

國立臺灣大學電機資訊學院資訊工程學系

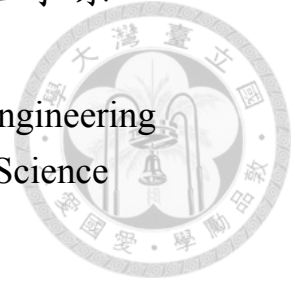
碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis



改善殘響環境中的自動語音辨識

Improving ASR in Reverberant Environments

廖彥綸

Yen-Lun Liao

指導教授：張智星博士

Advisor: Jyh-Shing Roger Jang, Ph.D.

中華民國 111 年 6 月

June, 2022





國立臺灣大學碩士學位論文
口試委員會審定書

改善殘響環境中的自動語音辨識

Improving ASR in Reverberant Environments

本論文係廖彥綸君（學號 R09922047）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 111 年 6 月 3 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

張智星

王新如

（指導教授）

林其翰

系主任

洪士瀨



誌謝

就讀台大資訊系研究所並不在我原本的就學規劃中。我於大四申請到了國外研究所時新冠肺炎疫情在全世界蔓延，經過再三考慮後我決定放棄國外的錄取機會，而在台大完成碩士學歷。雖然我擁有在大學部就讀的經驗，但碩士的學習和大學有許多不同之處，我有幸在許多人的幫助下讓碩士生活一步步進入正軌。

首先非常感謝指導教授張智星老師，提供許多指導，並且在口試練習中不耐煩的給予回饋，讓我得以發現研究上的缺點及設計更適切的實驗。同時，感謝炫均、淑芬等戰友，以及世展、政鷹、小龜、葉子等前輩，還有語音組的所有同學和學弟妹，在實驗陷入瓶頸時能得到大家的建議和想法，讓實驗進度可以持續推進。此外，感謝國家高速電腦中心提供計算與儲存資源，提供優質的硬體設備，讓實驗可以有效率地完成。

感謝高中朋友們，當我煩悶時總時能為我解憂，帶來生活中的調劑。最後感謝我的父母和兄弟，他們給我美好的家庭，讓我能無憂無慮進入研究所；他們支持我做許多的決定，帶給我背後的依靠；他們包容我的錯誤，讓我能不畏地再次奮鬥。感謝上述所有人的幫助。



摘要

本研究統合殘響去除和語音辨識系統，以增進語音辨識對具有殘響的語音訊號的辨識能力。當今常用語音辨識結合殘響去除的架構簡易地將通過殘響去除模型的訊號作為下階段語音辨識模型的輸入。這種方式雖然看似簡單明確，但存在殘響去除模型的輸出與語音辨識聲學模型的訓練資料特性不一致的問題，而可能使最後辨識的效果有所下降。

本文提出新的架構，並藉由四個面向改善原始的語音辨識系統：

1. 乾淨與殘響特性分類器
2. 殘響去除模型
3. 避免訓練及測試語音資料特性不一致
4. 模型組合

當訊號進入本架構後，使用各種不同的資料訓練，將進入針對語音品質的乾淨與殘響特性分類器，以讓訊號選擇適合的聲學模型。在此分類器中本研究探討 LCNN 搭配 MAX、AVG、SAP、ASP 等池化層的訓練架構的差異；根據分類器結果，被分類致殘響類別的語音訊號將經過殘響去除模型後，由利用相同性質的資料訓練的聲學模型辨識，避免特性不一致的問題。乾淨的語音訊號也會是更較適合的對應的方式辨識。最後使用模型組合方法，本研究提出 sentence-level fusion (SLF) 及 word-level fusion (WLF) 的組合方式以尋求更低的字元錯誤率 (CER)。

根據實驗結果，在自行生成的 aishell1 具有殘響的語音訊號測試集中使用 MetricGAN 去除殘響後，經過 tdnn 的聲學模型訓練中字元錯誤

率由原本的 15.26% 降低至 13.22 % ；Mixup 的資料擴增實驗在 triphone 的聲學模型中，和原始的訓練方式相比，字元錯誤率由 17.83% 降低至 17.34%；在自製的 aishell1 乾淨和含有殘響語音訊號的混合測試集中，使用 sentence-level fusion 或 word-level fusion 的模型混合方式的字元錯誤率為 7.23% ，相較於直接使用乾淨訊號的聲學模型 (CER = 9.12%) 或含有殘響的語音訊號的聲學模型 (CER = 7.85%) 有最多 20.72 % 的錯誤率減少。

關鍵字：自動語音辨識、殘響去除、生成對抗模型、資料擴增、動態規劃



Abstract

The emergence of reverberations usually corrupts the quality of indoor signals, giving rise to performance degradation in automatic speech recognition (ASR). To minimize the influence of reverberations, acoustic dereverberation models are established to pre-process the original signals before submitting them to ASR. This structure leads to an apparent improvement. However, the dereverberation model's output is inconsistent with the training dataset of ASR, resulting in the decline in the performance of ASR.

This paper refined the previous structure from four aspects:

1. signals classification
2. reverberation removal
3. data mismatch offset
4. string fusion

As soon as the audio stream was submitted to the proposed system, the reverberation classifier determined whether the signal was clean or reverberated. Depending on whether the signal was counted as reverberation, the system will submit the signal to a dereverberation model before sending it to ASR or will send the signal to ASR directly. The routine selection also helped the signal choose the most proper acoustic model (AM) whose establishment is trained using the audio stream with the corresponding acoustic label. Furthermore, this paper proposed sentence-level fusion (SLF) and word-level fusion (WLF) as methods to fuse the two results.

By dereverberating the signals with MetricGAN, the character error rate

(CER) had decreased from 15.26% to 13.22% in the rev-aishell1 test set with using tdnn acoustic model. With the help of mixup augmentation, CER decreased from 17.83% to 17.34% in triphone acoustic model. When SLF and WLF were applied, a CER of 7.23% was reached in the reverberant and clean aishell1 test set, achieving an improvement in the CER by 20.72% compared to the single model.

Keywords: automatic speech recognition, dereverberation, GAN, data augmentation, dynamic programming



Contents

誌謝	iii
摘要	iv
Abstract	vi
1 緒論	1
1.1 研究動機	1
1.2 研究貢獻	2
1.3 章節概述	2
2 文獻探討	4
2.1 語音辨識之背景知識	4
2.1.1 傳統 HMM 之 ASR 訓練方法	4
2.2 過去之殘響去除研究	7
2.2.1 基於演算法之殘響去除	7
2.2.2 基於神經網路之殘響去除	8
2.2.3 殘響去除模型的資料擴增	12
2.3 分類器及池化層之研究	13
2.3.1 LCNN	13
2.3.2 池化層	14
2.4 語音品質估計函數	16
2.4.1 PESQ	16
2.4.2 STOI	16

2.4.3	SRMR	17
2.4.4	CER	18
3	研究方法	19
3.1	殘響與乾淨特性分類器	20
3.2	殘響消除模型	22
3.2.1	用於語音的 Mixup 資料擴增	23
3.3	於聲學模型之資料擴增	24
3.4	模型結合 (字串組合)	24
3.4.1	Direct ensemble	26
3.4.2	Sentence-level fusion	26
3.4.3	Word-level fusion	27
4	語料介紹	32
4.1	語音辨識資料集	32
4.2	含有殘響的語音資料集	32
4.2.1	RIR_NOISES	33
4.3	資料生成	33
5	實驗設計與結果	35
5.1	實驗流程及訓練參數設定	35
5.1.1	聲學模型	36
5.1.2	語言模型	37
5.1.3	傅立葉轉換	38
5.1.4	殘響與乾淨特性分類器模型	38
5.1.5	Bi-LSTM 模型	39
5.1.6	MetricGAN 模型	39
5.2	效果評估方式	39
5.3	訓練設備規格	39
5.4	實驗一：基礎 ASR 訓練之結果	40
5.5	實驗二：殘響去除模型的分析及比較	41

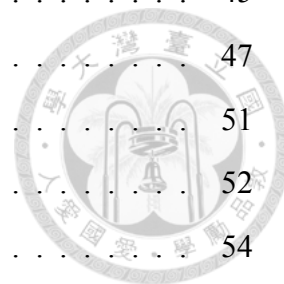
5.5.1	不同殘響去除模型(演算法)之間的比較	41
5.5.2	不同房間對於殘響去除模型及整體字元錯誤率之影響	43
5.5.3	應用 Mixup 資料擴增方法於殘響去除模型的效果	45
5.5.4	減少資料不一致的討論	46
5.6	實驗三：殘響與乾淨特性分類器訓練之探討	48
5.6.1	池化層比較	48
5.6.2	分類器結果分析	49
5.7	實驗四：探討模型結合演算法對字元錯誤率的影響	50
5.7.1	溫度法的分析	50
5.7.2	模型融合的分析	51
6	結論與未來展望	56
6.1	結論	56
6.2	未來展望	57
	Bibliography	58



List of Figures

2.1	差分特徵譜處理流程	5
2.2	LDA 特徵處理流程	5
2.3	以 HMM 建模聲學模型示意圖	6
2.4	使用 DNN 進行殘響去除示意圖	9
2.5	使用 LSTM 進行殘響去除示意圖	10
2.6	Mixup 示意圖	13
2.7	MFM 及 ReLu 比較圖 [1]	14
2.8	SAP 及 ASP 實作示意圖。左：SAP [2]，右：ASP [3]	16
3.1	原始殘響去除模型結合 ASR 架構	20
3.2	提出之殘響去除模型結合 ASR 架構	20
3.3	MetricGAN 訓練架構圖	23
3.4	時間維度維度橫向拼接 Mixup	23
3.5	頻率維度縱向拼接 Mixup	24
3.6	Sentence level fusion	26
3.7	Word level fusion	26
3.8	單詞對齊範例一	27
3.9	單詞對齊範例二	27
3.10	K-best LCS 範例	29
3.11	sausage 與 lattice 構圖的比較	30
5.1	ASR 基準線實驗結果	40
5.2	殘響去除對應評估方式比較圖	42
5.3	不同房間聲音品質綜合比較圖	44

5.4	Mixup 資料擴增對應評估方式比較圖	45
5.5	利用不同資料訓練之聲學模型於 CER 比較圖	47
5.6	溫度調整	51
5.7	不同模型組合方式比較圖	52
5.8	Model Fusion 實例	54





List of Tables

2.1	PE 參數對照圖	7
3.1	LCNN with ASP (frame = 400 ms)	21
3.2	路徑合併問題	27
3.3	WLF 參數對照圖	30
3.4	Sausage 與 WLF	31
4.1	Aishell1 統計	32
4.2	RIR_NOISES 模擬資訊	33
4.3	聲學模型使用資料分析	34
5.1	GMM-HMM 訓練參數	38
5.2	ASR 之基準線測試	40
5.3	殘響去除對應評估方式詳細結果	42
5.4	評估使用不同模型於不同房間之結果	44
5.5	Mixup 資料擴增對應評估方式詳細結果	46
5.6	利用不同資料訓練之 tri5 聲學模型的 CER 表	47
5.7	利用不同資料訓練之 tdnn 聲學模型的 CER 表	48
5.8	池化層比較	49
5.9	兩聲學模型的字元錯誤率比較圖	50
5.10	殘響與乾淨特性分類器與字元錯誤率對應表	50
5.11	不同模型組合方式比較表	52
5.12	Model Fusion 錯誤分析	53
5.13	Direct ensemble 和 SLF 比較	54

5.14 Direct ensemble 和 WLF 比較	54
5.15 SLF 和 WLF 比較	55





Chapter 1

緒論

本章節藉由分析現行 ASR 架構之缺點帶出本研究的動機、並簡述之後章節的概覽、及本研究的貢獻。

1.1 研究動機

語音辨識 (automatic speaker recognition, ASR) 為將人類所發出的聲音訊號轉換成可讀文字的過程。現今已有非常廣泛的應用，在合適的測試環境中，包括產生會議、影片、或日常使用的手機應用服務的逐字稿皆有辦法生成對應的結果。傳統上語音辨識分為兩個獨立訓練的模組：聲學模型 (acoustic model, AM) 及語言模型 (language model, LM)。雖然當今端對端 (end-to-end) 的語音辨識模型 [4] 也被提出並實作，但其效果更依賴資料的數量及與測試資料間的特性的一致性。在文本有限的情況下，傳統的方式依舊有所優勢。聲學模型和語言模型相輔相成達成語音辨識的任務，聲學模型透過特徵萃取和機率分布模型估計最能代表該訊號的音素序列；語言模型由統計大量文本的單詞前後的關係建構，在語音辨識運行時語言模型接收聲學模型辨識出的音素序列後回傳生成機率最大的文字串。

欲使語音辨識模型有更小的字元錯誤率需要克服許多挑戰，聲學模型及語言模型皆有必須克服的難題。如：語言模型傳統上使用 N-gram 觀察單詞前後不同詞出現的機率建立機率模型，由於只有前 N-1 個詞在辨識時被考慮，無法讓較遠的訊息提供資訊。之後的研究常使用變換器 (transformer) [5] 改善此問題，其中又以雙向編碼器表示技術 (bidirectional encoder representations from Transformers, BERT) 為最常見。

BERT) [6] 最為著名，並讓文字串能適切地生成。聲學模型主要面對的挑戰為錄音時的環境干擾，其中以殘響的影響最為普遍。殘響為在室內環境中當聲音源消失時訊號依舊短暫殘留於房中的現象。所有室內環境都有殘響出現，依照房間大小、形狀的差異，殘響對訊號產生不同程度的影響。殘響使得聲學模型辨識音素的困難性增加，嚴重的殘響甚至能讓辨識完全失準，於是讓 ASR 處理具有殘響的語音訊號是必要的研究。

一般架構中設計殘響去除模型並直接應用於 ASR 之前作為訊號前處理，達到統一消除測試訊號中的殘響的效果。然而殘響去除模型的輸出將不會應用為 ASR 聲學模型的訓練上，造成資料特性不一致的問題。本文提出應用於殘響去除和語音辨識的新架構，以四個面向改善現有的架構：

- 分別設計用於乾淨語音訊號及擁有殘響的語音訊號的聲學模型。
- 訓練用於殘響與乾淨特性的分類器，歸類輸入的訊號使其能選擇較適合的聲學模型進行辨識。
- 將通過殘響與乾淨特性分類器的訊號加入聲學模型的訓練中以消除資料不一致性。
- 本研究提出 sentence-level fusion (SLF) 和 word-level fusion (WLF) 的字串組合方式，藉由綜合考慮不同模型的輸出，追求更低的字元錯誤率。

1.2 研究貢獻

本研究的貢獻為下列數項：

1. 提出新的架構能有效降低在室內環境下的 ASR 字元錯誤率。
2. 分析數種殘響去除模型於不同語音品質評分函數的結果。
3. 提出 SLF、WLF 等用於文字串的模型組合方法。

1.3 章節概述

本文分為六個章節，規劃如下：

- 第一章為緒論：介紹本文的主題、研究動機與研究貢獻。
- 第二章為文獻探討：主要介紹討論不同的殘響去除模型和語音品質評分函數及使用到的相關模型。
- 第三章為研究方法介紹：詳細介紹本文提出的架構及使用的方法。
- 第四章為資料集說明：介紹研究中使用的資料集和生成方式。
- 第五章為實驗設計與結果：詳細介紹實驗的參數設計和架構，且呈現實驗結果和相互的比較關係。
- 第六章為結論與未來展望：總結實驗內容及結果，並且討論未來可能之改進。





Chapter 2

文獻探討

本章節將會回顧本研究將會使用的及相關的背景知識，內容包含使用傳統 HMM 之 ASR 訓練方法、過往對於殘響去除之研究、分類器的使用、語音品質估計函數。

2.1 語音辨識之背景知識

2.1.1 傳統 HMM 之 ASR 訓練方法

傳統上實作的 ASR 分為聲學模型和語言模型兩部分，聲學模型通常使用隱馬可夫傳遞鍊 (hidden markov model, HMM) 建模。抽取聲音特徵後，進一步經過隱馬可夫傳遞鍊估計機率最高的音素 (phone) 序列；語言模型傳統上以 N 元 (N-gram) 模型為主，以大量文本建模並統計文字前後的關聯性。ASR 結合兩模型生成逐字稿的結果，聲學模型將聲音估計不同機率對應的音素序列，語言模型接收預測的音素序列，計算產生各種文字序列的機率。之後的章節會針對聲學模型做調整，應對在殘響環境中的辨識，語言模型則為實驗的控制變因，以探討不同實驗下聲學模型的優劣。

聲學模型

音素為語言學家設計用以代表聲音的最基本單位，聲學模型旨在將聲音訊號轉換成可能的音素序列。常見的聲學模型中，訊號常先經過特徵抽取，利用語

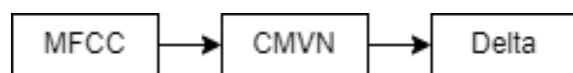


圖 2.1: 差分特徵譜處理流程



圖 2.2: LDA 特徵處理流程

音學的方式將聲音以更細緻的特徵表現，如：梅爾頻譜倒譜係數 (mel-frequency cepstral coefficients, MFCC)，每個音框時間維度的資訊轉換成頻率維度以獲取更細微的資訊後，再結合基頻特徵、和不依賴標記的無監督式的特徵轉換，如：在譜歸一化 (cepstral mean and variance normalization, CMVN)、頁框拼接及差分係數 (Delta) 的擴充 (圖 2.1) 作為 HMM 的輸入。

差分係數定義為計算音框前後的差分值為特徵，作為目前特徵的補充資訊。公式 2.1 為第 t 個音框的第 m 維特徵的一階差分計算方式，通常選擇 $W = 2$ 為鄰近音框，進行差分的計算。二階差分以相同公式則計算於一階的基礎上。頁框拼接為將前後許多頁框合併為一頁框的特徵，由於使用頁框拼接後特徵的維度將暴增為數倍，常搭配 LDA 等降維方式使用，如圖 2.2，LDA 為監督式的降維方式，為透過標記資訊訓練的轉換模型。CMVN 的目的是將特徵整體調整至正態分布，以避免訓練偏向部分特徵值，本研究使用的資料有語者標記，CMVN 的計算以個別語者為單位進行。

$$\Delta f(t, m) = \frac{\sum_{d=-W}^{d=W} d \times f(t + d, m)}{\sum_{d=-W}^{d=W} d^2} \quad (2.1)$$

圖 2.3 為 HMM 建模聲學模型示意圖。HMM 模型以發音階段不同建立不同的狀態 (state)，接收抽取後的特徵計算在狀態間轉移的機率。每個狀態中有不同的觀察 (observation) 機率模型評估在該狀態出現某音素的機率，傳統方法中高斯混合模型 (GMM) 是常用的觀察機率模型建模的方式。

深度學習的方法興起後帶給 ASR 相當大的改進。有許多研究直接以神經網路架構訓練聲學模型。但當前的較佳的聲學模型實作方式為：利用三音素模型作為起始系統，為神經網路的訓練提供對齊，再使用神經網路訓練。本文使用時延網路 (time delay neural network, TDNN) [7] 架構進行神經網路訓練。TDNN 在訓練

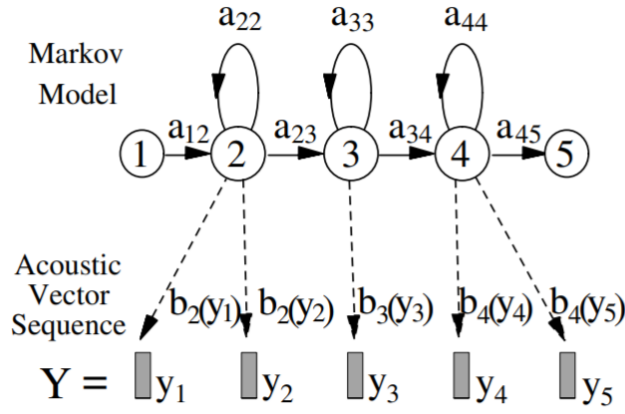


圖 2.3: 以 HMM 建模聲學模型示意圖， a_{ij} 為狀態 i 轉移至狀態 j 的機率，每個狀態有 $b_i(y_k)$ 的機率觀察到音素 k

時除了前層的特徵外，也會參考同層鄰近時間點的特徵向量，適合使用在時間變化上有影響的任務。

語言模型

聲學模型觀察 HMM 機率分布，得到可能代表訊號的音素序列，但僅憑藉聲學模型難以獲得低的字元錯誤率，尚需要語言學的知識協助辨識出正確的文字。傳統上最常使用的是 N-元 (N-gram) 模型，以上文資訊估算單字出現的條件機率。在 N-元模型，第 N 個出現的字只會受到前 N-1 個字的影響，公式 2.2 中 w_{i-k} 為單字 w_i 的前 k 個字。通過大量的語料蒐集讓單純考量單字前後文的方式形成統計上的意義。由於預測的文字和搜集語料的領域相關，通常使用於不同領域會有對應適合的語言模型。

$$P(w_i | w_{i-(N-1)}, \dots, w_{i-2}, w_{i-1}) = \frac{\text{count}(w_{i-(N-1)}, \dots, w_i)}{\text{count}(w_{i-(N-1)}, \dots, w_{i-1})} \quad (2.2)$$

N 過大會使 w_{N-1}, \dots, w_1 後出現 w_N 的可能性降低，而通常取 N=3 可以得到較佳的效果，此模型被稱為 tri-gram。當測試中總有單詞不存在於蒐集的訓練資料的情況，若未進行任何調整則將產生錯誤。實作上使用模型平滑處理此議題，讓每個詞出現機率不為零。除了 N-元模型，尚有使用神經網路的語言模型，如：NNLM [8] 及 RNNLM [9] 等...。語言模型為本實驗中的控制變因，本文僅使用固定的 3-元語言模型進行語音辨識，作為評估不同聲學模型或是測試資料的指標。

表 2.1: PE 參數對照圖

$s(k)$	時間 k 時的聲音源訊號
$h(k)$	時間 k 時的脈衝響應
$v(k)$	時間 k 時的外在噪音 (傳送過程中非週期性的部分)
G	聲音傳送的線性轉換過程
$x(k)$	時間 k 時的預測誤差
$F(G)$	整體訊號的預測誤差

2.2 過去之殘響去除研究

2.2.1 基於演算法之殘響去除

殘響去除的議題以資料的利用區分為三個主要研究方向。其一：使用乾淨音檔和具有殘響音檔的訓練對進行端對端監督式訓練。主要目的為調控轉換模型參數，讓含有殘響的語音訊號能通過模型乾淨訊號為目標學習。然而，同時蒐集乾淨和擁有殘響的語音在實際狀況不易且成本昂貴，常以人工方式在乾淨音檔中加入模擬的殘響替代。其二：僅使用具有殘響的訊號進行非監督式訓練，此方式大幅降低對資料的依賴性，使用生成對抗網路 (Generative Adversarial Network, GAN) 的架構時可使用此方式訓練，但由於缺乏學習目標，目前效果和監督式學習比略遜一籌。其三：以演算法針對音檔進行廣泛性的殘響去除。機器學習方式有較多的資料量要求，訓練時也需要更多的訓練時間和計算資源，傳統演算法方式能免除機器學習需求的大量資料、訓練時間和運算資源的限制。本章節會進行傳統方法及機器學習方法的回顧，方法間的效果比較會在之後的章節進行。

預測誤差法

預測誤差法 (prediction error, PE) 將聲音的傳送過程視為線性轉換，利用線性預測使得輸入 (具有殘響語音訊號) 和目標 (乾淨語音訊號) 的差距最小。表 2.1 為 PE 使用到的參數的表示 ($0 \leq k \leq T$)。公式 2.3 可以代表聲音傳輸過程。

$$y(t) = \sum_{\tau=0}^{T-1} h(\tau)s(t-\tau) + v(t) \quad (2.3)$$

在 PE 中聲音從聲音源至接收端被視為線性轉換過程 G 。由公式 2.4, 2.5 計算預

測誤差 $x(t)$ 。

$$\tilde{y}(t) = \sum_{\tau=0}^{T-1} G(\tau)s(t-\tau) \quad (2.4)$$

$$x(t) = y(t) - \tilde{y}(t) \quad (2.5)$$

殘響去除在此問題被視為預測誤差 (PE) 的最佳化問題，對 $x(t)$ 進行最佳化運算，計算和接收端誤差最小時值 $F_{PE}(G)$ (公式 2.6， $\|\cdot\|^2$ 表示向量正規化)，反推線性函數 G ，並可由 G 和接收的 $y(t)$ 推算出原始聲音源的 $s(t)$ 。

$$F_{PE}(G) = \sum_{t=0}^T \|x(t)\|^2 \quad (2.6)$$

加權預測誤差法

前小節所述的預測誤差法對 $F_{PE}(G)$ 進行最佳化，但之後研究顯示當語音訊號是非平穩的 (nonstationary) 過程時，不能保證最佳化後可以推算出適合的 G 。在平穩過程中任取一段區間，其平均值和變異數將與該區間任意平移後的計算得到的值相同，人類發出的訊號幾乎都不符合此規定。加權預測誤差法 (weighted prediction error, WPE) [10] 針對訊號強度設計時變方差 (time-varying variance) 為權重 $\lambda(t)$ 。

$$F_{WPE}(G) = \sum_{t=0}^T \frac{\|x(t)\|^2}{\lambda(t)} \quad (2.7)$$

人類發出的聲音訊號存在特大值或特小值使其成為並非平滑的分布，PE 在誤差計算時容易因為這些值的影響，而得到不適當的線性轉換。和 PE 相比，WPE 在每個時間點增加 $\lambda(t)$ 調整權重，使得每個時間段對誤差函數的重要性有相等的貢獻，這在殘響去除的任務中起到較好的效果。

2.2.2 基於神經網路之殘響去除

運算能力的躍進使神經網路廣泛運用於許多領域，模型架構、訓練方式等皆影響著訓練收斂的效果。本章節將回顧應用於殘響去除代表性的神經網路方式。

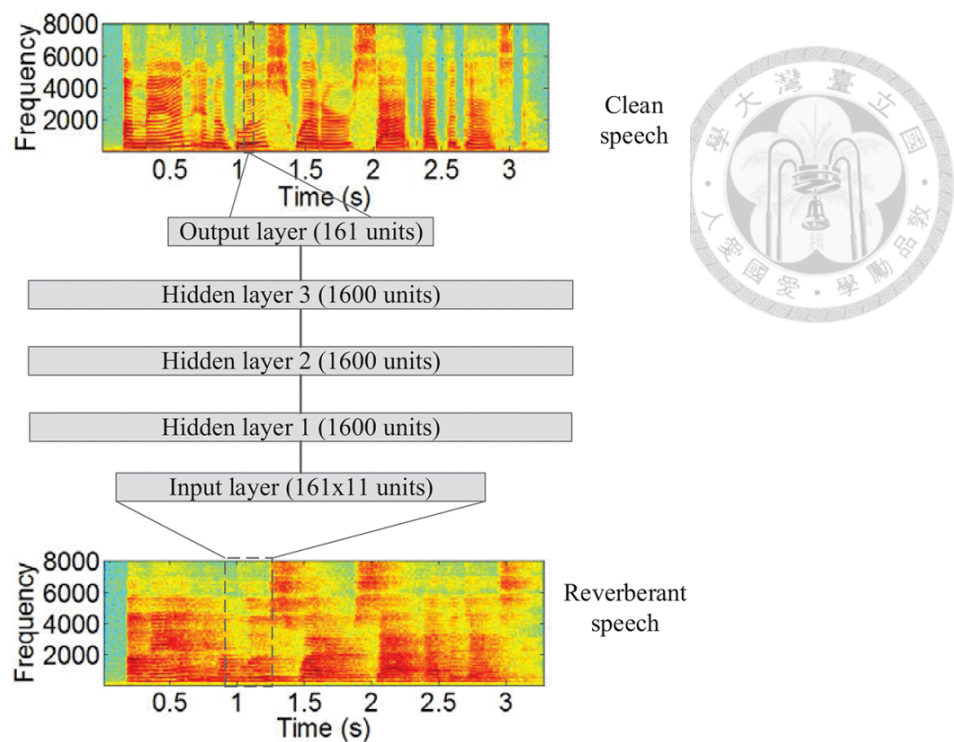


圖 2.4: 使用 DNN 進行殘響去除示意圖，輸入額外取前後各五行特徵作為額外資訊

DNN 模型

Han Kun 等人於 2015 年發表 [11] 利用 DNN (deep neural network) 模型對音訊進行殘響去除，使用 STFT (short-time fourier transform, 20 ms time frames, 10 ms time shift) 將音訊轉換成頻率維度作為輸入 DNN 模型的特徵，如圖 2.4 (引用自 [11])，除了當時時間點外輸入包含前後各五個時間點的音框作為額外資訊，以達到更好的訓練表現，隨後經過三層維度 1600 的隱藏層後以乾淨的訊號進行殘響去除的訓練，訓練時使用具有殘響和乾淨聲音當時時間點的 MSE (mean square error) 作為損失函數。

DNN 將音訊劃分為獨立的音框進行訓練並以附近的音框為額外資訊訓練，但在資料處理上擁有兩點明顯的缺陷：一、模型取得的資訊限縮於一定範圍內，無法得到更遙遠的資訊。二、音訊的前後端需要面對補值，讓訓練效果下降。此兩點互相抵觸，如果想要讓模型得以使用更遙遠的資訊訓練，則勢必需要增加輸入的範圍，也讓更多的音框面對到補值。

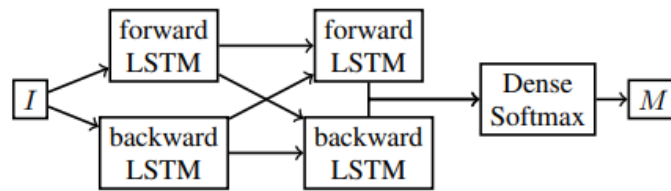


圖 2.5: 使用 LSTM 進行殘響去除示意圖，模型雙向且遞迴的接收新訊息訓練。

Bi-LSTM 模型

雙向長短期記憶（bidirectional long short-term memory，Bi-LSTM）應用於殘響去除上 [12] 讓模型能藉助於歷史上的資訊訓練，並解決了部分 DNN 遇到的問題。記憶單元和遞迴輸入的形式讓模型能夠使用更遠的資訊訓練；雖然補值的問題依舊存在，但雙向的架構讓訓練時至少有一個訓練方向是完整的資訊。Bi-LSTM 也成為殘響去除中熱門的解法。圖 2.5 (引用自 [12]) 為 Bi-LSTM 訓練的示意圖，此模型包含兩長短記憶單元記憶歷史資訊並處理音檔有多種長度的問題。和前述所有解法相比，所需要的訓練資料量和訓練時間大增，並擁有更高的硬體要求。

GAN 模型

結合生成對抗網路 (Generative Adversarial Network，GAN) 生成去除殘響之後的音訊 [13] 也取得有競爭力的結果。殘響去除 GAN 的鑑別器 (discriminator) 訓練判斷音訊含有殘響值的分數；生成器 (generator) 生成乾淨的訊號混淆鑑別器。為了增進生成器創造音訊的乾淨程度，研究針對架構及訓練方式有所改良。HiFiGAN [14] 建構多層次的生成網路：含有殘響的語音訊號經過 WaveNet [15] 和 PostNet [16] 後生成相對乾淨的訊號。鑑別器也分別針對頻率維度和時間為度的聲音訊號進行評分。MetricGAN [17] 改善生成器損失函數的選擇，藉由 PESQ (perceptual evaluation of speech quality)、STOI (short-time objective intelligibility) 等語音乾淨度評估指標 (後敘) 選擇更適當的生成模型；MetricGAN-U [18] 進一步使用非侵入式的 (non-intrusive) 評估函數，讓訓練得以選擇在非監督式的狀態下實現。

MetricGAN 模型

如普通的 GAN 模型，MetricGAN [17] 由鑑別器 (discriminator, D) 和生成器 (generator, G) 兩個模型組成。鑑別器以評估函數為訓練目標，接收任意音訊，並分類聲音的環境更接近乾淨或殘響；生成器接收含有殘響的音檔，並生成對應該音檔乾淨化的結果。最後生成器期望生成足夠乾淨的音訊以欺騙過鑑別器。GAN 的生成器以通用的損失函數 (如：MSE、L1Loss) 訓練，MetricGAN 改良鑑別器的訓練方式，使用判斷聲音乾淨的指標 (如：SRMR、STOI) 為損失函數，使鑑別器能以更符合讓聲音變清楚的目標評價訊號。

普通 GAN 模型鑑別器的損失函數 L_D 如公式 2.8 [19]：

$$L_D = -\left(\sum_{i=1}^m y_i \log(D(x_i)) + \sum_{i=1}^m (1 - y_i) \log(1 - D(x_i))\right) \quad (2.8)$$

其中 x_i 為第 i 筆輸入音訊、 y_i 為該音訊對應的類別 (乾淨音訊或由生成器產生去除殘響後的音訊)、 $D(x_i)$ 為 x_i 通過鑑別器的分類結果。損失函數借助最小化交叉熵 (cross entropy) 以分離兩者。然而該損失函數僅評估生成結果和目標結果對單一分類器的評估，未考慮到語音品質的衡量方式。

MetricGAN 以聲音品質評估函數 $Q(a, b)$ (如：PESQ、STOI 等) 為訓練目標，讓鑑別器得以更接近判斷聲音是否乾淨的真實情況。其中 b 為原始的乾淨聲音； a 為待評估的聲音，通常為相對於原始音檔在噪音環境下或進一步處理過後的訊號， $Q(a, b)$ 回傳基於 b 為標準， a 的聲音品質分數，鑑別器則以該結果為對象訓練。損失函數 $L_{MetricGAN_D}$ 如公式 2.9 [17] 改良。

$$\begin{aligned} L_{MetricGAN_D} = & (D(y, y) - Q(y, y))^2 \\ & + (D(x, y) - Q(x, y))^2 \\ & + (D(G(x), y) - Q(G(x), y))^2 \end{aligned} \quad (2.9)$$

y 代表乾淨訊號、 x 代表殘響環境下的訊號、 $G(x)$ 代表通過生成器的殘響去除功能後的聲音訊號，此三項和原始音檔的乾淨度評分構成 MetricGAN 鑑別器的損失函數，透過這三個方向鑑別器實現學習模擬目標評估函數的方式。生成器方面使用對抗式函數 (公式 2.10) [17]，讓鑑別器的輸出和評估函數的值靠近的方式訓

練，藉此混淆鑑別器。

$$L_{MetricGAN_G} = (D(G(x), y) - Q(G(x), y))^2 \quad (2.10)$$

實驗中訓練的學習對象使用預評估訊號和乾淨訊號成對的評估函數，也可以依情況不同，使用黑箱 (black-box) 評估或就聲音及頻譜狀態的非侵入式評估函數 (如：SRMR)，如此訓練過程可以成為非監督的學習 [18]，減少對成對訓練資料的依賴性，如此鑑別器的損失函數改寫成公式 2.11 [18]，生成器的損失函數改寫成公式 2.12 [18]。

$$L_{MetricGAN-U_D} = (D(x) - Q(x))^2 + (D(G(x)) - Q(G(x)))^2 \quad (2.11)$$

$$L_{MetricGAN-U_G} = (D(G(x)) - Q(G(x)))^2 \quad (2.12)$$

在非監督式的條件下，資料蒐集不需要依賴乾淨的語音訊號和具有殘響的語音訊號的配對也可訓練鑑別器。提供殘響消除模型另一種可能的研究方向。

2.2.3 殘響去除模型的資料擴增

神經網路模型藉由改變架構、特化訓練流程等... 方式提升模型訓練的結果。但對於神經網路，提升資料的數量是最顯而易見的方式，同時多變化的資料能顯示不同面想避免模型過擬和 (overfitting)。

Mixup 資料擴增

多數實驗中使用資料擴增需要藉助背景知識，例如：在圖片分類中常用的旋轉、鏡射擴增；在處理文字、數字辨識的題目時並不合適。Mixup [20] 為 2019 年在圖片訓練資料擴增的新想法，其效果已在物件辨識上被證明。透過不同類別的圖像彼此重新組合形成新的圖片，該圖片擁有組成的兩個類別各自的比例。圖 2.6 為期示意圖：紅藍兩色代表不同類別等大圖片，新的圖片包含兩者部分的特徵。若類別一 C_1 佔有新圖片 C_{mix} 比例中的 $\lambda (0 < \lambda < 1)$ ，類別二 C_2 佔有剩下的



圖 2.6: Mixup 示意圖：紅藍兩色代表不同類別等大的圖片，透過縱向結合 (左)；橫向結合 (右)

$1 - \lambda$ ，圖片間的關係表示為公式 2.13。

$$C_{mix} = \lambda C_1 + (1 - \lambda) C_2 \quad (2.13)$$

損失函數 L_{mix} 的計算也跟隨比例調整，公式 2.14 為更新後的損失函數。 L_{C_1} 和 L_{C_2} 分別為類別 C_1 ， C_2 的損失函數。

$$L_{mix} = \lambda L_{C_1} + (1 - \lambda) L_{C_2} \quad (2.14)$$

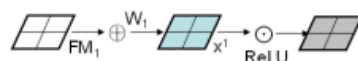
此圖像的資料擴增技巧也被應用在聲音訊號的訓練上，將聲音訊號經過頻率維度上的特徵抽取可以轉變成類似圖片的表示方法，包括語者反偽造偵測 [21]、聲音訊號分離等領域 [22] 使用過 mixup 的資料擴增技術。於是本文調整實作方式將其應用於殘響消除模型的訓練，實作方式將於章節 3.2.1 中描述。

2.3 分類器及池化層之研究

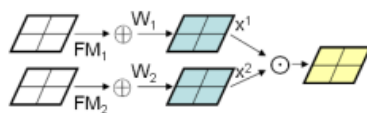
神經網路在分類的議題擁有傑出的表現，圖片分類器如一般的 CNN 模型至更進階的 VGG (visual geometry group) 網路和 ResNet (residual network) 利用更深的疊代或是殘差特徵訓練網路，雖然效果逐漸提升，但伴隨的往往是更大量的運算需求。在有限的計算資源中能得到相差不遠的效果為另一研究方向。

2.3.1 LCNN

LCNN (light convolutional neural network) [1] 最先使用於人臉辨識上，隨後也被應用於其他物件辨識。LCNN 提出的 MFM (max-feature map) 改善 CNN 常用的激



$$\text{ReLU: } h(x) = \max(0, x^1)$$



$$\text{MFM 2/1: } h(x) = \max(x^1, x^2)$$

圖 2.7: MFM 及 ReLu 比較圖。

勵函數 (activation function)：ReLU、PReLU 等容易造成資訊損失的現象，且達到中和大量資料可能存在錯誤標記的影響，同時減少模型參數數量。MFM 2.7 提出的方法中，本實驗使用其中兩層合一 MFM 的概念，兩層合一的 MFM 使用相鄰兩層架構的最大值取代激勵函數的輸出，讓模型得以綜合考慮相近兩層的資訊，使用 MFM 同時達到讓模型更輕量化的效果。

2.3.2 池化層

和圖片辨識不同，聲音相關的處理被期待的能兼容地處理各種不同長度的音訊。為了達成此目標，利用池化層 (pooling layer) 濃縮時間維度的資訊為可行方式。除了普通的 MAX (maximum pooling) 和 AVG (average pooling)，SAP (self-attentive pooling) [2] 和 ASP (attentive statistics pooling) [3] 的概念在後來的研究陸續被提出。

SAP

SAP (圖 2.8，左) 首先使用感知器 (perceptron) 抽取輸入特徵的隱藏層，實驗中使用如式 2.15 的單層感知器。

$$h_t = \tanh(Wx_t + b) \quad (2.15)$$

x_t 為單位時間之音框的輸入特徵， h_t 為其隱藏層特徵， W 及 b 為感知器的可訓練的參數。抽取完的 h_t 和另一可訓練向量 μ 以相似度計算估計在不同時間下音

框的重要性，公式 2.16 表示每個音框權重 (重要性) 的計算方式，且調用 softmax 算法進行正規化，讓 α_t 能代表對應的 $h_t(x_t)$ 的權重。最後可利用 $\sum_{t=1}^L \alpha_t x_t$ 作為池化結果。

$$\alpha_t = \frac{\exp(h_t^T \mu)}{\sum_{t=1}^L \exp(h_t^T \mu)} \quad (2.16)$$

ASP

ASP (圖 2.8，右) 使用與 SAP 相似的技巧計算每個音框的權重。圖中的 Attention model 為兩層的線性結構，公式 2.17 中 W_1, W_2, b_1, b_2 皆為模型的可訓練參數， W_2 和 b_2 的線性運算用於代替 SAP 中的相似度向量計算。公式 2.18 為經過 softmax 正規化的計算的過程， $\tilde{\mu} = \sum_{t=1}^L \alpha_t x_t$ 為加權平均值計算的結果。和 SAP 不同，除了加權平均值 ASP 同時計算加權標準差 (公式 2.19) 做為更多資訊的輸出。

$$e_t = W_2(\tanh(W_1 x_t + b_1)) + b_2 \quad (2.17)$$

$$\alpha_t = \frac{e_t}{\sum_{t=1}^L e_t} \quad (2.18)$$

$$\tilde{\sigma}_t = \sqrt{\sum_{t=1}^L \alpha_t x_t \cdot x_t - \tilde{\mu} \cdot \tilde{\mu}} \quad (2.19)$$

原文 [3] 認為統計上高階的訊息能在各種語音上需要使用池化層的任务達到更好的效果，於是 ASP 通過音框重要性的計算產生加權平均值與加權標準差，讓整體模型能達到更好的池化效果。

本文實作的殘響與乾淨特性分類器由 LCNN 進行改良，將於章節 3.1 詳細介紹；不同池化層的比較結果將記錄於章節 5.6.1。

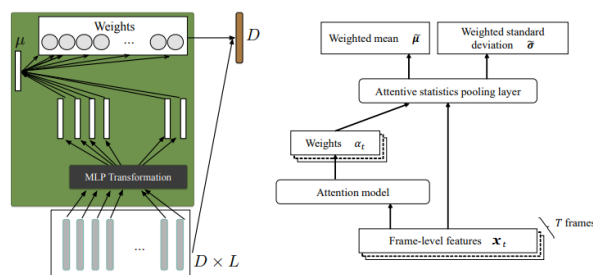


圖 2.8: SAP 及 ASP 實作示意圖。左：SAP [2] 將 D 維度 L 時間長度的特徵進行加權池化；右：ASP [3] 除了加權平均值外，考慮加權標準差作為更高維度的資訊。

2.4 語音品質估計函數

語音的品質有主觀認定的特性，並沒有絕對的判斷依據，根據評估面向不同，音訊強度、時序偏移、背景噪音、頻率及時間維度的清晰度等... 皆為衡量標準之一，以下介紹過去提出且本文使用的代表評估指標。

2.4.1 PESQ

PESQ [23] (perceptual evaluation of speech quality) 為過去常被使用的標準之一，2001 年被提出，主要為原始聲音和經過網路傳輸過後壓縮聲音的比較，在通訊、電信領域為常用的評估指標。PESQ 的評分考慮五種面向：一、時間延遲；二、訊號強度變化；三、頻率維度的對應關係；四、頻率校準度 (frequency warping)；五、壓縮程度。PESQ 輸出的分數介於 -0.5 至 4.5 之間，分數越大表示測試聲音越接近原始聲音。PESQ 的計算包含與原始聲音的比較，於是重視和原始聲音的相似度。並且對於聲音短片段靜音、聲音失真，在人類評斷中不敏感的指標都有嚴重扣分。之後也有研究發現 PESQ 可能有不同語言的差異。雖然擁有部分缺點，但作為早期的評估方式 PESQ 仍有其代表性。

2.4.2 STOI

STOI [24] (short-time objective intelligibility) 為評估吵雜環境下聲音可讀懂性的指標，輸出值的範圍為 0 至 1，越高表示聲音有越好的清晰度。假設 $X(n)$ 與 $Y(n)$ 為原始訊號和待測試訊號的第 n 個音框，首先利用公式 2.20 對測試資料進行均一化及極值截斷 (clip)，其中 α 為能量均一化參數，如公式 2.21。公式中的 β

為可調整參數，物理意義上表示信號失真比 (signal-to-distortion ratio, SDR) 的下限，根據後續研究 $\beta = -15$ 時能獲得較好的效果。



$$Y' = \max(\min(\alpha Y, X + 10^{-\beta/20} X), X - 10^{-\beta/20} X) \quad (2.20)$$

$$\alpha = \sqrt{\frac{\sum_n X(n)^2}{\sum_n Y(n)^2}} \quad (2.21)$$

訊號清晰度 $d(n)$ 被定義為乾淨訊號和欲測試訊號間的相關係數，令 \bar{X} 為所有乾淨訊號音框的平均 ($\bar{X} = \frac{1}{N} \sum_n X(n)$)； \bar{Y}' 則為欲測試訊號的音框平均 ($\bar{Y}' = \frac{1}{N} \sum_n Y'(n)$)，單個音框的訊號清晰度如公式 2.22。

$$d(n) = \frac{\sum_n (X(n) - \bar{X})(Y'(n) - \bar{Y}')}{\sqrt{\sum_n (X(n) - \bar{X})^2 \sum_n (Y'(n) - \bar{Y}')^2}} \quad (2.22)$$

最終 STOI 由所有音框清晰度的平均值決定 (公式 2.23)。

$$d = \frac{1}{N} \sum_n d(n) \quad (2.23)$$

欲使用 STOI 為評量指標時根據原文 [24] 實作，需將訊號統一調整採樣頻率至 10 kHz (文獻 2.20 規定)，使用長度 256 的 Hanning-window，50% 的重疊率進行傅立葉轉換。

2.4.3 SRMR

SRMR (speech to reverberation modulation energy) [25] 用能量比值評斷語音中含有殘響的比例，計算時使用 23 通道的 Gammatone filter [26] (模擬人耳聽覺特性的濾波器) 經過音框長度 256 ms，橫移長度 32 ms Hanning-window 得到頻譜，再經由 8 個調製濾波器 (modulation filterbank) 得到調製頻譜， E_{jk} 可表示為第 j 個通道的第 k 個調製頻譜，每個調製頻譜的能量以通道平均表示，公式 2.24 計算第 k 個調製頻譜的能量。

$$\bar{E}_k = \frac{1}{23} \sum_{j=1}^{23} E_{jk} \quad (2.24)$$

通常調製頻譜中越靠前的頻譜越能代表純粹人類語音的資訊，SRMR 利用語音和非語音類別能量的比值決定訊號的清晰度，計算公式如 2.25

$$SRMR = \frac{\sum_{k=1}^4 \bar{E}_k}{\sum_{k=5}^8 \bar{E}_k} \quad (2.25)$$

上述三評量方式中 PESQ 及 STOI 屬於對評估的方式，需要擁有原始乾淨訊號才能對目標評估訊號進行評估；SRMR 則是非侵入性的 (non-intrusive) 指標，只針對單一給定的訊號評估能量比例，不必給予乾淨音訊。

2.4.4 CER

以上介紹的指標分別從不同面向評估訊號清晰度，但皆涉及聲音學上的知識，一般民眾難以立即將這些指標聯想到實際應用的好壞，讓實驗缺乏推廣性。本研究提出的系統致力於針對 CER 進行改良，將不同訊號藉由語音辨識系統得到字元錯誤率 (character error rate, CER) 對人類會是更直觀的評估方式。

英文以 WER (word error rate) 為語音辨識的判斷指標，但華語句子需要經過斷詞，其可能依照不同辭典而異，於是 CER 以字元為最小的辨識單位是華語語音辨識常用的估計指標。辨識結果和目標句子之間的差異性可分為三種類別：插入 (insertion) 為辨識結果在目標句子中插入額外錯誤辨識字元的情況；刪除 (deletion) 為辨識結果缺少目標句子部分單字的情況；替換 (substitution) 表示目標句子的字元被辨識成另一個字元。以 S 表示替換的字元數、 D 表示刪除的字元數、 I 表示插入的字元數、 N 表示目標句子的總字元數。CER 的計算表示為公式 2.26。

$$CER = \frac{S + D + I}{N} \quad (2.26)$$

CER 借助 ASR 的辨識結果評分音訊，將整個 ASR 模型作為控制變因，輸入乾淨音訊、人工生成含有殘響的音訊、殘響去除後的音訊計算其錯誤率。雖然 CER 並未直接對音訊進行分析，但是更直觀的方式衡量殘響去除的好壞。



Chapter 3

研究方法

圖 3.1 為當今常用 ASR 結合殘響去除之架構圖。測試音訊首先經過特徵抽取後輸入至訓練完成的殘響去除模型，除去殘響的干擾，其結果再通過訓練好的 ASR 得到逐字稿。該架構看似簡潔但存在資料特性不一致的問題，在此架構中 ASR 的訓練主要以乾淨的聲音為主，然而測試時卻是通過殘響去除模型的訊號而非原始乾淨的音檔。雖然期望音訊能透過殘響消除模型更接近乾淨音訊，但聲音終究為模型所生成，其結構上與乾淨訊號的差異難以被忽略。結合前處理模型改善最後結果的研究已在其他領域被實作 [27]。本研究提出圖 3.2 的改良架構，研究分法主要以四個部分改進圖 3.1 的原始方式：

1. 在新架構中的實作殘響與乾淨特性分類器區分擁有殘響的聲音訊號和乾淨的聲音訊號，以便辨識過程中測試訊號能利用較適合的聲學模型辨識。
2. 實作且延伸章節 2.2.1 所提及的殘響消除模型，分析不同想去除方式的實作細節、殘響去除上的表現。
3. 建構可使用於殘響環境的聲學模型。在提出架構圖 3.2 中的上方路徑為分類器辨識測試音檔為含有殘響聲音訊號的路徑，其路徑使用的聲學模型會經過資料擴增的訓練，以減緩測試檔案和訓練資料間的資料不一致性。
4. 進行模型結合。除了單獨使用其一路徑，本研究嘗試以不同方式結合兩路徑以達到更佳的结果。

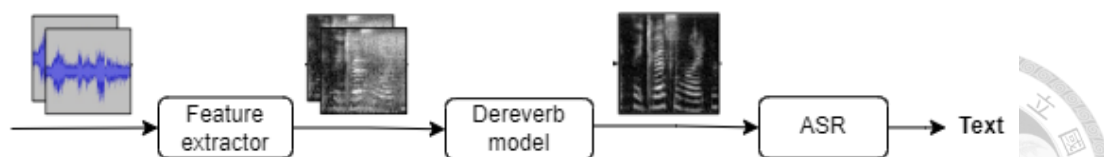


圖 3.1: 原始殘響去除模型結合 ASR 架構。可能產生資料處理不一致的問題

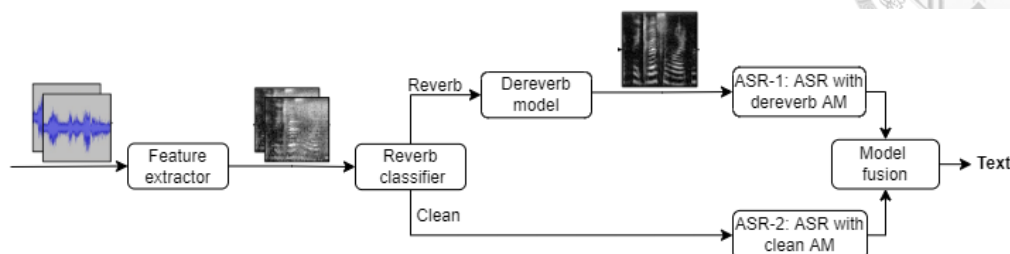


圖 3.2: 提出之殘響去除模型結合 ASR 架構。新的架構藉由分離含有殘響、乾淨音訊讓音訊能對應其適合的聲學模型，以消除資料特性不一致的問題。

章節 3.1 將介紹殘響與乾淨特性分類器使用的模型；章節 3.3 將提出適合殘響環境或去除殘響後使用的聲學模型實作方式；最後章節 3.4 將提出結合圖 3.2 兩路徑應用於 ASR 的演算法。

3.1 殘響與乾淨特性分類器

殘響與乾淨特性分類器訓練以判斷訊號的特性為乾淨類別或殘響類別，是二元分類的問題。訓練前常利用傅立葉轉換將音訊轉換成頻率維度，形成類似圖片的特徵。借助影像辨識優秀的處理手法提高訓練的效果，為音訊特徵處理得常見的手段。和圖片不同，處理聲音訊號的模型被期待能接受任意時間長度的輸入。相較於圖片可以利用縮放讓所有輸入端保持同等大小，音訊常使用池化層 (pooling layer) 壓縮時間維度的資訊，讓模型得以更有彈性的接收不定長度的音訊為輸入，平均池化 (AVG) 是最簡單的池化實作方式之一，視語音中每一個音框同等重要，實際上含有豐富的訊息的音框重要性將高於部分接近靜音的音框。於是將不同音框設定權重的方式也被陸續提出。依照本文架構，殘響與乾淨特性分類器為訊號特性區分的第一道關卡，會被所有測試訊號使用。於是本研究使用較輕型運算快速的 LCNN (light convolutional neural network) [1] 結合 SAP (self-attentive pooling) [2] 或 ASP (attentive statistics pooling) [3] 的池化方式作為判斷音訊受到殘響影響程度之模型。



表 3.1: LCNN with ASP (frame = 400 ms)

Type	Filter Size/Stride, Pad	Output size
STFT	-	1 x 40 x 802
Conv1	$5 \times 5/1, 2$	96 x 40 x 802
MFM	-	48 x 40 x 802
MaxPool	$2 \times 2/2$	48 x 20 x 401
Conv2	$1 \times 1/1$	96 x 20 x 401
MFM	-	48 x 20 x 401
Conv3	$3 \times 3/1, 1$	192 x 20 x 401
MFM	-	96 x 20 x 401
MaxPool	$2 \times 2/2$	96 x 10 x 201
Conv4	$1 \times 1/1$	192 x 10 x 201
MFM	-	96 x 10 x 201
Conv5	$3 \times 3/1, 1$	384 x 10 x 201
MFM	-	192 x 10 x 201
MaxPool	$2 \times 2/2$	192 x 5 x 101
Conv6	$1 \times 1/1$	384 x 5 x 101
MFM	-	192 x 5 x 101
Conv7	$3 \times 3/1, 1$	256 x 5 x 101
MFM	-	128 x 5 x 101
Conv8	$1 \times 1/1$	256 x 5 x 101
MFM	-	128 x 5 x 101
Conv9	$3 \times 3/1, 1$	256 x 5 x 101
MFM	-	128 x 5 x 101
MaxPool	$2 \times 2/2$	128 x 3 x 51
ASP	-	128 x 1 x 102

整體 LCNN 結合 ASP 的模型架構如表 3.1*，為了使訓練資料一致，訊號統一轉換至 16 kHz 的聲音訊號以 400 ms 為單位，先經由短時傅立葉轉換 (short-term fourier transform, STFT) 抽取維度 40 的特徵為梅爾頻譜 (mel-spectrogram) 通過擁有 9 層與 MFM 結合的卷積層後由 ASP 壓縮時間維度。之後可經由銜接全連接層進行進一步的訓練，表格 3.1 的 Output size 有三個參數，第一項為通道數 (channels) 第二項及第三項為頻譜圖的兩個維度。

3.2 殘響消除模型

在章節 2 中已介紹許多殘響去除的研究，其中包括利用傳統預測分析的 PE、WPE 及利用神經網路訓練模型的 DNN、LSTM、GAN。神經網路的方式利用大量的資料讓模型學習殘響去除的模式，且讓單個模型能應用於不同房間的殘響去除，讓此方法成為目前研究的主流。GAN 使用生成方式合成訊號，搭配專用於殘響去除的損失函數達到更好的效果。本段落將介紹在本文使用 MetricGAN 的模型設定。

章節 2.2.2 的研究顯示 GAN 模型得以利用非監督式學習完成模型的訓練。在非監督式的條件下，資料蒐集不需要依賴乾淨的語音訊號和含有殘響語音訊號的配對也可訓練鑑別器。然而在缺少學習目標的情況下，訓練模型的最終結果較有學習目標的監督式學習差。雖然非監督的訓練方式提供殘響消除新的可能性，然而本文以追求最佳表現的作法：利用人工方式從乾淨訊後加入殘響後，創立訓練對進行的監督式訓練。

本文使用的模型的架構如圖 3.3，鑑別器 (上) 和生成器 (下) 訓練時固定另一方的參數，輪流交互訓練。生成器使用兩層 200 為隱藏層的 Bi-LSTM 後銜接維度 300 使用 LeakyReLU 為激勵函數和維度 257 使用 sigmoid 為激勵函數的兩個全連接層。鑑別器使用二維的 kernel 大小為 5x5 的 CNN 模型，後連接平均池化層 (average pooling layer)，最後連接三層全連接層，分別為維度 50、10、1 (輸出)，兩層之間使用 LeakyReLU 為激勵函數。生成器使用 MSE 為損失函數，鑑別器則使用 PESQ 評斷乾淨程度。

*表格排版方式參考 Xiang Wu, A Light CNN for Deep Face Representation with Noisy Labels, 2018, p.4

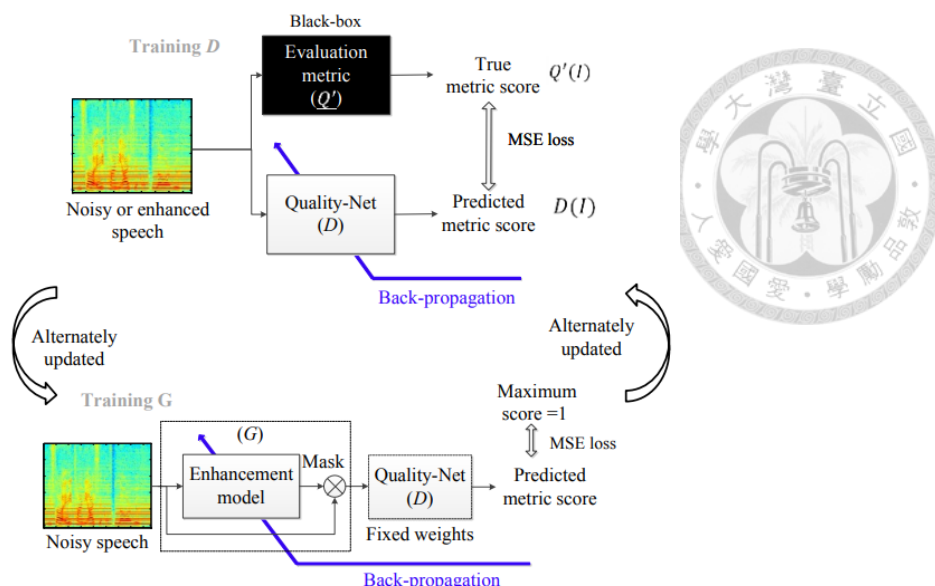


圖 3.3: MetricGAN 訓練架構圖：鑑別器 (上) 和生成器 (下) 交互訓練。

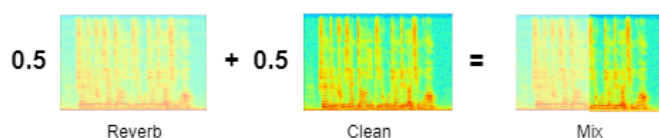


圖 3.4: 時間維度 Mixup：具有殘響語音訊號和乾淨語音訊號各佔合成訊號的一半，橫向結合為新訊號。

3.2.1 用於語音的 Mixup 資料擴增

Mixup 原先應用於圖形辨識，將其延伸至殘響消除的題目時需要考慮目標函數的更改。Mixup 在殘響消除時採用乾淨音訊和含有殘響音訊的組合，無論是何種類別皆是以乾淨訊號進行端對端學習，損失函數更改為公式 3.1。

$$L_{mix} = L_{C_1} = L_{C_2} \quad (3.1)$$

實作上分別在時間維度和頻率維度使用的 mixup 皆需要含有殘響和乾淨的語音訊號對為輸入。時間維度上的資料擴增：將兩訊號分為前半與後半，乾淨與具有殘響的語音訊號彼此混搭，組合成新的訓練訊號 (圖 3.4)。頻率維度的 mixup 則以縱向拼接頻譜 (圖 3.5)，兩種資料擴增的組合讓資料能體現更多面向。

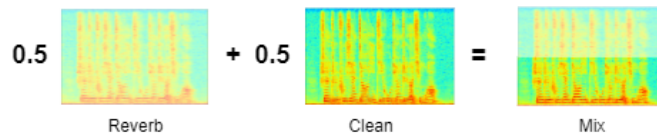


圖 3.5: 頻率維度 Mixup：具有殘響語音訊號和乾淨語音訊號各佔合成訊號的一半，縱向結合為新訊號。

3.3 於聲學模型之資料擴增

目前使用圖 3.1 的 ASR 系統。首先利用殘響去除模型降低音訊的殘響影響程度，使其成為相對更乾淨的訊號，再送至 ASR 模型辨識語音。然而 ASR 聲學模型的訓練不會涵蓋經過殘響去除模型的訊號，導致資料特性不一致的問題。本研究使用符合輸入音檔狀況的資料訓練聲學模型以減少資料特性不一致的影響。

實驗中將訓練及測試的訊號依照表 4.3 分類，章節 3.1 描述的 LCNN 殘響與乾淨特性分類器決定輸入音檔的類別為乾淨音訊或是含有殘響的語音訊號，並使用對應其特性的聲學模型訓練。例如：當分類器辨識測試音檔為乾淨時，聲學模型將選擇使用乾淨訊號訓練的模型；若判斷測試音檔為含有殘響的音檔，則利用殘響去除的模型去除測試音檔中的殘響，同時將訓練資料通過殘響去除的模型之後作為新的訓練資料，建立適合的聲學模型供其使用。章節 4 將進一步介紹資料的生成及使用方式。

3.4 模型結合 (字串組合)

模型結合是強化分類器結果常用的技巧，藉由組合各種方式訓練模型的答案，投票選出綜合機率更大的結果。最普通的方式是將多個模型的分類結果取眾數為最終預測；利用線性式組合，讓不同模型擁有相異的權重為進一步的作法。本實驗提出字串組合的方式，參考前文介紹加權比重的模型結合方式相似：藉由將針對去除殘響的聲學模型和乾淨聲學模型解碼的文字串結果重新組合產生字元錯誤率更低的文字串。按照提出架構圖 3.2 分為兩預測結果，實驗過程為以下步驟：

1. 將輸入訊號進入殘響與乾淨特性分類器 (章節 3.1 所介紹) 得到的分類結果除以 α (溫度法)。

2. 除以 α 之後的結果通過 softmax 函數得到輸入訊號屬於具有殘響語音訊號的機率 r 和乾淨訊號的機率 $1 - r$ ，以此機率作為對訊號品質的評分。
3. 利用乾淨訊號訓練的聲學模型 ASR-2 和固定的語言模型對輸入訊號解碼，得到分數最高的生成文字串為 p ， p 為文字串的整體分數，包含單詞 w_1 至 w_n 的個別單詞分數的總和，且 $P = p_{w_1} + p_{w_2} + \dots + p_{w_n}$ 。
4. 利用殘響去除後的訊號訓練的聲學模型 ASR-1 (章節 3.3 的實驗) 和固定的語言模型對輸入訊號解碼，得到分數最高的生成文字串為 q 。 q 為文字串的整體分數，包含單詞 w_1 至 w_n 的個別單詞分數的總和，且 $Q = q_{w_1} + q_{w_2} + \dots + q_{w_n}$ 。
5. 結合殘響與乾淨特性分類器的機率輸出，使用 Sentence-level fusion (SLF) 或 Word-level fusion (WLF) 混合模型輸出為預測結果。

步驟一在預測類別中添加 α 的除項使預測出兩邊的機率更靠近，類似於溫度調整手法 (temperature)，此方法可以避免分類器輸出過於極端的機率值，原始 softmax 中判斷句有殘響的機率輸出表示為公式 3.2， y_r 代表分類器輸出訊號特性為殘響的值； y_c 則是分類器輸出訊號為乾淨的值。依照分類器的運算過程， y_r 和 y_c 常有較大的差距，讓最終 softmax 的結果趨近於兩極化 (如：判斷語音訊號具有殘響特性的機率 0.999、乾淨訊號機率 0.001)。在此情況，後續聲學模型相關的計算將難以影響最終結果。

$$r = \frac{\exp(y_r)}{\exp(y_r) + \exp(y_c)} \quad (3.2)$$

透過設定 α 值作為分類器結果的除項，公式變化為 3.3，調整過後的值通過 softmax 計算讓輸出機率更接近，結果不會武斷的被決定，讓隨後提出的模型結合方式更有意義，不會在分類器階段直接決定結果。

$$r = \frac{\exp(y_r/\alpha)}{\exp(y_r/\alpha) + \exp(y_c/\alpha)} \quad (3.3)$$

步驟三及步驟四使用分數評估單詞或句子是否為合適的輸出。分數為聲學模型和語言模型各自對訊號進行的綜合評分，聲學模型以輸入訊號估計產生的音素串

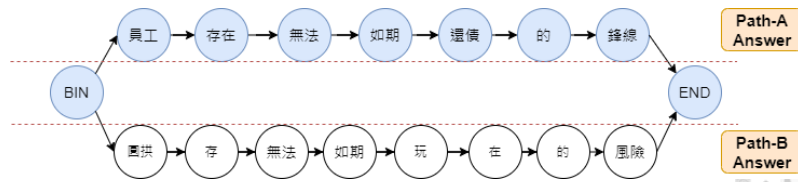


圖 3.6: Sentence level fusion：以整體文字串為單位選擇輸出結果

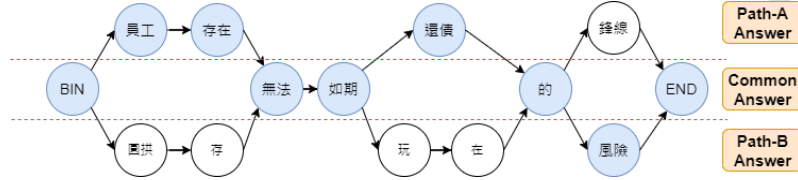


圖 3.7: Word level fusion：以單詞為單位選擇輸出結果

機率為輸出分數、語言模型以特定音素串產生文字串的統計機率為評分標準，以輸出單詞或整段訊號為單位輸出分數。根據架構圖 3.2，步驟三及步驟四將以聲學模型 ASR-2 進行乾淨訊號的評分，以及聲學模型 ASR-1 對具有殘響語音訊號的評分，並分別以 P ， Q 代表整體文字串的分數，文字串中的單詞 (w_1, w_2, \dots, w_n) 的分數以 $p_{w_1}, p_{w_2}, \dots, p_{w_n}$ 表示。

步驟五綜合考慮殘響與乾淨特性分類器對訊號品質的評分結合步驟三及步驟四語言模型評分，本研究提出三種模型組合方式，分別為 Direct Ensemble、Sentence level fusion (SLF)、Word level fusion (WLF)。

3.4.1 Direct ensemble

Direct ensemble 為只借助殘響與乾淨特性分類器進行模型選擇的方式，選擇結果直接依賴分類器的輸出，故依此命名。若殘響與乾淨特性分類器的結果判斷輸入訊號較接近乾淨訊號則直接選擇適合乾淨訊號的聲學模型的 ASR 作為整體系統的辨識結果，反之則選擇另一方適用殘響去除訊號的聲學模型為最終結果。

3.4.2 Sentence-level fusion

以降低字元錯誤率為目標，進一步考慮語音辨識模型的輸出文字的機率的結合方式，能更貼合需求。SLF 維持預測文字串的一致性。如圖 3.6，將由分類器產生的對乾淨訊號的語音品質評分 r 和由 ASR-2 的生成評分 P 及 ASR-1 的生成評分 Q 相乘後選擇分數更高的文字串為最終輸出。算法 3.4 描述路徑選擇的方式。

表 3.2: 路徑合併問題

文字串一	从而 和 启用 和 信心 的 案件
文字串二	从而 和 起 用户 信息 的 案件

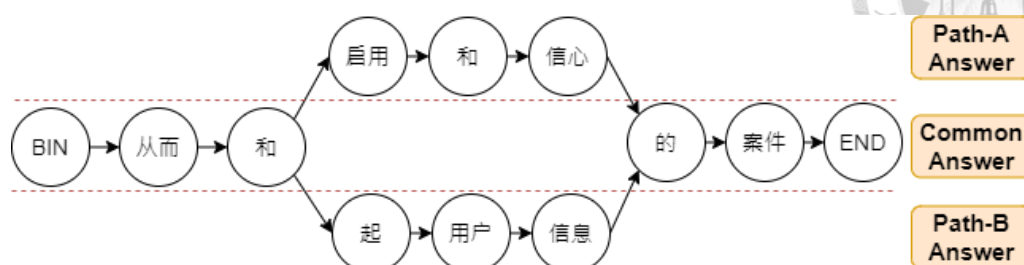


圖 3.8: 單詞對齊範例一：對應文字串一的第一個「和」。

$$\max(rP, (1-r)Q) \quad (3.4)$$

3.4.3 Word-level fusion

WLF 則將兩聲學模型辨識的文字串拆解成單詞。如圖 3.7，實作上將兩路徑的輸出文字串作圖且分為三區域。最上最下兩區域分別屬於兩文字串特有的單詞，中間區域為共有單詞。分類器產生的語音品質評分和 LM 的評分相乘為各單詞的分數，兩路徑辨識出相同的單詞時路徑合併為一，成為唯一選擇；如辨識結果為不同單詞時則選擇分數更高的一側。如何決定單詞合併為此方法的首要任務。

K-best LCS

Longest common subsequence (LCS) 可使用於此時尋找一個最長共同子序列當作合併的單詞。對於兩字串尋找任一個 LCS 的問題，以暴力法列舉所有子序列後比較，需要複雜度指數的計算時間。但通過動態規劃 (dynamic programming, DP)

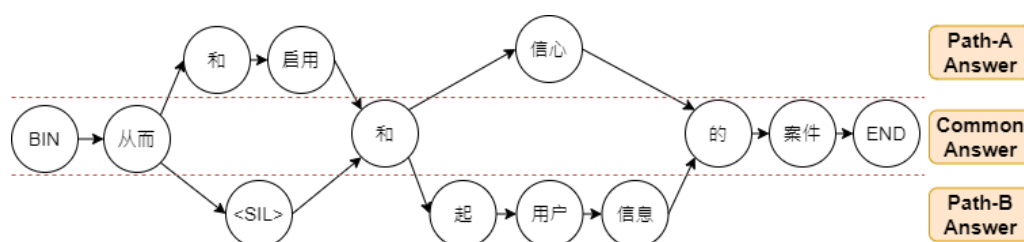


圖 3.9: 單詞對齊範例二：對應文字串一的第二個「和」。

加速，使複雜度壓縮在 $O(mn)$ (m 、 n 分別為兩文字串的單詞數目)。然而遇到如表格 3.2 的範例時，演算法僅能找出其中一只 LCS，圖 3.8 及圖 3.9 皆成為可能的結果。如要一次找出所有可能的 LCS，則複雜度為 $O(\frac{3}{n\pi}2.598^n)$ [28]，使得此作法無法處理較長文字串，如果產生的唯一個 LCS 恰好為組合上較差的子序列，則可能造成 WLF 的效果降低。本文提出較恰當的方式：尋找其中 K 個 LCS 後決定使用 K 個 LCS 中較佳的一個，以避免最差的 LCS 被選擇到。

本研究提出的 K-best LCS 方法參考傳統上尋找 K-best 最短路徑問題的解法變化。文獻 [29] 中 K-best 演算法以兩個方向簡化最短路徑問題：

1. 問題中不包含環的結構。
2. 圖形中邊長的費用只有 0 和 1。

除了無環的前題，同時排除圖形中負環的可能性，大幅簡化問題；費用較大的路徑可以拆解成數個 1 的連續路徑，假設圖形中只有 0、1 的值除了方便隨後的公式書寫，同時限制問題只在離散維度上進行討論，也排除路徑擁有浮點數的可能。令 s 為最短路徑邊數。針對 K-best 最短路徑問題的解法如下：

1. 利用 DP 之方法找出並保存最佳解，並設定 $k = 1$ 。
2. 輸出並移除費用最低的保存路徑，令其值為 $x^k = (x_1^k, x_2^k, \dots, x_n^k)$
3. 如果 $k = K$ ，表示已經得到 K 組最佳解，終止演算法；否則繼續。
4. 保留 x^k 中所有值為 0 的邊，依序翻轉剩下邊的 0、1 值建立一個新的子問題 (最多建立 $n - s$ 個新問題)。
5. 保存所有問題的解，設定 $k = k + 1$ ，返回步驟一。

此方法利用反轉每一條路徑值的方式，建立新的問題。以懲罰目前最佳解或是獎勵位在最佳解中的邊依序尋找可能的答案。若尋找最佳解的演算法需要 $O(c(n))$ 的時間複雜度，則如此作法需要複雜度 $O(Ksc(n))$ ，找出 K-best 解。本研究使用的 K-best LCS 借鏡演算法此方法，但為了符合範例進行了更改，做法如下：

1. 利用 DP 之方法找出並保存最佳 LCS，並設定 $k = 1$ 。

	B	D	A	D	E	B
A	0	0	1	1	1	1
B	1	1	1	1	1	2
D	1	2	2	2	2	2
B	1	2	2	2	2	3
E	1	2	2	2	3	3

	B	D	A	D	E	B
A	0	0	2	2	2	2
B	1	1	2	2	2	3
D	1	2	2	3	3	3
B	1	2	2	3	3	4
E	1	2	2	3	4	4

	B	D	A	D	E	B
A	0	0	1	1	1	1
B	1	1	1	1	1	2
D	1	2	2	3	3	3
B	1	2	2	3	3	4
E	1	2	2	3	4	4

圖 3.10: K-best LCS 範例：輸入字串分別為 BDADEB 和 ABDBE。左圖表示利用原始 DP 方法找到的答案；中圖及右圖為對深色字兩例分數後找尋到的 LCS。

2. 輸出最佳的 LCS，令其值為 $x^k = (x_1^k, x_2^k, \dots, x_n^k)$
3. 如果 $k = K$ ，表示以得到 K 組最佳解，終止演算法；否則繼續。
4. 對於一個不在最佳 LCS 當中對應到的字，在建立 DP 表格時增加其分數 1。
5. 如得到的 LCS 不再保存列中，保存它，並找尋下一個不在最佳 LCS 中的對應字，設定 $k = k + 1$ 。回到步驟三。如果已遍歷所有可能性，則回傳所有保存的值後，停止演算法。

除了第一次搜尋到的最佳 LCS，其餘尋找 LCS 的方式皆為獎勵其於對應到的字的情況，通過獎勵分數，可以驅使回溯時尋找不同路徑，並且只限制長度相同的 LCS 會被找到。圖 3.10 為一範例，輸入字串分別為 BDADEB 和 ABDBE。左圖表示利用原始 DP 方法找到的答案：BDE；在中圖深色的「A」為對應到的字段不再目前 LCS 中，於是給予分數上獎勵，隨後即尋找到另一條 LCS 路徑：ADE。右圖為給予深色「D」獎勵的情況，此時找尋到新路徑：BDE，雖然和原始 LCS 答案相同，但對應到的位置不同，在此研究中會產生相異的路徑合併方法。

WLF Algorithm

完成路徑合併後此問題可使用 DP 解決，表 3.3 為 DP 將使用到的參數的定義。遞迴式定義如公式 3.5 定義：

$$s(n_i) = LM(n_i) + \max(s(Pre(n_i))) \quad (3.5)$$

當目前點為單一輸入時，則記錄累加分數和自身分數的加總；如為多輸入端時

表 3.3: WLF 參數對照圖

n_i	每個單詞代表的點
$pre_{n_i,j}$	點 n_i 的第 j 個輸入
$LM(n_i)$	在 n_i 的語言模型分數 (已經過分類器分數加權)
$s(n_i)$	在點 n_i 的累積分數，其中 $s(BIN) = 0$
$Pre(n_i)$	指向 n_i 輸入點的集合

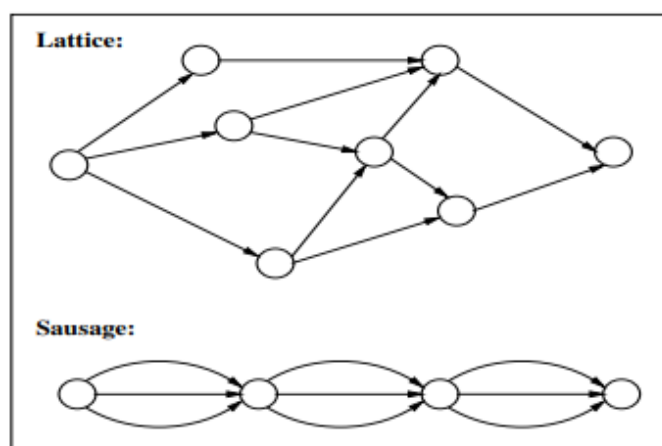


圖 3.11: lattice 為通常的構圖方式，sausage 法可建立更簡潔的構圖，圖片來源為 [30]。

則選擇最高一只加總自身分數後記錄至 DP 表。最終回傳分數 $s(END)$ ，並記錄填表的路徑以利回溯輸出結果。

WLF 和 sausage method 的比較

本文提出的 WLF 是建立在單詞單位的組合方式和傳統 ASR 使用的構圖方法 sausage 有共同點，於是在此分析兩者的差異性。

Sausage 以協助構圖的方式應用於 ASR [30]，相較於 lattice 的構圖方法，使用 sausage 會經過剪枝淘汰機率過小的組合，於是同時具備以時間點強制對其不同單詞和減少圖形大小的優點，圖 3.11 為普遍生成的兩構圖比較，分枝間以文字間の後設機率 (posterior probabilities) 表示連結。雖然可能會因為誤剪枝導致效果降低，但在部分任務中可以維持辨識效果並同時壓縮圖形。

Sausage 方法主要是在訓練的同時使用剪枝和對齊協助訓練，WLF 為訓練結束後利用選擇單詞更好的結果。其餘在使用時機上、使用功能等都有所不同。表 3.4 詳細列出兩者比較。



表 3.4: Sausage 與 WLF

方法 特性	Sausage	WLF
處理單位	單詞	單詞
使用時間	訓練時	訓練後
功能	剪枝、對齊	模型結合
分支的計算因子	後設機率	ASR 和訊號擁有殘響特性的評分



Chapter 4

語料介紹

本章節將介紹本文使用的語音資料集以及殘響訊號資料集，乾淨和殘響的語音訊號對將由人工方式生成。

4.1 語音辨識資料集

Aishell1 [31] 為中國希爾貝殼公司發布的語音辨識資料集，使用高保真麥克風在安靜的室內環境中錄製，包含中國不同口音的語者 400 名，語料總時長約 178 小時，語料子集分布如表 4.1 所示。為保持音檔的一致性，錄製後的音檔重新採樣至 16 kHz、16-bit 的 wave 格式。

4.2 含有殘響的語音資料集

殘響為聲音在環境中碰撞到置延遲被接收，產生當音源消失聲音能持續存在現象，在室內環境中尤其明顯。殘響讓接收端同時收到原音訊和延遲的具有殘響的

表 4.1: Aishell1 統計

資訊 子集	語者數	長度 (小時)	句數
Training	340	150	120,403
Development	40	18	14,329
Test	20	10	7,176

表 4.2: RIR_NOISES 模擬資訊

房間種類 \ 資訊	房間長寬	模擬房間數	每個房間模擬對話數
small	1m - 10m	200	100
medium	10m - 30m	200	100
large	30m - 50m	200	100

語音訊號，嚴重的殘響將使 ASR 的辨識效果降低，於是研究 ASR 系統如何處理具有含有殘響的音訊有其重要性。在監督式學習中，需要同時收集乾淨及具有殘響的聲音對以進行端對端訓練。然而，在現實環境中同時蒐集兩者的難度及成本高，於是通常以在乾淨的音訊資料上人工摻入殘響的方式實作資料集。

4.2.1 RIR_NOISES

RIR_NOISES [32] 資料集的內容包含模擬環境下的殘響及噪音，本文使其中有關殘響的子集 (simulated_rirs)，資料的數量以及特性如表4.2 所示，分為小房間、中房間、大房間，三種類。模擬房間的確切大小由表格 4.2 中標註房間長寬的範圍均勻抽取，房間高度則皆由 2m-5m 均勻抽取進行模擬，模擬時語者和接收端距離皆限制在不超過 5m 的距離。

4.3 資料生成

人工生成具有殘響的語音訊號資料為產生訓練對有效率的選擇。選擇 Aishell1 的乾淨音檔 $x(t)$ 及隨機抽取 RIR_NOISES 的模擬脈衝響應 $h(t)$ 進行卷積運算 (convolution) 以生成有殘響的語音檔 $y(t)$ 。如公式 4.1，生成含有殘響的語音 $y(t)$ 和原始乾淨語音 $x(t)$ 將成為訓練對，供監督式訓練進行端到端訓練。

$$x(t) \otimes h(t) = y(t) \quad (4.1)$$

公式中 \otimes 代表 convolution 運算，模擬聲音在房間中傳播至接收端的過程。以此方式由乾淨語音訊號一對一生成含有殘響的語音訊號。為了簡化表達，以聲音的種類分成表格 4.3 的表達方式，之後章節將以代號 A、B、C 代表訓練資料的使

表 4.3: 聲學模型使用資料分析

子集	標籤	特性	音檔數量
Training set	A	乾淨訊號	120,098
	B	含有殘響語音訊號	120,098
	C	殘響去除後的語音訊號	120,098
Test set	a	乾淨訊號	7,176
	b	含有殘響語音訊號	7,176
	c	殘響去除後的語音訊號	7,176

用方式，a、b、c 則是選擇使用的測試資料。

實驗中將依照情況不同使用不同的語料範圍，實驗與其對應的使用語料將於章節 5 提及實驗內容時說明。



Chapter 5

實驗設計與結果

章節 3 中介紹本文提出的結合殘響去除之 ASR 架構，本章節將針對架構提出的實驗細節設定、結果及針對觀察到現象的探討。

5.1 實驗流程及訓練參數設定

本研究包含圖 3.2 中四個組件的建構與改良：殘響與乾淨特性分類器、殘響去除模型、聲學模型、模型結合之演算法。四個模型或算法獨立訓練及實作。實驗設計內容如下：

- 實驗一：利用乾淨或具有殘響的 Aishell1 訓練語料進行基本 ASR 訓練測試字元錯誤率，以作為之後實驗的標準。
- 實驗二：分析及比較不同訓練方法的殘響去除模型在不同環境的效果，此實驗包含以下細項：
 - 殘響去除方式的比較。
 - 不同大小房間 (不同 IRs) 對殘響去除模型及整體字元錯誤率的影響。
 - 資料擴增方法應用於殘響去除模型的效果。
 - 將 Aishell1 訓練資料通過殘響去除模型後減少資料不一致性的分析。
- 實驗三：殘響與乾淨特性分類器的訓練方式及效果的探討。



- 池化層的比較。
- 分類器結果分析。
- 實驗四：探討字串結合演算法對字元錯誤率的影響。
 - 溫度法的效果。
 - 結合演算法間的比較。

實驗以提升語音辨識辨識率(降低字元錯誤率)為目標，利用結合殘響除去模型改進聲學模型和測試訊號的乾淨度。語音辨識中的語言模型在實驗中為控制變因：皆以 Aishell1 [31] 為訓練語料的文本，建立 3-元 (tri-gram) 語言模型。聲學模型以 GMM-HMM 的架構訓練三音素模型，聲學模型與語言模型的參數配置在不同實驗階段皆相同。本章節將介紹研究中使用技巧和訓練方式的參數設定。

5.1.1 聲學模型

特徵萃取階段

聲學模型通常利用特徵萃取將時間維度的訊號轉換成頻率維度為訊號的前處理。取音框長度為 25 毫秒，音框位移長度 (hop length) 為 10 毫秒為特徵抽取基本單位，運算時音框彼此以部分重疊增加轉換的穩定性；選取的音框經過快速傅立葉轉換 (fast Fourier transform, FFT) 將訊號轉換成頻率維度，生成功率譜 (power spectrogram)；應用梅爾濾波器 (20 個非線性分布的三角帶通濾波器，模擬人耳對頻率的感受) 取對數的功率譜後進行離散餘弦轉換 (discrete cosine transform, DCT) 得到特徵值的維度，特徵值的維度和 DCT 的計算設定有關，但通常保留前 12 個維度加上音框能量疊加的值為 13 維的梅爾倒頻譜 (mel-frequency cepstral coefficient, MFCC)。另外加上以頻率計算可靠度歸一化相關係數 (normalized cross correlation features, NCCF) 得到的包含濁音機率 (POV)、對數基頻 (pitch feature) 和基頻一階差分 (delta pitch feature) 的 3 維基頻特徵，共 16 維度的特徵將在特徵萃取階段以音框為單位產生。16 維度的特徵通過將倒頻譜平均值方差歸一化 (cepstral mean and variance normalization, CMVN) 以平衡特徵中各維度的比重。

而在使用 TDNN 訓練聲學模型時，則是使用 40 個梅爾濾波器組，再經過 DCT 取得 40 維高解析度之 MFCC (high resolution MFCC)。

GMM-HMM

本文以 GMM-HMM 實作聲學模型，其訓練過程借助著名語音辨識開源專案：Kaldi [33]。使用的模型、參數及對應的 Kaldi 腳本紀錄於表 5.1。訓練以單音素模型建構初始 gmm 模型，並利用強制對齊 (forced alignment) 取得每個音框當下代表的狀態，供後續聲學模型訓練使用。僅使用單音素模型結果不佳，三音素模型考慮相鄰音素訓練以降低字元錯誤率。將聲學模型提升至三音素將面臨參數量激增的問題，針對單音素需要建立 40 個模型，三音素考慮到前後的關係則需要建立 64,000 個模型 (40^3)，除了需要更大儲存空間、訓練時間外，增加過擬合風險。常見的解法為透過如決策樹 (decision tree) 的演算法，讓特徵相似的三音素共用一模型，決策樹的子葉數目如表 5.1 所示。tri1 為使用單音素模型對齊後建構的初始的三音素模型，以此模型為基準，疊代訓練更精細的模型；tri2 額外計算音框間的二階導數 (delta-delta) 特徵，作為補充特徵；tri3 使用線性判別分析 (linear discriminant analysis, LDA) 及最大似然線性變換 (maximum likelihood linear transform, MLLT) 降低 HMM 計算維度，並為每個說話者推導出一個獨特的轉換；tri4 及 tri5 加入說話者自適應訓練 (speaker adaptive training, SAT)，以語者為單位產生標準化的數據，估計音素引起的方差，減少說話者或錄音環境的影響。由於時間因素，本文部分實驗會使用 tri5 的模型進行比較，涉及到整體比較或較重要的實驗會使用下述的 TDNN 模型訓練。

TDNN

借助神經網路的方法訓練聲學模型是當今的趨勢之一，本文訓練使用的架構為因子分解時延神經網路 (time delay neural networks, TDNN)。本文的聲學模型使用 15 層 1536 維度的 TDNN 網路，在每層的連接中加入正規化再進入下一層。並使用 tri5 為對齊的資訊。初始學習率設定為 0.0005 每個 iteration 降低 0.0000025，訓練 138 個 iterations。

5.1.2 語言模型

本研究使用 Aishell1 [31] 為文本，以及其提供的辭典 (lexicon) 訓練的 3-元 (tri-gram) 語言模型。

表 5.1: GMM-HMM 訓練參數

模型	Kaldi 腳本	子葉個數	高斯模型數目
mono	train_mono.sh	-	-
align	steps/align_si.sh	-	-
tri1	train_deltas.sh	2,500	20,000
align	steps/align_si.sh	-	-
tri2	train_deltas.sh	2,500	20,000
align	steps/align_si.sh	-	-
tri3	train_lda_mllt.sh	2,500	20,000
align	steps/align_fmllr.sh	-	-
tri4	train_sat.sh	2,500	20,000
align	steps/align_fmllr.sh	-	-
tri5	train_sat.sh	3,500	100,000

5.1.3 傅立葉轉換

傅立葉轉換從時間維度上提取頻率維度的特徵，以下介紹的所有模型皆會使用傅立葉轉換，若未特別提及，使用傅立葉轉換時參照以下設定。首先將聲音訊號重新採樣至 16 kHz、音框大小 512、音框移動長度 128、音框使用 hamming window。

5.1.4 殘響與乾淨特性分類器模型

在本研究提出的架構中 3.2，殘響與乾淨特性分類器用於判斷訊號的特性較接近乾淨訊號或含有殘響的訊號，模型架構參照表格 3.1。以傅立葉轉換提取頻率維度的特徵 5.1.3，訓練時使用 Adam [34] 為優化器 (optimizer)；其 $\beta_1 = 0.5, \beta_2 = 0.9$ ，並設定 weight decay 為 0.01，學習率 (learning rate) 為 0.001，損失函數為 cross entropy，每個 epoch 減少為原來的 0.999 倍。使用 Aishell1 的乾淨 (A) 及具有殘響的語音資料 (B) 訓練及驗證，共訓練 80 個 epoches。

5.1.5 Bi-LSTM 模型

殘響去除的 Bi-LSTM 模型於章節 2.2.2 中提及，並作為其他神經網路方式的對照組。實驗中包含殘響語音訊號經過傅立葉轉換 5.1.3 後使用 2 層維度 500 的 Bi-LSTM 端對端訓練。訓練時使用 Adam [34] 為優化器；其 $\beta_1 = 0.9, \beta_2 = 0.999$ ，學習率設定為 0.001，損失函數為 minimum square error (MSE)，使用 Aishell1 的乾淨 (A) 及具有殘響的語音資料 (B) 訓練及驗證，共訓練 100 個 epoches。

5.1.6 MetricGAN 模型

MetricGAN 的架構已於章節 2.2.2 介紹，在生成器有殘響語音訊號經過傅立葉轉換 5.1.3，並以乾淨訊號為訓練目標。使用 Adam [34] 為優化器 (optimizer)；其 $\beta_1 = 0.9, \beta_2 = 0.999$ ，損失函數使用 PESQ，學習率 (learning rate) 設定為 0.00013，訓練 150 個 epoches。

5.2 效果評估方式

本實驗使用 PESQ [23]、STOI [24]、SRMR [25] 及 CER 評估訊號的乾淨程度。由於聲音的乾淨與否為主觀認定，評估指標旨於建立公正的評量方式得到聲音乾淨部的量化結果。本文使用評估方式相關的設定參照章節 2.4 的描述。

5.3 訓練設備規格

實驗包含純粹使用 CPU 訓練語言、聲學模型及涉及利用 GPU 加速機器學習的方式實作殘響去除系統。使用國家高速電腦中心台灣雲 (Taiwan computing cloud, TWCC) 的機器進行實驗 (型號：cm.super)，包含 GPU 型號：Nvidia Tesla V100 一個，及 CPU 型號：Intel(R) Xeon(R) Gold 6154 (四核心)。語音辨識使用 Kaldi 的架構訓練，殘響去除模型則主要利用 pytorch 撰寫、訓練。

使用 CPU 進行一次完整的語音辨識訓練花費約 20 小時，使用 GPU 加速訓練 Bi-LSTM 花費約 76 小時 (章節 5.1.5 描述之架構)，使用 GPU 加速訓練 MetricGAN 花費約 64 小時 (章節 5.1.6 描述之架構)。

表 5.2: ASR 之基準線測試

訓練資料	測試資料	CER (tri5)	CER (tdnn)
A	a	12.12%	6.55%
	b	24.27%	15.16%
B	a	13.47%	6.58%
	b	21.25%	10.32%
A+B	a	12.39%	6.51%
	b	21.25%	10.19%

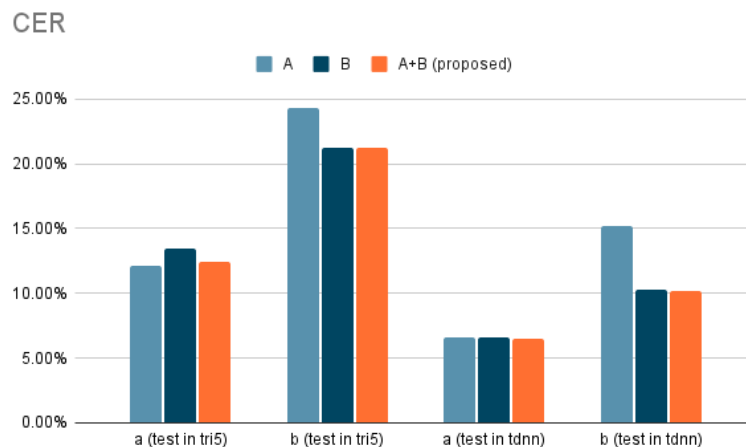


圖 5.1: ASR 基準線實驗結果：使用具有殘響的資料進行資料擴增助於整體字元錯誤率的下降

5.4 實驗一：基礎 ASR 訓練之結果

本實驗目的為分析 ASR 模型面對乾淨及含有殘響的語音訊號時的字元錯誤率，用於建立 ASR 的基準線以利於之後的分析，同時觀察資料擴增對 ASR 的影響。實驗中分別使用乾淨 (A)、含有殘響 (B) 及混合資料集訓練聲學模型，與對應的測試資料生成結果如圖 5.1。表 5.2 為其詳細資訊。訓練資料和測試資料所使用的代號依照表 4.3 定義。

在使用 tri5 的聲學模型時，僅使用乾淨訊號 (A) 訓練的聲學模型面對相同特性的測試資料 (a) 有最好的表現 (CER: 12.12%)，然而其對具有殘響的測試訊號 (b) 字元錯誤率為最高 (CER 高達 24.27%)；純粹使用殘響的語音訊號 (B) 訓練則擁有相反的現象，雖然能使在殘響環境下的辨識效果提升，但是會犧牲較多在乾淨環境的辨識率。綜合兩者訓練 (A+B) 整體而言有最好的結果。推斷增加訓練及測試

資料的一致性對於減少字元錯誤率有一定程度的幫助。

tdnn 的聲學模型使用神經網路能讓模型學習訓練資料中的隱藏特徵，其表現上普遍勝過對應的 tri5 聲學模型。且神經網路架構更全面的學習資料的特徵，讓使用 A+B 的訓練資料在不論 a 或 b 的測試項皆擁有最好的表現。加入具有殘響的語音訓練資料後，除了達到在具有殘響的語音訊號上 CER 的下降外，乾淨語音訊號的測試上也沒有表現的退步。和 tri5 模型最好的結果相比，表現最佳的 tdnn 模型在乾淨的語音測試資料上 CER 減少了 5.61%，在具有殘響的語音資料上 CER 減少了 11.06 %。

5.5 實驗二：殘響去除模型的分析及比較

殘響去除模型是本文提出的語音辨識模型架構的重點，優秀的殘響去除模型能夠大幅減少殘響環境下的字元錯誤率，從傳統演算法至機器學習的方法皆有針對此議題的研究。本實驗會對殘響去除模型進行一系列的討論，包括以下面向：

1. 不同殘響去除模型或演算法之間的比較。
2. 比較不同房間對於殘響去除模型及整體字元錯誤率之影響。
3. 應用 Mixup 資料擴增方法於殘響去除模型的效果。
4. 應用資料一致性於語音辨識之結果。

5.5.1 不同殘響去除模型 (演算法) 之間的比較

本實驗比較預測誤差法 (prediction error, PE)、加權預測誤差法 (weighted prediction error, WPE)、Bi-LSTM 及 MetricGAN 對於不同評估方式 (PESQ、STOI、SRMR、CER) 的結果，MetricGAN 分別測試使用 PESQ 或 STOI 為損失函數的情況。圖 5.2 為比較的結果，表格 5.3 為其詳細的數值。圖片及表格中的 CER 為只使用 Aishell1 乾淨聲音 (A) 訓練的聲學模型，整體語音辨識模型為控制變因，以輸入訊號的不同計算字元錯誤率。

相較於不採取殘響去除 (None) 的原始殘響環境中的訊號，經過 PE 處理的訊號在 STOI 及 SRMR 略有進步，但 PESQ 在測量效果大幅下降，CER 也有所降低。

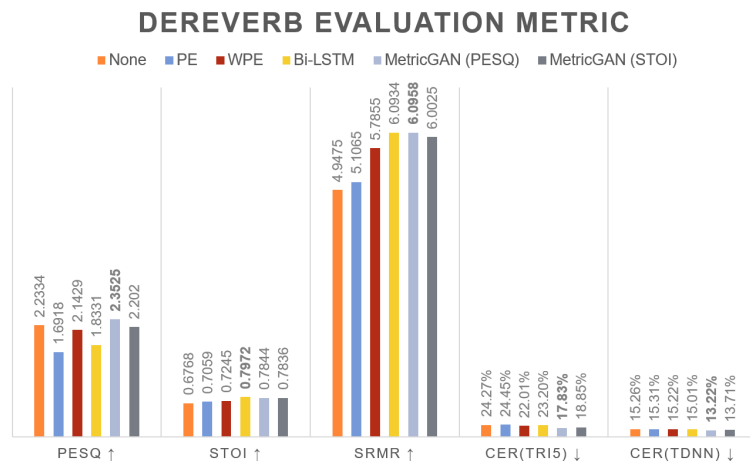


圖 5.2: 殘響去除對應評估方式比較圖：PE、WPE、Bi-LSTM、MetricGAN 對應 PESQ、STOI、SRMR、CER 的交叉比較。

表 5.3: 殘響去除對應評估方式詳細結果

殘響去除	評估指標				
	PESQ ↑	STOI ↑	SRMR ↑	CER (tri5) ↓	CER (tdnn) ↓
None	2.2334	0.6768	4.9475	24.27%	15.26%
PE	1.6918	0.7059	5.1065	24.45%	15.31%
WPE	2.1429	0.7245	5.7855	22.01%	15.22%
Bi-LSTM	1.8331	0.7972	6.0934	23.20%	15.01%
MetricGAN (PESQ)	2.3525	0.7844	6.0958	17.83%	13.22%
MetricGAN (STOI)	2.2020	0.7836	6.0025	18.85%	13.71%

可證明在殘響去除的題目上，如章節 2.2.1 提及，使用 PE 進行最佳化時需要訊號為平穩過程才能使問題等同於尋找合適的轉換矩陣，讓轉換矩陣能被訓練為好的殘響去除轉換。人類的語音幾乎皆不為平穩過程並不能等同於獲得合適的轉換矩陣。WPE 改良了 PE 的缺點，額外計算時變方差，並利用其作為均一化讓聲音訊號讓輸入的聲音訊號接近平穩過程，最佳化計算時能得到更好的結果。和 PE 相比 WPE 在各個評估指標皆有不小的進步；除了 PESQ 外，使用 WPE 後訊號的評估指標勝過原始訊號。隨後，機器學習的方法由於硬體的進步備受推崇，Bi-LSTM 為機器學習的熱門做法之一。此外，也有許多改良方式不斷被提出，本實驗測試基礎方式訓練 5.1.5 的 Bi-LSTM。Bi-LSTM 在 STOI 及 SRMR 的評估中有所進步，但在 PESQ 遜於 WPE；可解釋為如果神經網路作法未必能完全勝過傳統演算法，如欲使效果進一步提升，需要對網路架構或訓練方式有更多符合實

驗內容的調整。MetricGAN 在神經網路做法專門針對殘響去除調整模型的訓練方式，包括使用善於生成圖片、訊號的架構 (GAN 架構) 以及符合殘響去除的評估函數 (PESQ、STOI) 訓練。物理意義上 PESQ 著重聲音品質；STOI 注重聲音的清晰度，透過兩損失函數皆使訓練更為有效。以表格 5.3 的結果觀測，使用 PESQ 的效過較佳，之後的實驗將主要使用此模型。以 PESQ 為損失函數的 MetricGAN 擁有本文測試的殘響去除算法中最佳的結果，並且使使用 tri5 模型辨識的 CER 降低至 17.83%；使用 tdnn 模型辨識的 CER 降低至 13.22%。

以評估指標的角度分析結果，本研究關注殘響去除模型與 ASR 系統的結合後幫助字元錯誤率下降的目標。於是所有評量方式中，CER 被設定為最重要的評估指標。其餘指標中 PESQ 考慮訊號是否失真，比起模型創造的訊號，原始殘響環境的語音訊號容易獲得較高的評價，但對人類而言些許失真可能不影響判斷訊號的乾淨程度，且 PESQ 在分數計算上對靜默有較大的懲罰，在人類主觀認知上可能感覺不到偶有的 50 毫秒的靜音，但 PESQ 對其產生大幅扣分，於是比起模型合成的訊號，原始的殘響環境的語音訊號在 PESQ 計算上更有優勢。STOI 和 SRMR 的變化幅度則較為相似，他們分別定義頻譜的清晰度及能量為比較依據。但不論使用何種觀測方式皆無法直接和 CER 擁有直接的關聯，皆是以間接推估乾淨程度的方式評估是否能對 ASR 的辨識產生影響。

5.5.2 不同房間對於殘響去除模型及整體字元錯誤率之影響

本實驗分析在聲音在不同房間傳遞時受到殘響的影響。殘響在所有的室內環境中都會產生，但干擾程度存在差異。本實驗將室內空間依照表 4.2 的 Aishell1 分類方式將房間分為小房間 (寬度 1m 10m, small)、中房間 (寬度 10m 30m, medium)、大房間 (寬度 30m 50m, large)。此小節的比較在於不同房間的差異性，於是生成新的資料集。生成時 Aishell1 的訊號將與隨機 RIR_NOISES 的小房間訊號生成小房間的資料集；與隨機 RIR_NOISES 的中房間訊號生成中房間的資料集；與隨機 RIR_NOISES 的大房間訊號生成大房間的資料集。並以章節 5.5.1 提及的方式獨立訓練及測試不同環境下的表現。結果於長條圖 5.3 表示；為節省實驗時間，CER 的評估項在本實驗只以 tri5 的模型測試，但仍可觀測出其中的趨勢。表格 5.4 顯示結果的詳細數據。

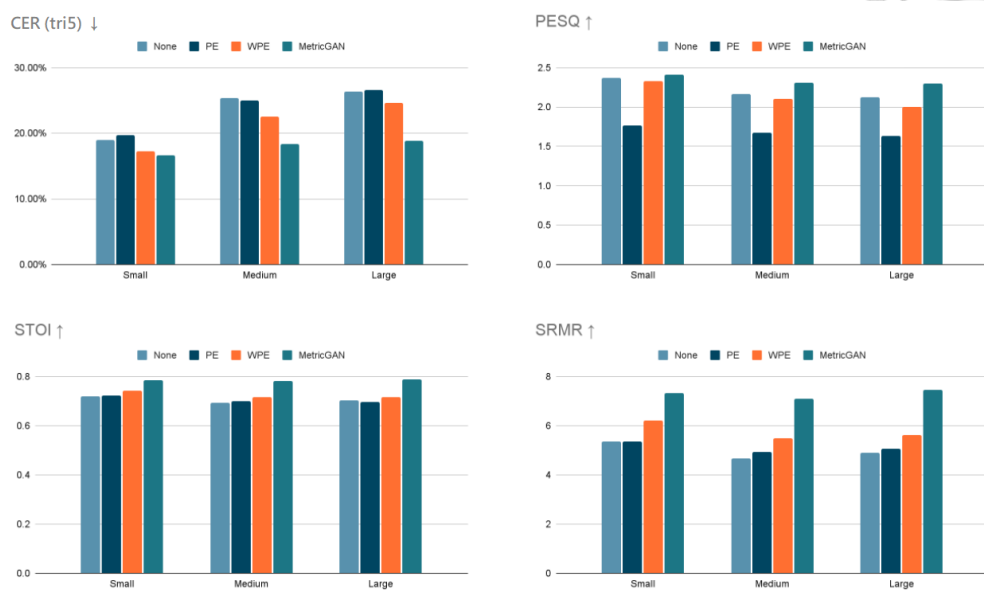


圖 5.3: 不同房間聲音品質綜合比較圖：使用 PESQ、STOI、SRMR、CER 為評量指標，對不同房間和做法進行分析。

表 5.4: 評估使用不同模型於不同房間之結果

殘響去除方法	房間大小	PESQ ↑	STOI ↑	SRMR ↑	CER (tri5) ↓
None	small	2.3700	0.7177	5.3456	19.01%
	medium	2.1625	0.6939	4.6786	25.31%
	large	2.1251	0.7014	4.8963	26.33%
PE	small	1.7688	0.7212	5.3443	19.73%
	medium	1.6758	0.6986	4.9174	25.04%
	large	1.6296	0.6974	5.0494	26.60%
WPE	small	2.3280	0.7418	6.2022	17.24%
	medium	2.1008	0.7156	5.4872	22.54%
	large	2.0039	0.7155	5.6304	24.61%
MetricGAN	small	2.4058	0.7840	7.3210	16.65%
	medium	2.3119	0.7810	7.0965	18.33%
	large	2.2950	0.7865	7.4604	18.81%

RESULT OF MIXUP AUGMENTATION

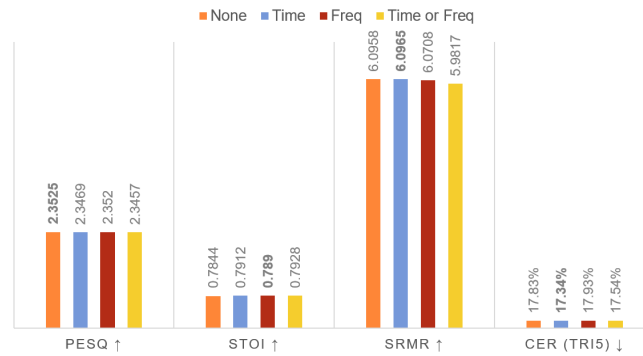


圖 5.4: Mixup 資料擴增對應評估方式比較圖：不使用資料擴增、時間維度的 Mixup、頻率維度的 Mixup 對應 PESQ、STOI、SRMR、CER 的交叉比較。

小房間的殘響對訓練的干擾較不顯著，在所有的評量指標，小房間相較於中型或大型房間皆得到更好的結果。訊號在中房間與大房間則有相近的表現，中房間在 PESQ 及 CER 來看表現較佳，大房間則是在部分 STOI 的觀測項和所有 SRMR 普遍有較佳的表現。

殘響去除方式中選擇 PE、WPE 及代表神經網路方法的 MetricGAN 比較以不同房間進行殘響去除的差異性，結果和章節 5.5.1 的分析相似。無論在何種類型的房間，使用 PE 並無法帶給訊號乾淨度顯著的提升或甚至有些下降；WPE 修補 PE 的缺點，使殘響去除效果提升；MetricGAN 則在各方面有最佳的表現。

5.5.3 應用 Mixup 資料擴增方法於殘響去除模型的效果

本實驗測試時間維度以及頻率維度的 Mixup 資料擴增以強化殘響去除模型的訓練結果。在本實驗 Mixup 資料擴增的生成是在時間階段產生，當模型將要使用一筆資料訓練時，資料將有一半的機率保持原樣進行訓練；而有一半的機率資料將進行 Mixup 取代原本的資料進行訓練。Mixup 在本文以三種方式被使用，時間維度 (Time) 的資料擴增、頻率維度 (Freq) 的資料擴增、以及隨機選取時間及頻率維度 (Time or Freq) 的方式。圖 5.4 為使用 Mixup 與評量方式的比較圖，表格 5.5 為詳細結果的表格。其中 CER 為只使用 Aishell1 乾淨聲音 (A) 訓練的聲學模型的語音辨識模型的辨識結果，由於實驗時間的限制，實作在 tri5 聲學模型，並於 CER 提供初步的比較。

使用 Mixup 擴增方式皆使 PESQ 下降，推測是 PESQ 觀察到資料間的非連續

表 5.5: Mixup 資料擴增對應評估方式詳細結果

Mixup 資料擴增	評估指標			
	PESQ ↑	STOI ↑	SRMR ↑	CER (tri5) ↓
None	2.3525	0.7844	6.0958	17.83%
Time	2.3469	0.7912	6.0965	17.34%
Frequency	2.3520	0.7890	6.0708	17.93%
Time or Frequency	2.3457	0.7928	5.9817	17.54%

性，然而對人類感官而言這缺陷是不明顯的，對品質的估計不一定會如 PESQ 嚴格。除了 PESQ 以外的觀測量，以時間維度進行 Mixup 擴增皆有些許提升，其中 CER 進步了約 0.5%；時間軸進行的資料擴增可視為頻率軸的橫向拼接，頻率維度的 Mixup 擴增使用了縱向拼接的方式，但並未取得好的結果，和完全不使用資料擴增的原始訓練方式相比，CER 小幅下降 0.1%。同時使用兩個擴增方式，和只使用時間維度的資料擴充相比也無明顯進步，除了 STOI 小幅提升外，並未提供其餘三項評估指標進步。

此實驗證明時間維度上的 Mixup 資料擴增對語音的訓練有一定的幫助，但其他利用於影像上的 Mixup 方式(如：斜線組合，以像素為單位組合)在語音上的有效性尚需更多驗證。

5.5.4 減少資料不一致的討論

本實驗使用和章節 5.4 中消除殘響測試與訓練訊號不一致性的方式，將原本的具有殘響的訓練訊號 (B) 通過 MetricGAN (章節 5.5) 的殘響消除模型產生新的訊號 (C) 加入訓練，利用此方法建立減緩資料不一致性後的訓練資料預期能得到更好的結果。圖 5.5 為整體的比較圖，其中代號依照表格 4.3 表示：A 為乾淨的訓練訊號、B 代表具有殘響的訓練訊號、C 代表原本 B 的資料利用殘響去除模型 (MetricGAN) 後的訓練訊號；a 代表乾淨的測試訊號、b 代表殘響的測試訊號、c 代表在 b 上利用殘響去除模型 (MetricGAN) 後的測試訊號。表格 5.6 及 5.7 為詳細的資訊，其中以殘響去除的訓練資料對原本在殘響環境的測試資料估計 CER 並不存在現實情況中使用，於是這部分的實驗不再討論範圍之中。

比較 ASR 在殘響去除過後的測試訊號和具有殘響的測試訊號的辨識結果，測

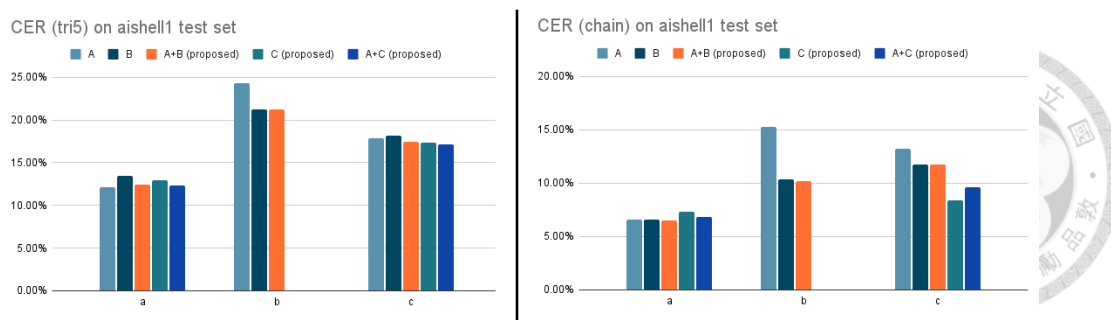


圖 5.5: 所有聲學模型於 CER 比較圖：提升訓練與測試資料的一致性助於字元錯誤率降低

表 5.6: 利用不同資料訓練之 tri5 聲學模型的 CER 表

訓練資料 測試資料	A	B	A+B (proposed)	C (proposed)	A+C (proposed)
a	12.12%	13.47%	12.39%	12.97%	12.35%
b	24.27%	21.25%	21.25%	-	-
c	17.83%	18.17%	17.42%	17.30%	17.14%

試訊號經過殘響去除後無論對應何種類的聲學模型皆有十分明顯的進步。使用特性相同的訊號進行資料擴增訓練的聲學模型，可以讓字元錯誤率更穩定地下降。但消除資料不一致性後需要讓模型對額外的資料進行擬合，對原始乾淨資料的字元錯誤率會略有上升。

在 tri5 的架構中，使用乾淨的語音語料 (A) 訓練的模型在乾淨語音測試資料中有最好的效果，若加上具有殘響語音的訓練資料一起訓練 (A+B)，反而會使使用在辨識乾淨語音訊號的結果為了擬和具有殘響的語音而稍微降低；若是在乾淨語音訊號的訓練上之上再增加殘響去除訊號 (A+C) 訓練後更能夠同時提升辨識含有殘響的語音訊號的能力。相較於使用 tri5 架構的聲學模型，使用 tdnn 架構的結合神經網路方式在辨識效果的比較上大幅提升。神經網路的訓練方式也被認為具有學習隱藏特徵的能力，在 A+B 的資料及結合訓練中在乾淨資料辨識上面也取得比單用 A 訓練稍好的成效。

當關注重點來到處理具有殘響的語音訊號，藉助殘響去除模型依舊有所優勢，在 tri5 回報的結果，A+C 訓練的模型擁有最好的效果，辨識含有殘響的語音訊號的能力達到 17.14% 的 CER。使用 tdnn 架構訓練後，單使用 C 訓練的模型雖然在乾淨的語音測試資料上表現最差，但在辨識具有殘響的語音資料效果最佳，其

表 5.7: 利用不同資料訓練之 tdnn 聲學模型的 CER 表

訓練資料 測試資料	A	B	A+B (proposed)	C (proposed)	A+C (proposed)
a	6.55%	6.58%	6.51%	7.31%	6.80%
b	15.26%	10.32%	10.19%	-	-
c	13.22%	11.77%	11.73%	8.39%	9.61%

CER 低達 8.39%。如果在只考慮辨識在殘響環境中的字句時，用 C 訓練的模型是最優秀的。

5.6 實驗三：殘響與乾淨特性分類器訓練之探討

本實驗的目的為從架構和結果分析乾淨與殘響訊號特性的分類器，該分類器可以區分輸入訊號的特性較接近乾淨訊號或含有殘響的訊號，讓使用本架構的訊號能選擇合適的聲學模型。

5.6.1 池化層比較

以乾淨和具有殘響的訊號 (A 和 B) 為分類器兩標記類別的訓練資料。當訊號被輸入至本文提出的架構時，作為分類音檔的第一步，所有音檔都將通過分類器，於是選擇使用輕巧、計算速度較快的 LCNN 為基本架構。LCNN 更改自 CNN，類似的模型應用於處理聲音訊號時，被期待模型能夠兼容處理不同時間長度的訊號。於是用池化的方式消除時間維度長度不一致的情況被考慮使用。表格 5.8 顯示在 LCNN 後分別連接 MAX、AVG、SAP、ASP 及銜接後面的全連接層的結果。由於不同的池化層影響輸出維度，參數數量的統計包括之後的全連接層。全連接層為維度 128 的隱藏層，向前與池化層連結，向後連接維度 2 (判斷特性接近殘響與乾淨的預測結果) 的輸出層。MAX 及 AVG 在池化層不需要訓練額外的參數，其參數來自於全連接層；SAP 及 ASP 訓練額外的參數計算加權數值，其中 ASP 以計算加權標準差增加輸出的維度，於是需求的參數量最多。

MAX 使用數值最大的音框代表全體音框的輸出，此方法過於偏頗，其模型擁有錯誤率最高；AVG 和 ASP 則有相近、最佳的結果；SAP 的表現不如預期，本

表 5.8: 池化層比較

池化層	錯誤率 (a)	錯誤率 (b)	總錯誤率	參數數量 (池化層 + 全連接層)
MAX	1.22%	5.46%	3.34%	66562
AVG	0.93%	2.07%	1.50%	66562
SAP	1.14%	3.53%	2.34%	83202
ASP	1.15%	1.50%	1.33%	148738

文對此現象有兩個推測：

1. 在訓練資料不充分的情況下 SAP 可能使池化結果著重於部分的音框 (SAP 原用於語者辨識訓練，訓練資料為 voxceleb2，大於 1600 小時)，如能加大資料量有利於驗證此看法的正確性。
2. AVG 再池化層中為較簡單的架構，AVG 平均考慮到每個音框視其為同等重要性，殘響在時間上是連續的特徵，殘響在一音框上的影響力會同時表現在其後續的音框。但 SAP 將音框視為不連續的區塊而著重部分音框，在殘響去除的題目中並不佔優勢，於是 AVG 在此題目擁有比 SAP 更佳的效果。ASP 計算更統計意義上更進階的數值：加權標準差，使其彌補 SAP 忽略部分音框的缺點，其結果能和 AVG 相比反而略勝一籌。

5.6.2 分類器結果分析

如章節 5.4 的結論，如僅使用單一路線決定最終辨識字串的方式，當訊號未使用合適的聲學模型會造成字元錯誤率的提升。此小節旨於分析選擇單一路線的正确程度對辨識結果的影響。本研究提出的架構圖包含兩個聲學模型，對應乾淨和具有殘響的訊號使用，其中 A+B 主要處理乾淨語音序號的辨識；C 為使用 MetricGAN 殘響去除過後的訊號訓練，處理擁有殘響的語音訊號辨識。整理如表格 5.9。本單元在 ASR 上的測試直接實驗於 tdnn 架構。

假設存在完美分類器能完全正確地分類特性為殘響和乾淨的訊號，則可以將辨識錯誤率壓到最低，但依照表格 5.10 的結果，即使利用完美分類器，字元錯誤率也只有些微的降低 (CER 從 7.497% 略進步至 7.470%)，使用章節 5.6.1 搭配池化層

表 5.9: 兩聲學模型的字元錯誤率比較圖

聲學模型	CER (a)	CER (c)
A+B	6.51%	7.31%
C	11.73%	8.39%



表 5.10: 殘響與乾淨特性分類器與字元錯誤率對應表

分類器選擇	CER (a+c)
完美分類器	7.470%
永遠分類特性為乾淨	9.120%
永遠分類特性為殘響	7.850%
LCNN + MAX	7.538%
LCNN + AVG	7.501%
LCNN + SAP	7.518%
LCNN + ASP	7.497%

的分類器也能得到接近的結果。追求分類器的正確性對結果影響甚微，且擁有過度擬合於特定情況的風險。下章節將討論模型結合演算法，通過使用多個模型期望讓字元錯誤率有更明顯的減少。

5.7 實驗四：探討模型結合演算法對字元錯誤率的影響

本實驗使用章節 3.4 描述的模型結合方式，綜合使用兩聲學模型的結果降低最終的字元錯誤率。首先溫度 (temperature) 法被用於避免分類器輸出過於武斷的結果，輸出的文字串將以結合分類器的結果或是進一步結合 ASR 評分的 sentence level fusion (SLF) 或 word level fusion (WLF) 演算法融合選擇兩路徑辨識出的文字串以達更佳的輸出。

5.7.1 溫度法的分析

溫度值的選擇 (依照章節 3.4 以 α 表示) 決定殘響與乾淨特性分類器兩輸出的接近程度。選用較大的 α ，會導致分類器的輸出的值將更接近，進而加重 ASR 的評分在決定最後文字串輸出的重要性；選用較小的 α 時，分類器的輸出差距較大，

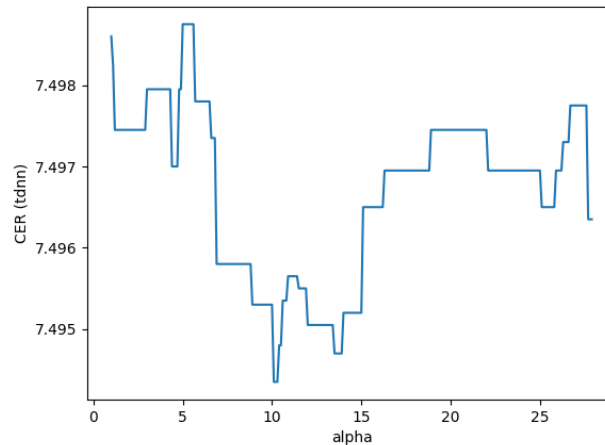


圖 5.6: 溫度調整：測試章節 3.4 中使用溫度法的不同 α 值導致的結果，使用 SLF 為結合方式。

最後文字串的輸出將更依賴分類器的輸出結果。

圖 5.6 為嘗試不同 alpha 值對字元錯誤率的測試結果，以 SLF 演算法為測試時的模型混合方式。實驗中 α 以每次增加 0.1 為一個選點，測試 $\alpha = 1$ 至 $\alpha = 28$ 。預期中溫度調整能找尋分類器和 ASR 評分的中適合的比例，和 CER 的作圖估計形成先降後升的趨勢，並在某溫度時擁有最佳解。在測試結果的曲線有所起伏，但在 α 位於 10-15 之間約略為低點。然而其造成的影響在小數點後三位數的範圍，並不顯著。本文推論形成浮動下降趨勢的原因：由於測試資料混合乾淨訊號和具有殘響的訊號的訓練資料各 7176 筆，每筆資料可能有各自適合的溫度值，對部分資料較小的 α 是適合的溫度，於是字元錯誤率對溫度的趨勢有所起伏。然而大部分資料藉由提升溫度讓 ASR 的分數能造成更大影響性，形成下降的趨勢。

5.7.2 模型融合的分析

此章節比較 SLF、WLF 及其他的模型組合方式的優劣和結果，使用 Aishell1 的乾淨及含有殘響訊號為測試資料。圖 5.7 為五種方式比較的結果，包含：僅使用乾淨訊號訓練的聲學模型、僅使用殘響去除訊號訓練的聲學模型、以分類器結果決定路徑 (direct ensemble)、sentence-level fusion (SLF) 及 word-level fusion (WLF)。表格 5.11 為比較後的詳細結果，ASR 中聲學模型使用 tdnn 的架構進行訓練。除了 CER 數值回報外，增加了錯誤下降率 (error reduction rate) 的資訊，以原始乾淨聲學模型為測試基準，比較實作其他實驗方式的進展。

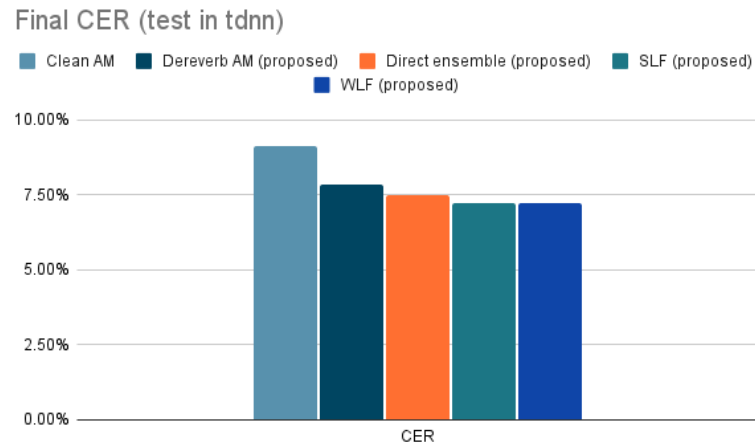


圖 5.7: 不同模型組合方式比較圖：使用模型組合方式比單一模型有更佳的结果，SLF 和 WLF 在實驗中有最好的效果。

表 5.11: 不同模型組合方式比較表

模型組合方式	CER (a+c)	Error reduction rate
Clean AM only (A)	9.12%	0.00%
Dereverb AM only (A+C)	7.85%	13.92%
Direct ensemble (depend on classifier only)	7.50%	17.76%
SLF	7.24%	20.61%
WLF (K-best LCS)	7.23%	20.72%

表 5.12: Model Fusion 錯誤分析

Ground true	如何 借力 拥抱 互联网 加 这 一 全新 变量
Direct ensemble	如何 界的 拥抱 互联网 加 这 以全 新 电量
SLF	如何 借力 拥抱 互联网 加 这 一 全新 变量
WLF (K-best LCS)	如何 借力 拥抱 互联网 加 这 以全 新 电量

表格 5.6 及 5.7 的結果說明，使用字串組合的方式相較於使用單一模型的語音辨識效果有明顯提升。單一模型的方法中，利用乾淨訊號訓練的聲學模型辨識含有殘響的訊號效果不佳；使用殘響去除後訊號和乾淨訊號的聲學模型在乾淨訊號的字元錯誤率略有提高，於是同時結合兩模型的優點為值得研究的議題。

直觀上結合兩路徑的方式為利用殘響與乾淨特性分類器的結果，分類器決定輸入訊號的類別為乾淨或殘響，依照類別的屬性選擇對應路線的聲學模型(分類器使用表格 5.8 中以 ASP 為池化層的模型)。由於室內環境皆會有殘響產生，原本被歸類於乾淨的音檔有一部分選擇屬於具有殘響訊號的路徑；另一方面，訊號中微弱的殘響對辨識影響不大，所以原本被歸類於特性是殘響的音檔有一部分選擇乾淨訊號的路徑。利用分類器決定選擇的路徑後字元錯誤率有所降低，然而該方式只考慮到分類器的評分，缺少考慮 ASR 辨識的評分，導致此方法對決定整體語音辨識系統仍有欠缺。

本文設計結合分類器預測和後續的 ASR 評分的方式：SLF、WLF，以期望達到更好的字元錯誤率。SLF 以兩路徑辨識的整體文字串為單位進行結合，模型將選擇兩路徑分數更高者而不只是分類器分數較高的，此改進讓結合 ASR 的選擇方式和系統的目標更一致，獲得較好的結果；WLF 將為字串拆解成單詞，以單詞為基本單位進行兩路徑的結合，利用動態規劃抽換單詞讓句子達到最高的分數，但同時冒著替換到明顯錯誤字詞的風險。按照章節 3.4.3 的算法組合最終的文字串，WLF 實作上需要先計算兩路徑相同的子序列，提出的方法為使用 K-best LCS 實作字詞結合。相較經驗法，K-best LCS 由執行多個可能的結合圖並選擇分數較佳的一個，達到更合理的輸出。使用 K-best LCS 的組合方式達到和 SLF 相似的結果；使用經驗法有時導致不合理的結合圖，而讓效果略微降低。

表 5.12 為使用模型組合的一實例，圖 5.8 為選擇方式的示意圖。Direct ensemble 僅依靠分類器的結果選擇路線，分類器在下方路線的分數為 -0.39 ，大

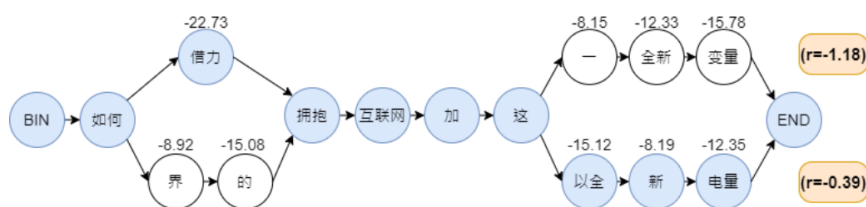


圖 5.8: Model Fusion 實例：三種模型組合方式將導致不同的輸出文字串。

表 5.13: Direct ensemble 和 SLF 比較

Method	Amount of files
Direct ensemble better	1223
SLF better	7173
Same effectiveness	5957

於上方路徑的 -1.18 ，於是 Direct ensemble 將直接選擇下路線的「如何界的拥抱互联网加这以全新电量」為輸出。使用 WLF 結合時將選擇每個分支的最佳(分數最高)的結果，第一個分岔點中「借力」的分數為 -22.73 高於「界的」的 $-8.92 - 15.08 = -24$ 分，於是選擇上方路徑；第二個分支，下路線的「以全新电量」分數合為 -35.66 大於上路線的 -36.26 。於是最後輸出為「如何借力拥抱互联网加这以全新电量」。SLF 則綜合考慮分類器的分數和語言模型辨識文字串的整體評分選擇路徑，最終選擇上方的路徑分數較高，於是「如何借力拥抱互联网加这一全新变量」為最終的輸出結果。在此範例中 Direct ensemble 的效果最差；WLF 藉由選擇部分正確路徑降低其字元錯誤率；SLF 達到完全正確的辨識效果。本文對三種方式進行統計的分析以比較其優劣。

測試音檔使用 Aishell1 的測試資料和其加上人工生成殘響訊號的人造音檔共 14352 筆。表格 5.13 為 Direct ensemble 和 SLF 的比較。共有 5957 筆資料兩者有同樣表現，其餘測試資料 SLF 累積勝過 Direct ensemble 5950 筆；表格 5.14 則是 Direct ensemble 和 WLF 的比較，WLF 累積勝過 5966 筆；SLF 和 WLF 結合方式

表 5.14: Direct ensemble 和 WLF 比較

Method	Amount of files
Direct ensemble better	1231
WLF better	7197
Same effectiveness	5925

表 5.15: SLF 和 WLF 比較

Method	Amount of files
SLF better	594
WLF better	602
Same effectiveness	13157



有非常接近的表現，WLF 的優點為可以用 DP 更正單一文字串的選擇，根據表格 5.15 的結果，有 602 個音檔因此得到較小的字元錯誤率。但其壞處為，WLF 可能破壞語言模型選擇單詞的邏輯，反而將正確單詞抽換成錯誤的情況，表格 5.15 顯示有 594 筆音檔面臨此情況。WLF 能小幅度提升結果，但 SLF 在模型結合已是有力的表現。



Chapter 6

結論與未來展望

第六章統整先前實驗的總結以及未來可實驗的研究方向。

6.1 結論

本文提出適用於殘響環境的語音辨識架構，以四個方面改善當前架構的缺陷：殘響與乾淨特性分類器、殘響去除模型、資料特性一致性、字串組合演算法。當本架構接收到測試訊號時，首先、使用 LCNN 搭配 ASP 池化層的殘響與乾淨特性分類器在乾淨與含有殘響語音訊號的 aishell1 資料集中以 1.33% 的錯誤率將訊號的特性分類為乾淨或殘響。

第二步透過殘響消除模型 (MetricGAN) 除去被分類器判斷為殘響特性訊號中的殘響。然而，通過殘響消除模型的訊號以普通、乾淨訊號訓練的聲學模型進行語音辨識無法達到最佳效果。同時本文透過實驗時間以及頻率維度的 mixup 資料增強提升殘響去除的效果。本實驗測試 ASR 效果時使用 tri5 的聲學模型。在此條件下，頻率維度增強使字元錯誤率減少 0.3%；時間維度的增強則有 0.5% 的字元錯誤率下降。

本文在第三步設計對應適用於殘響去除訊號的聲學模型：利用乾淨的訊號加上殘響去除後的訊號訓練的聲學模型進行辨識，相較於僅使用乾淨訊號的聲學模型，雖然在乾淨語音訊號 aishell1 測試資料集 CER 提升 0.80%，但形成擁有殘響語音訊號 aishell1 測試資料集 CER 3.34% 的減少。單一辨識方式無法完美符合所有訊號的需求，於是使用殘響與乾淨特性分類器區分資料特性後選擇用相同特性

訓練的聲學模型，和僅使用乾淨的聲學模型相比，擁有 17.76% 的錯誤下降率。

僅使用分類器的結果只達到有限的進步。考慮到語音辨識模型的評分，本文最終提出字串結合的方式：SLF、WLF。兩方法分別以串流或單詞為單位結合兩路徑辨識結果。WLF 在此實驗上略勝一籌達到和僅使用乾淨的聲學模型相比 20.72% 的錯誤下降率，並擁有明顯的進步。

6.2 未來展望

未來研究期望能透過減少訓練成本及增進辨識效果兩方面調整架構。相較於當前的語音辨識架構，本文提出的架構增加了殘響特性辨識器、兩路徑運算、及字串組合的運算量。未來期望能使用更輕量的模型、更快速的結合演算法、或共用兩路徑重複的參數以減少額外運算負擔。

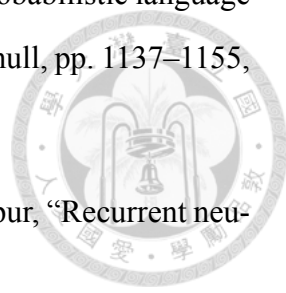
本文使用 kaldi 為基本架構研究語音辨識模型，但由於底層實作的方式 kaldi 難以兼容其他開發方式。kaldi 雖然在聲學模型有很好的表現，但只提供基礎的語言模型，語言模型近期開發的成果難以套用在當前的實作中。未來期望能將本文的想法用更有彈性的方式實作，以增加擴充能力。

本文使用 mixup 資料增強有效提升在 STOI、SRMR 面相的殘響去除能力，此能力反應至 CER 的進步上。然而，本文只測試時間以及頻率維度的 mixup，其他混和方式的效果有待未來時間驗證。此外，本文僅使用時間和頻率維度上的混合，其餘混合方式如斜向切割，像素網格切割等...，的效果值得未來探究。此外，在訓練殘響去除模型時，MetricGAN 利用將 PESQ、STOI 等聲音品質估計函式當作損失函數讓模型更適切地收斂。雖然有效提升殘響去除效果，但距離整體架構字元錯誤率降低的目標仍有些許出入，未來期望能將 CER 融入殘響去除模型的訓練中。



Bibliography

- [1] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," 2018.
- [2] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.
- [3] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.
- [4] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018.
- [7] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, Mar. 1989.

- 
- [8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model,” *The Journal of Machine Learning Research*, vol. 3, no. null, pp. 1137–1155, Mar. 2003.
- [9] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, “Recurrent neural network based language model,” in *INTERSPEECH*, 2010.
- [10] T. Yoshioka and T. Nakatani, “Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.
- [11] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.
- [12] W. Mack, S. Chakrabarty, F.-R. Stöter, S. Braun, B. Edler, and E. A. Habets, “Single-channel dereverberation using direct mmse optimization and bidirectional lstm networks,” in *INTERSPEECH*, 2018, pp. 1314–1318.
- [13] C. Donahue, B. Li, and R. Prabhavalkar, “Exploring speech enhancement with generative adversarial networks for robust speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.
- [14] J. Su, Z. Jin, and A. Finkelstein, “Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” *arXiv preprint arXiv:2006.05694*, 2020.
- [15] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.
- [16] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, “Natural tts synthesis by conditioning wavenet on

- mel spectrogram predictions,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.
- [17] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, “Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.
- [18] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, “Metricgan-u: Un-supervised speech enhancement/dereverberation based only on noisy/reverberated speech,” *arXiv preprint arXiv:2110.05866*, 2021.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [20] C. Summers and M. J. Dinneen, “Improved mixed-example data augmentation,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1262–1270.
- [21] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, “STC Antispoofing Systems for the ASVspoof2021 Challenge,” in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Counter-measures Challenge*, 2021, pp. 61–67.
- [22] A. Alex, L. Wang, P. Gastaldo, and A. Cavallaro, “Mixup augmentation for generalizable speech separation,” in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2021, pp. 1–6.
- [23] 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)*, vol. 2. IEEE, 2001, pp. 749–752.

- [24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.
- [25] T. H. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.
- [26] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, “Implementing a gammatone filter bank,” *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.
- [27] B. Safadi and G. Quénot, “Re-ranking by local re-scoring for video indexing and retrieval,” in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2081–2084.
- [28] R. I. Greenberg, “Bounds on the number of longest common subsequences,” *arXiv preprint cs/0301030*, 2003.
- [29] E. L. Lawler, “A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem,” *Management science*, vol. 18, no. 7, pp. 401–405, 1972.
- [30] G. Tür, J. H. Wright, A. L. Gorin, G. Riccardi, and D. Hakkani-Tür, “Improving spoken language understanding using word confusion networks.” in *Interspeech*, 2002.
- [31] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” 2017.
- [32] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

- [33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.