Unsupervised Speech Enhancementusing teacher-student

專題編號：111-CSIE-S011
執行期限：110 年第 1 學期至 111 年第 1 學期
指導教授：陳彥霖
專題參與人員： 108820038　陳力瑋

一、摘要

我們提出了一種靈感來自於 NyTT[1] 與 RemixIT[2] 的無監督語音增強的教師-學生框架。此方法除了域內噪音之外，只需要域外噪聲就可以訓練出用於語音增強的模型，與之前的方法不同之處在於此前的方式多半還是需要使用到域外的乾淨語音。

此外，我們還發現在推斷的時候，使用教師增強過後，再用學生增強一遍，可以使最後產出語音的語音質量感性評價(PESQ)上升許多。

二、緣由與目的

語音增強(SE)旨在通過去除語音信號中的噪聲來提高音頻質量。SE 模塊從嘈雜的輸入語音中去除噪聲，從而提高識別結果。目前 SE 開發及其應用的成功主要依賴於包含配對的干淨和嘈雜語音的大量訓練數據。傳統上，嘈雜語音通常是通過混合乾淨的語音和噪聲來合成的。然而，由於乾淨的語音和噪聲主要是在專業音頻中錄製的，因此在現實世界的場景中很難收集到。為了緩解這個問題，我們提出了一個訓練框架，只需要噪聲和未配對的嘈雜語音。

三、方法

我們的方法是使用 RemixIT 框架，並使用 NyTT 預訓練的模型作為初始教師模型。我們提出了六種訓練學生模型的方法，其中三種是 NyTT ，另外三種是清潔目標訓練(CTT)。 CTT 是一種常見的有監督的語音增強模型訓練方法，它只是將噪聲語音作為輸入，將乾淨的語音作為目標。

在我們的方法中，我們發現在推理過程中同時使用教師和學生模型可以提高性能。我們使用教師模型進行第一階段的增強，然後將結果反饋給學生模型進行第二階段的增強。在嘗試了所有的學生模型作為第二階段後，我們發現第二階段推理的結果會比只使用教師模型和學生模型的情況要好。此外，我們還將使用教師模型作為第一階段與第二階段進行了比較，雖然它的表現比只使用教師模型好，但不如使用學生模型作為第二階段好。

四、實驗

我們使用 VoiceBank-DEMAND[18]作為域內噪聲數據集，它是原始 NyTT 論文訓練集的一部分[11]。訓練集包括 28 個發言人(11,572 個語料)，有四個信噪比(15、10、5 和 0dB)。測試集由 2 個說話人(824 個語料)組成，有四個信噪比(17.5、12.5、7.5 和 2.5dB)。

而我們使用 CHiME3[19] 背景作為 OOD 噪聲數據集，這也是 NyTT 原始論文的訓練集的一部分。

在訓練學生模型時，我們使用 VoiceBank-DEMAND-CHiME3 作為輸入信號，它由 VoiceBank-DEMAND 和 CHiME3 混合組成，使用隨機選擇的信噪比在 -5 到 5dB 之間。

在訓練時我們使用 l1 loss ，並且用 500 迭帶訓練教師模型，而在使用 Static teacher 時，也是選用訓練了 500 迭帶的模型來做測試，而在使用 Exponentially moving average teacher 時，使用訓練了 20 迭帶的模型來做測試， 20 迭帶是從 RemixIT 這篇論文而來。

Static teacher: 教師模型始終是初始模型。只有 $S^{(0)}$ 從一開始訓練， $S^{(k)}, k > 0$

從 $S^{(k-1)}$ 延續訓練。

Exponentially moving average teacher: 對於每個歷時，用最新的學生模型和當前的教師模型的加權和來替換教師模型 $\mathrm{T}^{(\mathrm{k}+1)} = \mathrm{S}^{(\mathrm{k})} + (1-\gamma)\mathrm{T}^{(\mathrm{k})}$ ，$\gamma = 0.01$

五、 結果

OOD 噪聲與組成 in-domain 噪雜語音的噪聲越相近，對於效能的表現越好。而 Static teacher 訓練出來的 model 可以看到雖然效果都輸 baseline 但通過 2-stage 推斷之後幾乎都超越了 baseline，而 Exponentially moving average teacher 訓練出來的 model 可以看到明顯效果比 Static teacher 更接近 baseline，且在 2-stage 推斷之後也都超越 baseline。甚至到達 2.36 左右的 PESQ。

參考文獻

[1] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, “Noisy-target Training: A training strategy for DNN-based speech enhancement without clean Speech,” in *Proc. EUSIPCO*, 2021, vol. 2021-Augus, pp. 436–440. doi: 10.23919/EUSIPCO54536.2021.9616166.

[2] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, “RemixIT: Continual self-training of speech enhancement models via bootstrapped remixing,” *EEE J. Sel. Top. Sig. Proc.*, vol. 14, no. 8, pp. 1–12, 2022, doi: 10.1109/JSTSP.2022.3200911.