The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

AaWLoss: An Artifact-aware Weighted Loss Function for Speech Enhancement
用於語音增強之僞影感知加權損失函數

En-Lun Yu, Kuan-Hsun Ho, Berlin Chen
National Taiwan Normal University
{enlunyu, jasonho610, berlin}@ntnu.edu.tw

摘要

語音增強（Speech Enhancement, SE）系統不僅能夠提升語音的聽覺品質，還可以與自動語音辨識系統（Automatic Speech Recognition, ASR）相結合，從而增強 ASR 在噪聲環境下的強健性。然而，單通道 SE 可能會產生對 ASR 辨識不利的僞影，進而導致 ASR 的識別錯誤。最近的研究表明，通過引入新的 SE 損失函數 NAaLoss，對模型進行微調，能夠有效減少模型產生僞影的效果。然而，該方法仍然存在潛在的錯誤假設。因此，在本研究中，我們通過深入分析該方法並進行大量實驗和案例分析，尋找其內部的潛在問題。爲此，我們提出了改進後的新損失函數 AaWLoss。經過修正和優化，AaWLoss 成功解決了 NAaLoss 在相同設置下可能喪失抑制噪聲條件僞影功能的缺點。此外，AaWLoss 在抑制乾淨條件下的僞影能力達到了巔峰水平，甚至使經過增強的乾淨語音具備了有利於 ASR 辨識的資訊。

Abstract

The Speech Enhancement (SE) system not only enhances the perceptual quality of speech but also make the ASR performance robust in noisy enviornments when integrating with ASR systems. However, single-channel SE may generate detrimental artifacts to ASR recognition, leading to recognition errors. Recent research indicates that by introducing the novel SE loss function NAaLoss and fine-tuning the model, the generation of artifacts can be effectively reduced. Nonetheless, this approach still needs to be revised in its underlying assumptions. Therefore, we extensively analyze this method in this study and conduct numerous experiments and case studies to identify the inconsistencies. To address this, we propose an improved loss function, AaWLoss. AaWLoss successfully resolves the potential loss of noise-condition artifact suppression inherent in NAaLoss under the same settings through modifications and optimizations. Furthermore, AaWLoss achieves peak performance in suppressing artifacts under clean conditions, even adding information beneficial for ASR recognition to the enhanced clean speech.

關鍵字：單通道語音增強、強健性自動語音辨識、僞影處理

Keywords: single-channel speech enhancement, noise-robust speech Recognition, processing artifacts

1 緒論

近年來，隨著類神經網路技術的進步，語音增強（Speech Enhancement, SE）方法已經取得了顯著的發展。這些方法通過學習並建模乾淨語音與噪聲語音之間的複雜關係，極大地提升了在聽覺指標上的表現。然而，這些 SE 方法不僅僅局限於提升音訊的聽覺感知，另一個同樣重要的應用領域是與自動語音辨識（Automatic Speech Recognition, ASR）系統的結合。這種結合能夠賦予前端的 SE 方法在面對噪聲、混響等聲學干擾時更強大的強健性 (Robustness)。雖然一些基於波束形成 (Beamforming) (Heymann et al., 2016; Erdogan et al., 2016; Boeddeker et al., 2018) 等多通道技術的語音增強方法已經在這方面取得了成功 (Barker et al., 2015, 2018)，然而，由於這些方法需要使用麥克風陣列，因此如何開發一種能夠在單通道環境下有效賦予 ASR 強健性的 SE 方法仍然是一個值得深入討論和研究的重要議題。

儘管許多研究已經證明單通道的語音增強對減少噪音對語音訊號的影響非常有幫助，但同時也存在可能產生多餘的僞影（artifacts）和失眞的風險 (Menne et al., 2019; Chen et al., 2018; Fujimoto and Kawai, 2019; Iwamoto et al., 2022)。這些問題在後續的

ASR 系統的特徵抽取階段可能導致一些錯誤。舉例來說，語音增強可能會改變原始語音的時間結構或持續時間特性，進而造成 ASR 在辨識過程中出現詞語或音素的錯位，從而影響 ASR 系統的整體性能。

由於僞影的產生取決於所使用的 SE 模型以及輸入訊號的特性，因此要找到一個一致的定義來描述僞影是相當困難的。有一項致力於解決這個問題的研究採用了正交投影的誤差分解方法 (Iwamoto et al., 2022; Vincent et al., 2006)。該方法通過將訊號投影到語音與噪聲的正交子空間中，以分析訊號的組成，進而獲得僞影的成分。然而，這種方法所基於的假設有時可能不太精確，因爲使用正交投影的前提是噪聲與乾淨語音之間必須是互相獨立的。而這種前提在存在談話性噪聲（Babble Noise）等具有語音特徵的噪聲干擾時可能不成立。另一方面，SE 產生僞影的原因可能來自於 SE 與 ASR 訓練目標之間的差異。儘管聯合訓練 (Chen et al., 2018; Menne et al., 2019; Hu et al., 2023) 與資料擴增技術 (Fujimoto and Kawai, 2019; Tan and Wang, 2020) 已被用來解決這個問題，但並非所有情況下都能對 ASR 系統進行修改。因此，對於基於 SE 的強健 ASR 系統而言，是否具備減少影響辨識結果的僞影的能力，顯得格外關鍵。

通常，SE 的訓練目標函數旨在最小化估計乾淨語音與目標乾淨語音之間的差距 (Braun and Tashev, 2020; Xu et al., 2014)。儘管這樣的目標函數長期以來被廣泛用來有效提升目標乾淨語音的聽覺指標，但它並未充分考慮到僞影的存在。值得注意的是，即便聽覺指標的改善可以帶來正面效益，相關研究 (Hu et al., 2023) 也已經指出，聽覺指標與 ASR 的性能並不總是有著絕對的相關性。換句話說，在串聯 ASR 的 SE 系統中，如何定制一個考慮到僞影的目標函數，以使 SE 的訓練目標更加符合 ASR 任務的需求，是一個值得探討的議題。我們近期的一項研究 (Ho et al., 2023) 提出了一個具有僞影概念的 SE 目標函數，稱爲 NAaLoss。該研究在實驗中顯示，使用 NAaLoss 來訓練 SE 模型後，在串聯 ASR 後表現得更加出色。然而，該項研究並未深入分析其所提出的三個目標函數元件，且其所依據的假設仍然存在一些與實際不符之處，導致最終訓練出的模型實際上未能同時達到所有元件的目標。

本研究深入的探討目標函數 NAaLoss。透過消融實驗，我們分析了不同目標函數元件對模型的實際影響，以及可能導致結果的原因。在實驗分析的基礎上，我們確認了原始的 NAaLoss 確實存在假設上的潛在問題。我們排除了其中的錯誤假設，並加入了對兩種僞影情況的加權估算，從而提出了一個經過優化的目標函數，稱爲 AaWLoss。在最終的實驗中，我們證實 AaWLoss 相對於 NAaLoss 更能有效地實現去除乾淨語音條件下的僞影。此外，AaWLoss 所需要的訓練迭代次數，相較於 NAaLoss 更加符合模型微調的使用情境。

2 NAaLoss 簡介

NAaLoss 在針對 SE 與 ASR 串聯的情況下爲僞影提出了四項定義，1) 僞影會降低 ASR 的詞語錯誤率 (Word Error Rate, WER) 表現；2) 僞影無法反應在聽覺或可理解性的指標上；3) 僞影是由 SE 模型所產生，且會隨著 SE 模型的替換而也所變化；4) 僞影是對原始 SE 輸入的某種訊號失眞，並透過公式來表達上述所定義的僞影。

噪聲語音 $x \in \mathbb{R}^T$ 可以由 $x = y + z$ 所組成，其中 $y \in \mathbb{R}^T$ 爲目標乾淨語音，$z \in \mathbb{R}^T$ 爲干擾的噪音。我們將 $f(.)$ 設爲 SE 模型，θ 爲僞影。NAaLoss 根基於以下三個假設：

1. $f(y) = \theta_c + x$；SE 模型在輸入爲乾淨語 y 時，輸出包含乾淨條件僞影 θ_c 與乾淨語音 y。
2. $f(x) = \theta_m + \tilde{z} + x$；在理想情況下，SE 模型在輸入爲噪聲 x 時，輸出噪聲條件僞影 θ_m、殘餘噪音 $\tilde{z}$、以及乾淨語音 y。
3. $f(z) = \tilde{z}$；將噪音輸入進 SE 模型後的結果爲殘餘噪音。

根據以上三個假設，我們可以推算出乾淨條件僞影與噪聲條件僞影的估計公式分別爲 $\theta_c = f(y) - y$ 與 $\theta_m = f(x) - f(z) - y$。

NAaLoss 以 SE 經常用的目標函數，計算估計的乾淨語音與目標乾淨語音的差距作爲首個元件 $\mathcal{L}_{\text{estim}} = \text{dist}(f(x), y)$；消除僞影的目標函數元件 $\mathcal{L}_{\text{deatf}} = \sum_i \text{dist}(\theta_i, 0), i \in c, m$ 將乾淨條件僞影與噪聲條件僞影進行加總。再根據假設 3. 估計未知噪音的方法，作爲另一個目標函數元件 $\mathcal{L}_{\text{ignor}} = \text{dist}(f(z), 0)$。總體來說，NAaLoss 如下列式子所示：

$$\mathcal{L}_{\text{NAa}} = (1 - \alpha - \beta)\mathcal{L}_{\text{estim}} + \alpha\mathcal{L}_{\text{deatf}} + \beta\mathcal{L}_{\text{ignor}}$$

其中，α 與 β 爲權衡三個元件的超參數，在 NAaLoss 原始的設定中爲 $\alpha = \beta = 0.1$。

3 從 NAaLoss 到 AaWLoss

3.1 NAaLoss 的消融實驗

爲了分析 NAaLoss 的實際效用，我們對 NAaLoss 的三個元件進行消融實驗。

3.1.1 實驗設置

消融實驗在 NAaLoss 所使用的基準數據集 VoiceBank-DEMAND (Valentini-Botinhao et al., 2016) 上進行，數據集的相關資訊在 4.1詳細說明。在模型選擇上，我們延續了 NAaLoss 所使用的 MANNER-small (Park et al., 2022)，作爲 SE 模型的架構。訓練的方式參照了 NAaLoss 的實驗結果，以微調預訓練模型參數的方式進行。同樣地，我們使用了兩種 ASR 系統來對乾淨語音 y 和噪聲語音 x 進行辨識。CCT-AM 是一個使用乾淨語音進行訓練的聲學模型（Acoustic Model, AM），而 MCT-AM 則是使用受到噪音干擾的噪聲語音進行訓練的 AM。這意味著相對於 CCT-AM，MCT-AM 更具有強健性。

3.1.2 實驗基準線

消融實驗的實驗基準線 (Baselines) 如表 1. 所示。表中的第一列表示將乾淨語音和噪聲語音直接輸入 ASR 系統中，第二列則表示使用 MANNER 作爲前端的 ASR 系統，第三列和第四列則分別表示在 MANNER 前端的基礎上使用 NAaLoss 進行微調。CCT-AM 和 MCT-AM 欄位中的數值表示詞語錯誤率（Word Error Rate, WER），可用於評估 ASR 的辨識能力；PESQ（Perceptual Evaluation of Speech Quality）欄位則用來評估與語音的聽覺品質。

我們在重現實驗過程中觀察到，在約 20 次訓練迭代後，模型已經達到一定的擬合程度。爲了更有效地進行消融實驗，我們將訓練迭代次數從 NAaLoss 原先使用的 350 次調整爲 20 次，同時保持其他實驗設定不變。這樣的調整允許我們在較短的時間內獲得有意義的結果，同時仍然能夠評估模型性能的變化。

3.1.3 結果分析

表 2. 包含了各個元件分別運行與個別消除的實驗結果。在參與的元件欄位中，被標記的元件的權重爲 1，這與 NAaLoss 文獻中使用的加權損失方式有所不同。

我們的發現是，作爲常用的 SE 損失函數，$\mathcal{L}_{\text{estim}}$ 可以有效地提升噪聲環境下的 ASR 性能。然而，$\mathcal{L}_{\text{deatf}}$ 和 $\mathcal{L}_{\text{ignor}}$ 在單獨作爲損失函數運行時並未帶來明顯的改善效果。同時，在消除 $\mathcal{L}_{\text{estim}}$ 的情況下，模型無法改善 WER。這是因爲在這三個元件中，只有 $\mathcal{L}_{\text{estim}}$ 能夠針對噪聲語音與乾淨語音之間的誤差進行最小化，因此 $\mathcal{L}_{\text{estim}}$ 對於噪聲的強健性能提升具有最直接的幫助，因此它是不可或缺的元件。

另一方面，$\mathcal{L}_{\text{deatf}}$ 在與 $\mathcal{L}_{\text{estim}}$ 共同作用下，將乾淨語音引入模型，如預期地減少了乾淨語音條件僞影 θ_c，同時對噪聲的強健性能也帶來了一定的提升。

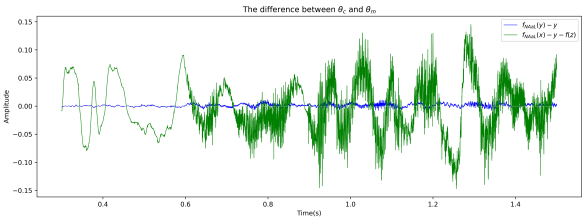


圖 1. 不同條件僞影的差異。乾淨條件僞影 (藍) 通常遠少於噪聲條件僞影 (綠)

3.2 改進方法

3.2.1 NAaLoss 的潛在問題

$\mathcal{L}_{\text{ignor}}$ 單獨運作時，在消融實驗中，使用具有強健性的 MCT-AM 進行辨識，其輸入噪聲時 WER 的退步幅度相比其他元件單獨運行時更大。我們認爲這是由於 $\mathcal{L}_{\text{ignor}}$ 的假設具有某些不夠貼切實際情形的地方。

正如在第 2 章節中所提到的，$\mathcal{L}_{\text{ignor}}$ 的假設是當 SE 模型輸入爲噪音時，理想情況下應該去除掉噪音並不輸出任何內容。然而，在實際情況中，不論是基於映射 (Mapping-based) 或遮罩 (Masking-based)SE 模型，模型所採取的方法並不是從輸入中減去噪音，而是通過映射或遮罩的方式來強調語音訊號。這意味著，即使在噪音輸入至 SE 模型時，模型仍然會從噪音中提取出類似於語音的訊號。由於我們所使用的 MANNER 無法獲得輸入類型的相關資訊，因此要求模型在特定情況下停止提取任何可能是語音的資訊，如同 $\mathcal{L}_{\text{ignor}}$ 所要求的，是相當牽強的，可能會混淆模型訓練時的目標。我們的實驗也確認並證實，無論我們如何訓練 SE 模型，都無法在保持 ASR 的強健性能的同時，使其在輸入噪音情況下產生接近無聲音的輸出 $f(z)$。

此外，另一個可以改進的是對噪聲條件僞影 θ_m 的估算方式。NAaLoss 的假設將噪聲條件僞影 θ_m 視爲與殘餘噪音 $\tilde{z}$ 不相交的兩種訊號。我們認爲，僞影應該同時考量到殘餘噪音 $\tilde{z}$ 對 ASR 可能帶來的危害，提升去僞影對於 ASR 的有利之處。

在消融實驗過程中，我們還觀察到 $\mathcal{L}_{\text{estim}}$ 的數值通常遠高於 $\mathcal{L}_{\text{deatf}}$ 和 $\mathcal{L}_{\text{ignor}}$，這暗示了 (Ho et al., 2023) 設定的權重使得提出的 $\mathcal{L}_{\text{deatf}}$ 和 $\mathcal{L}_{\text{ignor}}$ 在訓練中的影響並不顯著。因此，爲了更明確地分析 $\mathcal{L}_{\text{deatf}}$ 是否對模型有正面影響，我們排除了 $\mathcal{L}_{\text{ignor}}$，並將 $\mathcal{L}_{\text{deatf}}$ 的權重分別設置爲 1、10、500 進行遞增，結果如表 3 所示。

Method	The weights in NAaLoss			Input	CCT-AM	MCT-AM
	$\mathcal{L}_{\text{estim}}$	$\mathcal{L}_{\text{deatf}}$	$\mathcal{L}_{\text{ignor}}$			
-	-			y	5.04	4.86
				x	23.76	8.32
MANNER	-			y	5.28	4.91
				x	7.37	6.62
NAaL 350 epochs	0.8	0.1	0.1	y	5.17	4.88
				x	6.83	6.41
NAaL 20 epochs				y	5.31	4.99
				x	7.07	6.62

表 1. 實驗基準線。所有方法皆使用 MANNER 的 SE 模型架構，並輸入乾淨語音 y 與噪聲語音 x；底線標記的 WER 表示其超越第二列 MANNER 的結果。

Components used			Input	CCT-AM	MCT-AM	PESQ
$\mathcal{L}_{\text{estim}}$	$\mathcal{L}_{\text{deatf}}$	$\mathcal{L}_{\text{ignor}}$				
✓	✓	✓	y	5.26	5.17	4.06
			x	7.25	6.54	3.06
✓			y	5.41	4.94	4.26
			x	7.04	6.57	3.12
	✓		y	5.91	5.26	4.29
			x	9.76	6.86	2.56
		✓	y	5.78	5.68	3.67
			x	9.76	8.32	2.64
✓	✓		y	5.23	5.04	4.13
			x	7.26	6.57	3.09
✓		✓	y	5.41	5.01	4.09
			x	7.23	6.59	3.10
	✓	✓	y	5.33	5.12	4.13
			x	7.62	6.87	3.01

表 2. 消融實驗。被使用 (✓) 的損失函數元件之權重皆設爲 1；底線 標記的 WER 表示其超越 MANNER 的結果。

$\mathcal{L}_{\text{deatf}}$ weight	Input	CCT-AM	MCT-AM	PESQ
1	y	5.23	5.04	4.13
	x	7.26	6.57	3.09
10	y	**5.14**	4.96	4.34
	x	7.31	6.55	3.00
500	y	**4.93**	**4.77**	4.60
	x	34.33	8.98	2.06

表 3. $\mathcal{L}_{\text{deatf}}$ 於不同領導地位的結果。所有結果的 $\mathcal{L}_{\text{estim}}$ 權重皆設爲 1。底線 標記的 WER 表示其超越 MANNER；紅色字體標示的 WER 表示低於未經處理的噪音。

隨著 $\mathcal{L}_{\text{deatf}}$ 權重的增加，模型在乾淨條件下降低僞影的效果增強，然而在噪聲條件下的僞影反而變得更加嚴重。尤其是當 $\mathcal{L}_{\text{deatf}}$ 和 $\mathcal{L}_{\text{estim}}$ 之間的尺度差距逼近時，生成的增強語音就像未經處理的噪聲。對此，我們認爲這主要是由圖 1 所呈現的情況所引起的，θc 往往遠小於 θ_m，這使得模型更傾向於減少 θ_m 以利優化。這樣的結果證實了 θ_m 因包含殘餘噪音 $\tilde{z}$，可能混淆模型訓練目標，導致模型在減少殘餘噪音以最小化這一損失的同時，擴大了 $\tilde{z}$ 錯誤假設所帶來的不良影響。

3.2.2 僞影感知加權損失函數 AaWLoss

我們認爲，爲了去除錯誤假設導致模型在訓練中可能出現的問題，從而提升目標函數的合理性，刪除 $\mathcal{L}_{\text{ignor}}$，並且從 θm 的估算公式中刪除殘留噪音 $\tilde{z}$ 是必要的。基於這一考量，我們修正了原本的假設，並提出了一個改進後的損失函數 AaWLoss：

$$\mathcal{L}_{\text{AaW}} = \mathcal{L}_{\text{estim}} + \alpha \mathcal{L}_{\text{wdeatf}}$$

其中，$\theta_c = f(y) - y, \theta_m = f(x) - y$，$\alpha$ 是控制元件權重的超參數。

刪除這個錯誤的假設能夠有效地解決損失函數導致模型在噪聲條件下難以有效抑制僞影的問題。同時，這也意味著殘留噪音將被納入 θ_m 的估算中。爲了使損失函數更符合 SE 模型在噪聲下的實際應用場景，我們基於 $\mathcal{L}_{\text{deatf}}$，對 θc 和 θ_m 增加了一種動態的加權方式，使得：

$$\mathcal{L}_{\text{wdeatf}} = (1-\gamma)\theta c + \gamma \theta_m$$

其中，$\gamma = \frac{\|f(x)-y\|}{\|f(x)-y\|+\|f(y)-y\|}$。如此一來，損失函數將會根據當前 SE 模型的輸出，計算 θ_c 和 θ_m 在整個輸出中的比例，然後動態調整 θ_c

α	Input	CCT-AM	MCT-AM	PESQ
1	y	5.36	4.98	4.10
	x	7.18	6.60	3.09
10	y	**5.09**	4.99	4.21
	x	7.28	6.52	3.08
500	y	**5.04**	**4.81**	4.14
	x	7.05	6.73	2.94

表 4. 使用 AWaLoss 的結果。底線 標記的 WER 表示其超越 MANNER；**粗體**標示的 WER 表示超越 NAaLoss 的效能。

和 θ_m 在損失函數中的權重。這樣的調整使得模型在降低噪聲條件下的僞影爲首要目標的同時，也能減少乾淨條件下的僞影生成。

4 實驗

4.1 實驗設置

爲了比較 AaWLoss 的效果，我們在廣泛使用的開源數據集 VoiceBank-DEMAND 上進行了一系列實驗。該數據集的訓練集包括了 28 位語者錄製的共 11,572 個語句，並且使用了 DEMAND 資料集中的 10 種不同類型噪音，以 0、5、10 和 15 dB 的信噪比（Signal-to-Noise Ratio, SNR）進行混合。測試集則包括了兩位語者錄製的共 824 個語句，並且分別在 2.5、7.5、12.5 和 17.5 dB 的信噪比下混合噪音。此外，我們從訓練集中選取約 200 個語句作爲驗證集，所有語音資料的採樣率均爲 16 kHz。

4.2 實驗結果

爲了對比 AaWLoss 的效能，我們使用了在表 3. 中所設定的 α 值，並對訓練模型進行了 20 次迭代的微調。實驗結果如表 4. 所示。我們可以觀察到，AaWLoss 確實如預期般解決了 NAaLoss 中因錯誤假設而導致的噪聲條件性能下降的問題。

隨著 α 值的增加，模型在乾淨條件下的僞影明顯得到改善，不僅在 CCT-AM 上達到了與直接輸入乾淨語音相當的性能，甚至在 MCT-AM 上還展現出超越乾淨語音的辨識能力。在微調 20 次迭代的條件下，AaWLoss 抑制噪聲條件僞影的表現，也能使 CCT-AM 在不具備強健性的前提下超越 NAaLoss 並接近使用原始損失函數微調的改善幅度。這表明 AaWLoss 能夠在保持原有損失函數在噪聲強健性上的優勢，同時爲模型在乾淨條件下的去僞影甚至增強辨識能力方面提供額外的優勢。

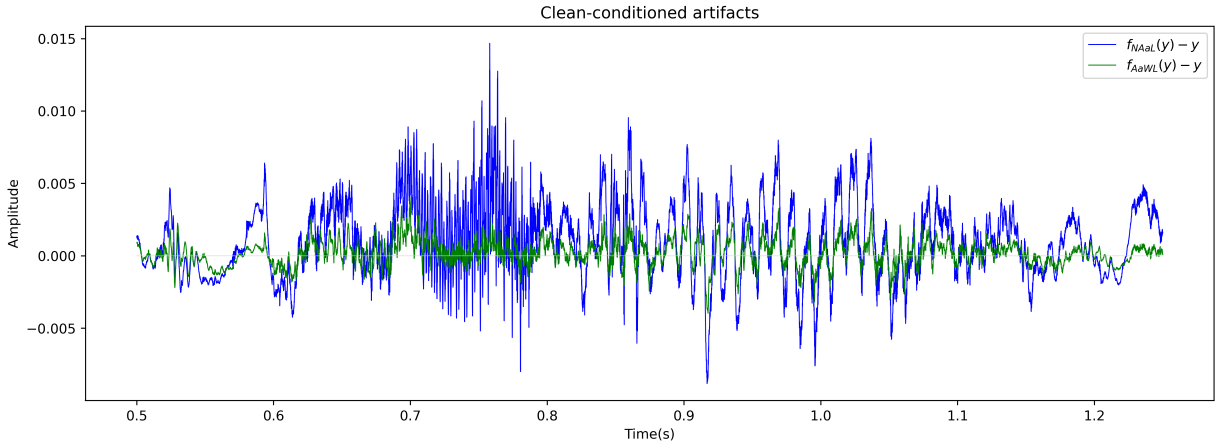
爲了更深入了解 AaWLoss 對模型的實際效果，我們選擇了某些特定實驗案例，對實驗結果進行進一步的分析。在圖 2. 中，我們展示了 AaWLoss 和 NAaLoss 在相同案例上產生的乾淨條件僞影 θ_c 的波形圖和聲譜圖。從圖 2a. 中，我們可以觀察到，透過使用 AaWLoss 進行微調後的 SE 模型，幾乎不會產生乾淨條件下的僞影。此外，從圖 2b. 中我們也可以看出，AaWLoss 所產生的微弱僞影相對於 NAaLoss 而言更加與語音發聲的資訊無關。這些結果表明 AaWLoss 在減少乾淨僞影方面的功能遠超過 NAaLoss，這對於提升模型的實際應用價值具有重要意義。

圖 3. 中的聲譜圖分別展示了使用 AaWLoss 微調後的增強語音 $f_{AaWL}(x)$ 和噪聲條件僞影 θ_m。圖中的方框標示了輕擦音 /f/ 的發聲範圍。根據 (Chen) 的研究，輕擦音 (Voiceless fricatives) /f/ 通常在共振峰（Formants）方面不太明顯，且在 3000 至 4000Hz 之間會有高頻湍流（High Frequency Turbulence）。然而，在 $\alpha = 1$ 和 $\alpha = 10$ 的情況下，我們可以觀察到發音受到了輕微的破壞，原本應該存在的發音部分被誤認爲噪音，並修飾成類似濁爆破音 /b/ 的靜默期（Stop Gap），以至於 ASR 辨識錯誤。相反地，當 $\alpha = 500$ 時，我們可以看到模型保留了應有的聲音資訊。這種現象可能是由於 $\mathcal{L}$wdeatf 在整個損失函數中所佔的影響力較小，使得 $\mathcal{L}_{\text{estim}}$ 所提供的去噪能力導致了聲音中重要資訊的消失。然而，當 $\mathcal{L}_{\text{wdeatf}}$ 的影響力增強時，模型更能保留這些關鍵的音訊資訊，這有助於在將增強語音與 CCT-AM 進行串接時，減少受到僞影的干擾，提高整體辨識效果。

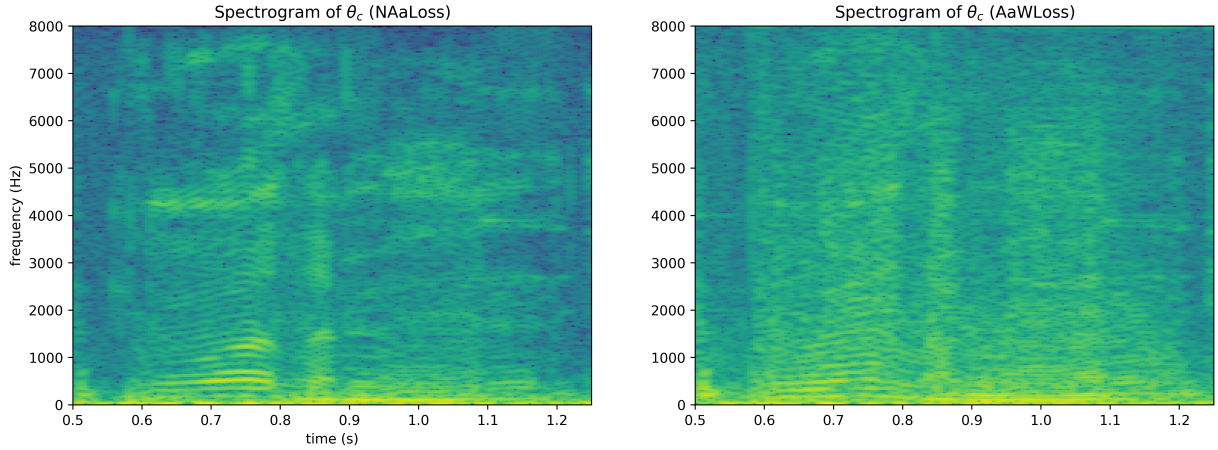
另外一個值得注意的觀察是 PESQ 和 WER 之間的關聯性。對於不具強健性的 CCT-AM 來說，我們可以觀察到具有最佳辨識結果的 SE 模型在 PESQ 表現方面卻相對較差。這進一步驗證了許多研究的結果，認爲 PESQ 與 ASR 的辨識表現之間並沒有絕對的關聯性。

5 結論

本研究延續先前 NAaLoss 的研究概念，並進一步進行消融實驗，深入探討其實際效能。透過這些實驗，我們發現了 NAaLoss 中存在一些不全的假設，並提出了一個更加合理且符合實際應用情境的新型損失函數，稱之爲 AaWLoss。我們的實驗結果清楚顯示，相對於 NAaLoss，AaWLoss 在僅需小於原來迭代次數 $\frac{1}{15}$ 的情況下，就能夠在去除乾淨條件僞影方面超越甚至接近完美的效果，同時還具有對噪聲僞影的抑制能力。透過案例分析，我們也證實了 AaWLoss 解決了傳統 SE 損失函數可能導致 ASR 辨識錯誤的問題，並使得語音



(a) NAaLoss(藍色訊號) 與 AaWLoss(綠色訊號) 產生的乾淨條件僞影 θ_c 波形圖



(b) NAaLoss(左) 與 AaWLoss(右) 產生的乾淨條件僞影 θ_c 聲譜圖

圖 2. 比較 NAaLoss 與 AaWLoss 產生之乾淨條件僞影 θ_c。(a) 爲波形圖；(b) 爲聲譜圖。

增強在保護 ASR 辨識特徵的同時取得更好的聽覺指標。

References

Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2015. The third 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines.

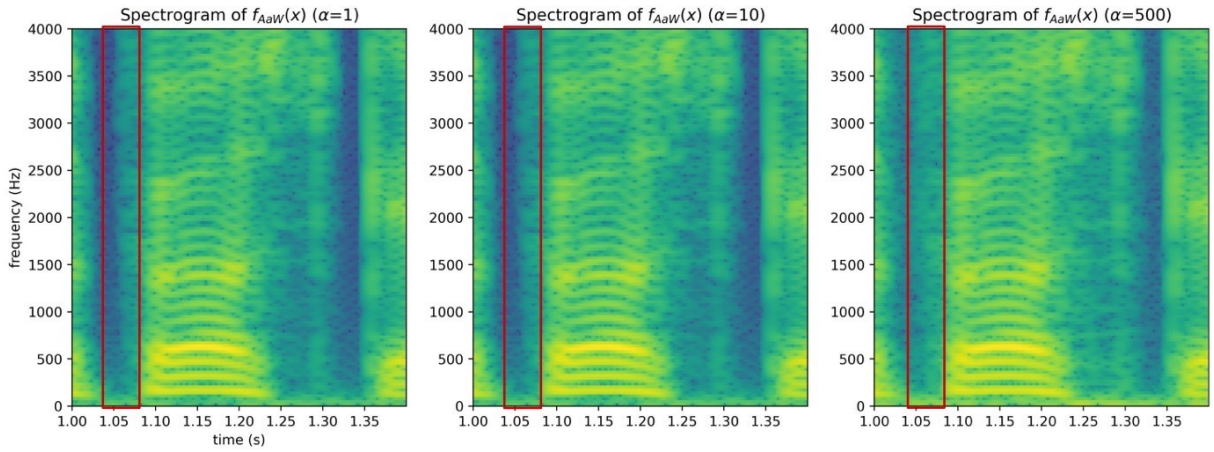
Christoph Boeddeker, Hakan Erdogan, Takuya Yoshioka, and Reinhold Haeb-Umbach. 2018. Exploring practical aspects of neural mask-based beamforming for far-field speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6697–6701.

Sebastian Braun and Ivan J. Tashev. 2020. A consolidated view of loss functions for supervised deep learning-based speech enhancement. *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pages 72–76.

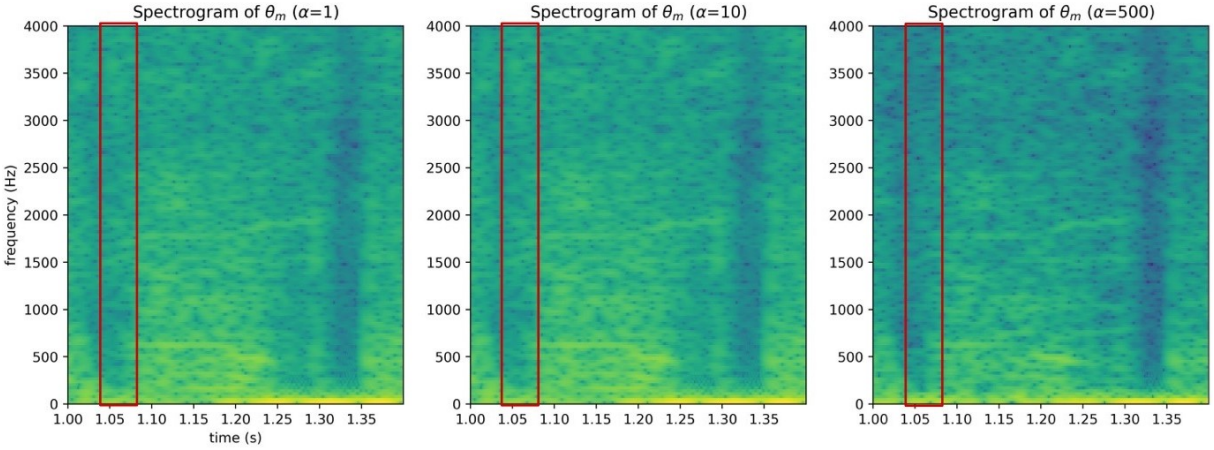
H. C. Chen. 3.2. acoustic aspects of consonants. Accessed: August 22, 2023.

Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, and Shinji Watanabe. 2018. Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline.

Hakan Erdogan, John R. Hershey, Shinji Watanabe, Michael I. Mandel, and Jonathan Le Roux. 2016. Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks. In *Proc. Interspeech 2016*, pages 1981–1985.



(a) AaWLoss 在 α 爲 1(左)、10(中)、500(右) 時的增強語音聲譜圖



(b) AaWLoss 在 α 爲 1(左)、10(中)、500(右) 時的噪聲條件僞影聲譜圖

圖 3. 比較不同設定下，AaWLoss 產生之噪音條件僞影。該案例正確內容爲"FIRST"，α 爲 1(左)、10(中) 時卻被辨識爲"BEST"。紅色方框標記爲音素/f/的發音範圍，唯有 $\alpha = 500$(右) 保留了/f/的發音特徵。

Masakiyo Fujimoto and Hisashi Kawai. 2019. One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features. pages 486–490.

Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. 2016. Neural network based spectral mask estimation for acoustic beamforming. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200.

Kuan-Hsun Ho, En-Lun Yu, Jeih weih Hung, and Berlin Chen. 2023. Naaloss: Rethinking the objective of speech enhancement.

Yuchen Hu, Chen Chen, Ruizhe Li, Qiushi Zhu, and Eng Siong Chng. 2023. Gradient remedy for multi-task learning in end-to-end noise-robust speech recognition.

Kazuma Iwamoto, Tsubasa Ochiai, Marc Delcroix, Rintaro Ikeshita, Hiroshi Sato, Shoko Araki, and Shigeru Katagiri. 2022. How bad are artifacts?: Analyzing the impact of speech enhancement errors on asr.

Tobias Menne, Ralf Schlüter, and Hermann Ney. 2019. Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust asr. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6660–6664.

Hyun Joon Park, Byung Ha Kang, Wooseok Shin, Jin Sob Kim, and Sung Won Han. 2022. Manner: Multi-view attention network for noise erasure.

Ke Tan and DeLiang Wang. 2020. Improving robustness of deep learning based monaural speech enhancement against processing artifacts. In

ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6914–6918.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2016. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA Speech Synthesis Workshop*, pages 146–152.

E. Vincent, R. Gribonval, and C. Fevotte. 2006. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469.

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2014. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68.