_2')) + Sim(p_2, sg(z_1'))}{2} \quad (1)$$

Eq.1 中的 sg 代表 stop gradient，因此在反傳遞過後，只有 Encoder 與 Predictor 會依靠梯度對參數進行更新，而 Encoder′ 則是以 Momentum 的方式更新自身權重：

$$\theta_{E'} = \tau\theta_{E'} + (1-\tau)\theta_E \quad (2)$$

Eq.2 中的 $\theta_{E'}$ 與 $\theta_E$ 分別表示 Encoder′ 與 Encoder 的參數。

### 2.2. SimSiam

SimSiam 與 BYOL 最大的差異是移除了 Momentum Encoder，除此之外的流程如 Fig.2 所示皆與 BYOL 相同：

$$Loss_{CL} = -\frac{Sim(p_1, sg(z_2)) + Sim(p_2, sg(z_1))}{2} \quad (3)$$

## 3. Method

本專題使用的模型結構如 Fig.3 所示，是由一組 Autoencoder 構成的模型，並使用 SI-SDR[4] 作爲 Speech Enhance-

[1]在 Fig.1 與 Fig.2 中的 ~ 代表 Stop Gradient，因此梯度並不會通過該區段向前更新。

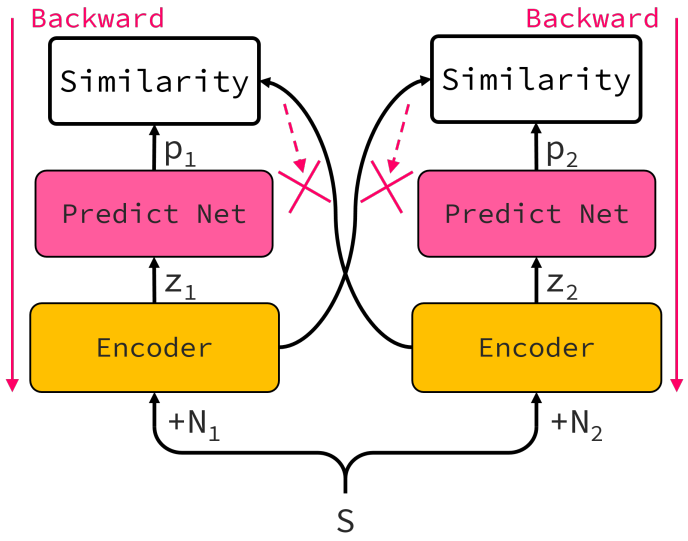


Figure 2. SimSiam 的算法架構。將混合不同噪音的語音輸入 Encoder 得到 $Z_1$ 與 $Z_2$，在將 $Z_1$ 與 $Z_2$ 輸入 PredictNet 後獲得 $P_1$ 與 $P_2$，$P_1$ 與 $P_2$ 會各自將 $Z_2$ 與 $Z_1$ 做爲目標並計算差距，以此更新 Encoder-PredictNet 的參數。

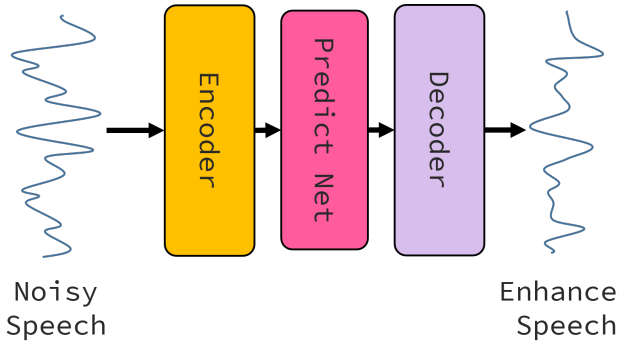


Figure 3. 語音增強使用的 Auto Encoder。期望模型能將受雜訊污染的語音還原成乾淨語音。

ment 任務的損失函數：

$$Loss_{SE} = 10log_{10}(\frac{\left\|\frac{\hat{s}^T s}{\|s\|^2}s\right\|^2}{\left\|\frac{\hat{s}^T s}{\|s\|^2}s - \hat{s}\right\|^2}) \quad (4)$$

在 Eq.4 中的 $s$ 與 $\hat{s}$ 分別表示乾淨語音與去噪後的語音。

配合 Contrastive Learning 進行訓練時，語音 $S$ 會混合 $N_1$ 與 $N_2$ 兩個不同的噪音，在代入 Autoencoder 中得到 $z$、$p$ 以及 $\hat{s}$ 後計算 $Loss_{CL}$ 與 $Loss_{SE}$，最後在權重相加作爲本次研究的損失函數：

$$Loss_{Mix} = Loss_{CL} + 0.1Loss_{SE} \quad (5)$$

## 4. Experiments

本專題使用的噪音資料集是由 20 種不同類型的背景噪音所組成，總共有 100 個音檔的 Nonspeech [5]。而語音資料集則是選用 TIMIT [6]，TIMIT 具有 6300 句語音，這些語音包含美國八個地區共 630 人所念出的 10 個指定句子。

在訓練時的噪聲語音是將 4120 句 TIMIT 語音與 75 個 Nonspeech 雜訊以 -10, -5, 0, 5, 10 這三種 SNR 混合產生的，

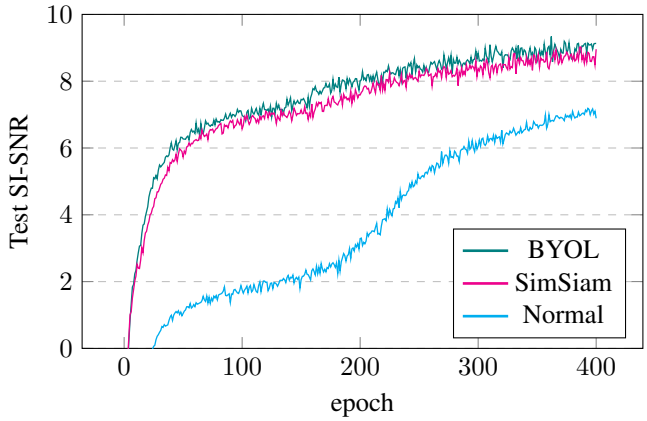


Figure 4. 使用 BYOL、SimSiam 的模型與未使用 Contrastive Learning 方法的 Normal 進行比較。可以發現 Contrastive Learning 在前期可以快速的提升模型的效能。

Table 2. 各模型對噪聲語音進行去噪後的平均評估結果

| Model | Evaluation Metrics | | |
|---|---|---|---|
| | PESQ | STOI | SI-SNR |
| Noisy | 1.813 | 0.764 | 0.001 |
| Normal | 2.273 | 0.814 | 7.146 |
| BYOL | 2.392 | 0.844 | 9.174 |
| BYOL round | 2.461 | 0.858 | 9.378 |
| **BYOL round(100 s)** | **2.474** | **0.861** | **9.526** |
| SimSiam | 2.374 | 0.84 | 8.884 |
| **SimSiam round** | **2.472** | **0.861** | **9.529** |

其取樣頻率爲 16K Hz，擷取時間約爲 1s (16384 個樣本點)。測試則分別使用 500 與 25 個語音及噪音以 -7.5, -2.5, 2.5, 7.5 SNR 進行混合，並擷取約 2s (32768 個樣本點) 進行測試並計算其 PESQ[7]、STOI[8] 及 SISDR[4] 作爲效能評估指標。

使用 SGD 作爲優化器，learning rate、SGD momentum 與 weight decay 分別設爲 0.05、0.9 與 0.0001，使用的 Batch Size 是 256，BYOL 的超參數 $\tau$ 設定是 0.99。

實驗中訓練了使用 BYOL、SimSiam 以及未使用 CL 的 Normal 這大三類模型，並依據訓練過程中的 Loss 切換方式，可區分爲從頭到尾皆使用 Mix Loss 的基本模型、每 50 個 epoch 就在 Mix 與 SE Loss 之間切換的 Round，與只有初期 50 個 epoch 使用 Mix Loss 後都使用 SE Loss 的 Pretrain。

在 Fig.4 中比較 CL 方法的基本模型與 Normal 的收斂差異，BYOL 與 SimSiam 在初期就以極快的速度提升效能。在 Fig.5 則顯示基本模型與切換 Loss 的方法差異，不論 BYOL 或是 SimSiam 皆呈現出在中期切換 Loss 可以加快模型收斂速度，但 SimSiam pretrain 在訓練後期明顯出現 Overfitting 的問題，這代表在後期繼續使用 CL Loss 進行一定程度的約束是有必要的。

爲了測試 Contrastive Learning 在資料量不足情況是否也能夠提升模型效能，在 Fig.6 的實驗中將測試資料與訓練資料互換，BYOL 與 SimSiam 的實驗結果依舊具備比 Normal 更優秀的表現。這表示即使只有少量資料，Contrastive Learning 依然能快速找出比一般方法更好的特徵。

Table.2 記載了不同模型在三種評估指標下的結果。與 Normal 相比，所有使用了 CL 的方法皆有明顯的效能提升。而不全程使用 CL Loss 約束的 Round 模型，則具備著比基礎模型更優秀的評估結果。在不同 SNR 噪音下的表現，分別記錄在 Table.3、Table.4 與 Table.5。

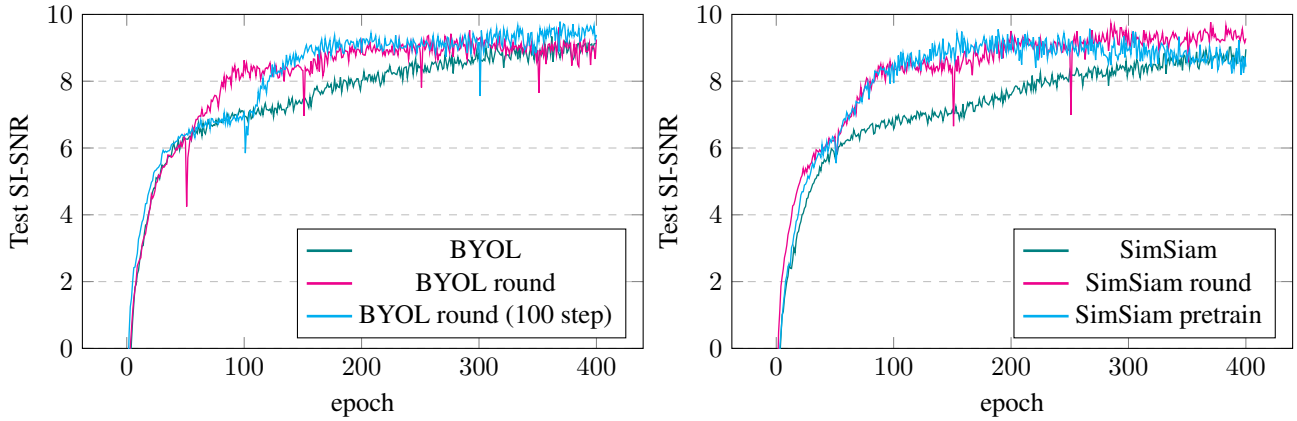


Figure 5. BYOL 與 SimSiam 的實驗結果。若在中後期減少 CL Loss 的約束力，反而能使模型效能提升更加迅速。但也可從 SimSiam pretrain 的實驗也觀察到，若完全移除 CL Loss 會導致模型發生 overfitting。

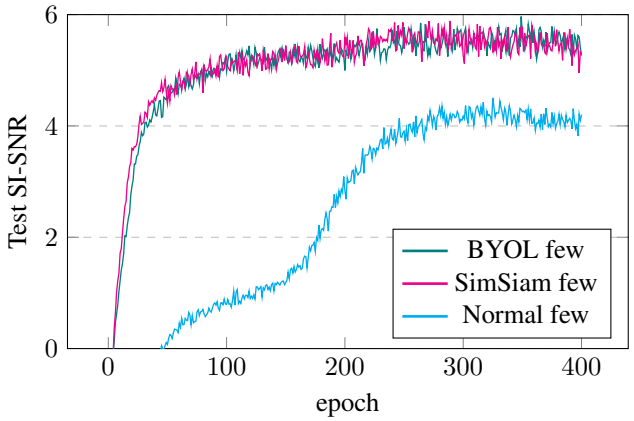


Figure 6. 改爲使用測試集進行訓練、訓練集進行測試。即使是在只有少量資料的情況，Contrastive Learning 依然比只使用 Speech Enhancement Loss 的方法具有更高的效能。

Table 3. 各模型針對混和不同 SNR 雜訊進行去噪後的 PESQ 評估結果

| Model | -7.5 | -2.5 | 2.5 | 7.5 |
|---|---|---|---|---|
| | PESQ | | | |
| Noisy | 1.337 | 1.644 | 1.971 | 2.3 |
| Normal | 1.826 | 2.138 | 2.438 | 2.688 |
| BYOL | 1.875 | 2.253 | 2.59 | 2.851 |
| BYOL round | 1.904 | 2.3 | 2.671 | 2.97 |
| **BYOL round(100 s)** | **1.913** | **2.308** | **2.683** | **2.991** |
| SimSiam | 1.873 | 2.24 | 2.563 | 2.82 |
| **SimSiam round** | **1.937** | **2.317** | **2.672** | **2.962** |

Table 4. 各模型針對混和不同 SNR 雜訊進行去噪後的 STOI 評估結果

| Model | -7.5 | -2.5 | 2.5 | 7.5 |
|---|---|---|---|---|
| | STOI | | | |
| Noisy | 0.643 | 0.728 | 0.809 | 0.878 |
| Normal | 0.702 | 0.793 | 0.859 | 0.904 |
| BYOL | 0.734 | 0.826 | 0.889 | 0.928 |
| BYOL round | 0.746 | 0.841 | 0.904 | 0.942 |
| **BYOL round(100 s)** | **0.75** | **0.844** | **0.906** | **0.944** |
| SimSiam | 0.729 | 0.822 | 0.885 | 0.926 |
| **SimSiam round** | **0.753** | **0.845** | **0.905** | **0.942** |

Table 5. 各模型針對混和不同 SNR 雜訊進行去噪後的 SI-SDR 評估結果

| Model | -7.5 | -2.5 | 2.5 | 7.5 |
|---|---|---|---|---|
| | SI-SDR | | | |
| Noisy | -7.497 | -2.498 | 2.503 | 7.498 |
| Normal | 2.611 | 6.065 | 8.972 | 10.935 |
| BYOL | 3.677 | 7.785 | 11.281 | 13.951 |
| BYOL round | 3.396 | 7.772 | 11.615 | 14.728 |
| **BYOL round(100 s)** | **3.457** | **7.859** | **11.784** | **15.004** |
| SimSiam | 3.544 | 7.508 | 10.913 | 13.572 |
| **SimSiam round** | **3.583** | **7.935** | **11.751** | **14.847** |

## 5. Conclusion

本次將 Contrastive Learning 運用於 Speech Enhancement 的研究結果可發現，前期使用 CL Loss 對中間的隱藏特徵進行高強度的約束，能提升模型的收斂速度。然而若在中後期繼續讓 CL Loss 維持較高的權重反而會阻礙模型提升效能。但這並不代表 CL Loss 就完全不重要，相反的，保持一定程度的 CL 約束能有效避免 Overfitting 的發生。

未來將使用兩種以上的雜訊與語音進行混和，以驗證在更加複雜的資料擴增下使用 Contrastive Learning，能否讓 Speech Enhancement 的效能有更大幅度的提升。並研究具有自適應性的 CL Loss 與 SE Loss 權重調整方法，以減少人工調整超參數的必要性。

## References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[2] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020.

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020.

[4] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR - half-baked or well done? *CoRR*, abs/1811.02508, 2018.

[5] Guoning Hu and DeLiang Wang. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2067–2079, 2010.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993.

[7] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2, 2001.

[8] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

Table 6. Encoder 結構

| | |
|---|---|
| DSB(1, 32) | Rearrange(bz 1 (L 64) → bz 64 L) |
| DSB(32, 64) | Conv1d(64, 128, 3) |
| DSB(64, 128) | |
| Concat() | |
| Conv1d(256, 128, 1) | |
| Main Block(128, 256, 9, 16, 8) | |
| Main Block(128, 256, 9, 16, 8) | |
| Main Block(128, 256, 9, 16, 8) | |
| Main Block(128, 256, 9, 16, 8) | |

Table 7. Predictor 結構

| |
|---|
| Main Block(128, 256, 9, 16, 8) |
| Main Block(128, 256, 9, 16, 8) |

Table 8. Decoder 結構

| |
|---|
| Conv1d(128, 256, 9, 16, 8) |
| Main Block(128, 256, 9, 16, 8) |

# A. Model