Contrastive Learning for Speech Enhancement

郭品辰
F14066127
NCKU SNAME

黃仁鴻
P76094169
NCKU CSIE

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，對比學習應用於語音增強任務的可行性將會是本次專題的研究主軸。

然而，如同 SimCLR [1] 等等 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，像是在 SimCLR 的論文中，實驗所使用的 batch size 就到達了 4096。

爲了因應實驗環境的硬體限制，本專題將會基於 BYOL [2] 與 SimSiam [3] 這兩種且在小型 batch size 也有良好成效的 CL 方法進行實驗，並與監督式語音增強方法進行比較。

2. System framework

本專題使用的模型結構如 Figure 1. 所示，是由 Encoder 與 Decoder 兩區塊組合而成，其中負責抽取語音特徵 Encoder 區塊將會使用 BYOL 與 SimSiam 進行預訓練。在進行 CL 訓練時，語音 S 會混合 N_1 與 N_2 兩個不同的噪音後，輸入 Encoder 取出語音的特徵向量，並用 Figure 2. 與 Figure 3. 所述的方法更新其內部權重[1]。

[1]在 Figure 2. 與 Figure 3. 中的 X 符號代表 Stop Gradient，因此梯度並不會通過該區段向前更新。

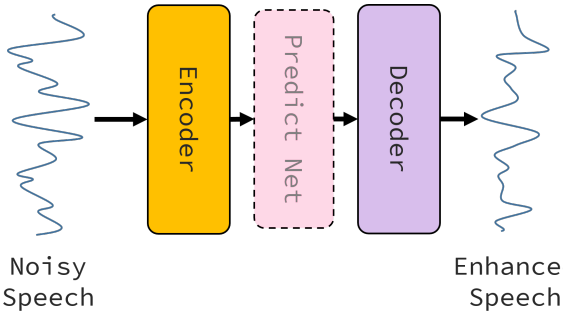

Figure 1. 語音增強使用的 Auto Encoder。期望模型能將受雜訊污染的語音還原成乾淨語音。

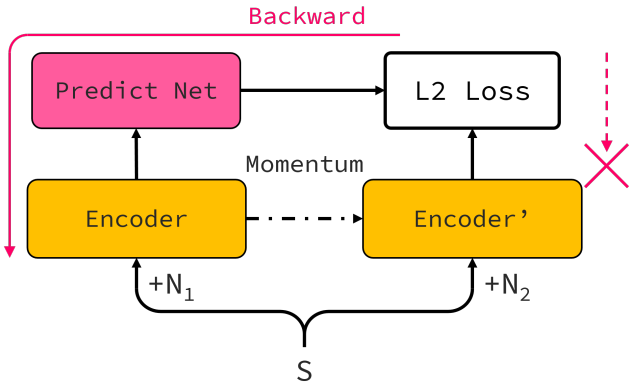

Figure 2. BYOL 的算法架構。Encoder-PredictNet 會將 Encoder' 取出特徵做爲目標，計算 L2 Loss 後更新 Encoder-PredictNet 的參數，Encoder' 則會把自身目前參數與更新後 Encoder 做加權平均當新參數。

在使用 CL 預訓練完 Encoder 後，便會將其串接上 Decoder，將乾淨的語音當成目標來進行語音增強任務的學習。更詳細的實驗內容在 Expected results 中。

3. Expected results

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 [4]。而語音資料集則是選用 TIMIT [5]，TIMIT 具有 6300 句

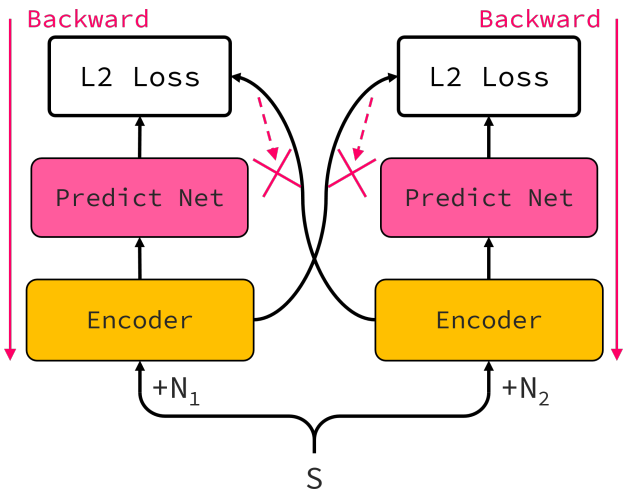

Figure 3. SimSiam 的算法架構。將混合不同噪音的語音輸入 Encoder 得到 Z_1 與 Z_2，在將 Z_1 與 Z_2 輸入 PredictNet 後獲得 P_1 與 P_2，P_1 與 P_2 會各自將 Z_2 與 Z_1 做爲目標並計算差距，以此更新 Encoder-PredictNet 的參數。

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 TIMIT 與 Nonspeech 以 -5, 0, 5 這三種 SNR 混合產生的。

本專題預計將會進行以下幾項實驗：

1. 直接對 AutoEncoder 訓練 Speech Enhancement 任務。
2. 基於 BYOL 對 Encoder 與 Bottleneck 進行預訓練後凍結參數，然後接上 Decoder 訓練 Speech Enhancement 任務。
3. 基於 SimSiam 對 Encoder 與 Bottleneck 進行預訓練後凍結參數，然後接上 Decoder 訓練 Speech Enhancement 任務。
4. 將 BYOL 訓練得到的參數作爲初始權重後，對 AutoEncoder 訓練 Speech Enhancement 任務。
5. 將 SimSiam 訓練得到的參數作爲初始權重後，對 AutoEncoder 訓練 Speech Enhancement 任務。
6. 基於不同數量的樣本進行上述實驗，以測試各種方法的泛化能力。

最終將比較上述不同方法所訓練得到模型的 PESQ[6]、STOI[7] 與 SISDR[8]。

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] 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.

[5] 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.

[6] 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.

[7] 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.

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