國立臺灣大學電機資訊學院電機工程學系

碩士論文

Department of Electrical Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

基於語音基石模型之語者自動分段標記系統

Improved Speaker Diarization Based on Speech Foundation Models

李高迪

Ko-Tik Lee

指導教授: 李宏毅 博士

Advisor: Hung-Yi Lee Ph.D.

中華民國 113 年 1 月

January, 2024

摘要

當前，語者自動分段標記系統 (speaker diarization) 主要運用三種方法：階段性、端到端和端到端-階段性混合系統。端到端系統在某些資料集上顯著優於其他方法，引起廣泛關注。然而，這種系統可能在實際應用中面臨泛用性限制，而階段性系統的潛力可能被低估。與此同時，近期的語音基石模型 (speech foundation model) 在多項語音任務中表現出色，顯示出其廣泛應用的潛力。然而，在語者自動分段標記方面，對其應用尚未深入探討。

因此，本研究旨在將語音基石模型應用於語者自動分段標記相關任務，進行性能比較並進行表現基準化。同時，針對階段性系統存在的問題，提出了改進方法，例如具緩衝區意識的話語開始點偵測和聚類純化，顯著提升了其性能。最後，透過域外評估方法，證實了端到端-階段性混合系統的泛用性問題，並提出了改進方法。本論文改進後的階段性和端到端-階段性混合系統在多個資料集上實現了與最先進技術相當甚至更優越的表現。

關鍵字：語者自動分段標記、語音基石模型

Abstract

Currently, speaker diarization systems primarily employ three methods: incremental, end-to-end, and hybrid incremental end-to-end systems. The end-to-end approach has shown significant superiority over other methods in certain datasets, garnering widespread attention. However, this system might face limitations in real-world applications, potentially underestimating the potential of incremental systems. Simultaneously, recent advancements in speech foundation models have showcased outstanding performance across multiple speech tasks, indicating their broad applicability. Nevertheless, their application specifically in speaker diarization remains insufficiently explored.

Therefore, this study aims to apply speech foundation models to tasks related to speaker diarization, conducting performance comparisons and standardization. Additionally, addressing issues present in incremental systems, proposed enhancements such as collar-aware speech onset detection and cluster outlier handling significantly improved their performance. Finally, through out-of-domain evaluations, the limitations of the hy-

brid systems were confirmed, along with proposed solutions for improvement. The refined incremental and hybrid systems in this paper achieved comparable or even superior performance to state-of-the-art methods across multiple datasets.

Keywords: speaker diarization, speech foundation model

目次

圖次

表次

第一章　導論

1.1　研究背景與動機

語者自動分段標記 (speaker diarization, SD) 是語音處理領域的關鍵任務，其主要目標是從語音中識別語者以及其說話片段。過去，它主要用作自動語音辨識系統 (ASR) 的前處理步驟，但後來也在其他多個應用領域取得顯著進展 [69]，包括語音導航、內容檢索以及更高階的人機互動情境。

語音基石模型 (speech foundation model) 亦是近年備受關注的研究方向，其目標是以單一大型模型同時完成各式各樣的語音任務。過去，語音基石模型已被證實在自動語音辨識 [11]、語者驗證 [13] 及聲音事件偵測 [30] 等下游任務取得卓越的成果。然而，在語者自動分段標記下游任務中，過去的研究都只有將語音基石模型應用到特定電話對話情景 [13]、人工生成的數據 [97]、或只有應用到系統的一小部分 [48]。因此，為了彌補此研究領域的不足，本論文將探討語音基石模型在語者自動分段標記系統的應用。

語者自動分段標記系統主要可分成階段性、端到端及端到端-階段性混合系統三種。早期的系統通常都使用由不同模組構成的階段性系統：先使用語音活性偵測 (voice activity detection, VAD) 偵測語音中有人聲的的片段，再透過提取語者特徵向量 (speaker embedding) 以及聚類 (clustering)，識別每個片段的語者。基於聚

類的傳統階段性系統通常隱含地假設每個語音片段都只有一位語者，難以應對有重疊語者的語音。後來，有研究者提出端到端系統 EEND [25]，直接最小化辨識錯誤 (diarization error rate, DER)，並在處理重疊語音的機制更為直覺。但是，端到端系統存有一些缺陷，例如無法處理過長的音檔 [36]，以及無法輕易處理具有任意數量說話者的語音 [101]。最後，有研究者提出端到端-階段性混合系統，在局部語音區塊 (chunk) 使用端到端模型並結合全域語者聚類 [44] [43]，在享有端到端模型優勢的同時，亦能透過全域聚類解決端到端系統的缺陷。爾後不少相關研究 [13] [71] 都採用類似做法。

近年來，階段性系統逐漸不再是研究的焦點，主要原因有三。首先，階段性系統在語者分割 (speaker segmentation) 階段存有準確率與時間精準度的平衡問題 [69]。其次，階段性系統一般被認為在處理重疊語音的表現不佳。最後，階段性系統在效率上亦不及端到端-階段性混合系統 [81]，不利於真實世界部署。然而，隨著語者特徵向量新架構的提出 [93] 以及聚類演算法的進步 [82]，階段性系統的性能存有被低估的可能。因此，本論文將重新審視並評估階段性系統每個模組的性能。

端到端及端到端-階段性混合系統雖然在諸多方面均被認為優於傳統階段性系統，但卻存有域泛化性 (domain generalization) 不佳的隱憂 [69]。原始 EEND 論文發現模型有過度擬合 (overfitting) 到資料分佈的現象，但後續其眾多變形 [36] [37] [43] 並未質疑系統的泛化性問題。與本論文同期的研究 [71] [28] 有嘗試將其提出之模型同時用在多個資料集訓練，以測試其泛化能力，但均未完整討論系統的域外 (out-of-domain) 泛化能力。因此，本論文將針對端到端-階段性混合系統的域外泛化能力作詳細探討。

本論文的最終研究動機是要將語者自動分段標記系統用於真實世界部署，因

此需要在系統準確率 (尤其是域外泛化能力) 及模型推論速度上取得平衡。部分語音基石模型雖然性能優秀，但因參數量巨大難以被使用 [1]。因此，本論文也會訂定基準來評估不同參數量的語音基石模型在語者相關任務中的表現，以找出準確率和效率兼具的最佳模型，並將其應用到語者自動分段標記系統中。

基於以上，本論文提出以下研究問題：

- 語音基石模型提供語者自動分段標記系統多大的幫助？
- 哪一個語音基石模型最適合被用於語者自動分段標記系統真實世界部署？
- 在系統性能上，階段性系統是否明顯低於端到端-階段性混合系統？
- 在域外泛化能力上，階段性及端到端-階段性混合系統的表現分別如何？

1.2 研究方法

本論文有兩大研究方向，第一個方向是應用語音基石模型到語者自動分段標記系統及其相關下游任務。

首先，為驗證語音基石模型帶來之性能提升，本論文在實驗中會比較使用語音基石模型前 (亦即使用傳統模型)，以及使用語音基石模型後的表現。

其次，本論文欲比較不同語音基石模型在不同參數下的表現，以找出最適合用於語者自動分段標記系統的模型；然而，開源的語音基石模型數量繁多，無法在所有實驗逐一比較。因此，本論文會先利用基準衡量十二個開源模型的表現，並從中挑選一個高資源模型、一個中資源模型及一個低資源模型做代表，去完成其他實驗。

最後，在應用模型到下游任務的順序上，本論文採逐層深入探討的方式：先

從被廣泛認可的 SUPERB 基準 [97]，挑選語者相關任務作評估；再討論不同系統通用的任務，如重疊語音偵測、語者切換點偵測及語者特徵向量聚類等；最後討論在階段性系統及端到端-階段性混合系統之表現。

本論文的另一研究方向包括探討不同模型及系統的泛化能力。

深度學習模型的評估需要同時考慮其在特定領域（域內, in-domain）和未知領域（域外, out-of-domain）的表現能力。域內評估強調模型在已知領域的表現，而域外評估則考慮模型對於新領域的適應性。綜合這兩種評估方式能確保模型不僅在熟悉領域表現出色，同時也能適應未知領域的挑戰，提升其實用價值。因此，本論文在比較模型及方法的優劣時，有時候會同時報告域內表現及域外表現，以呈現其泛化能力。

由於本論文在進行訓練模型時，很常使用由多個資料集組成的「複合資料集」，以提升系統表現，因此需要特別釐清域內評估及域外評估的定義。圖 1.1 比較了本論文進行域內評估及域外評估時的流程。

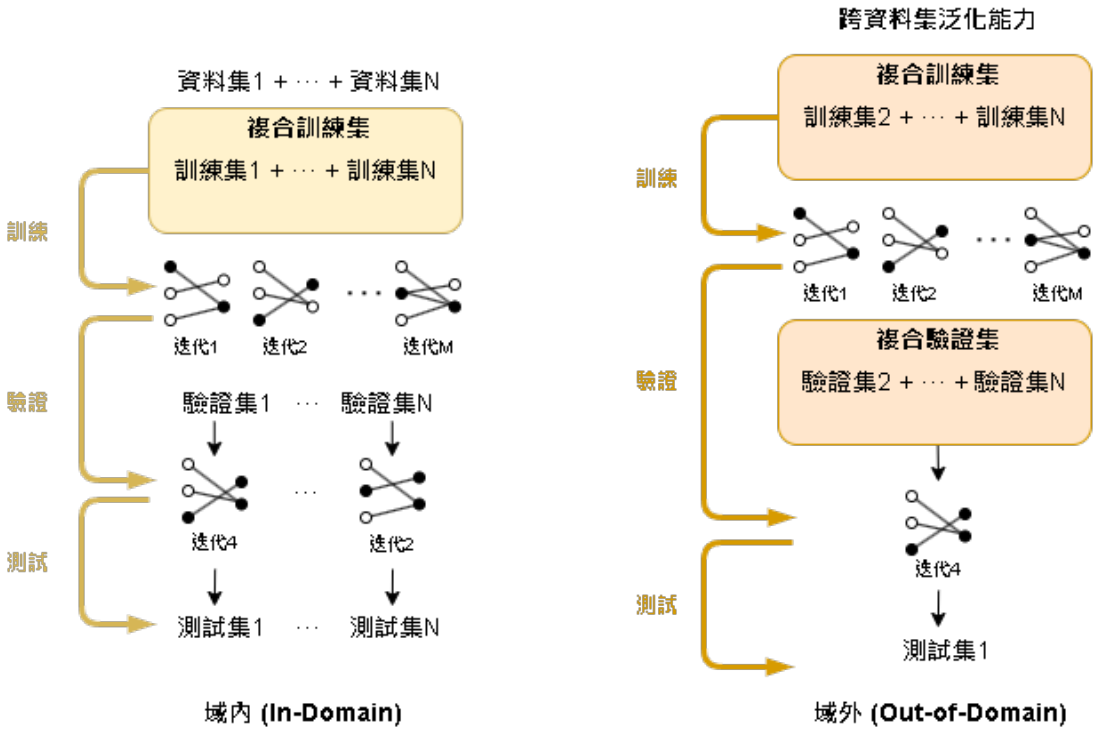

圖 1.1: 域內評估及域外評估在訓練、驗證及測試階段之示意圖

在域內評估中，假設使用了 N 個資料集 $D = \{D_1, D_2, ..., D_N\}$，訓練階段會利用整合的全部訓練集 $Train = Train_1 + Train_2 + ... + Train_N$ 進行模型訓練，得出模型迭代 $I = \{Iter_1, Iter_2, ..., Iter_M\}$。測試階段會在每個資料集上，逐一利用驗證集 $Val = \{Val_1, Val_2, ..., Val_N\}$ 和測試函數 $Score$ 尋找最佳的模型迭代 $bestIter$，並報告其在對應的測試集 $Test = \{Test_1, Test_2, ..., Test_N\}$ 上的表現。

而在域外評估中，同樣使用了 N 個資料集 $D = \{D_1, D_2, ..., D_N\}$，然而需要進行 N 次訓練。對於每個資料集 $\forall D_i \in D$，使用除了該資料集 D_i 以外的複合訓練集 $\bigcup_{j \neq i} Train_j$ 進行模型訓練，得出模型迭代 $I = \{Iter_1, Iter_2, ..., Iter_M\}$。在測試階段，使用除了對應的驗證集 Val_i 以外的複合驗證集 $\bigcup_{j \neq i} Val_j$，並利用測試函數 $Score$ 尋找最佳的模型迭代 $bestIter$，並報告其在對應的測試集 $Test_i$ 上的表現。

1.3 主要貢獻

本論文有以下三點主要貢獻:

- 應用語音基石模型到語者自動分段標記系統及其相關下游任務，並在多個資料集取得與最先進方法相若或更優的表現

- 提出階段性系統各個模組之改進方法，如具緩衝意識之話語開始點偵測，聚類純化等，大幅提升性能

- 研究階段性系統及端到端-階段性混合系統之跨領域泛化能力，發現端到端-階段性混合系統仍有泛化能力不足的問題，並可透過移除系統限制或使用語音基石模型改善

1.4 章節安排

以下為本論文的章節安排:

- 第二章: 介紹相關背景知識，如語音基石模型、語者自動分段標記系統以及其評估方法與資料集
- 第三章: 探討語音基石模型用於語者下游任務之表現，以及階段性系統各個模組之改進方法，最後探討整體階段性系統之性能及泛化能力
- 第四章: 探討端到端-階段性混合系統之性能及泛化能力，最後綜合比較不同系統的性能
- 第五章: 論文總結與未來展望

第二章　背景知識

2.1　語音基石模型

2.1.1　簡介

基石模型 (foundation model) [3] 是指任何在廣泛數據上訓練並可適應於多種下游任務 (downstream task) 的模型。此概念最初盛行於自然語言處理領域，後來逐漸推廣至電腦視覺及語音處理領域。

在語音領域，語音基石模型主要分為三類: 基於自監督式學習 (self-supervised learning) 之自監督式語音模型、自動語音辨識系統 (ASR) 及語音大型語言模型 (LLM)。前兩者更常用於判別式任務 (如分類問題)；而語音大型語言模型則更常用於生成式任務，在語者自動分段標記任務上的應用可能相對不那麼直觀。因此，本節將主要討論自監督式語音模型及自動語音辨識系統兩種語音基石模型。

此外，本節亦會介紹本論文用到的 SUPERB 基準 [97]，它常被用於評估自監督式語音模型在語音下游任務的表現。

2.1.2　自監督式語音模型

自監督學習的目標是在不需依賴人類標註的情況下，從大量未標記的資料中學習具有區分性的特徵 [31]。目前最先進的方法 [112] 常使用自監督學習進行預訓

練 (pretrain)，然後在下游任務中進行監督式訓練以微調模型 (finetune)。這種方式往往能夠用更少的標註資料獲得更出色的表現。

自監督式語音模型 (self-supervised speech model) 按照預訓練目標主要可分成生成式 (generative)、對比式 (contrastive) 及預測式 (predictive) [65]，其中對比式模型及預測式模型被證實在下游任務具有更好的表現 [97]。表 2.1 按照預訓練目標列舉本論文用到的幾種自監督式語音模型。

模型	類別	預訓練目標
Wav2vec2 [2]	對比式	對比損失 (constrative loss)
HuBERT [40]	預測式	掩蔽區域預測 (masked prediction loss)
WavLM [13]	預測式	掩蔽區域預測 + 降噪 (denoising)

表 2.1: 本論文所使用的的自監督式語音模型種類

以下以 HuBERT 模型作為示例簡介自監督式語音模型及其預訓練方式:

HuBERT 模型 (如圖 2.1) 主要由卷積神經網絡 (CNN) 及轉換器編碼器 (transformer encoder) 組成。其中 CNN 編碼器可以從語音波型提取有用特徵並降採樣，而轉換器編碼器則能利用自注意力機制 (self-attention)，捕捉輸入序列不同位置的相互依賴關係，讓輸出的結果更精確。

HuBERT 在進行預訓練前，會先使用 K-Means 演算法對從語音提取的 MFCC 特徵進行聚類 (clustering)，以無監督的方式提供離線預測目標。接著在預訓練階段，以掩蔽區域預測損失訓練模型：先掩蔽輸入序列的一部分，讓模型透過其相鄰序列的資訊，推論出被掩蔽區域的預測目標。這樣的預訓練目標能幫助模型學習提取更有用的特徵。最後，由於 MFCC 特徵的表現能力有限，預訓練的目標會在訓練一輪後，改由訓練好的模型提供，採迭代優化的方式進一步優化模型。

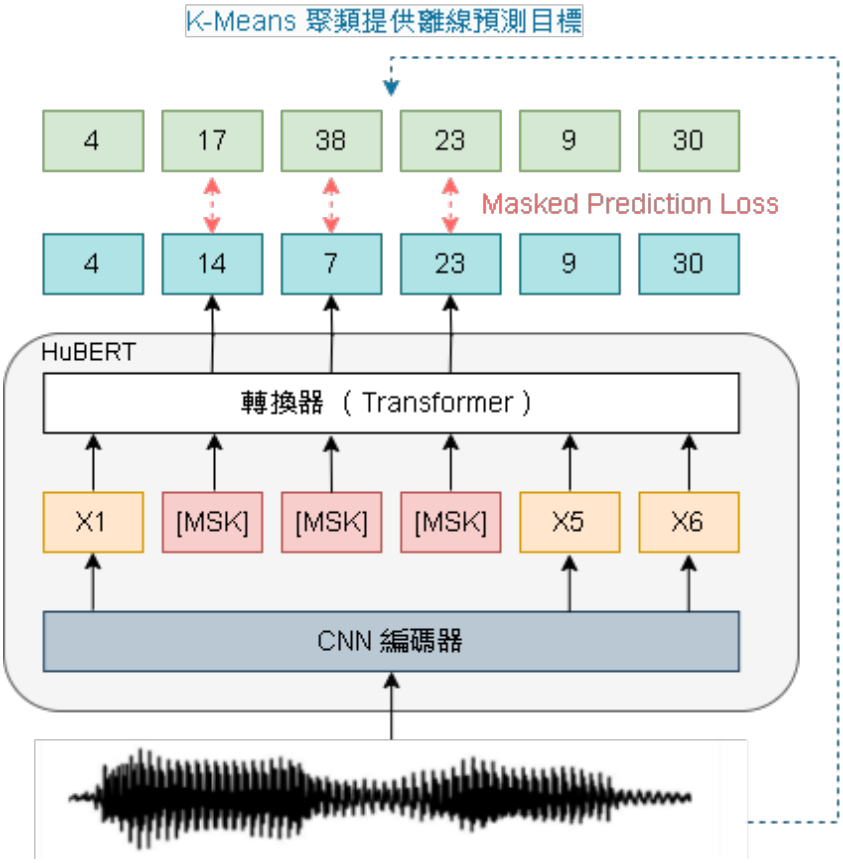

圖 2.1: HuBERT 模型示意圖

2.1.3 自動語音辨識系統 Whisper

過去的自動語音辨識系統僅被用於一項任務：將語音轉換為文字。然而，隨著 Whisper 模型 [74] 的出現，自動語音辨識系統現在也可視為語音基石模型。

Whisper 模型 (如圖 2.2) 使用 Seq2Seq (序列對序列) 架構 [88] 進行語音轉錄 (transcribe) 與語音翻譯 (speech translation) 多任務學習。模型輸入為從語音提取的梅爾頻譜 (mel-spectrogram)，先利用轉換器編碼器提取特徵，再以交叉注意力 (cross-attention) 的方式傳遞給轉換器解碼器 (transformer decoder) 進行下一個標記預測 (next-token prediction)。解碼器每次預測下一個標記時，其輸入包括與多任務學習相關的標記和過去的輸出。

Whisper 一共使用 680000 小時多語言資料上進行弱監督式學習，大幅超越以

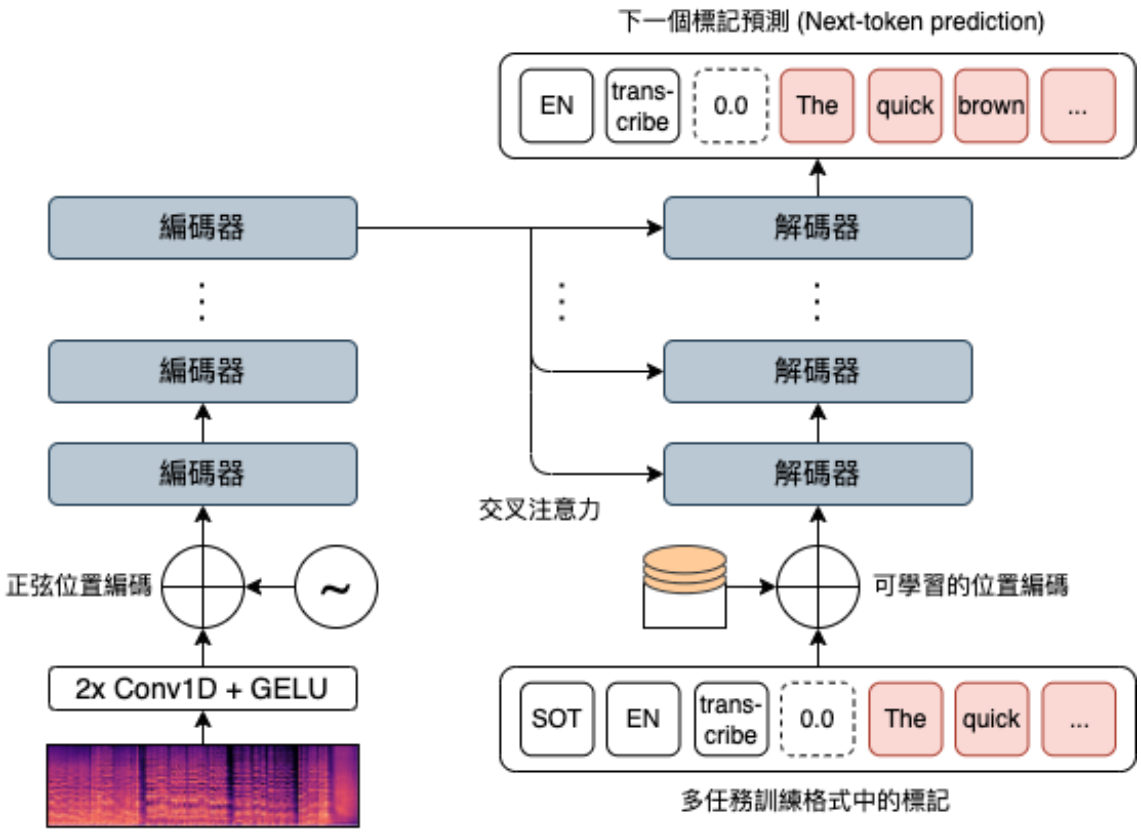

圖 2.2: Whisper 模型示意圖

往自動語音辨識系統使用的數據量，並在自動語音辨識取得最先進的結果。不少研究因此對 Whisper 預訓練模型的能力感興趣，將其用在不同下游語音任務上。例如 [30] 利用 Whisper 編碼器輸出在聲音事件偵測上取得最先進的表現；[102] 證實 Whisper 編碼器輸出在低資源的情況下，在語義下游任務有優秀的表現，卻相對不擅長語者任務。[12] 則發現直接以 Whsiper 解碼器輸出進行下游任務效果一般，需要進一步微調 Whisper 才能得出好表現。基於上述研究結果，本論文採用 Whisper 編碼器作為語音基石模型用於下游任務。

2.1.4 SUPERB 基準

隨著語音基石模型的廣泛應 c 用，對於各模型能力的基準化變得至關重要。SUPERB 基準 [97] 將十個語音相關下游任務分成內容、語者、語義及副語言四類，並為每個任務訂定統一的模型架構、資料集、訓練方法及評估方法。

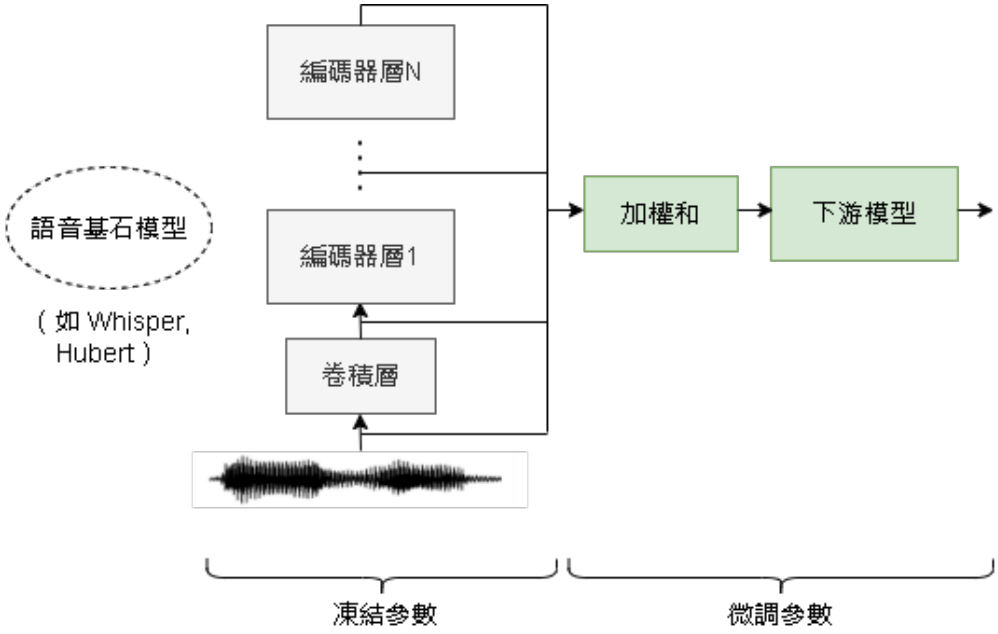

圖 2.3: 語音基石模型用於下游任務通用架構

SUPERB 基準總共評估了十四個自監督式語音模型的表現，也成為隨後提出模型的比較基準 [13] [102]。

圖 2.3 為 SUPERB 基準統一使用的模型架構。在進行下游任務訓練時，語音基石模型 (左) 的參數會被固定，並另外添加參數量較少的下游模型 (右) 進行微調，以準確測量模型在預訓練階段習得的知識。由於語音基石模型最好的特徵並不一定在模型的最後一層，因此訓練時會將模型所有中間層的特徵，以一個可學習的加權和模組整合，以提升表現。

下游模型的選擇與下游任務有關：對於比較簡單的任務，如語者辨識、情緒辨識，會使用單層全連接層 (即常見的線性評估方法); 而對於比較困難的任務，如自動語音辨識、語者驗證、語者自動分段標記，則使用長短期記憶模型 (LSTM) 或 x-vector 等深度神經網路。

本論文在使用語音基石模型的實驗中，均使用圖 2.3 的下游任務通用架構，並將於 3.1 節討論各模型在 SUPERB 語者任務的結果。

2.2 階段性系統

2.2.1 簡介

傳統上，語者自動分段任務使用多個模組組成的階段性系統 [69][6]：先使用語音活性偵測找出語音中有人聲的片段；接著進行語者切割 (或語者切換點偵測)，將人聲片段切成多個只有一個語者說話的片段；隨後使用語者特徵向量抽取每個片段的語者特徵；最後以聚類決定語者數量及每個片段所屬的語者。圖 2.4 為傳統階段型系統的流程圖。

以下各小節將介紹階段性系統的每個模組：

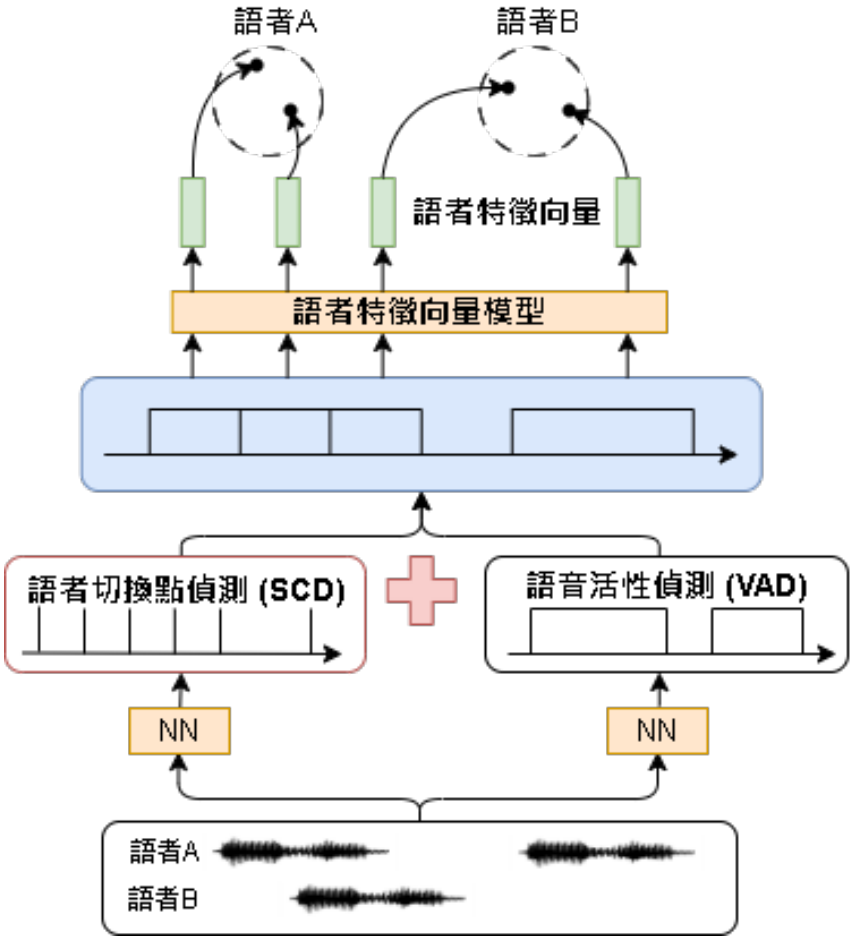

圖 2.4: 階段性系統示意圖

2.2.2 語音活性偵測 (VAD)

語音活性偵測的目標是要找出語音中有人聲的部分。在深度神經網路流行前，VAD 通常使用 MFCC 等聲音特徵加上簡單分類器或高斯混合模型；後來則逐漸被深度神經網路取代 [64] [86]。雖然 VAD 是整體系統最不可或缺的部分，但近年甚少文獻單獨提及。而部分文獻甚至會報告在採用 VAD 真實標記 (Oracle VAD) 下繼續完成其他模組的結果 [53]。

2.2.3 重疊語音偵測 (OSD)

重疊語音偵測的目標是要找出語音兩個或以上語者重疊的部分。傳統階段性系統假設每個語音片段只有一個語者，因此無法直接處理重疊語音。有部分文獻為階段性系統額外添加重疊語音偵測模組，透過後處理的方式整合。

常見重疊語音偵測後處理的方法有兩種。第一種方法是 Nearest-2 (見圖2.5a)，直接使用重疊部分附近的兩位語者的標記作為重疊部分的標記。此方法被證實在擁有完美重覆語音偵測標記的情況下能大幅提升表現 [67]，而且是一個難被輕鬆超越的基準方法 [5]。另一種方法是指定第二可能語者 (見圖2.5b)，直接使用聚類結果顯示第二可能的語者作為重疊部分的標記 [8] [98]

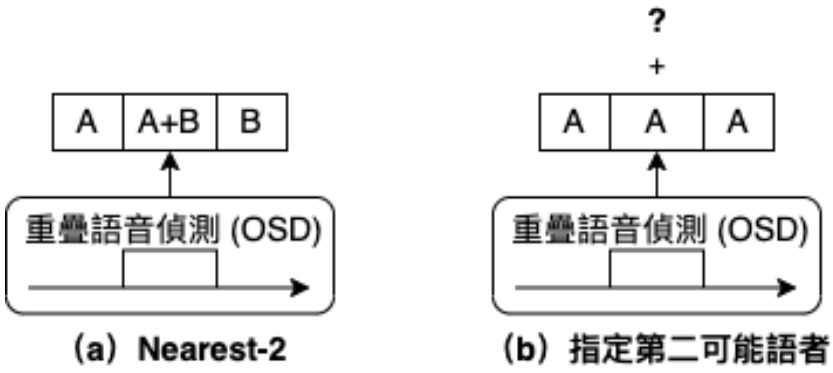

圖 2.5: 重疊語音偵測兩種後處理方法示意圖

隨著端到端系統的發展，重疊語音偵測被隱含在模型優化的目標中。然而，各系統最後的準確率仍高度依賴模型在重疊語音的表現，因此本論文將在 3.2 評估不同模型在重疊語音偵測上的結果。

2.2.4 語者切換點偵測 (SCD)

語者切換點偵測最初的目標是要找出兩位不同語者之間的語音切換邊界 [104]，但隨著重疊語音的偵測受到重視，其定義演變為包括每個語者發言的開始及結束 [39] [48]。換句話說，不僅偵測兩位不同語者之間的轉換，還偵測語者與無人聲之間，或是一位語者與多位語者之間的邊界。

目前常見的語者切換點偵測方法，可分成兩大類：

1. 基於時間之語者切換點偵測

   直接以二元分類目標預測語者切換點並分割人聲片段 [104] [105]。此方法比較直接，但卻意外地存有準確率不足的問題，至今很少被實際用於語者分段標記系統。文獻對此現象意見不一：如 [69] 推測片段長短的不一致降低了語者特徵向量的穩定性，從而影響表現；而 [99] 推測語者切換點任務需要文字資訊的輔助，並提出將語音資訊與文字基石模型特徵結合的方法；[42] 則認為進行語者切換點偵測時需考慮其人工標註錯誤，並提出具緩衝意識語者切換點偵測方法。

2. 基於文字之語者切換點偵測

   一般語音辨識系統的訓練目標只有文字，若能直接訓練於有語者切換標註的文字，則有兩個好處：模型在預測切換點時能直接考慮文字資訊，以及預測的切換點能直接與語音辨識的結果整合。有一系列研究 [112] [100] 使用基於自動語音辨識系統的方法並取得優秀表現，但目前此類方法仍未普及，可能

原因有二：其表現源自使用超過 10 萬小時內部資料集 [112]，且衡量語者切換點的標準遠較一般系統寬鬆，較難直接用於語者分段標記系統。

由於上述語者切換點偵測方法存有的缺陷，許多階段性系統仍以統一長度切割 (uniform segmentation) 的方式，直接將人聲部分等分為相同長度的片段，如 0.4 秒 [110]、2 秒 [29] 等，再將片段用於語者特徵向量抽取。此方法簡單並穩定，但卻存在準確率與時間精確度的平衡：使用太短的片段導致語者資訊不足，使用太長的片段則無可避免出現片段內超過一位語者的情況。後來 VBx [53] 使用長度 1.5 秒、跨步 0.25 秒的重疊窗口切割，緩解了上述問題；但其時間精確度仍有限制，且頻密的切割窗口會在語者特徵向量提取階段花費更長的運算時間。

綜上所述，語者切換點偵測的表現是目前階段性系統的瓶頸，因此本論文將在 3.3 節詳細討論其改進方法。

2.2.5 語者特徵向量提取

語者特徵向量提取的目標，是要從語音片段中提取可以通過相似度度量 (如餘弦相似度) 去量化語者相似程度的向量。以往，語音特徵向量多數使用基於高斯混合模型的 i-vector。隨著深度神經網路流行，陸續出現 d-vector、x-vector 的架構：先利用語者辨識任務訓練模型，並直接以模型輸出線性層前的隱藏層輸出作為語者特徵向量。目前最常見提取語者特徵向量的方法是基於時延神經網路（TDNN）的 x-vector [84]。它首先處理語音特徵的幀級信息，然後通過統計池化將這些特徵聚合到片段層級。隨後有許多基於 x-vector 的模型架構與損失改進方法。如 ECAPA-TDNN [21], CAM++ [93] 等。

2.2.6 語者聚類

常用的聚類演算法，如 K-means 聚類、譜聚類 (spectral clustering)、聚集層次聚類等 (agglomerative hierarchical clustering)，都常被用於語者特徵向量聚類。聚類衍生的問題主要有兩點：如何挑選聚類演算法，以及如何決定語者數量。

早期 K-means 是最普遍被使用的聚類演算法，後來有文獻 [94] [110] 比較不同聚類演算法的表現，發現譜聚類的表現明顯優於 K-Means。最近的文獻 [44] [71] 則更常使用聚集層次聚類，逐步把相似的向量合併成更大的群組。

然而，不同聚類演算法通常以不同準則決定語者數量。如 K-Means 常以手肘法 (elbow method) [63] 估算群數，但此方法已被 [80] 證實存有謬誤，常常出現難以判斷的情況。譜聚類常見以特徵值差值搜索（eigengap heuristic）[89] 決定群數，但存有運算複雜度太高及無法處理短音檔的問題 [95]。而目前最常被用到的聚集層次聚類，需要訂正停止準則，此超參數需要根據不同資料集作出調整 [69]，且傾向高估語者數量 [95]。此外，輪廓係數 (或輪廓分數 silhouette score) 最近被廣泛用在聚類表現分析 [82]，並應用到只基於語者特徵向量的語者自動分段標記系統 [50] [51]；但目前仍沒有文獻直接比較其表現，值得更深入的探討。

最後，聚類的表現常會受到異常值 (outlier) 的影響 [58]，在聚類前或聚類的過程中移除異常值通常可以改善表現。有文獻提出使用語者特徵向量的 l2-norm (大小) 去衡量特徵向量的品質 [49]，並使用二階段聚類的方式，先移除低品質資料點進行聚類，找出語者數量及群心，接著再推論其他資料點。不過由於特徵向量的 l2-norm 受片段長度影響，此方法需要額外使用資料集驗證集微調參數。

本論文將會在 3.4 節評估輪廓係數及不同聚類方法在各資料集上的表現，並探討更直接的方法去進行聚類異常值異除 (本論文稱為聚類純化)。

2.3 端到端系統

2.3.1 EEND

端到端系統的代表作 EEND [25] 在 2019 年由 Yusuke Fujita 提出。EEND 將語者自動分段標記視作多標籤分類問題：假設一段語音有兩個語者，則分別對應到模型的兩個輸出頻道 (output channel)，並以二元交叉熵損失 (binary cross entropy loss) 優化。由於語者跟輸出頻道可能存有不同的排列，因此訓練時需要應用置換不變訓練 (utterance-level permutation invariant training, 以下稱 u-PIT) [107]。在計算損失前，先列舉所有可能的排列 (排列的總數為 N! ，N= 輸出頻道數量)，逐一計算二元交叉熵損失，並取所有排列中的最小值作為最後損失。

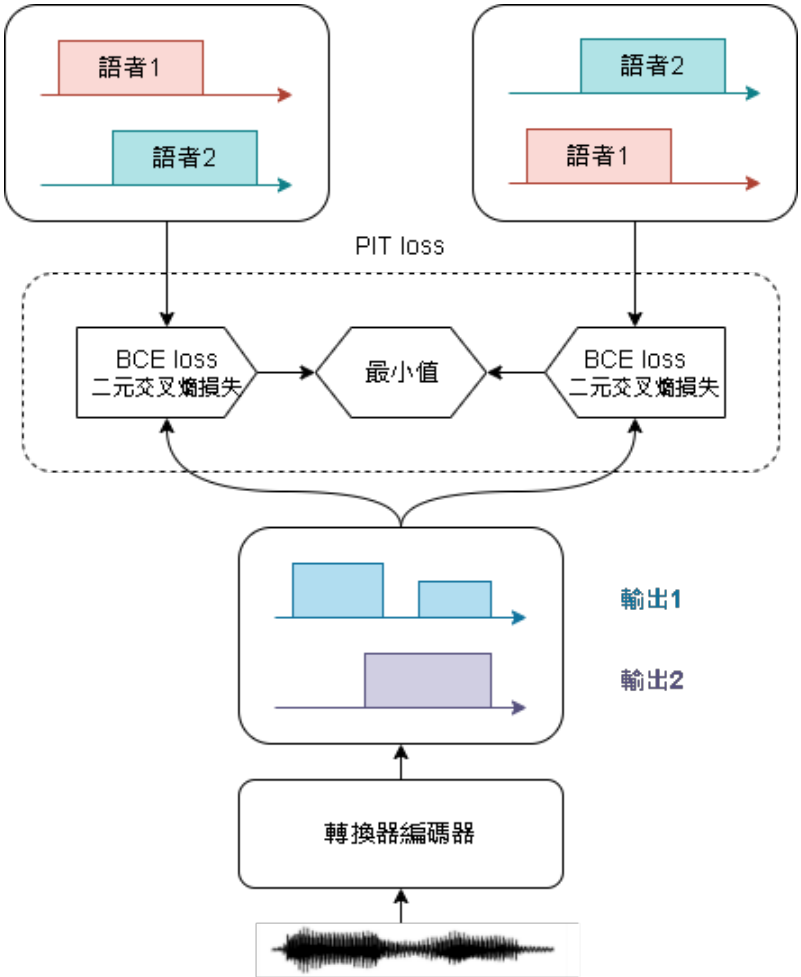

圖 2.6: 端到端系統 EEND 置換不變訓練 (u-PIT) 示意圖

圖 2.6 為 EEND (語者數 =2）使用 u-PIT 訓練的流程圖，其中左上及右上是正確標註的兩種排列方式。從圖中可看出輸出 1 及輸出 2 與左上的標註最接近，因此最後在 u-PIT 訓練時會取左上標註與輸出計算損失。

以下列出 u-PIT 公式

$$J_{\text{PIT}} = \frac{1}{TC} \min_{\phi \in \text{perm}(C)} \sum_{t} \text{BCE}(l_t^{\phi}, z_t)$$

其中 J_{PIT} 為置換不變訓練目標函數；T 為樣本序列長度；C 為語者數量；ϕ 為對於給定長度為 T 的樣本序列的所有可能排列的集合。$\text{BCE}(l_t^{\phi}, z_t)$ 為二元交叉熵損失函數，其中 l_t^{ϕ} 是模型輸出的第 t 個時間步對排列 ϕ 的預測結果；z_t 是實際標註。

EEND 使用雙向長短期記憶 (BiLSTM) 模型，並進行兩步驟訓練：先訓練在模擬不同重疊比例的生成音檔，再微調到真實資料集上。EEND 在模擬生成音檔上獲得很好的表現，但在真實資料集 CallHome [73] 上表現反而不及當時的階段性模型。其續作 SA-EEND [26] 及 Conformed-based EEND [62] 分別將雙向長短期記憶模型更換成更具表達力的轉換器編碼器及 Conformer 編碼器，並成功在真實資料集達到超越階段性模型的表現。

EEND 作者在論文提出以端到端模型取代階段性模型的兩個優點，包括能夠直接處理重疊語音，以及直接優化辨識錯誤。然而，論文在實驗中亦發現模型容易過擬合到資料的重疊語音比例。

2.3.2 EEND 變形

雖然端到端系統 EEND 成功在 CallHome 資料集達到超越階段性系統的表現，但是它仍存有不少限制。以下討論它的三個主要問題，以及目前文獻的相關改進方法。

1. 無法處理具有任何數量說話者的語音

   在使用 u-PIT 訓練時，輸出頻道的數目定義了模型最大可支援語者的數目。相比之下，階段性系統能在聚類階段靈活地決定語者數量。

   為解決這個問題，EEND-EDA [36][35] 提出基於編碼器-解碼器的架構及吸引子 (attractors) 的概念 (見圖 2.7)。代表輸出語者數量的吸引子會在解碼器一直被生成，直到吸引子存在的概率低於閾值 (threshold)。接著，將吸引子乘以 EEND 輸出的幀級語者特徵向量，以計算每個語者說話的區域。

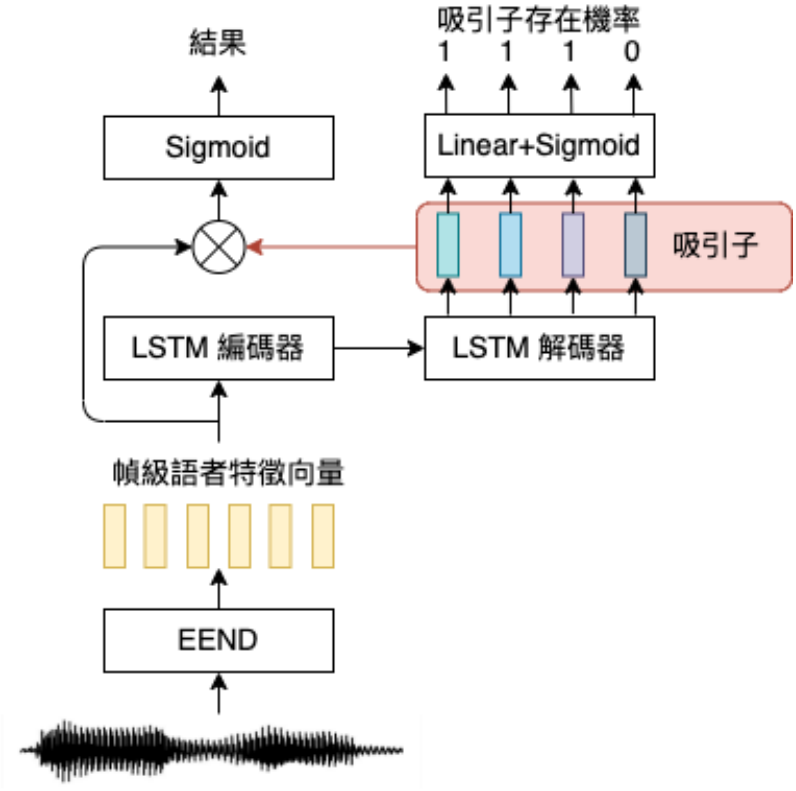

圖 2.7: 端到端系統 EEND-EDA 示意圖

後續，有不少文獻基於 EEND-EDA 提出進一步改善，如 EEND-NA [27]。然而，大部分文獻都只應用在模擬生成音檔及 CallHome 資料集。一直到近期，才逐漸有文獻 [7] [78] 將端到端模型應用到更多資料集上。

此外，雖然 EEND-EDA 部分解決了任意數量說話者的問題，但此類不使用無監督聚類的端到端系統，仍受其訓練資料限制，無法有效輸出比訓練資料更多的語者數量 [37]。

2. 無法處理過長的語音

現實會議的長度通常達半小時或更長，若直接全部輸入到模型，會出現速度變慢及記憶體不足的問題。為解決這個問題，[101] 在 SA-EEND 的基礎上提出先把長音檔切成小區塊 (chunk)，再以比較緩衝區的方式去整合全部區塊；而 BW-EDA-EEND [32] 則在 EEND-EDA 的基礎上，以因果編碼器 (causal transformer) 及區塊間循環 (recurrence) 的方式去提升速度及減少延遲。雖然以上改動會稍為降低模型準確率，但卻局部消除端到端系統用於現實系統的障礙。

3. 泛化能力不足

EEND 論文曾提及其可能存在的泛化能力問題，但後來甚少文獻討論。其中最接近的是本論文的同期論文 [28]，論文中使用適配器 (adapter) 作領域自適應訓練 (domain adaptation training)，以改善 EEND-EDA 模型泛化能力。論文使用多個資料集訓練的結果顯示，若在測試階段時沒有添加適配器提示測試集所屬之領域，準確率會大幅下降。此結果印證原本 EEND-EDA 存有泛化能力不足的問題。

雖然端到端系統仍存有上述問題，但其表現已被證實遠超階段性系統，亦為未來的重要研究方向。

2.4 端到端-階段性混合系統

2.4.1 EEND-VC

端到端-階段性混合系統的代表作是 EEND-VC [44] [43]，它提出在局部語音區塊使用端到端模型，並在全域使用語者聚類的方法。目標是在享有端到端系統處理重疊語音能力的同時，亦能透過全域語者聚類解決上一節提到端到端系統的問題。在訓練階段，EEND-VC 處理局部語音區塊的方法與 EEND 類似，同樣地使用 u-PIT 方法並優化二元交叉熵損失。它們有主要兩點差異：

首先，EEND-VC 假設局部語音區塊的語者數量有限，希望使用更少的輸出頻道解決無限語者的問題。在 EEND 的架構下，若語音共有五個語者 ($S_{global} = 5$)，則模型需要五個輸出頻道。而 EEND-VC 則限制局部區塊的最大語者數，如規定在 $T = 15s$ 的區塊內最多只有三個語者 ($S_{local} = 3$)

其次，EEND-VC 額外在模型添加全連接層，預測區塊中每個頻道的語者特徵向量，以供全域聚類時使用。在訓練時，模型會為每個輸出頻道的所有時間點預測幀級特徵向量 (frame-level embedding)，並按頻道語音活性加權平均成一個特徵向量 $\hat{\mathbf{e}}_s$，然後採用 [109] 提出的損失函數，拉開不同語者的特徵向量，同時拉近屬於同一語者的特徵向量。具體做法是定義一個可學習的全域語者特徵字典 E，其中 E_m 為每個全域語者的特徵向量 ($m = S_{global}$)。接著定義局部特徵向量與全域特徵向量的距離為 $d(\hat{\mathbf{e}}_s, E_m) = \alpha\|E_m - \hat{\mathbf{e}}_s\|^2 + \beta$，其中 $\alpha > 0, \beta$ 為可學習純量。最後每個區塊中的特徵向量損失函數為 $L_{\text{speaker}} = \frac{1}{S_{\text{Local}}} \sum_{s=1}^{S_{\text{Local}}} l_{\text{speaker}}(\sigma_s, \hat{e}^s)$，其中 $\sigma_i = [\sigma_{i,1}, \ldots, \sigma_{i,S_{\text{Local}}}]$ 是對應到 u-PIT 最小損失排列的語者索引，而 $l_{\text{speaker}}(\sigma_s, \hat{\mathbf{e}}_s) = -\ln\left(\frac{\exp(-d(E_{\sigma_s}, \hat{\mathbf{e}}_s))}{\sum_{m=1}^{M}\exp(-d(E_m, \hat{\mathbf{e}}_s))}\right)$ 是 [109] 提出的損失函數。最終的損失函數，是所有語音區塊 u-PIT 二元交叉熵損失及特徵向量損失函數的加總。

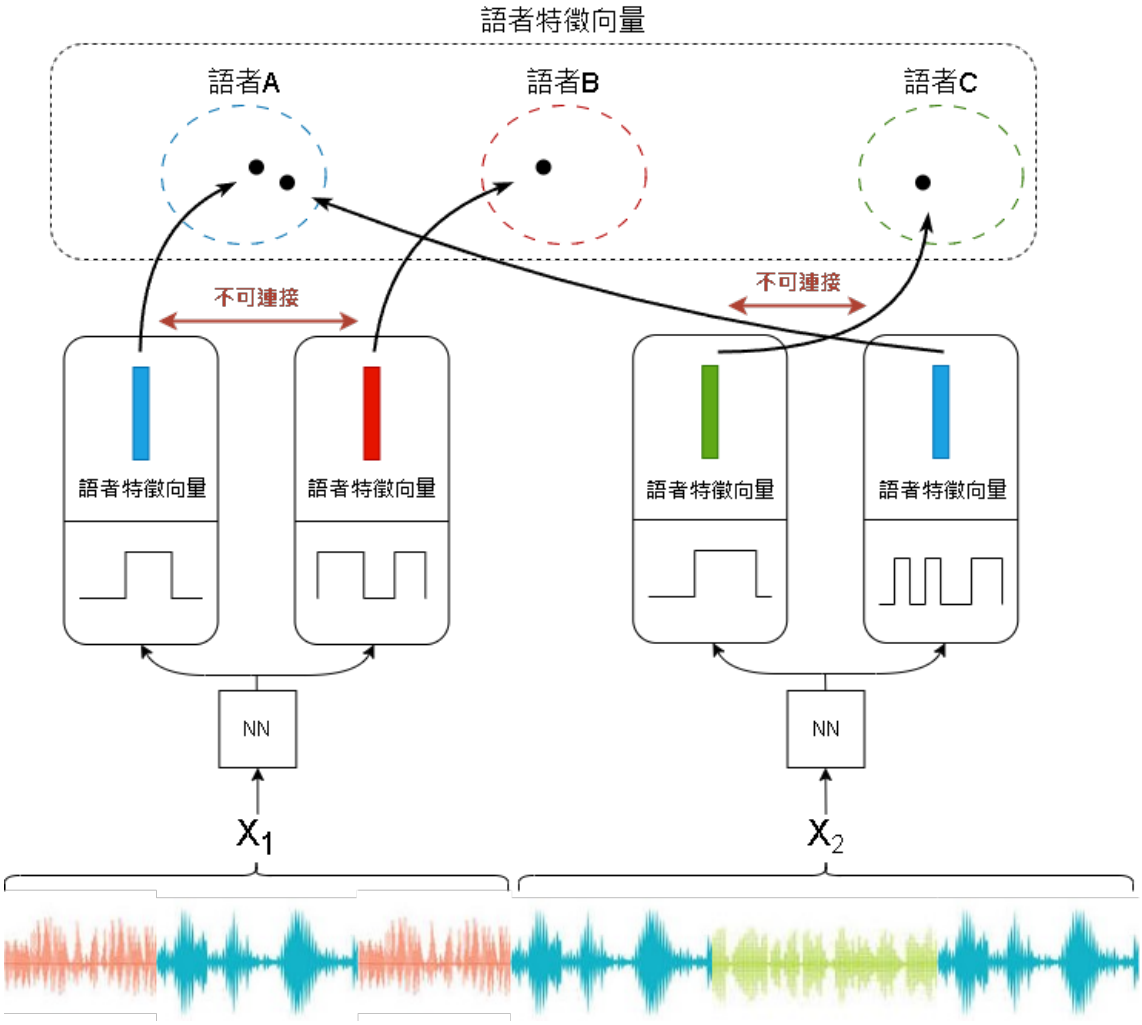

圖 2.8: 端到端-階段性混合系統 EEND-VC 示意圖

在推論階段，EEND-VC 在收集所有局部語音區塊的預測結果後，會對所有語者特徵向量進行約束聚類 (constrained clustering)，以決定全域語者數量 (S_{global})。其中，其聚類限制源自局部語音區塊的預測結果：在局部被分辨到不同語者的片段，在全域聚類時不應被分到同一語者。因此，聚類時會對局部區塊不同的語者特徵向量添加「不可連接」限制，以確保全域聚類的結果不違反局部預測的結果。圖 2.6 為 EEND-VC 的流程圖。

EEND-VC 分別測試了不同的約束聚類演算法，如 COP-KMeans [91]、約束譜聚類及約束聚集層次聚類，並發現約束聚集層次聚類的表現最好，且表現優於不使用約束聚類。最後，實驗結果顯示 EEND-VC 特別擅長處理長音檔以及語者數

量偏多的音檔，表現大幅優於 EEND 、其變形 EEND-EDA，以及傳統階段性系統。

後來，在基於 EEND-VC 混合系統的概念下，EEND-EDA 亦提出變形 EEND-GLA [38]，用無監督的方式對模型生成的吸引子聚類，並得出明顯優於 EEND-EDA 的表現 (尤其在語者數量偏多的情況)。從此以後，端到端-階段性混合系統成為熱門研究方向，後來的文獻 [81] 在比較各個系統亦得出 EEND-VC 性能最優的結論。

然而，應用端到端-階段性混合系統於現實系統仍有兩個問題待解決：

1. 局部語音區塊語者數量限制

   EEND-VC 方法中限制局部語音區塊的最大語者數量，例如在 50 秒的區塊只有三個語者說話 ($S_{local} = 3$)。然而此限制常在真實資料集中被打破 [9] [108]，10 秒的區塊內可能存有四個或以上語者說話。這個問題有兩個可能解法，其一是增加 S_{local} 的數量以滿足限制，然而 u-PIT 訓練在輸出頻道過多時存有問題 [85]，在實務中難以被運用；其二是縮短語音區塊的長度，如 [71] 只使用 5 秒的區塊 ($S_{local} = 3$)。但是，使用較短區塊時，會因片段邊界的上下文資訊不足而導致準確率有所下降 [43] [45]。

2. 泛化能力

   EEND-VC 在約束聚類中探討了兩種設定，一種是直接使用約束演算法添加限制；另一種是利用策略法則添加軟性限制 (soft constraint)，先以一般聚類決定全域語者數量 S_{global}，再計算特徵向量與全域語者群心 (cluster centroid) 的距離，按距離大小逐一將局部預測的語者分配到全域語者。EEND-VC 發現在部分聚類演算法下，約束演算法的表現沒有明顯優於策略法，顯示約束演算法可能存有泛化問題。然而，文中沒有進一步探討移除更多限制下的表

現，且後續文獻亦未有特別研究端到端-階段型混合系統的泛化表現。因此，本論文會在 4.2 節及 4.3 節，藉討論 EEND-VC 移除更多限制下的表現，探討混合系統的泛化能力。

2.4.2 Graph-PIT-EEND-VC

上一節提到，基於 u-PIT 方法的 EEND-VC 可能存有局部語音區塊語者數量限制，後來 Graph-PIT-EEND-VC [45] 嘗試以 Graph-PIT [90] 取代原有的 u-PIT，並在模擬生成音檔中取得明顯較好的表現。

其中，Graph-PIT 是置換不變訓練方法 u-PIT 的變形 (見圖 2.9)。在 u-PIT 方法下，局部語音區塊最多只能有 K 個語者，並分別對應到模型的輸出頻道 $(k = 1, 2, 3)$。Graph-PIT 沒有限制局部語音區塊的語者數量，而是限制在同一時間內，同時說話的語者數量 (以圖中例子，最多兩個語者同時說話)。

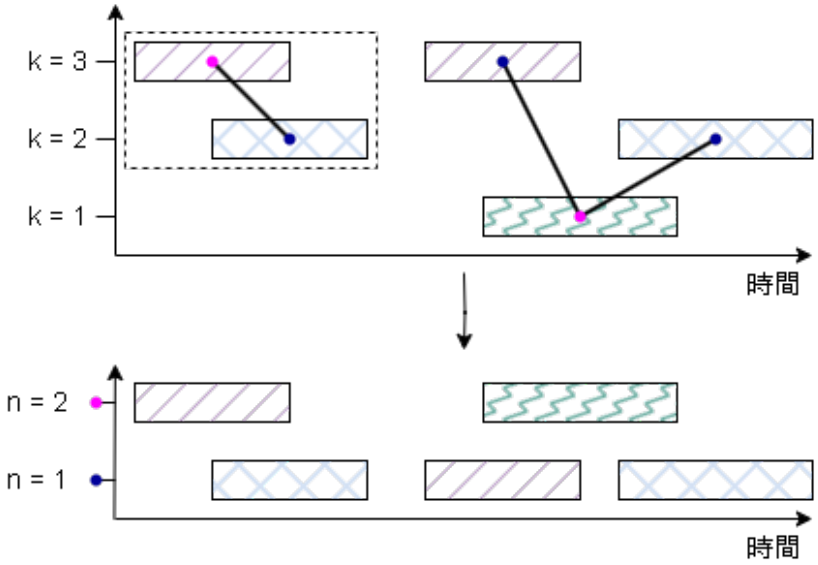

圖 2.9: 置換不變訓練方法 u-PIT 及其變形 Graph-PIT 示意圖

Graph-PIT 置換不變訓練方法基於有名的圖著色問題 (graph coloring problem)，設有圖 $G = (V, E)$，其中頂點 $V = \{1, \ldots, U\}$，U 為語者的每次發言；邊 $E = \{\{u, v\} : \forall u, v \in V, u \neq v$ 如s_u 及s_v時間上重疊$\}$。在 Graph-PIT 置換不變訓練

時，會解上述的圖著色問題並得出所有可能的排列，並逐一以模型輸出比較找出最小值，作為最後損失。

Graph-PIT-EEND-VC 方法以 Graph-PIT 取代 u-PIT，並應用在 EEND-VC。由於現實中甚少會有三個語者同時說話的情況 [113]，因此 Graph-PIT 只需使用兩個輸出頻道，並且可以使用更長的局部區塊長度 (理論上可無限延伸)。最後，Graph-PIT-EEND-VC 在模擬生成語音中，大幅領先受 u-PIT 限制影響，被迫使用較短局部片段長度的 EEND-VC。然而，在真實資料 CallHome 中，Graph-PIT-EEND-VC 的表現與 EEND-VC 大致相若，未能取得明顯優勢。此結果顯示 Graph-PIT 方法可能存有適應困難，或泛化能力不足的隱憂。因此，本論文將在 4.1 節嘗試解釋此現象並提出可能的解法。

總結以上對三種系統的介紹，端到端及端到端-階段性系統均是文獻研究的重點，但它們仍存在一些待解決問題。而歷史悠久的階段性系統因性能欠佳而逐漸不被關注，但本章討論亦顯示其仍有許多值得研究的部分。本論文會於第三章先討論階段性系統各模組的表現及可能改進方式，並在 3.5 節討論完整階段性系統的表現。隨後，在第四章討論端到端-階段性混合系統的泛化能力問題及可能改進方式，並在 4.4 節綜合比較三種系統的表現。在不同章節的討論中，亦會頻繁地使用本章一開始介紹的語音基石模型。

2.5 評估指標

$$\text{DER} = \frac{\text{SER (語者錯誤)} + \text{FA (誤報)} + \text{Miss (漏報)}}{\text{Total Speech (總說話長度)}}$$

語者自動分段標記任務常見的評估指標是 Diarization Error Rate (以下簡稱 DER)，公式如上。

DER 由三個部分組成，其中 SER (語者錯誤, speaker error rate）表示語音被錯誤歸屬給不正確語者的時間量，FA (誤報, false alarm) 表示非語音區域被錯誤歸屬給某位語者的時間量 (或單一語者語音區域被錯誤歸類為發現重疊語音的時間量)，Miss (漏報, miss detection) 表示語音沒有被歸屬給任何語者的時間量。Total 總說話長度則為所有語者說話時間的加總 (包括重疊的時間，因此總說話長度有可能長於語音長度)。

近年來亦有文獻提出 CDER [16]、JER [77]、BER [60] 等指標改善原有的 DER，但大部分文獻仍以報告 DER 為主，因此本論文只採用 DER 評估標準。

此外，DER 亦可分成完整、公正及寬容三種評估方法，統計於表 2.2。較早期的文獻通常採用公正評估方法，在每個語者切換點前後 0.25 秒劃出緩衝區，在評估時忽略緩衝區內的錯誤，以彌補資料集人工標註的誤差。近期的 DIHARD 挑戰及文獻 [77] [53] 則較傾向採用完整評估方法。因此本論文將主要採用完整評估方法，並在部分結果補充公正及寬容評估方法，以利與其他文獻比較。

評估方法	緩衝區大小	重疊語音評估
完整 (Full)	0s	是
公正 (Fair)	0.25s	是
寬容 (Forgiving)	0.25s	否

表 2.2: 比較 DER 常見的評估指標

2.6 資料集

表 2.3 列出本論文會使用的所有資料集。本論文挑選資料集的原則為只採用開源資料集，並儘量涵蓋不同來源及語言，以利適應真實世界的變化。

資料集名稱	語言	來源	訓練/驗證/測試時數	語者數	重疊語音佔比
Aishell4 [24]	中	會議	108/-/12	4-8	19%
AliMeeting [108]	中	會議	105/4/10	2-4	42%
RAMC [103]	中	聊天	150/10/21	2	0%
AMI-IHM [9]	英	會議	81/10/9	3-5	13%
AMI-SDM [9]	英	會議	80/10/9	3-5	13%
Voxconverse [17]	英	Youtube	20/-/44	1-21	4%
MSDWild [59]	多語言	Youtube	69/11/-	2-4	13%

表 2.3: 本論文使用的資料集概覽

由於不同文獻在使用資料集時會採用不同設定，因此以下逐一介紹本論文的設定：(括號內為本論文數據表格呈現時使用之縮寫)。

AISHELL-4 (Ai4) [24] 包括 120 小時 8 通道圓形麥克風陣列會議錄音，本論文跟隨 [71] 只使用通道一錄音。由於此資料沒有提供驗證集，因此本論文使用訓練集最後 20 則錄音作驗證用，並使用官方測試集。

AliMeeting (Ali) [108] 包括 120 小時會議錄音，分別使用近距離麥克風與八通道麥克風陣列錄製。本論文跟隨 [71] 只使用陣列的通道一錄音並跟隨官方訓練/驗證/測試集。

MagicData-RAMC (RAMC) [103] 跟隨官方訓練/驗證/測試集。

AMI-IHM / AMI-SDM (AMI-I / AMI-S)[9] 包括 100 小時的會議錄音，分別使用近距離與遠場麥克風進行錄製。其中 IDM (independent headset mix) 為近距離麥克風的混合，而 SDM (single distant microphone) 則只使用遠場麥克風陣列中第一個

的通道一，在辨識難度上明顯高於 IDM。本論文完全跟隨 [53] 在此資料集訂定的新設定。

Voxconverse (Vox) [17] 使用 Voxconverse (版本 0.3)。由於資料集只分成驗證集及測試集，因此本論文切分其驗證集作訓練及驗證，並以測試集作測試。有部分論文 [5] 切分的方式不一樣，因此不能直接比較表現。

MSDWild (MSD) [59] 分成 Few-Talker-Train、Many-Talker-Val 及 Few-Talker-Val，分別用作訓練集、驗證集及測試集。

最後，本論文在大部分實驗中均使用由 AISHELL-4、AliMeeting、AMI-IHM、Voxconverse、MSDWild 五個資料集組合而成約 360 小時的「複合訓練集」。資料集 AMI-SDM 及 MagicData-RAMC 則只於與其他文獻方法比較時，用於微調模型及測試。

第三章　階段性系統

3.1　語音基石模型用於 SUPERB 基準

3.1.1　簡介

本論文在 2.1 節介紹語音基石模型 HuBERT、Whisper，以及它們在不同下游任務的優秀表現。在用於語者自動分段系統相關任務前，本節將先使用 SUPERB 基準評估各模型於語者相關下游任務的表現，以基準化不同模型的能力。

3.1.2　實驗方法

模型名稱	模型參數	預訓練資料集
Sincnet [76]	*42K*	/
Whisper Tiny （Encoder）[74]	8M	680,000 小時多語言 ASR 數據 [74]
Whisper Base （Encoder）[74]	20M	680,000 小時多語言 ASR 數據 [74]
Distill Hubert [10]	23M	960 小時英文數據（LibriSpeech [68]）
Whisper Small （Encoder）[74]	87M	680,000 小時多語言 ASR 數據 [74]
Hubert Base [40]	90M	960 小時英文數據（LibriSpeech [68]）
WavLM Base+ [13]	90M	94,000 小時英文數據 [13]
Whisper Medium （Encoder）[74]	305M	680,000 小時多語言 ASR 數據 [74]
Hubert Large [40]	317M	60,000 小時英文數據（Libri-Light [41]）
Chinese Hubert Large [87]	317M	10,000 小時中文數據（WenetSpeech train L [111]）
WavLM Large [13]	317M	94,000 小時英文數據 [13]
Wav2vec2-xls-r-300m [18]	317M	436,000 小時多語言數據 [18]
Whisper Large v2 （Encoder）[74]	634M	680,000 小時多語言 ASR 數據 [74]

表 3.1: 本論文使用的語音基石模型，按參數量排列

表 3.1 列出本論文使用的十三個上游模型，並按參數量分成六組。

首先，在傳統模型的選擇上，本論文選用表中第一行的 Sincnet（非語音基石模型）。Sincnet 基於有可學習參數的帶通濾波器及卷積神經網路，並被廣泛用於語者辨識及先進的語者自動分段標記系統 [71]。結合了數位訊號處理和深度神經網路的獨特特性，使其具有出色的泛化能力。因此，本論文選擇以它作為傳統模型的代表。

接著，在語音基石模型的選擇上，主要選擇自監督式語音模型 HuBERT 及自動語音辨識模型 Whisper（只使用編碼器）。為驗證模型參數的因素，亦採用其不同大小版本，其中 Hubert 為 Distill Hubert, Hubert Base, Hubert Large；Whisper 為 Whisper Tiny / Base / Small / Medium / Large v2（Encoder）。為驗證模型訓練資料及語言的因素，亦採用 Chinese Hubert Large 及 Wav2vec2-xls-r-300m。最後，亦納入目前在 SUPERB 排行榜表現最佳的 WavLM Base+ / Large。

語音基石模型在進行下游任務訓練時，參數會固定，並提取各層的隱藏特徵作加權和，只有加權和及下游模型為可訓練參數。唯一的例外是 Sincnet：因為模型沒有經過預訓練，因此下游任務訓練時容許更新其上游模型參數。

3.1.3 實驗設定

表 3.2 列舉本節採用的 SUPERB 下游任務及其下游模型架構。其中語者辨識、語者驗證、語者自動分段標記均屬於語者相關任務。值得注意的是此處的語者自動分段標記任務，只使用 LibriMix 兩語者模擬生成音檔訓練而非真實資料，而其訓練方法則使用 EEND 中提及的 u-PIT。

此外，下游任務訓練的超參數調校完全依照基準規定。在十二個語音基石模

任務	下游模型架構
語者辨識（SID）	平均池化層 + 單層全連接層
語者驗證（ASV）	x-vector [84]
語者自動分段標記（SD）	單層長短期記憶模型
情緒辨識（ER）	平均池化層 + 單層全連接層

表 3.2: 本論文使用的 SUPERB 下游任務

型中，Distill Hubert / Hubert Base / Hubert Large 三個模型的結果直接取自 SUPERB 論文。另外，WavLM Base+ / WavLM Large 雖然其論文 [13] 有公佈在 SUPERB 的結果，然其採用較 SUPERB 基準寬鬆的超參數調校，因此為公平比較，本論文不直接採用其結果。

3.1.4 實驗結果分析

模型	SID Acc ↑	ASV EER ↓	SD DER ↓	ER Acc ↑
Sincnet [76]	19.80	11.44	10.67	48.9
Whisper Tiny （Encoder）[74]	46.60	7.80	5.75	64.30
Whisper Base （Encoder）[74]	61.28	7.61	4.96	66.37
Distill Hubert [10]	73.54	8.55	6.19	63.02
Whisper Small （Encoder）[74]	66.2	6.42	3.73	68.75
Hubert Base [40]	81.42	5.11	5.88	64.92
WavLM Base+ [13]	*86.01*	***4.50***	*3.63*	68.65
Whisper Medium （Encoder）[74]	81.92	5.83	3.40	70.72
Hubert Large [40]	90.33	5.98	5.75	67.6
Chinese Hubert Large [87]	95.64	5.07	3.95	67.2
WavLM Large [13]	***96.10***	*4.95*	*3.65*	70.62
Wav2vec2-xls-r-300m [18]	89.81	5.12	3.80	68.54
Whisper Large v2 （Encoder）[74]	83.67	6.01	**3.15**	**71.04**

表 3.3: 比較不同語音基石模型用於 SUPERB 下游任務之表現

各語音基石模型在 SUPERB 下游任務之表現如表 3.3 所示。以下就幾個方向分析結果：

1. 整體表現比較

首先，傳統模型 sincnet 在所有任務上的表現明顯劣於語音基石模型。其次，在比較 Hubert（以及其變形 WavLM）和 Whisper 編碼器兩種語音基石模型時，發現在語者辨識及語者驗證任務上，Hubert 稍微優於 Whisper；然而，在語者自動分段標記及情緒辨識方面，Whisper 的表現則略優於 Hubert。這與 [102] 的研究結果一致，他們發現 Whisper 編碼器在語義任務上表現良好，但在語者相關任務上表現較差。此外，這也驗證了其他文獻中的觀點，即語者自動分段標記 [99] 和情緒辨識任務 [57] 都假設模型需要有良好的語言能力。

2. 語者資訊線性可分性

比較相近參數的 Whisper Base 編碼器與 Distill Hubert 在語者辨識及語者驗證任務上的表現，發現在使用線性評估標準的語者辨識任務上，Distill Hubert 準確率比 Whisper Base 高出 12%；但在使用 x-vector 作下游模型的語者驗證任務上，Whisper Base 反而有較好的表現。此結果顯示，Whisper 編碼器在語者任務不一定比 Hubert 差，只是其語者資訊隱藏在模型深處。本論文推測兩個可能的原因：一是模型預訓練目標的不一致，二是線性評估標準的不穩定性 [55]。

3. 語言及訓練資料

儘管四個下游任務皆使用英文資料集，但對比 Hubert Large 和 Chinese Hubert Large 的表現後發現，使用 10,000 小時中文資料預訓練的 Chinese Hubert Large 明顯優於使用 60,000 小時英文資料預訓練的 Hubert Large。Chinese Hubert Large 的訓練資料來源包括 YouTube 和 Podcast，覆蓋各種錄製場景、背景噪音和說話方式；而 Hubert Large 則是使用有聲書作為訓練資料。本論文推測，與語言差異相比，預訓練資料來源的多樣性可能才是影響預訓練模型能力的最大關鍵。

3.2 語音基石模型用於語音活性偵測及重疊語音偵測

3.2.1 簡介

語音活性偵測（下稱 VAD）及重疊語音偵測（下稱 OSD）均為語者自動分段標記系統最基礎的部分。本節將以多個開源資料集，基準化各語音基石模型在這兩個下游任務的域內 (In-Domain) 及域外 (Out-of-Domain) 表現。

3.2.2 實驗方法

VAD 和 OSD 都是簡單的二元分類問題，可以透過二元交叉熵損失函數進行優化。雖然文獻中有提到將兩者合併成三類別分類問題以提升 OSD 的準確率 [98]，但為了獨立評估它們的表現，本節使用多任務訓練的方式，在同一個深度神經網路中同時預測 VAD 和 OSD。

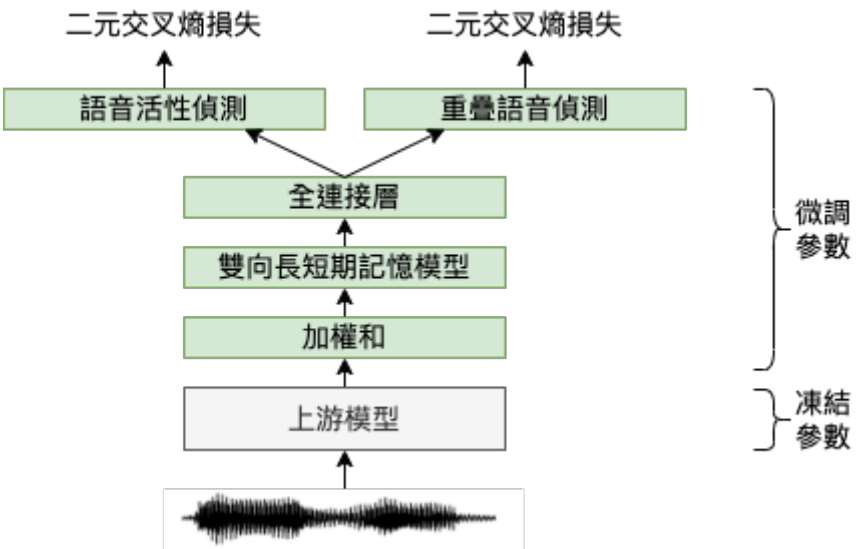

圖 3.1: 語音活性偵測及重疊語音偵測模型架構

本節採用了 SUPERB 基準中將語音基石模型應用到下游任務的方法，並選擇了單層雙向長短期記憶模型（BiLSTM）作為下游模型。在這種架構中，BiLSTM

的輸出通過單層全連接層後，再經過兩個全連接層作為輸出頭（prediction head），分別用於 VAD 和 OSD 的輸出。圖 3.1 展示了下游模型的結構。

在 VAD 任務中，常見的評估方法包括誤報率 (FA)、漏報率 (Miss)，以及在不同偵測閾值下的 ROC 曲線下面積 [5] [23]。針對語者自動分段系統，VAD 帶來的錯誤可透過錯誤率（= FA + Miss）進行評估 [5]。在 OSD 任務中，常見的評估方法是精確度和召回率的調和平均數，即 F1-Score [106]。在評估 VAD 及 OSD 的域內表現時，會從 0.25 到 0.75 的閾值範圍內，每 0.05 步進行搜索，並找出對應的最佳閾值所得的分數；而在評估域外表現時，則統一報告 0.50 閾值下的分數。

3.2.3 實驗設定

在訓練階段，模型會隨機從複合訓練集中抽取長度為 7 秒的語音區塊進行訓練。而在測試階段，測試集中的每個音檔會被切分成 5 秒長的區塊，並在每個區塊的左右各延伸 1 秒作為上下文，以提升模型的表現。

至於超參數設置，出於效率和一致性的考量，本論文大多數實驗均使用相同的超參數設定。未特別提及的情況下，後續章節中的所有實驗都將遵循以下超參數設置或搜索範圍：

- 學習率（learning rate）：1.0e-3, 1.0e-4
- 批次大小（batch size）：32
- 優化器（optimizer）：AdamW
- 下游模型隱藏層大小：128（包括加權和輸出、BiLSTM 及全連接層）
- 訓練步數：200,000
- 驗證步數：10,000

此外，為了強化模型的穩健性和泛化能力，本論文後續所有實驗在訓練時都採用了基於背景嘈音和房間聲學衝激響應（room impulse response）的資料增強

（data augmentation）。背景嘈音部分使用 musan 資料集 [83]，並以 0.25 的機率添加嘈音（信噪比範圍在 0 至 15 之間），同時以 0.25 的機率添加音樂（信噪比範圍在 5 至 15 之間）。另外，房間聲學衝激響應則使用 Room Impulse Response and Noise Database [46]，並以 0.25 的機率進行添加。

3.2.4 實驗結果分析

VAD 域內表現	Ai4	Ali	AMI-I	MSD	Vox	**平均 ↓**
Sincnet	4.0	3.3	6.7	7.5	6.0	5.5
Hubert Base	4.2	3.4	6.0	6.1	5.2	5.0
Distil Hubert	3.8	3.1	6.0	6.2	5.3	4.9
Whisper Tiny Enc	3.9	3.1	5.8	5.8	5.1	4.8
Hubert Large	3.8	3.3	5.6	6.0	4.7	4.7
Wavlm Base Plus	4.0	3.2	5.6	5.9	4.7	4.6
Whisper Base Enc	3.9	3.2	5.6	5.5	5.0	4.6
Wav2vec2-xls-r-300m	3.7	3.0	5.6	5.7	**4.5**	4.5
Wavlm Large	3.7	**2.8**	5.2	5.7	4.8	4.5
Whisper Small Enc	3.7	2.9	5.2	5.5	5.0	4.4
Whisper Medium Enc	**3.6**	3.0	**5.1**	**5.3**	4.8	4.4
Hubert Large Chinese	3.7	2.9	**5.1**	5.5	**4.5**	**4.3**

表 3.4: 比較不同模型用於語音活性偵測下游任務之表現，按在五個資料集的平均表現從最低到最高排列; 評估指標為錯誤率（= FA + Miss）

VAD 域外表現	Ai4	Ali	AMI-I	MSD	Vox	**平均 ↓ (域內-域外) ↓**
域內（In-Domain）						
Sincnet	4.0	3.3	6.7	7.5	6.0	5.5
Miss/FA	1.5/2.5	1.0/2.3	2.8/3.9	2.7/4.8	1.8/4.1	
Whisper Tiny Enc	3.9	3.1	5.8	5.8	5.1	4.8
Miss/FA	1.2/2.7	1.1/2.1	2.9/2.9	2.2/3.6	1.6/3.5	
Hubert Chinese Large	3.7	2.9	5.1	5.5	4.5	4.3
Miss/FA	1.7/1.9	1.6/1.3	2.7/4.1	2.1/3.4	1.8/2.7	
域外（Out-of-Domain）						
Sincnet	8.2	4.4	13.2	9.0	6.7	8.3 (-2.8)
Miss/FA	6.7/1.5	1.5/2.9	10.5/2.7	1.3/7.7	1.2/5.5	
Whisper Tiny Enc	7.6	4.3	13.0	5.9	6.1	7.5 **(-2.5)**
Miss/FA	6.6/1.0	1.8/2.5	11.7/1.3	3.7/2.2	2.6/3.5	
Hubert Chinese Large	7.2	4.1	12.5	5.7	5.4	6.9 (-2.6)
Miss/FA	5.8/1.3	1.9/2.2	10.6/1.8	3.5/2.2	2.8/4.1	

表 3.5: 比較不同模型在語者活性偵測下游任務上，域內情景與域外情景之表現差異; 評估指標為 F1-Score

首先討論 VAD 任務的結果 (域內：表 3.4，域外：表 3.5)。域內表現顯示，在相同參數下，Whisper 系列模型稍微優於 Hubert 系列模型，而 Sincnet 的表現最差。然而，在域外表現方面 (以域外與域內的差值衡量)，不同模型的差距並不大。這可能顯示 VAD 是一個相對較簡單的任務，容易取得合理的域外泛化能力。

OSD 域內表現	Ai4	Ali	AMI-I	MSD	Vox	**平均 ↑**
Sincnet	53.3	74.2	69.6	55.4	51.9	60.9
Distil Hubert	58.1	79.9	77.6	65.6	60.5	68.3
Hubert Base	59.9	80.0	78.2	66.0	60.8	69.0
Whisper Tiny Enc	63.3	81.1	77.1	69.2	64.9	71.1
Hubert Large	62.5	82.0	80.3	68.2	62.8	71.2
Wavlm Base Plus	64.3	82.7	81.2	70.8	62.8	72.4
Whisper Base Enc	65.5	82.6	79.4	71.4	64.5	72.7
Wav2vec2-xls-r-300m	62.4	85.6	82.4	72.8	63.8	73.4
Wavlm Large	66.8	84.9	**82.8**	72.8	63.3	74.1
Whisper Small Enc	67.8	84.5	81.5	74.4	63.2	74.3
Hubert Large Chinese	**68.4**	**86.3**	82.7	74.7	63.2	75.1
Whisper Medium Enc	**68.4**	85.3	81.8	**75.3**	**65.0**	**75.2**
SOTA	58.0 [106]	81.6 [106]	80.4 [54] 79.2 [48]		57.5 [106]	

表 3.6: 比較不同模型用於重疊語音偵測下游任務之表現，按在五個資料集的平均表現從最低到最高排列; 評估指標為 F1-Score

OSD 域外表現	Ai4	Ali	AMI-I	MSD	Vox	**平均 ↑ (域外-域內) ↑**
域內 (In-Domain)						
Sincnet	53.3	74.2	69.6	55.4	51.9	60.9
Whisper Tiny Enc	63.3	81.1	77.1	69.2	64.9	71.1
Hubert Chinese Large	68.4	86.3	82.7	74.7	63.2	**75.1**
域外 (Out-of-Domain)						
Sincnet	47.5	69.0	46.1	48.8	46.6	51.6 (-9.3)
Whisper Tiny Enc	61.4	77.0	63.9	63.5	63.4	65.8 (-5.3)
Hubert Chinese Large	66.7	83.3	68.8	70.1	61.7	**70.5 (-4.6)**

表 3.7: 比較不同模型在重疊語音偵測下游任務上，域內情景與域外情景之表現差異; 評估指標為 F1-Score

接下來討論 OSD 任務的結果 (域內：表 3.6，域外：表 3.7)。在域內表現上，Whisper 系列模型的表現優於 Hubert 系列模型，而 Sincnet 的表現則最差。至於域外表現，Chinese Hubert Large 及 Whisper Tiny 編碼器明顯優於傳統模型 Sincnet。

OSD 是明顯比 VAD 困難的任務，此結果顯示在較具挑戰的任務上，大型語音基石模型具有更強的泛化能力。此外，在所有資料集上報告的結果均優於最先進的技術。值得注意的是，雖然 [54] 和 [48] 使用了語音基石模型 Wavlm 和 Wav2vec2，但僅提取模型最後一層特徵進行下游任務訓練，因此效果稍差。

最後，討論不同資料集在泛化能力上的差異。在 VAD 和 OSD 任務中，不同模型在 AMI-I 資料集上的域外表現明顯低於域內表現。其中在 VAD 任務，錯誤率從約 6% 大幅提高至約 13%，變差了約一倍。在 OSD 任務，F1 Score 也下降了約 15% - 20% (絕對數值)。而在 Alimeeting，MSDWild，Voxconverse 等其他資料集上，模型的表現差異較小，VAD 和 OSD 任務的表現下降僅約 1% 和 5%。這種現象可能有兩種解釋：一是 AMI-I 資料集的難度較高，導致模型泛化能力下降；二是 AMI-I 資料集的標註方式與一般資料集有顯著不同，將許多背景語音也標記為說話，以及有很多錯誤標註的部分 [47]，導致了語音標記上的模糊不清。而根據本論文的觀點，更傾向支持第二種解釋。

總括而言，語音基石模型改善了傳統模型在語音活性偵測任務上的表現，而在重疊語音偵測任務上，其表現及泛化能力的提升更為顯著。由於 VAD 及 OSD 的錯誤率會直接反映在所有語者自動分段系統的錯誤率 DER 上，因此選擇合適的語音基石模型是重要的方向。

3.3 語音基石模型用於語者切換點偵測

3.3.1 簡介

語音模型在語者切換點偵測方面表現不佳，因此階段性系統通常仍採用統一長度切割或重疊窗口切割的方式切割人聲片段 [53]。若能直接使用語者切換點進行分割，不僅能提升切割邊界的精準度，還能大幅降低提取語者特徵向量的運算

負擔。因此，本節將討論改進語者切換點偵測的方法。

3.3.2 具緩衝區意識之語者切換點偵測

語者切換點偵測的目標是找出語音中所有發言的開始點及結束點，通常被視為二元分類問題。然而，由於語者切換點的稀疏性，此任務存有類別不平衡的問題。過去的研究嘗試解決這個問題，主要是透過增加正標籤數量的方法。例如，有些文獻像是 [104] 將切換點周圍約 0.1 到 0.2 秒的區域都標記為正標籤；而 [48] 則以切換點為中心，在正負 0.2 秒間線性下降至零，同時使用均方誤差損失進行優化。

不過，近期有研究 [42] 提出了不同的觀點，認為語者切換點的問題可能源自標註的不準確性，因此提出一個類似連結時序分類損失 (CTC) 的目標函數，使得在緩衝區範圍內，存在且僅存在一個語者切換點。(見圖 3.2c)

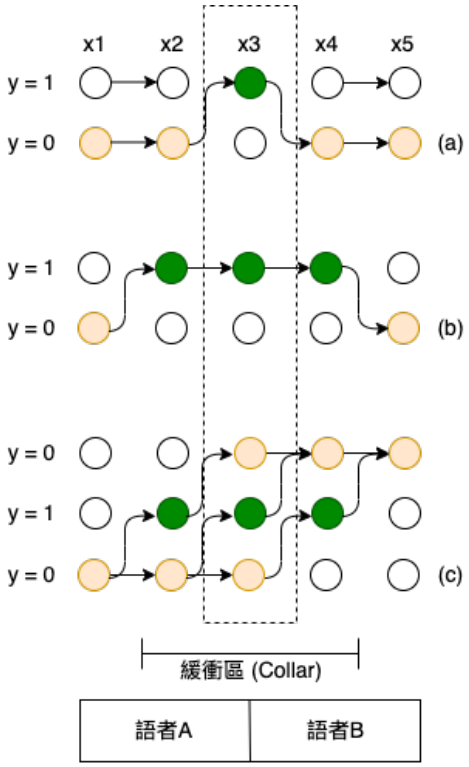

圖 3.2: 在語者切換點任務上，(a) 無緩衝區 (b) 增加正標籤數量 (c) 具緩衝區意識三種不同切換方式之示意圖

設 $\hat{y}$ 是模型輸出，Z 是所有真實語者切換點的集合，則一般的二元交叉熵損失為 $L(\hat{y}, Z) = -\sum_{i=1}^{N} (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$。對每個語者切換點 $z_i \in Z$ 而言，緩衝區的定義為 $C_i = \{x | z_i - c < x < z_i + c\}$，其中 $c > 0$ 為緩衝區大小。接著，定義超集合 $S(Z) = \{\{z_1, ..., z_N\} | z_i \in C_i \forall i \in \{1, ..., N\}\}$。假設真實標記序列為 $[0, 0, 1, 0, 0]$ 且 $c = 2$，則超集合 $S(Z) = \{[0, 1, 0, 0, 0], [0, 0, 1, 0, 0], [0, 0, 0, 1, 0]\}$，亦即在緩衝區範圍內，只存在且僅存在一個語者切換點的所有序列的集合。最後，具緩衝意識之二元交叉熵損失為所有語者切換點序列二元交叉熵的加總：

$$L_{\text{collar}}(\hat{y}, Z) = -\log \sum_{Z_0 \in S(Z)} e^{-L(\hat{y}, Z_0)}$$

實際使用這個目標函數進行訓練時，可能會遇到無法收斂的問題，這是由於真實資料集中的語者切換點有時出現得比較頻繁。當兩個語者切換點之間的距離少於緩衝區大小的兩倍時，如模型預測的切換點位置落在兩個緩衝區之間，按照上述目標函數定義，其損失會成為負值，引導模型往錯誤方向優化。為了解決這個問題，需要對目標函數的定義進行簡單的修改。先設 Z 為時間上有序集合，並修改緩衝區定義為 $C_i = \{x | max(z_i - c, \frac{z_{i-1} + z_i}{2}) < x < min(z_i + c, \frac{z_i + z_{i+1}}{2})\}$，以確保緩衝區沒有重疊。

雖然修改後目標函數後模型可以收斂，但在切換點密集的情況，緩衝區大小會變小，導致模型學習困難。有其他文獻 [48] 會在訓練時，忽略同一語者者兩次發言之間的短暫間隙（少於一秒），以防模型變得過於敏感。然而，這做法無法處理兩語者緊接著說話的情況，且去除部分切換點會造成模型學習上的混亂。

本節提出簡單的解決方法：將語者切換點分成話語開始點及話語結束點。在正常對話下，通常不會頻密地出現話語開始點或話語結束點，因此這做法可以大幅減緩緩衝區重疊的問題 (參見圖 3.3)。

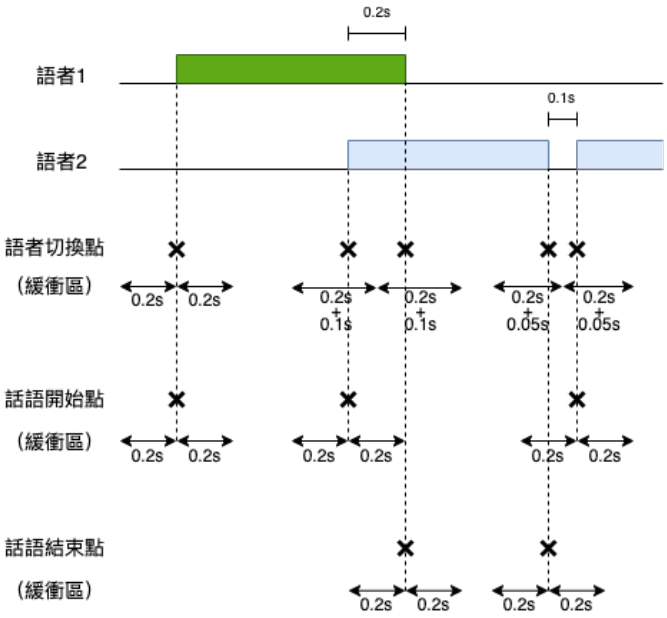

圖 3.3: 具緩衝區意識之語者切換點，話語開始點及話語結束點偵測示意圖

3.3.3 實驗方法

為驗證上一節提及的方法，將進行兩個對照實驗，一是比較使用無緩衝區、增加正標籤數量、具緩衝意識三種偵測方法的表現，二是比較使用話語開始＋結束點取代語者切換點的表現。本節沿用 3.2 節多任務訓練的方式，同時進行語音活性偵測、重疊語音偵測、話語開始點偵測及話語結束點偵測下游任務。模型架構如圖 3.4。

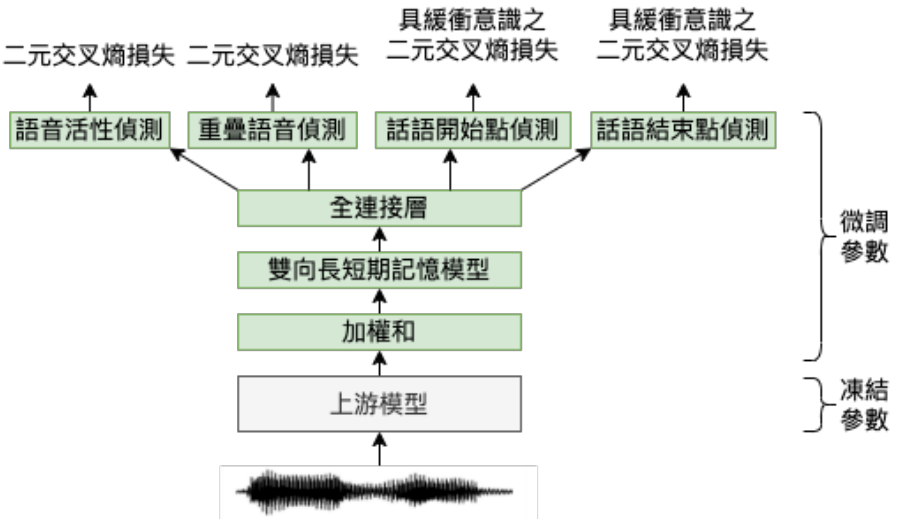

圖 3.4: 語者切換點偵測模型架構

語者切換點任務常見的指標包括 F1-Score 及 Purity-Coverage F1 [4]。由於語者切換點的稀少性，在使用傳統的 F1-Score 衡量時，通常會訂定一個誤差容許範圍。如容許範圍為 0.2 秒，只要預測的語者切換點在真實切換點的前後 0.2 秒內即算作正確。其中，預測切換點及標註點之間是一對一的關係。為避免發生不同文獻使用不同容許範圍的問題，[4] 提出不需訂定誤差容許範圍的 Purity 及 Coverage 評估方法，以及它們的調和平均數 Purity-Coverage F1。

本論文在進行初步實驗時，發現使用 Purity-Coverage F1 評估標準與最終系統 DER 的關聯性遠比 F1-Score 小，因此仍採用 F1-Score 作主要評估標準 (誤差容許範圍 0.2 秒)，並在本節的最後補充使用 Purity-Coverage F1 評估的結果。

3.3.4 實驗結果分析

	Ai4	Ali	AMI-I	MSD	Vox	**平均 ↑ （域內-域外）↑**
Sincnet						
域內（In-Domain）						
（a）無緩衝區	44.3	46.5	49.2	47.3	41.8	45.8
（b）增加正標籤數量（+-200ms）	55.9	52.3	60.2	59.2	47.2	54.9
（c）具緩衝區意識（+-200ms）	56.3	58.2	58.5	58.7	48.0	**55.9**
域外（Out-of-Domain）						
（a）無緩衝區	11.2	24.4	27.6	24.5	6.5	18.8（-27.0）
（b）增加正標籤數量（+-200ms）	34.7	33.7	40.9	50.8	16.8	35.4（-19.5）
（c）具緩衝區意識（+-200ms）	53.1	39.1	47.0	54.0	24.7	**43.6 (-12.3)**
Whisper Tiny Enc						
域內（In-Domain）						
（a）無緩衝區	54.4	60.8	68.2	71.4	61.2	63.2
（b）增加正標籤數量（+-200ms）	57.3	61.9	69.7	72.8	65.0	**65.4**
（c）具緩衝區意識（+-200ms）	57.1	62.4	68.8	73.1	64.5	65.2
域外（Out-of-Domain）						
（a）無緩衝區	32.0	41.8	52.2	50.9	53.6	46.1（-17.1）
（b）增加正標籤數量（+-200ms）	48.8	60.7	55.4	62.3	57.9	57.0（-8.4）
（c）具緩衝區意識（+-200ms）	55.3	61.7	56.8	67.6	62.4	**60.8 (-4.4)**

表 3.8: 比較於話語開始點偵測下游任務中，訓練時使用（a）無緩衝區（b）增加正標籤數量（c）具緩衝區意識，共三種不同偵測方法，不同模型在域內情景與域外情景之表現差異; 評估指標為 F1-score（容許偏差 +-200ms）

首先，表 3.8 比較了在話語開始點偵測任務上，三種不同切換方式的表現。在域內表現上，增加正標籤數量與具緩衝區意識方法的表現，意外地大致相若，但皆優於不使用任何緩衝區。在域外表現上，三個方法的差距明顯變大：具緩衝區意識方法明顯優於增加正標籤數量，而無緩衝區更次之。此結果顯示，增加正標籤數量雖然改善了不平衡分類的問題，但並沒有考慮到不同資料間標註的差異；相比之下，具緩衝區意識方法直接考慮到標註的模糊性，因此大幅提升模型的域外泛化能力。

	Ai4	Ali	AMI-I	MSD	Vox	**平均↑**	**（域內-域外）↑**
Sincnet							
域內（In-Domain）							
（a）語者切換點	41.2	52.7	56.2	52.7	41.8	48.9	
（b）話語開始＋結束點	50.1	54.8	56.4	54.2	43.4	**52.0**	
（開始點/結束點）	56.0/44.3	57.6/50.1	59.5/50.1	58.1/50.3	47.5/39.6	55.7/46.9	
域外（Out-of-Domain）							
（a）語者切換點	25.5	46.9	40.1	42.8	14.9	34.0	（-14.9）
（b）話語開始＋結束點	31.2	48.0	38.5	44.5	22.7	**37.0**	（-15.0）
（開始點/結束點）	39.0/20.1	51.9/42.3	48.2/27.6	53.0/36.3	28.3/16.9	44.1/28.6	（**-11.6**/-18.3）
Whisper Tiny Enc							
域內（In-Domain）							
（a）語者切換點	44.0	61.8	64.0	67.6	58.4	59.2	
（b）話語開始＋結束點	53.7	61.3	64.6	67.3	58.7	**61.1**	
（開始點/結束點）	59.5/48.6	64.3/55.5	68.7/58.6	71.9/63.2	63.5/54.8	65.6/56.1	
域外（Out-of-Domain）							
（a）語者切換點	36.4	55.0	51.9	55.8	51.0	50.0	（-9.2）
（b）話語開始＋結束點	42.6	57.3	51.8	59.1	53.2	**52.8**	（-8.3）
（開始點/結束點）	55.0/29.3	61.9/51.4	57.3/42.2	67.0/51.3	61.8/46.0	60.6/44.0	（**-5.0**/-12.1）

表 3.9: 比較於語者切換點下游任務中，訓練時使用（a）語者切換點訓練目標（b）話語開始點及結束點訓練目標，不同模型在域內情景與域外情景之表現差異; 評估指標為 F1-score（容許偏差 +-200ms）

接下來，表 3.9 比較了使用話語開始點＋話語結束點取代語者切換點的表現。在域內表現上，話語開始＋結束點的表現已明顯優於使用語者切換點，兩者相距約 2-4%。另外，話語開始點偵測的表現大幅領先話語結束點，兩者相距高達 9-10%。在域外表現上，上述的表現差異更被進一步放大。此結果印證了前面提到，語者切換點緩衝區重覆所導致的準確率問題，並帶來一個重要的發現：模型更擅長於預測話語開始點。此現象在未來值得更深入的探討。

另外，表 3.10 補充了使用 Coverage 及 Purity 評估標準的表現，結果與前兩項

實驗一致，且在同樣使用舊版 AMI 資料集的設定下，優於 [48] 報告的表現。

	Cov ↑	Pur ↑	F1 ↑
Whisper Tiny Enc			
話語開始 + 結束點（具緩衝區意識）	92.94	90.54	91.73
話語開始 + 結束點（增加正標籤數量）	93.02	90.59	91.77
語者切換點 （增加正標籤數量）	91.56	89.82	90.70
Wav2vec2-xls-r-53			
話語開始 + 結束點（具緩衝區意識）	93.36	91.08	**92.21**
話語開始 + 結束點（增加正標籤數量）	93.24	90.88	92.04
語者切換點 （增加正標籤數量）	92.22	90.43	91.40
語者切換點 [48]	91.93	90.59	91.26

表 3.10: 比較於語者切換點下游任務中，不同模型及方法在 AMI （pyannote version）[70] 資料集使用評估指標 Coverage、Purity 及 Coverage Purity F1-Score 的表現

	Ai4	Ali	AMI-I	MSD	Vox	**平均** ↑
Sincnet	56.3	58.2	58.5	58.7	48.0	55.9
Distil Hubert	60.0	62.0	66.1	64.9	54.1	61.4
Hubert Base	60.9	63.6	67.6	66.2	57.3	63.1
Hubert Large	62.5	64.1	71.0	69.3	60.8	64.2
Whisper Tiny Enc	57.1	62.4	68.8	73.1	64.5	65.2
Wavlm Base Plus	62.5	66.3	70.8	70.8	60.1	66.1
Wav2vec2-xls-r-300m	62.2	69.0	73.2	74.1	63.7	68.4
Whisper Base Enc	60.8	67.3	71.2	75.5	67.7	68.5
Wavlm Large	64.9	68.4	**75.7**	74.8	64.4	69.6
Hubert Large Chinese	**66.5**	69.4	73.9	75.2	64.9	70.0
Whisper Small Enc	63.0	69.0	74.6	77.5	**68.2**	70.5
Whisper Medium Enc	63.6	**69.7**	75.6	**78.3**	68.1	71.1

表 3.11: 比較不同模型用於話語開始點偵測下游任務之表現，按在五個資料集的平均表現從最低到最高排列；評估指標為 F1-Score （容許偏差 +-200ms）

最後，表 3.11 補充了不同模型用於話語開始點偵測的基準。結果顯示 Whisper 系列模型明顯優於 Hubert 系列模型，且領先的幅度更甚於上一節的語音活性偵測及重疊語音偵測基準。因此可合理推測，在各個語者自動分段標記系統相關下游任務上，話語開始點偵測最需要文字資訊的輔助，其次是重疊語音偵測，而語音活性偵測的需求則更少。

3.4 語者特徵向量提取及聚類

3.4.1 簡介

語者特徵向量聚類是階段性系統及端到端-階段性混合系統 (下稱混合系統) 的最後一個部分，決定最終的語者數量，對整體結果有極大的影響。然而，在文獻中，其表現很少被單獨衡量。通常，對於不同聚類方法的比較，都是基於在聚類之前的其他特定系統和方法。因此，在本節中，我們直接使用各資料集的真實標註，將語者特徵向量提取和聚類視為獨立討論的議題。

3.4.2 聚類純化

在討論如何使用真實標註評估表現之前，本小節先探討聚類純化的概念，以建立後續討論的基礎。

在進行聚類前先找出資料點的異常值，通常可以改進聚類的表現 [58]。在語者自動分段標記任務中，[49] 發現語者特徵向量的大小 (l2-norm) 與品質有關聯，因此提出二階段聚類：先移除低品質的片段進行聚類，然後將這些被移除的片段與首次聚類的群心計算相似度，並歸屬到最接近的語者中。然而，本論文在進行初步實驗時，發現使用 l2-norm 過濾異常值的效果並不穩定，因為其在不同的語者特徵向量和資料集上表現不一致。相反，本節提出使用兩個極簡單 (embarrassingly easy) 但效果穩定的準則：

1. 移除短句

   語者特徵向量在短句上，通常因語者資訊不足而表現不佳 [72] [14]。雖然後來有不少方法 [34] 嘗試改進在短句的表現，但在 2 秒以下片段提取的語者

特徵向量，其分佈仍異於長句。然而，現時的主流系統 [43] 均沒有對短句作特別處理，其中，VBx [53] 更直接統一採用 1.5 秒短片段切割。

2. 移除重疊

語者特徵向量在訓練時一般沒有特別考慮重疊語者的情況，一直到近期才有文獻提出專門處理重疊語音的特徵向量 [19]。因此，從重疊語音提取的語者特徵向量，可合理地被視作異常值。然而，文獻對於移除重疊對系統表現的影響沒有一致的看法 [22]，且現時主流系統 [53] [43] 在聚類時皆不會移除重疊部分。

本節將進行實驗比較聚類前先移除重疊部分及不同長度短句對結果的影響。

3.4.3 真實標註評估方法

接下來本小節討論如何使用真實標註進行評估。由於語者特徵向量聚類是階段性系統及混合系統共同擁有的部分，因此本節的實驗會主要分成兩種模擬情景：第一種是「端到端-階段性混合系統模擬」，用以模擬完美混合系統 + 不完美語者聚類的大致表現；第二種是「階段性系統模擬」，用以模擬完美階段性系統 + 不完美語者聚類可達到的表現。本節除了希望透過這兩種實驗情境評估語者聚類的表現外，也期望能間接比較階段性系統和混合系統的極限表現。

由於這兩種模擬情景分別在不同的程度上利用了資料集的真實標註，以下定義它們使用標註的方式：(可參考下頁圖 3.5 的例子)

語者自動分段標記任務中，真實標註的形式為有序集合 G：

$$G = \{(t_{\text{開始}_1}, t_{\text{結束}_1}) : \text{語者}_1, (t_{\text{開始}_2}, t_{\text{結束}_2}) : \text{語者}_2, \ldots, (t_{\text{開始}_n}, t_{\text{結束}_n}) : \text{語者}_n\}$$

其中，$t_{\text{開始}_i}$ 和 $t_{\text{結束}_i}$ 分別代表第 i 段標註的開始和結束時間，語者$_i$ 是在這段時間

內的語者。

在混合系統模擬中，本節參考的方法是 2.4.2 曾介紹的 Graph-PIT-EEND-VC，其模型的預測目標包含重疊語音不能被歸類到同一輸出頻道的「不可連接限制」。根據此方法，可修改真實標註的定義為 $Hybrid = \{(t_{開始_1}, t_{結束_1}), \ldots, (t_{開始_n}, t_{結束_n})\}$ (移除了語者標註)，並添加不可連接限制：

$\forall i, j \quad$ 如果 $\quad (1 \leq i < j \leq n) \quad$ 且 $\quad (t_{開始_i} < t_{結束_j}) \wedge (t_{結束_i} > t_{開始_j}) \quad$ 則 $\quad (語者_i \neq 語者_j)$

值得注意的是，在一般的混合系統中，模型會同時預測語者特徵向量並用於語者聚類。本節改以外部語者特徵向量模型提取特徵向量，因此表現可能稍有落差。(註: 本論文在第 4.2 節將討論以外部語者特徵向量取代混合系統模型預測的特徵向量，並驗證了本節實驗並不會低估混合系統的表現)

最後，在階段性系統模擬情景，真實標註定義為語音活性偵測、重疊語音偵測、話語開始點偵測三個任務的目標。其中，話語開始點會被用於進一步分割語音活性偵測的標註。

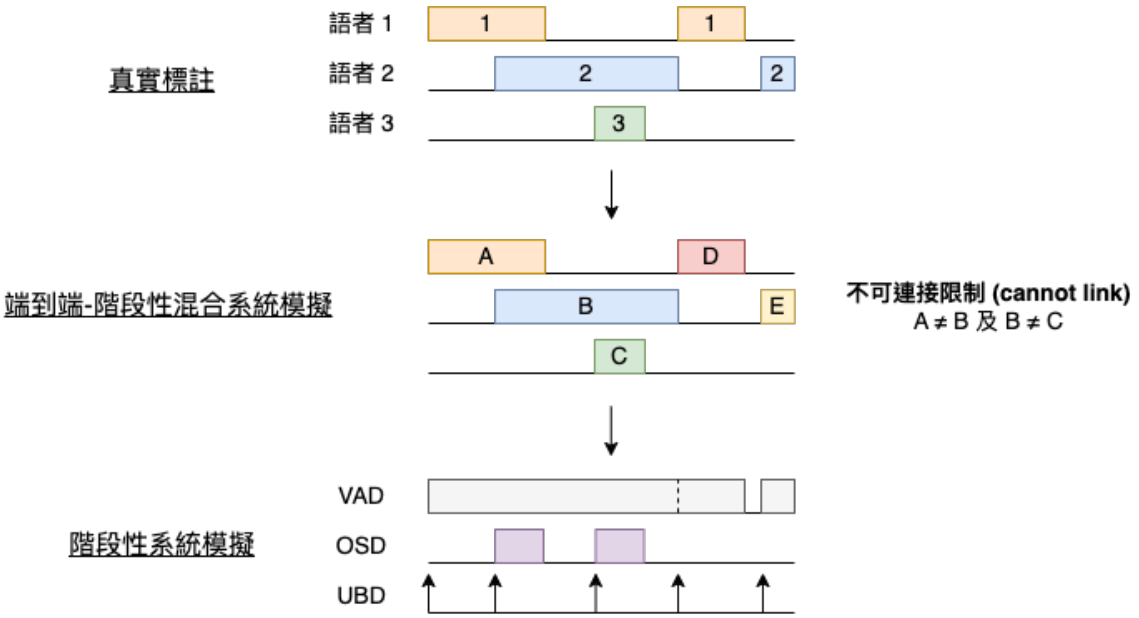

圖 3.5: 從真實標註到擁有部分標註的「端到端-階段性混合系統模擬」及「階段性系統模擬」示意圖

3.4.4 實驗方法

上一小節定義了兩種模擬情景使用真實標註的方法。接下來，本節將闡述如何進一步利用真實標註進行語者特徵向量提取及聚類。值得注意的是，以下描述的過程，將會在後續章節討論完整系統時繼續沿用。圖 3.6 為以下三點過程的綜合示意圖。

1. 聚類純化

   首先，將真實標註的每個片段分成非重疊部分及重疊部分，其中非重疊部分會實際用於語者特徵向量提取與聚類，重疊部分則成為非重疊部分的「附屬片段」，其語者標記會由非重疊部分決定。若片段只有重疊部分，或片段的非重疊部分短於閾值 (見圖中特例)，則會將整個片段當作「重疊片段」，在最後階段才處理。處理重疊語音後，會將短於某個閾值的片段標為「短句」，長於的片段標為「正常片段」。其中，片段的長度只包含非重疊部分。

2. 語者特徵向量提取

   接著，使用語者特徵向量模型提取所有正常片段、重疊片段及短句的語者特徵向量。本論文採用基於 x-vector 方法的 CAM++ 開源模型 [93] [92] 進行實驗，因為它在效率和準確率方面表現出色。

3. 二階段聚類

   最後，透過對語者特徵向量聚類，找出每個片段的語者標記。聚類分成兩個階段。第一階段是一般的無監督聚類，僅使用「正常片段」的特徵向量進行，以確保聚類過程不受異常值的影響。第二階段則利用了第一階段聚類中每個語者的群心，將所有「重疊片段」和「短句」的特徵向量分配到距離最近的語者群中。

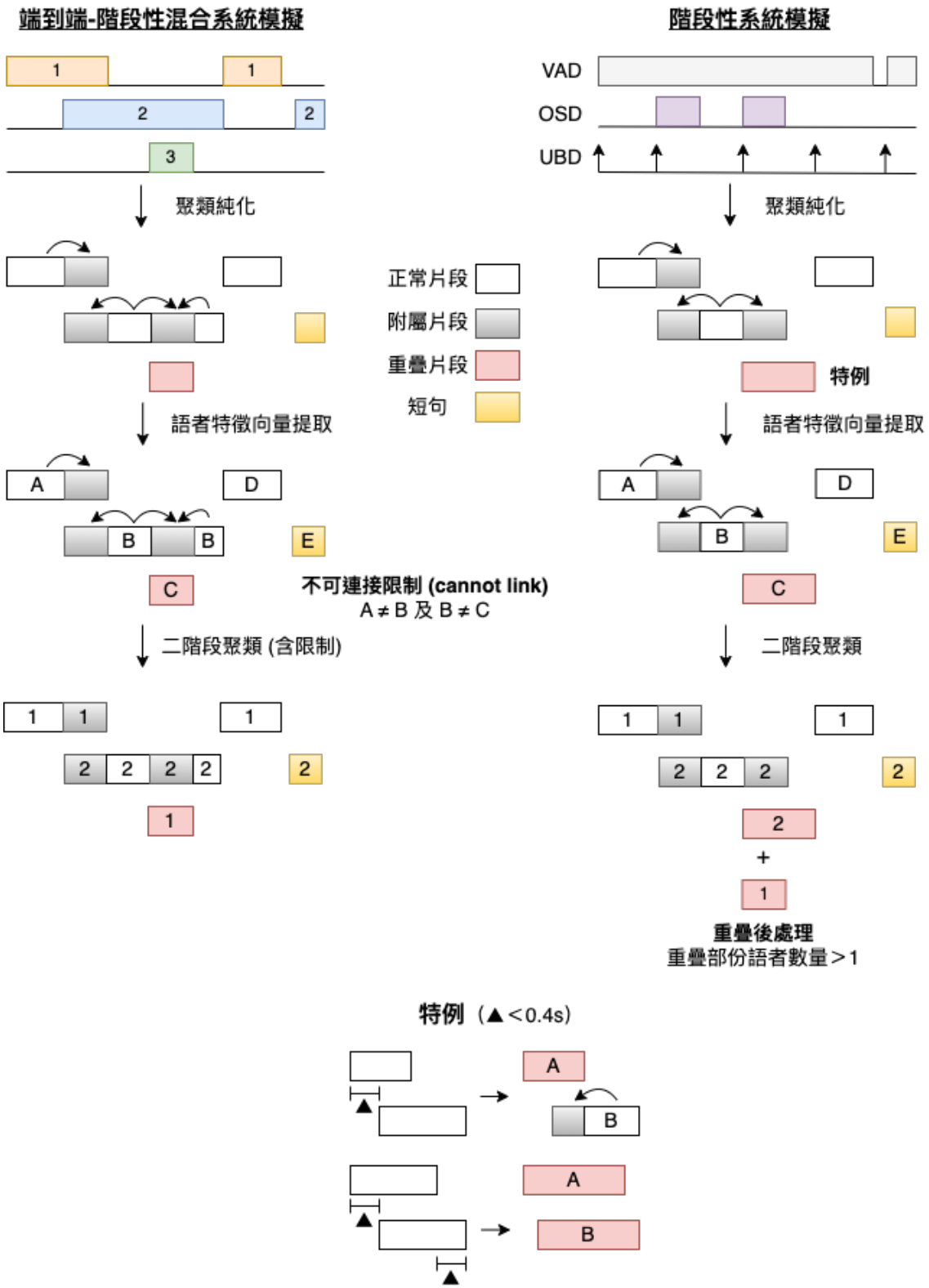

圖 3.6: 端到端-階段性混合系統模型及階段性系統模擬之過程示意圖

以上描述的過程同時適用於混合系統及階段性系統，以下討論混合系統獨有的「不可連接限制」，以及可小幅度改善階段性系統表現的「重疊後處理」。

根據 3.4.3 小節的定義，混合系統的「不可連接限制」定義為時間上重疊的片段不能被分配到相同的語者。有兩種添加限制的方式：一是約束聚類，二是使用策略法。以約束層次聚類為例，在聚類之前先計算所有特徵向量的相似度，然後將不可連接的特徵向量對的相似度手動調整到極大的數值，再進行一般層次聚類，就能確保所有不可連接的片段不會被歸類到同一語者。而策略法則是按照片段長度由大到小排列，並逐一指定語者。如出現違反限制的情況，則指定為下一可能的語者。在某些情況下，策略法無法滿足不可連接的限制條件。

在階段性系統中，假設重疊語音偵測的預測準確，則重疊部份應該要指定兩個語者。當重疊片段只有被指定一位語者時，需要進行重疊後處理。若該重疊片段的語者特徵向量已經被提取，那麼直接將第二接近的語者加入預測中。若重疊片段的語者特徵向量尚未提取，則在提取語者特徵向量後，將其分配給除了原先預測的語者以外最接近的語者。

3.4.5 聚類演算法

在語者聚類問題上，有兩個主要變因：語者數量決定及聚類演算法，並有許多可能的組合。此處將語者聚類視為兩個步驟，先討論決定語者數量的方法，再討論在既定語者數量下，不同演算法的表現。

在決定語者數量方面，目前文獻常見做法是訂定一個閾值 (停止準則)，決定兩個群是否屬於同一個語者 [71]。如兩個群的距離小於閾值，則它們會被合併。然而此停止準則需要根據不同資料集訂定，影響系統泛化能力。另一種做法是使用輪廓分數評估聚類結果的好壞，並取在不同語者群數下，輪廓分數的極大值作

為最終語者數量。本節實驗將比較在各資料集上，使用停止準則及輪廓分數決定語者數目的表現。其中，停止準則使用兩個群心的餘弦距離，而在輪廓分數上，則採餘弦距離及歐氏距離準則預測語者數量，並選擇其中較大的值作為最終預測。在挑選演算法方面，本節實驗將比較三種常見聚類演算法 (K-Means，譜聚類及聚集層次聚類) 在各資料集上的表現。

3.4.6 實驗結果分析

	Ai4	Ali	AMI-I	AMI-S	MSD	Vox	平均 ↓
端到端-階段性混合系統模擬	5.3	18.1	9.4	12.3	15.1	4.6	10.8
＋移除重疊	4.5	12.0	7.4	8.7	13.4	4.1	8.4
＋不可連接限制（約束聚類）	4.2	9.6	4.7	7.3	10.2	3.9	6.7
＋不可連接限制（策略法）	4.0	7.7	5.0	6.6	10.0	3.7	**6.2**
階段性系統模擬	5.5	20.4	10.4	13.5	15.5	4.8	11.7
＋移除重疊	4.9	15.2	9.3	10.4	12.8	4.3	9.5
＋重疊後處理	4.7	13.5	8.6	9.8	10.2	4.0	**8.5**

表 3.12: 比較端到端-階段性混合系統模擬及階段性系統模擬，在不同方法下的表現變化（均排除 3 秒以下短句）；評估指標為 DER （完整）

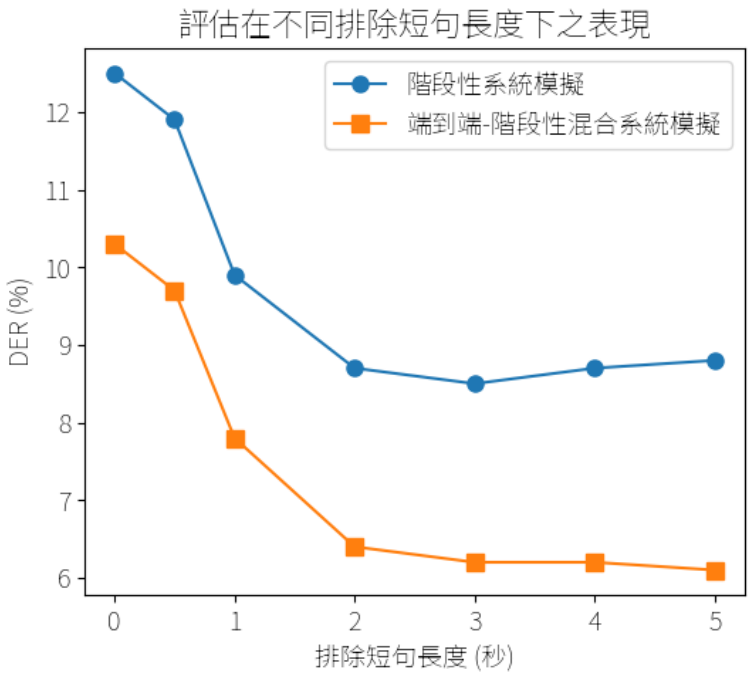

圖 3.7: 比較端到端-階段性混合系統模擬及階段性系統模擬，在不同排除短句長度下的表現變化；評估指標為 DER （完整）

表 3.12 及圖 3.7 為本節進行「端到端-階段性混合系統模擬」及「階段性系統模擬」的主要實驗結果。以下針對結果作出分析。

首先，在兩個系統的比較上，混合系統的表現優於階段性系統。其中，混合系統模擬的最好 DER 為 6.2%，優於階段性系統模擬約 2%。這個結果顯示混合系統有潛力表現更好，不過它與階段性系統的差距並不算很大。

其次，進行聚類純化 (移除重疊語音及短句) 可以大幅改善系統表現。移除重疊後，兩個系統的準確率都提高了約 2%，而排除短句長度則帶來了接近 4% 的提升，驗證了這種方法的有效性。然而，值得留意的是，排除短句後的系統表現，直到排除秒數達到 3 秒才到達最佳水平。這表明短於 2 秒的語音片段都會影響聚類表現。不過，若排除短句的長度過長，可能會導致聚類資料點不足，或某些語者幾乎沒有長語音片段，進而降低辨識準確度。

另外，階段性系統中，重疊後處理提升約 1% 表現；在混合系統中，不可連接限制帶來約 2% 的提升，且使用策略法施加限制的表現稍優於更嚴格的約束聚類。這兩種方法都以語者特徵向量和群心距離作為辨識重疊部分的標記依據，因此暗示了特徵向量在某種程度上，具有能夠同時識別兩個語者身份的潛力。而在重疊的語音中，通常只有其中一位語者的聲音佔主導地位，使用約束聚類嚴格施加限制，有時候反而會導致聚類結果變差。

為了深入研究兩種模擬中辨識錯誤的來源，表 3.13 進一步列出了使用公正及寬容 DER 指標的表現。首先，比較公正指標與寬容指標的表現。在寬容指標下，兩個系統的表現僅相差 0.3%；然而在公正指標下，這個差距增加到了 1.7%。這兩種指標主要區別在於是否針對重疊語音進行評估，因此可以推斷出混合系統相對於階段性系統的優勢幾乎完全在於對重疊語音的處理上。特別是在 Alimeeting 和 AMI 資料集上，這兩者的差異尤其明顯。這兩個資料集經常出現三個語者同

時說話的情況，而這種情況在現實生活中極為罕見，除非是在夾雜笑聲等情況下[79]。其次，寬容指標下的錯誤率主要來自語者數量辨識錯誤，在實驗中這種情況不常見，但只要出現一次就會明顯提高錯誤率。最後，完整指標與公正指標的表現差異則主要源自短句 (及重疊短句) 被歸類到錯誤的語者。

	Ai4	Ali	AMI-I	AMI-S	MSD	Vox	平均 ↓
端到端-階段性混合系統模擬							
完整	4.0	7.7	5.0	6.6	10.0	3.7	6.2
公正	2.9	7.3	3.7	5.0	8.5	3.1	5.1
寬容	1.3	1.8	0.5	1.3	6.1	2.3	2.2
階段性系統模擬							
完整	4.7	13.5	8.6	9.8	10.2	4.0	8.5 (+2.3)
公正	3.5	11.1	6.6	7.9	8.5	3.3	6.8 (+1.7)
寬容	1.4	1.8	0.8	2.0	6.4	2.4	2.5 (+0.3)

表 3.13: 比較端到端-階段性混合系統模擬及階段性系統模擬，在完整、公正及寬容 DER 評估指標下的表現

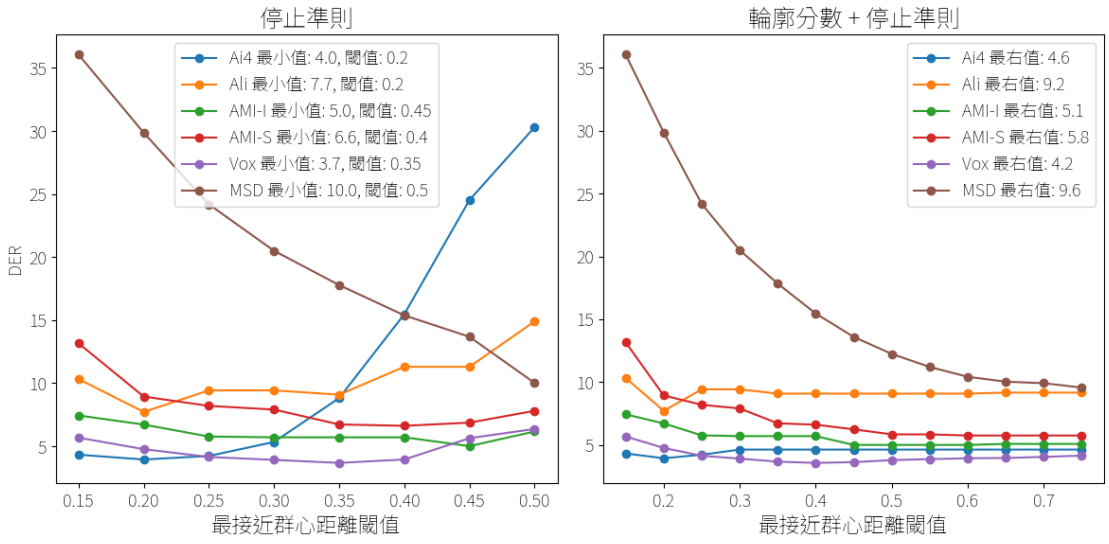

圖 3.8: 比較在聚集層次聚類中，使用停止準則及輪廓分數訂定語者數量的表現 (端到端-階段性系統模擬)；評估指標為 DER（完整）

這一節的另一目標是評估聚類演算法的表現。圖 3.8 對比了使用停止準則和輪廓分數來決定語者數量的結果。從左圖可以看出，在僅使用停止準則時，不同資料集需要不同的閾值才能達到最佳效果，且沒有一個能夠適用於所有資料集的通用閾值。右圖展示了結合輪廓分數和停止準則的結果，顯示使用輪廓分數能夠

在所有資料集上獲得不錯的表現。(註：右圖中，隨著閾值增加，聚類結果受到輪廓分數的影響越大，最右邊的數值代表僅使用輪廓分數的結果)

	Ai4	Ali	AMI-I	AMI-S	MSD	Vox	平均↓
K-Means 聚類	5.4	13.7	**8.6**	9.5	**10.2**	4.4	8.6
譜聚類	5.3	13.6	**8.6**	**9.2**	10.4	5.3	8.7
聚集層次聚類	**4.7**	**13.5**	**8.6**	9.8	**10.2**	**4.0**	**8.5**

表 3.14: 比較不同聚類演算法在階段性系統模擬的表現；評估指標為 DER（完整）

最後，表 3.14 比較了在使用輪廓分數決定語者數量後，不同聚類演算法的表現。結果顯示，三個演算法在模擬情景下，其平均表現皆在誤差範圍內，因此可得出聚類表現主要受決定語者數量的方法影響，而非聚類演算法。由於聚集層次聚類的速度最快，K-Means 聚類次之，而譜聚類最慢，所以本論文剩下的實驗，都將採用聚集層次聚類結合輪廓分數。

3.5 語音基石模型用於階段性系統

3.5.1 簡介

本章之前已經討論了階段性系統的各個主要部分。這一節將綜合各部分之改進並補充細節，最後在實驗中探討其整體表現。圖 3.9 為本章提出，經改進後的完整階段性系統的示意圖。

3.5.2 實驗方法

本章主要從三個方向改進階段性系統：

首先，在模型架構方面，使用了語音基石模型替代傳統模型，並且利用多任務訓練同時處理語音活性偵測、重疊語音偵測和話語開始點偵測。其次，在語者

分割方法方面，採用了具有緩衝區意識的話語開始點，取代了原有的語者切換點或統一長度分割方式。最後，在語者特徵向量聚類方面，採用了二階段聚類的方法。先移除重疊和短句進行聚類，然後使用聚類推論方式將移除的片段重新加回，並進行重疊後處理。在這個過程中，聚類採用了輪廓係數來決定語者的數量。整體來說，這個流程結合了多項技術，以改進階段性系統的性能。

本節實驗將比較不同模型在域內和域外情景的表現，並進行切除研究，分析每項改進對最終表現的影響。此外，亦會在 4.4 節與其他系統一同進行運算成本分析。

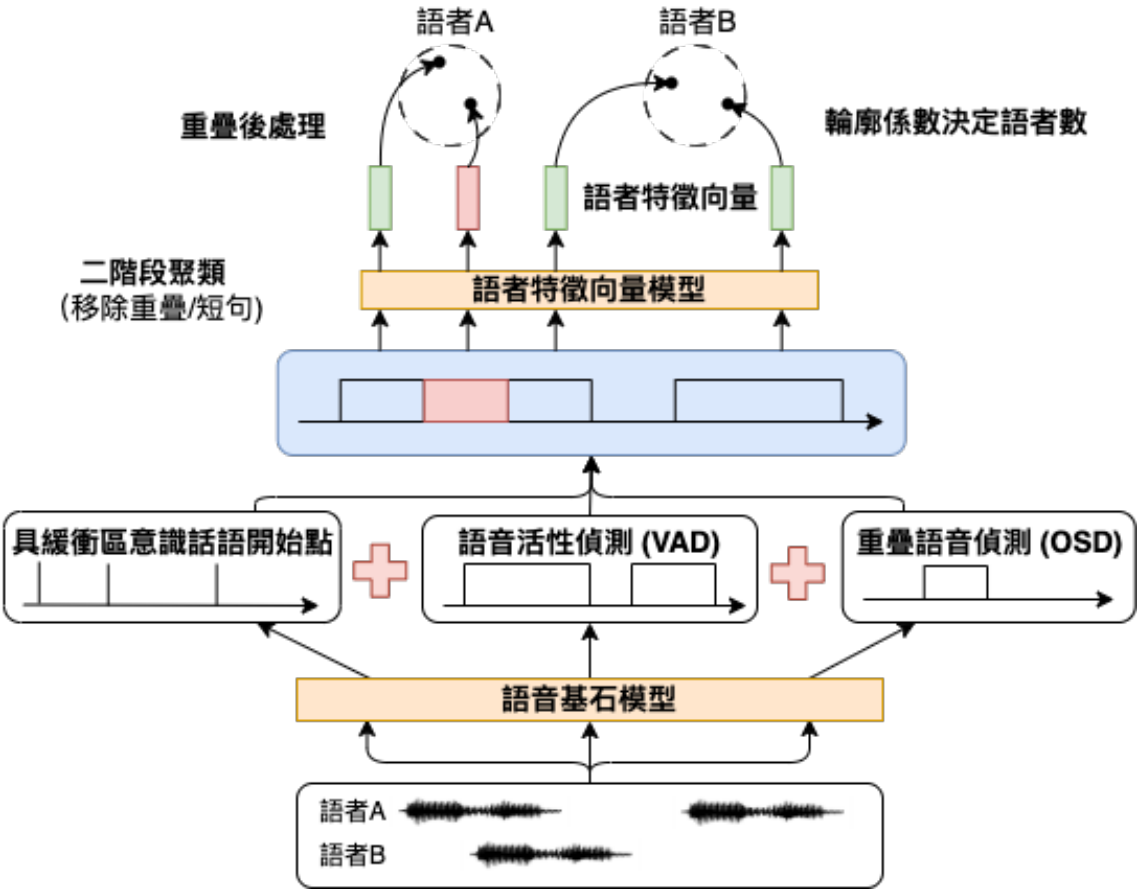

圖 3.9: 完整階段性系統示意圖

3.5.3 實驗設定

接下來補充模型訓練及推論時的一些細節：

有別於 3.2.3 節使用 7 秒長度區塊訓練，本節實驗改以 20 秒長度區塊訓練，以小幅提升表現。在測試階段，先將語音切成 20 秒長度區塊，並在每個區塊的左右各延伸 1 秒作上下文。最後的預測結果為所有區塊預測結果的黏貼，不需進行特殊處理。

在評估域內表現時，會對三個任務，VAD、OSD、UBD (話語開始點偵測) 的閾值進行網格搜索 (grid search)，並報告最佳的 DER 表現。其中 VAD / OSD 搜索範圍 0.3 到 0.7，UBD 搜索範圍 0.5 到 0.9，每 0.1 步進行搜索。而在評估域外表現時，則 VAD / OSD 統一使用 0.5 閾值，UBD 統一使用 0.7 閾值。

此外，有別於使用真實標記模擬的情景，在推論階段，話語開始點不一定在時間上完全對齊 VAD 及 OSD 的預測，直接用於切割會錯誤切割出許多短片段。由於訓練時使用的緩衝區大小為 ±0.2 秒，所以推論時會將話語偵測點與其 ±0.2 秒內的 VAD 及 OSD 片段邊界對齊，將 VAD 及 OSD 的邊界修改至話語開始點。這裏假設了一點：以話語開始點切分的邊界比 VAD 及 OSD 的邊界更為精準，原因在於話語開始點偵測具有緩衝意識，而 VAD 及 OSD 任務的邊界不具「緩衝意識」。在進行切除研究時，此對齊過程稱作「推論時跟隨話語開始點時間」。圖 3.10 為對齊過程之示意圖。

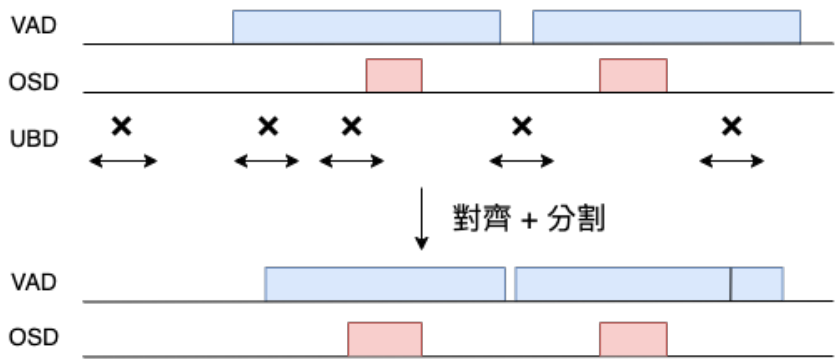

圖 3.10: 推論時對齊語音活性偵測、重疊語音偵測及話語開始點偵測結果之示意圖

	Ai4	Ali	AMI-I	MSD	Vox	**平均↓**	**（域外-域內）↓**
Sincnet							
域內（In-Domain）	15.1	25.6	22.4	23.9	11.9	19.8	
域外（Out-of-Domain）	18.7	27.4	28.3	26.7	13.4	22.9	（+3.1）
Whisper Tiny Enc							
域內（In-Domain）	13.5	21.6	19.3	21.3	9.9	17.1	
域外（Out-of-Domain）	15.5	22.8	23.3	24.0	10.8	19.1	（+2.0）
Chinese Hubert Large							
域內（In-Domain）	**12.8**	**20.8**	**19.0**	**20.4** *(13.2)*	**9.8** *(5.6)*	**16.6**	
域外（Out-of-Domain）	14.0	21.0	22.3	21.1	10.6	17.8	**(+1.2)**
階段性系統 SOTA	15.8 [52] 16.1 [75]	23.5 [75] 28.8 [52]	19.9 [5] 22.4 [52]	*(16.9)* [52]	11.1 [20] *(*6.1) [52]		

表 3.15: 比較改進後的階段性系統，在不同模型在域內情景與域外情景之表現差異; 評估指標為 DER （完整），括號斜體為 DER （公正）

3.5.4 實驗結果分析

表 3.15 比較在改進後的階段性系統中，不同模型的域內及域外表現。首先，在域內表現方面，語音基石模型在 DER 上比 Sincnet 進步約 2-3%。而在域外表現方面，語音基石模型的域內域外表現差距僅約 1-2%，亦優於 Sincnet 的 3%。此結果顯示階段性系統的泛化能力相當優秀，而語音基石模型能進一步提升其泛化能力。最後，跟現時階段性系統 SOTA 比較 (通常使用 VAD + VBx [53] + OVD)，使用 Sincnet 的系統在 Aishell4 及 MSDWild 資料集表現更優，而使用語音基石模型的系統則在所有資料集上均大幅優於最先進模型。

最後，表 3.16 列舉本節對改進後系統進行的切除研究。其中，最重要的改進是以具緩衝意識話語開始點偵測取代語者切換點。其次則是在聚類前先移除重疊部分及短句。而推論時跟隨話語開始點時間帶來約 0.3% 的進步，驗證了具緩衝意識話語開始點的精確度。

	Ai4	Ali	AMI-I	MSD	Vox	平均↓	
Whisper Tiny Enc	13.5	21.6	19.3	21.3	9.9	17.1	
- 話語開始點取代語者切換點	21.4	34.3	32.3	29.0	15.2	26.4	(+9.1)
- 話語開始點緩衝意識	16.1	24.0	22.3	25.0	12.3	20.0	(+2.7)
- 聚類移除短句	17.2	25.9	25.0	24.8	13.0	21.2	(+3.9)
- 聚類移除重疊	16.5	29.0	24.1	26.3	10.3	21.2	(+3.9)
- 重疊後處理	13.8	22.6	20.2	22.3	10.8	17.9	(+0.6)
- 推論時跟隨話語開始點時間	13.9	21.4	19.6	22.0	11.1	17.5	(+0.3)

表 3.16: 階段性系統各項改進方法之切除研究（Ablation Study）

3.6 本章總結

本章有兩個主要討論的議題：一是語音基石模型於語者分段標記及其相關下游任務的應用，二是階段性系統不同模組的改善方法。

在第一個議題中，本章首先在 3.1 節使用公開的 SUPERB 基準評估各模型表現，接著在 3.2 節及 3.3 節以域內及域外評估方式，在語音活性偵測、重疊語音偵測、話語開始點偵測等任務上基準化不同模型表現。最後，在 3.5 節將其應用到完整的階段性系統。本章的結果顯示了語音基石模型在相關任務上卓越的表現及泛化能力，同時也指出了不同任務性質與模型能力之間的關聯。

在第二個議題中，本章首先在 3.2 節討論了最基礎的語音活性偵測和重疊語音偵測任務，接著在 3.3 節詳細討論語者切換點任務，發現模型在辨識話語結束點方面表現不佳，因此提出以具有緩衝意識的話語開始點偵測取代語者切換點偵測。隨後在 3.4 節運用真實標註評估的方法，衡量並改進了語者聚類的表現，並比較端到端-階段性混合系統及階段性系統的性能極限差距。綜合了以上的改進，本論文在 3.5 節提出的改進階段性系統展現了相當優異的表現。

第四章　端到端-階段性混合系統

4.1　語音基石模型用於 EEND-VC

4.1.1　簡介

端到端-階段性混合系統是目前文獻上相當受重視的的研究方向，然而目前仍幾乎沒有相關研究探討其泛化能力。本節將應用語音基石模型到混合系統的代表作：EEND-VC 以及其變形 Graph-PIT-EEND-VC，並觀察它們在不同資料集的域內及域外表現。

4.1.2　通用模型架構

本節首先定義基於語音基石模型的混合系統通用模型架構，以利後續討論。從圖 4.1 中可見，使用語音基石模型的方法大致與前一章相同，只有下游模型的預測目標有所不同。混合系統的預測目標主要被分為三個部分：多頻道語音活性偵測、多頻道話語開始點偵測以及多頻道語者特徵向量。在本章後續討論中，所有對混合系統的修改都將包含在這個通用架構之上。這可能包括某些部分的移除或新增。

首先，混合系統的首要目標是多頻道語音活性偵測。不同於第三章僅有單一

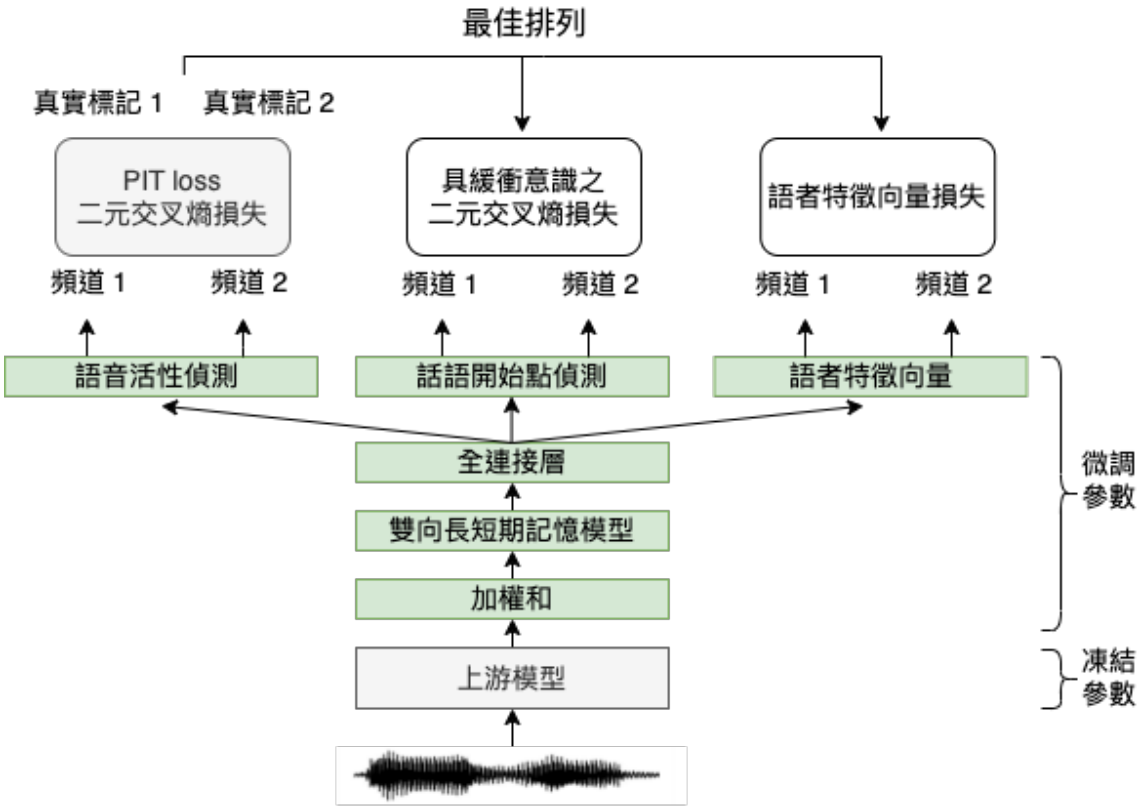

圖 4.1: 端到端-階段性混合系統通用模型架構

輸出頻道的語音活性偵測，混合系統同時偵測多個頻道的語音活性，並透過施加某種「限制」，將所有語音片段分配到各個頻道。例如，基於 u-PIT 的 EEND-VC 限制每個輸出頻道對應一位語者；而基於 Graph-PIT 的 Graph-PIT-EEND-VC 則限制了時間上有重疊的語音片段不得出現在同一頻道。在訓練過程中，各頻道被視為相等，透過置換不變訓練方法，尋找所有符合限制的真實標記排列，計算模型輸出的二元交叉熵損失，選取最小損失作為最終損失。

其次，第二個目標是多頻道話語開始點偵測。第三章將話語開始點定義為所有語音片段的開始，但在此，每個頻道將有其獨立的話語開始點，即語音活性偵測所對應頻道中語音片段的開始點。在訓練階段，目標排列與語音活性偵測一致。換言之，透過置換不變訓練方法找到的最佳排列不僅用於語音活性偵測，同時用於話語開始點偵測。

最後，第三個目標是多頻道語者特徵向量。混合系統除了預測頻道中語音活

性外，同時估算每個時間點的幀級語者特徵向量 (frame-level embedding)。每個語音片段的片段層級特徵向量 (utterance-level embedding) 則是幀級特徵向量按照其語音活性加權平均。在訓練過程中，同樣使用置換不變訓練方法找出的最佳排列，並利用在 2.4.1 節介紹的特徵向量損失函數進行優化。

4.1.3 實驗方法及設定

在定義通用模型架構後，接下來討論本節實驗使用的 EEND-VC 及 Graph-PIT-EEND-VC 方法。本節在進行實驗時，儘量在方法上保持與文獻一致，但由於原本方法的一些缺陷，因此以下討論時亦會詳述對原本方法作出的修改。

EEND-VC 使用 u-PIT 方法，其限制是每個輸出頻道對應一位語者。在符合此限制的情況下，頻道中的每個片段均隸屬同一語者，因此不需要使用話語開始點進一步分割語音片段。因此，EEND-VC 只有用到上述架構的第一個及第三個目標。在實際應用中，考慮到真實資料可能在五秒內含有三個或以上的語者，以及 u-PIT 在過多頻道時可能變得不穩定，使用 EEND-VC 時通常會選擇五秒的區塊並限制輸出頻道數為三個 [71] [45]，以確保模型輸出與實際情況符合。

而 Graph-PIT-EEND-VC 使用 Graph-PIT 方法，其限制是重疊的語音片段不得出現在同一頻道，亦即兩個重疊語音片段中存在一「不可連接」限制。在此限制下，會出現一種最壞情況：兩個在時間上相鄰的非重疊語音片段，在推論階段剛好被分配到同一個頻道上。如圖 4.2 中，第一個頻道中的片段 1 及片段 2 因時間上過於接近，測試時被合併成片段 A。因此，Graph-PIT-EEND-VC 在訓練時額外添加話語開始點預測目標，進一步分割語音片段，以解決上述問題。本章的實驗中，會將話語開始點偵測的二元交叉熵損失，替換成於 3.3 節發現效果明顯較好的具緩衝意識之二元交叉熵損失。此外，由於在同一時間有三位語者說話的情況

極為罕見，因此 Graph-PIT 方法只需要兩個輸出頻道。

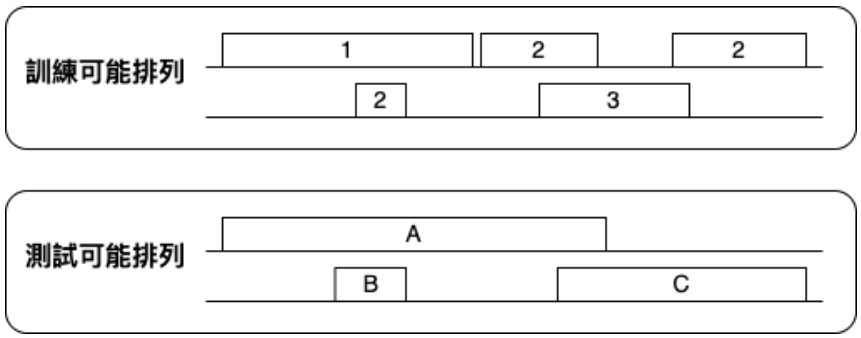

圖 4.2: Graph-PIT 推論時排列出現的最壞情況示意圖

然而，本節在實際使用 Graph-PIT 方法訓練時，發現除了時間上接近的片段會被合併，在時間上重疊的語音片段也很常被分配到同一頻道。如圖 4.2 中，片段 2 及片段 3 雖然在時間上重疊，推論時被合併成片段 C。在觀察訓練過程 Graph-PIT 的排列方式時，進一步發現，模型傾向在滿足不可連接限制的前提下，將大部分語音片段置於同一個頻道上，出現頻道不平衡 (channel imbalance) 的問題，削弱模型表現。本論文推測，Graph-PIT-EEND-VC 應用此方法在模擬生成語音訓練時，因為模擬語音重疊比例較高的緣故，所以頻道不平衡的問題較不顯著。然而，在現實情況下，不同資料集有顯著不同的重疊比例，導致 Graph-PIT 方法出現泛化能力不足的問題。

因此，本節簡單修改 Graph-PIT 的訓練目標，以解決頻道不平衡問題，(稱為 Modified Graph-PIT)。除了限制重疊的語音片段不能放置在同一頻道外，進一步限制在一定時間範圍內 (如 1 秒)，隸屬於不同語者的語音片段不能放置於同一頻道。(參圖 4.3)。與原本的 Graph-PIT 相比，Modified Graph-PIT 大幅增加了不可連接片段的數量，有效解決了在真實資料上頻道不平衡的問題。相較於 u-PIT，Modified Graph-PIT 更注重局部的語者切換，因為 u-PIT 在整個區塊中要求不同語者被放置於不同頻道，而 Modified Graph-PIT 僅在局部區域內施加這樣的限制。

Graph-PIT 在置換不變訓練時，會解圖著色問題並得出所有可能的排列。

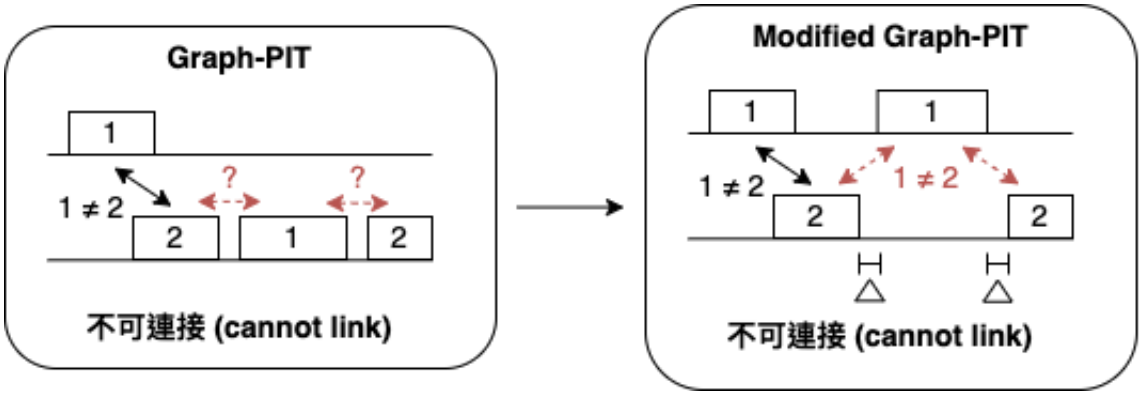

圖 4.3: Graph-PIT 改進前及改進後模型限制示意圖

Modified Graph-PIT 的目標亦是解圖著色問題：

設有圖 $G = (V, E)$，其中頂點集合 V 由每個語者發言的開始時間 Ub 和結束時間 Ue 構成。邊集合 E 由以下條件形成：對於所有 $u, v \in V$ 且 $u \neq v$，若 Ue_u 和 Ub_v 的時間差小於一個閾值、語者 s_u 不等於語者 s_v，且在 Ub_v 和 Ue_u 之間沒有其他語者 s_w 的時間片段存在，則存在邊 $\{u, v\}$。

在實際訓練時，要施加以上時間差限制，可簡單透過在 Ue_u 的最後補零的方式。而移除時間差內有其他片段的邊 $\{u, v\}$，則避免了出現無法滿足條件的狀況。在訓練的時候，時間差閾值訂為 1 秒；而在測試的時候，時間差閾值則訂為 0.2 秒。

接下來討論推論時的區塊組合問題。EEND-VC 或 Graph-PIT-EEND-VC 都會在推論時將語音切割成小區塊，分別預測結果後組合在一起。然而，它們都沒有對區塊邊界作特別處理，因此在區塊邊界很容易出現短片段。由於語者特徵向量不擅長處理短片段，所以當區塊長度變短時，準確率會有所下降。為解決這個問題，[71] 採用了間隔 0.5 秒的滑動窗口，以確保每個時間點都擁有充足的上下文。然而，使用頻密的滑動窗口會降低系統的效率。為平衡準確率和效率，本節採增加上下文的方法，先在區塊左右各填充 1 秒，並在提取特徵向量時，同時提取同一語者在上下文的延伸部分。圖 4.4 比較了一般分塊處理、滑動窗口及增加上下

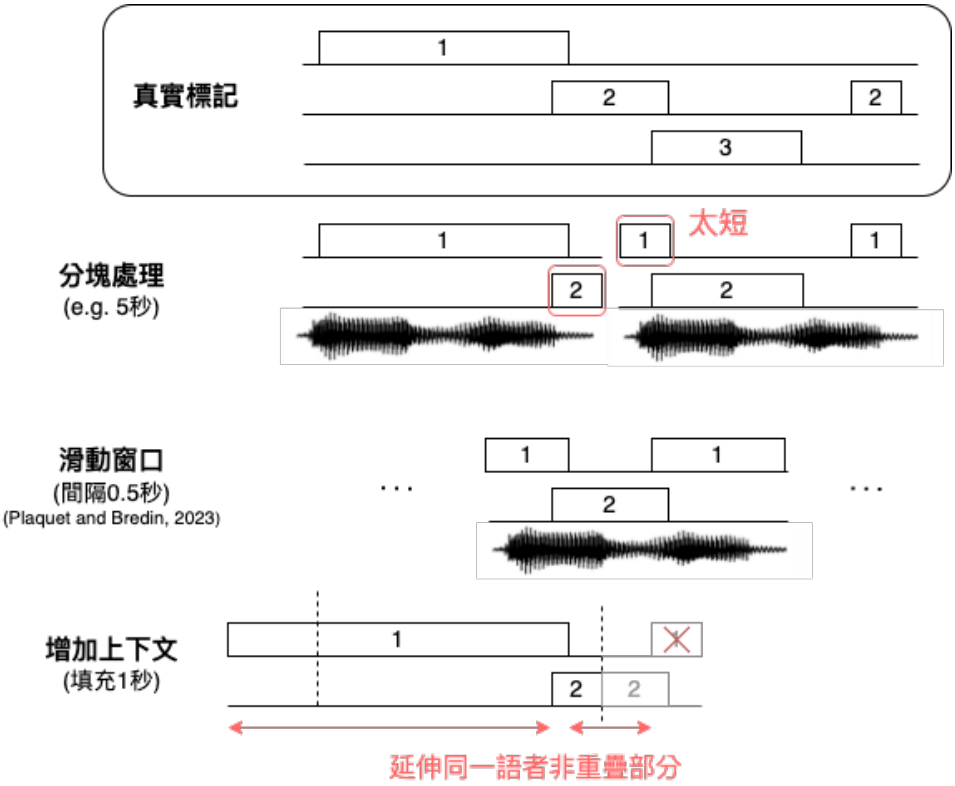

圖 4.4: 比較端到端-階段性混合系統處理區塊邊界的三種方法

文三種方法。

最後，本節在進行語者特徵向量提取與聚類時，會大致跟隨 3.4 節提出的三個步驟：先移除重疊部分及短句，再提取語者特徵向量，最後進行二階段聚類。唯一的區別，在於在本節，語者特徵向量的預測會跟隨 EEND-VC 論文直接採用系統預測的多頻道語者特徵向量，而非使用外部語者特徵向量模型。

以上討論了基於 u-PIT、Graph-PIT 及本節提出 Modified Graph-PIT 置換不變訓練方法的混合系統，接下來將比較三種方法在使用不同語音基石模型下的域內表現及域外表現。

4.1.4 實驗結果分析

表 4.1 為本節實驗的綜合結果，以下從三個方面分析：

	Ai4	Ali	AMI-I	MSD	Vox	**平均 ↓**	**(域外-域內) ↓**
Sincnet							
域內 (In-Domain)							
u-PIT	18.5	35.1	**21.0**	34.2	14.6	24.7	
Graph-PIT	20.7	**33.0**	24.8	38.1	18.6	27.0	
+ Modified	**18.1**	33.4	24.5	**32.7**	**14.4**	**24.6**	
域外 (Out-of-Domain)							
u-PIT	25.2	34.2	35.2	42.7	18.2	31.1	(+6.4)
Graph-PIT	28.6	**33.9**	34.9	44.0	20.0	32.3	(+5.2)
+ Modified	**22.4**	34.2	**31.6**	**36.5**	**16.5**	**28.2**	**(+3.6)**
Whisper Tiny Enc							
域內 (In-Domain)							
u-PIT	14.9	22.6	**18.1**	**22.1**	12.2	**18.0**	
Graph-PIT	15.3	22.5	20.2	24.1	13.9	19.2	
+ Modified	**14.6**	**22.3**	19.8	22.5	**11.7**	18.2	
域外 (Out-of-Domain)							
u-PIT	**18.0**	24.6	26.6	26.0	16.1	22.3	(+4.3)
Graph-PIT	22.0	**23.3**	26.6	29.8	15.4	23.4	(+4.2)
+ Modified	19.2	23.4	**25.2**	**25.0**	**13.5**	**21.3**	**(+3.1)**
Hubert Large Chinese							
域內 (In-Domain)							
u-PIT	12.3	**19.4**	**16.6**	19.0	11.1	**15.7**	
Modified Graph-PIT	**12.1**	20.0	18.2	**18.7**	**10.9**	16.0	
域外 (Out-of-Domain)							
u-PIT	15.8	20.9	22.2	21.1	14.7	18.9	(+3.3)
Modified Graph-PIT	**14.9**	**20.3**	**21.5**	**19.7**	**12.7**	**17.8**	**(+1.8)**

表 4.1: 比較端到端-階段性混合系統中，不同模型使用 u-PIT、Graph-PIT 及 Modified Graph-PIT 方法在域內及域外情景的表現；評估指標為 DER（完整）

首先，基於 Modified Graph-PIT 方法的系統，在大部分資料集上表現均優於 Graph-PIT。Graph-PIT 方法不管是在域內還是域外情景均表現不佳，明顯劣於 u-PIT 方法的結果，顯示 Graph-PIT 的目標在真實資料上泛用能力不足，可能存有方法上的缺陷。而 Modified Graph-PIT 則解決了這個問題，在域內情景表現與 u-PIT 方法相當。

其次，基於 Modified Graph-PIT 方法的系統在域外的泛化能力優於 u-PIT。u-PIT 與 Modified Graph-PIT 的主要差別在於其限制的強弱，u-PIT 的限制遍及整

個區塊，而 Modified Graph-PIT 的限制只影響局部。由此可見，施加太嚴格的限制有機會導致模型過擬化到資料分佈，影響跨領域泛化能力。

最後，觀察不同上游模型的表現，發現不同模型之間的表現差距，遠比階段性系統各模型之間的差距高。3.5 節階段性系統在 Sincnet、Whisper Tiny 編碼器及 Hubert Large Chinese 的平均域內 DER 分別為 16.6%、17.1% 及 19.8%，最大及最小值差距 3.2%。此處混合系統的表現，除 Hubert Large Chinese 稍優於階段性系統，Whisper Tiny 表現反而略差，而 Sincnet 更明顯較差。不同模型的準確率最大及最小值差距高達 9.1%。在域外表現，差距更進一步放大。這結果顯示混合系統的預測目標過於困難，只有高資源語音基石模型才能取得合理的表現。

總括而言，本節討論的混合系統在表現上仍略差於階段性系統，而其中一個可能的原因是模型施加不適當的限制或預測目標。因此，後面的兩節將分別探討混合系統在移除語者特徵向量預測目標及移除連接限制後的表現。

4.2 移除語者特徵向量預測目標

4.2.1 實驗方法

端到端-階段性混合系統，如 EEND-VC、EEND-GLA 等，其預測目標通常包括語者特徵向量。這種做法避免了額外使用外部語音特徵向量模型，因此提高了效率。但是，卻有泛化能力不足的隱憂，最先進方法 [71] 亦改採外部特徵向量取代之。因此，本節將會探討不同模型及方法，在以外部特徵向量模型替代特徵向量預測目標的表現及泛化能力。其中，EEND-VC 方法在移除語者特徵向量預測目標後，只剩下多頻道語音活性偵測預測目標，而 Graph-PIT-EEND-VC 則剩下多頻道語音活性偵測及多頻道話語開始點偵測預測目標。

4.2.2 實驗結果分析

表 4.2 列出了本節的實驗結果，以下按照不同參數量模型的表現進行分析：

在低資源模型 Sincnet 上，改用外部語者特徵向量帶來顯著的表現提升，不管在 u-PIT 方法還是 Modified Graph-PIT 方法均進步約 4%。而在域外情景，帶來的表現提升進一步增加到 5-6%。在中資源模型 Whisper Tiny 編碼器上，域內及域外的表現提升則縮減到 1% 及 2%。最後在高資源模型 Hubert Large Chinese，改用外部語者特徵向量沒有帶來域內的表現提升，但仍在域外情景帶來約 1% 提升。

從以上分析可見，將語者特徵向量納入混合系統模型預測目標通常會削弱模型的表現。針對這現象本論文提出兩個可能原因。

第一，用於語者自動分段標記任務的數據量遠遠少於語者特徵向量模型的訓練數據量。本論文已經收集了網路上大部分開源資料集，總時長達 500 小時，但與語者特徵向量模型使用的 3000 多小時語者資料集相比 [56] [66]，仍顯得有限。

第二，語者分段所用的模型架構未必適合語者驗證任務。目前最先進的語者特徵模型多使用時延神經網路 TDNN 搭配卷積神經網路；而在語者自動分段標記任務中，由於其需要預測序列，常使用如長短期記憶模型、轉換器等架構。然而，將通用轉換器架構應用到語者驗證任務可能會遇到效率不足的問題。舉例來說，使用大型轉換器 (語音基石模型 Wavlm Large) 結合基於 TDNN+CNN 的下游模型，固然在語者驗證任務上取得最先進的表現 [13]。但是，這種出色的結果建基於龐大的轉換器模型。相對地，使用參數量較少的 Wavlm Base 模型再加上 TDNN+CNN，相對於傳統的特徵向量模型，並未呈現明顯的優勢。從電腦視覺領域的圖像分類任務中也可得到一些啟示，一些研究 [61] 指出，通用轉換器架構的模型效率不及卷積神經網路。因此，改進轉換器架構可能是未來的重要方向。

	Ai4	Ali	AMI-I	MSD	Vox	**平均 ↓**	**(域外-域內) ↓**
Sincnet							
域內 (In-Domain)							
u-PIT	18.5	35.1	21.0	34.2	14.6	24.7	
＋外部語者特徵向量	15.2	**27.2**	**20.8**	26.5	**12.1**	**20.4**	
Modified Graph-PIT	18.1	33.4	24.5	32.7	14.4	24.6	
＋外部語者特徵向量	**14.6**	27.8	23.9	**25.0**	12.5	20.8	
域外 (Out-of-Domain)							
u-PIT	25.2	34.2	35.2	42.7	18.2	31.1	(-6.4)
＋外部語者特徵向量	20.8	**28.9**	29.5	33.0	15.3	25.5	(-5.1)
Modified Graph-PIT	22.4	34.2	31.6	36.5	16.5	28.2	(-3.6)
＋外部語者特徵向量	**17.5**	29.5	**28.2**	**29.7**	**13.9**	**23.8**	**(-3.0)**
Whisper Tiny Enc							
域內 (In-Domain)							
u-PIT	14.9	22.6	**18.1**	22.1	12.2	18.0	
＋外部語者特徵向量	14.0	21.0	18.8	**19.8**	10.3	**16.8**	
Modified Graph-PIT	14.6	22.3	19.8	22.5	11.7	18.2	
＋外部語者特徵向量	**13.7**	**20.9**	19.8	20.5	**10.1**	17.0	
域外 (Out-of-Domain)							
u-PIT	18.0	24.6	26.6	26.0	16.1	22.3	(-4.3)
＋外部語者特徵向量	16.6	22.0	24.7	24.5	13.0	20.2	(-3.4)
Modified Graph-PIT	19.2	23.4	25.2	25.0	13.5	21.3	(-3.1)
＋外部語者特徵向量	**16.1**	**21.1**	**23.6**	**23.5**	**11.4**	**19.1**	**(-2.1)**
Hubert Large Chinese							
域內 (In-Domain)							
u-PIT	12.3	**19.4**	**16.6**	19.0	11.1	**15.7**	
＋外部語者特徵向量	12.5	19.8	18.1	18.8	**10.2**	15.9	
Modified Graph-PIT	12.1	20.0	18.2	18.7	10.9	16.0	
＋外部語者特徵向量	**11.9**	**19.4**	18.7	**18.5**	10.4	**15.7**	
域外 (Out-of-Domain)							
u-PIT	15.8	20.9	22.2	21.1	14.7	18.9	(-3.3)
＋外部語者特徵向量	14.6	**19.7**	21.7	20.2	11.8	17.6	(-1.7)
Modified Graph-PIT	14.9	20.3	21.5	19.7	12.7	17.8	(-1.8)
＋外部語者特徵向量	**13.8**	19.8	**20.9**	**19.2**	**11.0**	**16.9**	**(-1.2)**

表 4.2: 比較端到端-階段性混合系統中，不同模型使用 u-PIT 及 Modified Graph-PIT 方法，在使用語者特徵向量預測目標及外部語者特徵向量的情況下，於域內及域外情景的表現；評估指標為 DER（完整）

4.3 移除連接限制

4.3.1 簡介

端到端-階段性混合系統通常在置換不變訓練的架構上，在訓練和推論時對模型輸出施加限制。然而，根據 4.1 節的結果，在推論時施加嚴格的限制可能會導致域外表現的下降。因此，本節將會逐一移除 u-PIT 方法及 Graph-PIT 方法的連接限制，研究連接限制對表現的影響。

4.3.2 實驗方法及設定

首先，分別定義 u-PIT 及 Modified Graph-PIT 方法的連接限制和移除限制的方式。

u-PIT 的限制定義為每個輸出頻道對應一位語者。這個限制可以被分解為兩個部分（見圖4.5）。第一個是不可連接限制：不同輸出頻道的語者不能被對應到同一語者。第二個是必須連接限制：同一輸出頻道的語者必須被分配到同一語者。其中，必須連接限制不僅限制同一頻道內的兩個片段必須歸屬於同一語者，還意味著頻道內同一片段只能屬於一個語者。

當同時應用所有限制時，每個頻道分別對應到一個語者特徵向量，並根據頻道的語音活性平均值，從最高到最低逐一指定語者，指定方式是根據語者特徵向量與全域語者群心的相似度。當僅移除不可連接限制時，不同頻道可以分配到同一語者，因此直接將每個頻道的語者特徵向量與全域群心計算距離，然後分配到最接近的語者。然而，僅移除必須連接限制在實務上無法實現，因為當每個頻道都有多個語者時，所有頻道的語者數加總很容易超過全域語者數。最後，在移除

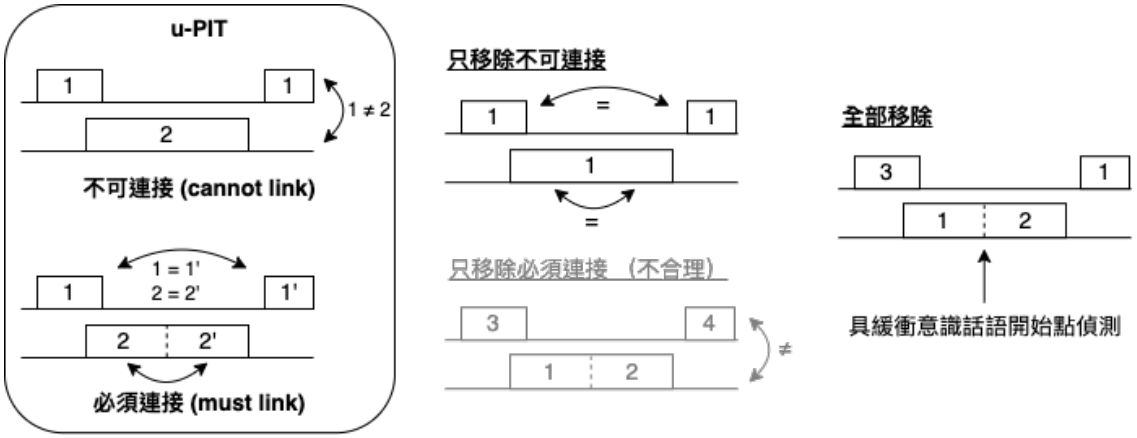

圖 4.5: u-PIT 置換不變訓練方法在推論時的連接限制及移除限制方式

所有限制下，每個頻道的每個片段都被視為獨立的，且片段還會進一步被具緩衝話語開始點分割，並逐一按距離分配到最接近的語者。

Graph-PIT 的限制定義為重疊語音片段不能對應到同一個頻道，4.1 節提出的 Modified Graph-PIT，將其限制延伸至時間差小於一個閾值，不同語者的片段不能對應到同一個頻道。應用上述不可連接限制時，根據片段的語音活性平均值，從最高到最低逐一指定語者，指定方式是根據語者特徵向量與全域語者群心的相似度。當移除不可連接限制時，則每個片段獨立地按距離分配到最接近的語者。此外，本節跟隨 Graph-PIT-EEND-VC 的做法，不管是在移除不可連接前，還是在移除不可連接後，均使用具緩衝意識話語開始點進一步分割片段 (也就是說，原始 Graph-PIT 可被視作已移除必須連接限制)。

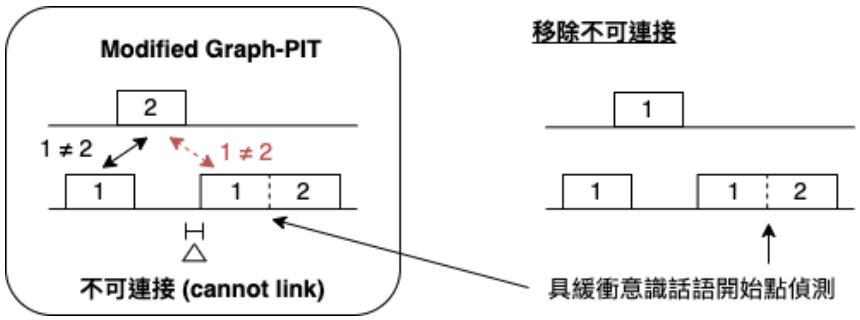

圖 4.6: Modified Graph-PIT 置換不變訓練方法在推論時的連接限制及移除限制方式

值得注意的是，在移除連接限制後，重疊語音有機會只被分配到一個語者。

因此，本節會採用於 3.4.3 節階段性系統用到的「重疊後處理」方法，直接將第二接近的語者加入預測中。此外，由於 4.2 節的結果證實了使用外部語者特徵向量的表現優於模型預測的特徵向量，因此本節的實驗均使用外部語者特徵向量。

4.3.3 實驗結果分析

表 4.3 列出了本節的實驗結果，以下進行分析：

首先，在域內表現上，移除連接限制對系統的表現並未有太大幫助，反而可能導致準確率下降。特別是在高資源模型下，下降的幅度比低資源模型更為明顯。由此可見，模型的連接限制對於在特定領域內的泛化能力是有利的。而準確率下降的原因很大程度上是因為語音片段被更細分，進而影響了語者特徵向量的品質。

然而，在跨領域的情境中，移除連接限制卻稍微提升了系統的性能，特別是在低資源模型上，準確率提升約 1-2%。這顯示出某個資料集適用的連接限制未必適用於其他資料集。在實際應用中，會導致有兩個語者的片段被合併成一個的情況。移除連接限制雖然解決了兩個語者片段誤合併的問題，但也可能造成同一語者的語音被切分得太短，進而影響語者特徵向量的品質。準確率的變化同時受到這兩個因素的影響，但明顯誤合併對系統的負面影響比片段過短更為嚴重。

最後，綜合比較 u-PIT 及 Modified Graph-PIT 方法的表現。在域內情景，u-PIT 和 Modified Graph-PIT 的表現在移除限制前後均相距不大。但在域外情景，移除限制前的 u-PIT 稍差於 Modified Graph-PIT，而兩者差距在移除限制後非常接近。事實上，移除限制後的 u-PIT 及移除限制後的 Graph-PIT 在片段的處理方式上非常相似，唯一的區別是 u-PIT 使用三個輸出頻道，而 Graph-PIT 使用兩個輸出頻道。

	Ai4	Ali	AMI-I	MSD	Vox	平均	(移除前後變化)
Sincnet							
域內 (In-Domain)							
u-PIT	15.2	27.2	**20.8**	26.5	12.1	20.4	
- 不可連接	15.6	28.1	22.0	26.7	12.5	21.0	(+0.6)
- 必須連接	14.9	**26.6**	21.7	25.6	**12.0**	**20.2**	(-0.2)
Modified Graph-PIT	**14.6**	27.8	23.9	**25.0**	12.5	20.8	
- 不可連接	14.8	28.4	25.0	25.9	12.5	21.3	(+0.6)
域外 (In-Domain)							
u-PIT	20.8	28.9	29.5	33.0	15.3	25.5	
- 不可連接	20.5	26.8	27.8	31.0	14.7	24.2	(-1.3)
- 必須連接	19.0	**26.6**	27.3	**28.6**	14.0	23.2	(-2.3)
Modified Graph-PIT	17.5	29.5	28.2	29.7	13.9	23.8	
- 不可連接	**16.9**	27.7	**26.5**	28.8	**13.4**	**22.7**	(-1.1)
Whisper Tiny Enc							
域內 (In-Domain)							
u-PIT	14.0	21.0	**18.8**	**19.8**	10.3	**16.8**	
- 不可連接	14.6	21.5	20.7	20.6	10.5	17.6	(+0.8)
- 必須連接	14.5	21.7	21.0	20.4	10.3	17.6	(+0.8)
Modified Graph-PIT	**13.7**	**20.9**	19.8	20.5	**10.1**	17.0	
- 不可連接	14.1	21.8	21.6	22.0	**10.1**	17.9	(+0.9)
域外 (In-Domain)							
u-PIT	16.6	22.0	24.7	24.5	13.0	20.2	
- 不可連接	16.4	20.9	23.8	24.1	13.0	19.6	(-0.5)
- 必須連接	**15.7**	21.2	23.2	**22.2**	12.1	**18.9**	(-1.3)
Modified Graph-PIT	16.1	**21.1**	23.6	23.5	11.4	19.1	
- 不可連接	15.9	21.3	**23.0**	22.9	**11.2**	**18.9**	(-0.3)
Hubert Large Chinese							
域內 (In-Domain)							
u-PIT	12.5	19.8	**18.1**	18.8	**10.2**	15.9	
- 不可連接	14.0	20.1	19.7	20.2	10.4	16.9	(+1.0)
- 必須連接	14.2	19.9	20.4	20.4	11.0	17.2	(+1.3)
Modified Graph-PIT	**11.9**	**19.4**	18.7	**18.5**	10.4	**15.8**	
- 不可連接	13.6	20.6	20.1	19.5	10.7	16.9	(+1.1)
域外 (In-Domain)							
u-PIT	14.6	**19.7**	21.7	20.2	11.8	17.6	
- 不可連接	14.9	19.8	21.0	20.7	11.8	17.6	(0.0)
- 必須連接	14.6	20.0	**20.1**	20.2	11.5	17.3	(-0.3)
Modified Graph-PIT	**13.8**	19.8	20.9	**19.2**	11.0	**16.9**	
- 不可連接	14.0	20.1	20.9	20.3	**10.9**	17.2	(+0.3)

表 4.3: 比較端到端-階段性混合系統中，不同模型使用 u-PIT 及 Modified Graph-PIT 方法並使用外部語者特徵向量下，移除限制前及移除限制後在域內及域外情景的表現；評估指標為 DER（完整）

4.4 綜合表現比較

4.4.1 實驗設定

本節將綜合比較上章提出之階段性系統，本章修正之端到端-階段性混合系統，以及文獻其他方法的表現。其中，階段性系統會直接沿用 3.5 節域內評估的結果，並額外新增在 AMI-S 及 RAMC 資料集的結果。至於端到端-階段性混合系統，由於 4.2 節及 4.3 節的實驗結果顯示在域內評估時，添加連接限制通常有稍好的表現，因此本節會使用在沒有移除連接限制下的結果。此外，為彌補資料集標註的不一致性，本節進一步在 AMI 資料集及 Voxconverse 上微調模型以得到更好的結果。

4.4.2 實驗結果分析

以下根據表 4.4 結果分析：

首先，混合系統在平均表現上優於階段性系統約 1%-1.5%，此結果符合 3.4 節模擬兩個系統表現差距的極限 (2.3%)。其中，表現差異主要來自重覆語音佔比較高的 AMI 資料集及 MSDWild 資料集，而在幾乎沒有重疊語音的 Voxconverse 及 MagicData-RAMC 資料集，兩個系統的表現大致相同。

此外，基於 Hubert Large Chinese 的階段性系統表現，與基於 Whisper Tiny 編碼器的混合系統表現相若。此結果顯示，與改進系統架構相比，改進語音基石模型的能力可能是同樣重要的方向。

最後，與各系統最先進技術比較。無論是基於 Whisper Tiny 編碼器，還是基於 Hubert Large Chinese，本論文改進之兩個系統的表現，在大部分資料集上

	Ai4	Ali	AMI-I	AMI-S	MSD	Vox	RAMC	**平均**
Whisper Tiny Enc								
階段性系統	13.5	21.6	19.3	22.4	21.3	9.9	13.1	17.3
端到端-階段性混合系統	12.9	20.9	17.1	19.8	19.6	10.1	13.4	16.3
Hubert Large Chinese								
階段性系統	12.8	20.8	19.0	21.2	20.4	9.8	**11.4**	16.4
端到端-階段性混合系統	**11.9**	**19.4**	15.4	**17.6**	**18.5** *(11.7)*	**9.6** *(5.3)*	11.6	**14.9**
SOTA (截至 2023 年 12 月)								
階段性系統	15.8 [52]	23.5 [75]	19.9 [5]	23.7 [75]	*(16.9)* [52]	11.1 [20]	18.2 [52]	
	16.1 [75]	28.8 [52]	22.4 [52]	34.6 [52]		*(6.1)* [52]	19.9 [103]	
端到端-階段性混合系統	13.2 [71]	23.3 [71]	18.0 [71]	22.0 [71]	27.1 [71]	10.4 [71]		
端到端系統	31.3 [52]	26.3 [52]	**13.0** [15]	19.5 [33]	*(14.6)* [52]	9.9 [78]	13.6 [78]	
			16.8 [33]	24.6 [78]			14.4 [7]	

表 4.4: 各系統表現綜合比較; 評估指標為 DER（完整），括號斜體為 DER（公正）

均明顯優於最先進技術。除了在 AMI-I 資料集，端到端系統 AED-EEND-EE + Conformer [15] 取得更好的表現。

雖然本論文提出之系統在各資料集上都表現優異，甚至超越了目前最先進的技術，但卻有一個潛在的問題：語音基石模型通常擁有極多的參數，這在實際應用中可能會帶來困難。接下來的討論將針對這個問題進行探討。

4.4.3 運算成本分析

本小節將探討本論文介紹各系統的運算成本，並將其與文獻中表現最佳的階段性系統及混合系統進行比較。

表 4.5 統計了不同模型的運算成本，其中實時率 (real-time factor，下稱 RTF) 使用 CPU (Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz) 以單線程 (single thread)

方式測量，以反映系統用在真實世界部署的效率。

模型	RTF
Sincnet	0.003
Whisper Tiny Enc	0.018
Hubert Large Chinese	≈1
下游模型 (BiLSTM)	0.002
Sincnet + BiLSTM	0.005
Whisper Tiny Enc + BiLSTM	0.02
CAM++	0.022

表 4.5: 不同模型運算成本

比較本論文大部分實驗使用的三個語音基石模型：Sincnet，Whisper Tiny 編碼器及 Hubert Large Chinese，發現 Hubert Large Chinese 因參數量龐大難以被實際用於真實系統。至於 Sincnet 及 Whisper Tiny 編碼器，其運算成本相對較低，在加上下游模型後 RTF 分別僅 0.005 及 0.02，適合用於真實世界部署。另外，本論文使用的語音特徵向量模型 CAM++ RTF 也僅約 0.02。

不同系統通常包括兩個模型：用於語者分割的語音基石模型 (包括 VAD，OSD，話語開始點偵測，u-PIT 目標，Graph-PIT 目標)，以及語者特徵向量模型。語者分割模型會比較 Sincnet 及 Whisper Tiny 編碼器 (**代號 Seg**)。至於語者特徵向量模型，雖然其他文獻通常使用參數更多的模型 (如 VBx [53] 使用 Resnet101 模型，Plaquet & Bredin, 2023 [71] 則使用 ECAPA-TDNN [21])，但為方便比較，以下運算成本分析假設其他系統都使用效率更高的 Cam++ 模型 (**代號 Emb**)。

接下來逐一討論不同系統的運算成本：(見表 4.6)

- 階段性系統 (3.5節)

  此系統使用同一模型預測 VAD、OSD、話語開始點並切割成語者片段 (Seg)。由於其移除重疊片段的機制，因此在一般情況下每個時間點都只會提取一次語者特徵向量 (Emb)。在推論時，20 秒的區塊會延伸 1 秒作上下文，因此系統最後運算成本為 1.1 x Seg + Emb (1.1 = 22 / 20)。

系統	運算成本	估算 RTF (Sincnet)	估算 RTF (Whisper)
階段性系統 (3.5)	1.1 x Seg + Emb	**0.028**	**0.044**
階段性系統 SOTA (VAD+VBx+OSD) [53] [5]	1.1 x Seg + 6 x Emb	*0.138*	
EEND-VC (4.1)	1.4 x Seg	0.007	0.028
EEND-VC + 外部特徵向量 (4.2/4.3)	1.4 x (Seg + Emb)	0.038	0.059
Graph-PIT-EEND-VC (4.1)	1.1 x Seg	0.006	0.022
Graph-PIT-EEND-VC + 外部特徵向量 (4.2/4.3)	1.1 x (Seg + Emb)	**0.03**	**0.046**
混合系統 SOTA (Plaquet & Bredin, 2023) [71]	10 x (Seg + Emb)	*0.27*	

表 4.6: 不同系統運算成本分析

- 階段性系統 SOTA (VAD+VBx+OSD) [53] [5]

  此系統使用 VAD 偵測人聲片段 (Seg)，接著使用 1.5 秒重疊窗口切割人聲片段 (每 0.25 秒一步)，接著提取每個重疊窗口片段的語者特徵向量 (6 x Emb)，最後以 OSD 偵測重疊語音進行後處理 (Seg)。原有系統在 VAD / OSD 亦使用重疊窗口切割方式以提升表現，但也可以改採延伸上下文的方式代替。另外，VAD 及 OSD 亦可共用模型。因此，估算系統最後運算成本為 1.1 x Seg + 6 x Emb。

- EEND-VC (4.1節) + 外部語者特徵向量 (4.2/4.3節)

  原始 EEND-VC 在語者分段 (u-PIT 目標) 時會同時預測語音特徵向量，因此只需使用分段模型 (Seg)。然而，在實務中受 u-PIT 限制，通常使用 5 秒區塊長度，並延伸 1 秒作上下文，因此系統運算成本為 1.4 x Seg (1.4 = 7 / 5)。若改用外部語者特徵向量，則運算成本上升至 1.4 x (Seg + Emb)。

- Graph-PIT-EEND-VC (4.1節) + 外部語者特徵向量 (4.2/4.3節)

  Graph-PIT-EEND-VC 放鬆了 u-PIT 的限制，因此可使用較長區塊長度，如使用 20 秒區塊，並延伸 1 秒作上下文。在不使用外部語者特徵向量的情況，運算成本為 1.1 x Seg；反之則為 1.1 x (Seg + Emb)。

- 端到端-階段性混合系統 SOTA (Plaquet & Bredin, 2023) [71]

  此系統使用 EEND-VC u-PIT 目標進行語者分段，但使用大量重疊窗口切割的方式提升表現 (5 秒重疊窗口切割，每 0.5 秒一步) (10 x Seg)，接著使用外部特徵向量提取每個語者片段 (10 x Emb)，因此估算系統最後運算成本高達 10 x (Seg + Emb)。

總結以上各系統運算成本分析，有三點發現。

首先，基於語者切換點 (或話語開始點) 切割語音片段的方法，其運算成本明顯低於基於重疊窗口切割的方法。傳統階段性模型因語者切換表現不佳，轉而使用重疊窗口切割，雖然彌補了準確率不足的問題，卻大幅提高了運算成本。相反地，運用具緩衝意識話語開始點偵測方法，在提升表現的同時，能夠確保每個語音片段只提取一次特徵向量，從而在效率上占有優勢。

其次，基於 Graph-PIT 的方法運算成本低於基於 u-PIT 的方法。在第 4.1 節改進的 Modified Graph-PIT 目標在各資料集中取得與 u-PIT 目標一致或更優的表現。由於 Modified Graph-PIT 放鬆了 u-PIT 的連接限制，因此可以使用更長的區塊，從而減少為延伸上下文而被重覆預測的語音片段。

最後，比較使用 Sincnet 及 Whisper Tiny 編碼器作為語者分段上游模型的性能。儘管基於 Whisper Tiny 的系統運算成本約高出 50%，但卻明顯提升了準確率和泛化能力。這種系統特別適合應用於對準確率要求更高的場景。此外，Whisper Tiny 在情緒辨識和聲音事件偵測方面表現出色，並且在與解碼器結合後能夠進行自動語音辨識，具有廣泛的應用潛力。

總括而言，本論文提出改進後的階段性系統及混合系統，在準確率和運算效率方面都優於以往的方法，尤其適合用於真實世界部署。

4.5 本章總結

本章主要討論了端對端的階段性混合系統在泛化能力方面的問題。

首先，本章在 4.1 節將語音基石模型應用於 EEND-VC 和 Graph-PIT-EEND-VC，指出了 Graph-PIT 方法存在的問題，並提出了改良版本 Modified Graph-PIT。接著，在 4.2 和 4.3 節中，通過移除語者特徵向量和連接限制，改進了混合系統的表現和泛化能力。最後，在 4.4 節中綜合比較了不同方法的表現和運算成本。

研究結果顯示，最初提出的混合系統 EEND-VC 和 Graph-PIT-EEND-VC 存在表現和泛化能力不足的問題。這些方法施加了過多的限制，導致系統的穩健性下降。本章發現，採用外部語者特徵向量以及推論時移除連接限制等簡單方法，都能大幅提升系統的泛化能力。另外，使用高資源語音基石模型也有助於在施加限制時保持一定的泛用性。不過，這也帶來了運算成本過高的問題。

最後，經由本章改良後的混合系統無論在域內還是域外的表現上，都稍微領先於本論文提出的階段性系統，且優於大部分最先進的方法。

第五章　結論與展望

5.1　研究貢獻與討論

本論文首先應用多個語音基石模型到語者自動分段標記系統相關基準，如SUPERB 基準及本論文提出的重疊語音基準，印證了其優秀表現及泛用能力。接著逐一分析階段性系統的每個模組，並提出方法改善性能，使得過去被認為落後的階段性系統亦能媲美端到端系統的表現。此外，本論文是首次在這個領域廣泛使用域外評估的方式，來驗證各系統和模型的泛化能力。最後，本論文改進後的系統，在多個資料集中獲得與最先進技術相當或更優的表現，且在效率上滿足現實世界部署。

然而，本論文在研究方法上可能仍存有某些限制：

首先，在比較傳統模型及語音基石模型時，本論文只有報告一種傳統模型的表現，亦沒有報告與語音基石模型參數量相若的傳統模型。在初步實驗階段，本論文曾嘗試其他模型架構，如 [62] 中使用的 Conformer 架構。然而在使用複合訓練集訓練時，這些複雜的模型均無法收斂到合理的準確率；反之，參數量極低的 Sincnet [76] 更容易在訓練時收斂。由於無法確定收斂問題是基於模型、資料還是訓練超參數，因此沒有把這些模型納入考慮。即便如此，基於使用語音基石模型的系統在多個資料集取得優於其他文獻的表現，仍足夠推論出語音基石模型之優

越性。

另外，在研究域泛用性的部分，部分文獻會在域外評估時施加更嚴謹的限制，如排除所有類似領域的資料集，而非只排除要測試的資料集；甚至有文獻在域外評估時只允許使用單一訓練集 [96]。本論文在初步實驗時，發現若只使用單一資料集訓練，模型的泛化能力會再明顯降低。由於這樣的實驗設定會導致不同次訓練間存在著巨大的訓練變異，且無法彌補不同資料集標註方式的不一致性，較難清楚地比較方法差異，因此沒有採用。

接下來，本論文逐一討論第一章提出的研究問題：

1. 語音基石模型提供語者自動分段標記系統多大的幫助？

語音基石模型對語者自動分段標記系統的幫助主要呈現在準確率及泛用性的提升。首先在準確率上，語音基石模型在重疊語音基準、話語開始點偵測基準及語者自動分段標記系統的最終表現均明顯優於一般模型以及其他文獻。其次在泛用性上，語音基石模型在各任務上域外與域內情景的準確率差異遠比一般模型低。此外，語音基石模型在進行下游任務訓練時，固定預訓練模型的參數並以加權和抽取特徵往往能得到最好的表現。

2. 哪一個語音基石模型最適合被用於語者自動分段標記系統真實世界部署？

Whisper Tiny 編碼器最適合用於語者自動分段標記系統真實世界部署。在準確率上它的表現優於參數更多的 Hubert 模型，在效率上其 CPU 推論速度滿足部署成本考量。此外，Whisper Tiny 編碼器同時在情緒辨識任務及聲音事件偵測任務上有優秀的表現，亦能在加上解碼器後同時輸出自動語音辨識結果，符合語音基石模型同時處理多項任務的目標。

3. 在系統性能上，階段性系統是否明顯低於端到端-階段性混合系統？

傳統上一般認為階段性系統在準確率及效率皆不佳，但經過改良後的階段性系統準確率明顯提升。在重疊語音佔比較低的資料集上，階段性系統表現與端到端-階段性混合系統一致；只有在重疊語音佔比很高的資料集上，端到端才具有優勢。在推論速度上，階段性系統與端到端-階段性混合系統的差異主要在於外部語者特徵向量的使用。在不使用外部語者特徵向量的情況下，端到端-階段性混合系統的推論速度明顯較快，但在域外泛化能力較差，需要使用較強的基石模型才能擁有與階段性系統一致的域外泛化能力。因此，可得出階段性系統在效率上不亞於端到端-階段性混合系統的結論。

4. 在域外泛化能力上，階段性及端到端-階段性混合系統的表現分別如何？

在不修改原本系統的定義下，階段性系統的泛化能力優於端到端-階段性混合系統。在移除模型限制後，端到端-階段性混合系統的域外泛化能力與階段性系統大致相同。然而，在使用域內評估時，移除模型限制有可能會降低模型表現。

5.2 未來展望

本論文有三個未來發展方向:

1. 提升系統效率：語者特徵向量仍是階段性相關系統之推論速度瓶頸，若能汰除則能提升 2 倍推理速度。語音基石模型過去已被證實在語者驗證下游任務中表現優秀，可以在添加下游模型後直接用作語者特徵向量。然而其訓練方式與語者自動分段標記系統不太一致，需要進一步研究。

2. 延伸至端到端系統：語音基石模型於端到端-階段性混合系統取得不錯的表現，因此應用到端到端系統及以域外評估探討其泛化能力是重要的研究。但

是端到端模型架構複雜且有眾多變形，因此並未包括在本論文的討論範圍。

3. 結合聲音事件偵測：在聚類前先移除短句及重疊語音，可以達到聚類純化的效果並提升系統表現；然而在真實世界情景仍有更多干擾聚類的因素，如背景音樂、唱歌、笑聲等，結合聲音事件偵測有機會提升系統的穩健性。

總括而言，本論文詳細探討語音基石模型在語者自動分段標記系統上的應用，並以域外評估方式研究階段性及端到端-階段性混合系統各個模組的表現，類似的研究方法和技術仍有更多可以探索的方向。

參考文獻

[1] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts,

A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2022.

[4] H. Bredin. pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 2017.

[5] H. Bredin and A. Laurent. End-to-end speaker segmentation for overlap-aware resegmentation, 2021.

[6] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill. pyannote.audio: neural building blocks for speaker diarization, 2019.

[7] S. J. Broughton and L. Samarakoon. Improving end-to-end neural diarization using conversational summary representations, 2023.

[8] L. Bullock, H. Bredin, and L. P. Garcia-Perera. Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection, 2019.

[9] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In S. Renals and S. Bengio, editors, Machine Learning for

Multimodal Interaction, pages 28–39, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[10] H.-J. Chang, S. wen Yang, and H. yi Lee. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert, 2022.

[11] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S. wen Yang, Y. Tsao, H. yi Lee, and S. Watanabe. An exploration of self-supervised pretrained representations for end-to-end speech recognition, 2021.

[12] V. Chemudupati, M. Tahaei, H. Guimaraes, A. Pimentel, A. Avila, M. Rezagholizadeh, B. Chen, and T. Falk. On the transferability of whisper-based representations for ”in-the-wild” cross-task downstream speech applications, 2023.

[13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6):1505–1518, Oct. 2022.

[14] W. Chen, J. Huang, and T. Bocklet. Length- and Noise-Aware Training Techniques for Short-Utterance Speaker Recognition. In Proc. Interspeech 2020, pages 3835–3839, 2020.

[15] Z. Chen, B. Han, S. Wang, and Y. Qian. Attention-based encoder-decoder end-to-end neural diarization with embedding enhancer, 2023.

[16] G. Cheng, Y. Chen, R. Yang, Q. Li, Z. Yang, L. Ye, P. Zhang, Q. Zhang, L. Xie, Y. Qian, K. A. Lee, and Y. Yan. The conversational short-phrase speaker diarization (cssd) task: Dataset, evaluation metric and baselines, 2022.

[17] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman. Spot the conversation: speaker diarisation in the wild. In Interspeech, 2020.

[18] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

[19] T. Cord-Landwehr, C. Boeddeker, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach. Frame-wise and overlap-robust speaker embeddings for meeting diarization, 2023.

[20] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset. Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation, 2021.

[21] B. Desplanques, J. Thienpondt, and K. Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In Interspeech 2020, interspeech 2020. ISCA, Oct. 2020.

[22] J. Diliberto, C. Pereira, A. Nikiforovskaja, and M. Sahidullah. Speaker diarization with overlapped speech, 2021.

[23] Z.-C. Fan, Z. Bai, X.-L. Zhang, S. Rahardja, and J. Chen. Auc optimization for deep learning based voice activity detection. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6760–6764, 2019.

[24] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario, 2021.

[25] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe. End-to-end neural speaker diarization with permutation-free objectives, 2019.

[26] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe. End-to-end neural speaker diarization with self-attention, 2019.

[27] Y. Fujita, T. Komatsu, R. Scheibler, Y. Kida, and T. Ogawa. Neural diarization with non-autoregressive intermediate attractors, 2023.

[28] I. Fung, L. Samarakoon, and S. J. Broughton. Robust end-to-end diarization with domain adaptive training and multi-task learning, 2023.

[29] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree. Speaker diarization using deep neural network embeddings. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4930–4934, 2017.

[30] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass. Whisper-at: Noise-robust automatic speech recognizers are also strong audio event taggers. In Proc. Interspeech 2023, 2023.

[31] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao. A survey on self-supervised learning: Algorithms, applications, and future trends, 2023.

[32] E. Han, C. Lee, and A. Stolcke. Bw-eda-eend: streaming end-to-end neural speaker diarization for a variable number of speakers. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, June 2021.

[33] M.-K. He, J. Du, Q.-F. Liu, and C.-H. Lee. Ansd-ma-mse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:1561–1573, 2023.

[34] Y. He, Z. Kang, J. Wang, J. Peng, and J. Xiao. Voiceextender: Short-utterance text-independent speaker verification with guided diffusion model, 2023.

[35] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia. Encoder-decoder based attractors for end-to-end neural diarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:1493–1507, 2022.

[36] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, 2020.

[37] S. Horiguchi, S. Watanabe, P. Garcia, Y. Takashima, and Y. Kawaguchi. Online neural diarization of unlimited numbers of speakers using global and local attractors. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:706–720, 2023.

[38] S. Horiguchi, S. Watanabe, P. Garcia, Y. Xue, Y. Takashima, and Y. Kawaguchi. Towards neural diarization for unlimited numbers of speakers using global and local attractors, 2021.

[39] M. Hruz and M. Hlaváč. Lstm neural network for speaker change detection in telephone conversations, 09 2019.

[40] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[41] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve,

A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7669–7673, 2020. https://github.com/facebookresearch/libri-light.

[42] J. Kalda and T. Alumäe. Collar-aware training for streaming speaker change detection in broadcast speech, 2022.

[43] K. Kinoshita, M. Delcroix, and N. Tawara. Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech, 2021.

[44] K. Kinoshita, M. Delcroix, and N. Tawara. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds, 2021.

[45] K. Kinoshita, T. von Neumann, M. Delcroix, C. Boeddeker, and R. Haeb-Umbach. Utterance-by-utterance overlap-aware neural diarization with graph-pit, 2022.

[46] 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), pages 5220–5224, 2017.

[47] M. Kunešová, M. Hrúz, Z. Zajíc, and V. Radová. Detection of overlapping speech for the purposes of speaker diarization. In Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings 21, pages 247–257. Springer, 2019.

[48] M. Kunešová and Z. Zajíc. Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0. In ICASSP 2023 - 2023 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, June 2023.

[49] N. Kuzmin, I. Fedorov, and A. Sholokhov. Magnitude-aware probabilistic speaker embeddings. In The Speaker and Language Recognition Workshop (Odyssey 2022), odyssey 2022. ISCA, June 2022.

[50] Y. Kwon, H. S. Heo, J. Huh, B.-J. Lee, and J. S. Chung. Look who's not talking, 2020.

[51] Y. Kwon, J. weon Jung, H.-S. Heo, Y. J. Kim, B.-J. Lee, and J. S. Chung. Adapting speaker embeddings for speaker diarisation, 2021.

[52] F. Landini, M. Diez, T. Stafylakis, and L. Burget. Diaper: End-to-end neural diarization with perceiver-based attractors, 2023.

[53] F. Landini, J. Profant, M. Diez, and L. Burget. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks, 2020.

[54] M. Lebourdais, T. Mariotte, M. Tahon, A. Larcher, A. Laurent, S. Montresor, S. Meignier, and J.-H. Thomas. Joint speech and overlap detection: a benchmark over multiple audio setup and speech domains, 2023.

[55] J.-H. Lee, D. Yoon, B. Ji, K. Kim, and S. Hwang. Rethinking evaluation protocols of visual representations learned via self-supervised learning, 2023.

[56] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang. Cn-celeb: multi-genre speaker recognition, 2021.

[57] Y. Li, Z. Zhao, O. Klejch, P. Bell, and C. Lai. Asr and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition, 2023.

[58] H. Liu, J. Li, Y. Wu, and Y. Fu. Clustering with outlier removal, 2019.

[59] T. Liu, S. Fan, X. Xiang, H. Song, S. Lin, J. Sun, T. Han, S. Chen, B. Yao, S. Liu, Y. Wu, Y. Qian, and K. Yu. MSDWild: Multi-modal Speaker Diarization Dataset in the Wild. In Proc. Interspeech 2022, pages 1476–1480, 2022.

[60] T. Liu and K. Yu. Ber: Balanced error rate for speaker diarization, 2022.

[61] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention, 2023.

[62] Y. C. Liu, E. Han, C. Lee, and A. Stolcke. End-to-end neural diarization: From transformer to conformer. In Interspeech 2021, interspeech 2021. ISCA, Aug. 2021.

[63] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In 2018 International Seminar on Application for Technology of Information and Communication, pages 533–538, 2018.

[64] S. Mihalache, I.-A. Ivanov, and D. Burileanu. Deep neural networks for voice activity detection. In 2021 44th International Conference on Telecommunications and Signal Processing (TSP), pages 191–194, 2021.

[65] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaloe, T. N. Sainath, and S. Watanabe. Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing, 16(6):1179–1210, Oct. 2022.

[66] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman. Voxceleb: Large-scale speaker verification in the wild. Computer Science and Language, 2019.

[67] S. Otterson and M. Ostendorf. Efficient use of overlap information in speaker diarization. In 2007 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 683–686, 2007.

[68] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210, 2015.

[69] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan. A review of speaker diarization: Recent advances with deep learning, 2021.

[70] J. Patino, R. Yin, H. Delgado, H. Bredin, A. Komaty, G. Wisniewski, C. Barras, N. Evans, and S. Marcel. Low-Latency Speaker Spotting with Online Diarization and Detection. In Odyssey 2018, The Speaker and Language Recognition Workshop, pages 140–146, Les Sables d'Olonnes, France, June 2018.

[71] A. Plaquet and H. Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In INTERSPEECH 2023. ISCA, aug 2023.

[72] A. Poddar, M. Sahidullah, and G. Saha. Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics, 7(2):91–101, 2018.

[73] M. Przybocki and A. Martin. Nist speaker recognition evaluation (ldc2001s97). In New Jersey: Linguistic Data Consortium, 2001, 2000.

[74] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[75] D. Raj, D. Povey, and S. Khudanpur. Gpu-accelerated guided source separation for meeting transcription, 2023.

[76] M. Ravanelli and Y. Bengio. Speaker recognition from raw waveform with sincnet, 2019.

[77] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman. The second dihard diarization challenge: Dataset, task, and baselines, 2019.

[78] L. Samarakoon, S. J. Broughton, M. Härkönen, and I. Fung. Transformer attractors for robust and efficient end-to-end neural diarization, 2023.

[79] E. A. Schegloff. Overlapping talk and the organization of turn-taking for conversation. Language in Society, 29(1):1–63, 2000.

[80] E. Schubert. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. ACM SIGKDD Explorations Newsletter, 25(1):36–42, June 2023.

[81] L. Serafini, S. Cornell, G. Morrone, E. Zovato, A. Brutti, and S. Squartini. An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings, 2023.

[82] K. R. Shahapure and C. Nicholas. Cluster quality analysis using silhouette score. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pages 747–748, 2020.

[83] D. Snyder, G. Chen, and D. Povey. Musan: A music, speech, and noise corpus, 2015.

[84] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5329–5333, 2018.

[85] H. Tachibana. Towards listening to 10 people simultaneously: An efficient permutation invariant training of audio source separation using sinkhorn's algorithm. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, June 2021.

[86] S. Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2021.

[87] TencentGameMate. Chinese Speech Pretraining GitHub Repository. https://github.com/TencentGameMate/chinese_speech_pretrain, 2022. Accessed: October 2, 2023.

[88] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[89] U. von Luxburg. A tutorial on spectral clustering, 2007.

[90] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach. Graph-PIT: Generalized permutation invariant training for continuous separation of arbitrary numbers of speakers. In Interspeech 2021. ISCA, aug 2021.

[91] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 577–584, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[92] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[93] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking, 2023.

[94] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno. Speaker diarization with lstm, 2022.

[95] Q. Wang, Y. Huang, H. Lu, G. Zhao, and I. L. Moreno. Highly efficient real-time streaming and fully on-device speaker diarization with multi-stage clustering, 2023.

[96] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh. Learning to diversify for single domain generalization, 2023.

[97] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee. Superb: Speech processing universal performance benchmark, 2021.

[98] J. weon Jung, H.-S. Heo, Y. Kwon, J. S. Chung, and B.-J. Lee. Three-class overlapped speech detection using a convolutional recurrent neural network, 2021.

[99] J. weon Jung, S. Seo, H.-S. Heo, G. Kim, Y. J. Kim, Y. ki Kwon, M. Lee, and B.-J. Lee. Encoder-decoder multimodal speaker change detection, 2023.

[100] J. Wu, Z. Chen, M. Hu, X. Xiao, and J. Li. Speaker change detection for transformer transducer asr, 2023.

[101] Y. Xue, S. Horiguchi, Y. Fujita, S. Watanabe, and K. Nagamatsu. Online end-to-end neural diarization with speaker-tracing buffer, 2021.

[102] H. Yang, J. Zhao, G. Haffari, and E. Shareghi. Investigating pre-trained audio encoders in the low-resource condition, 2023.

[103] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, L. Xie, and Y. Yan. Open source magicdata-ramc: A rich annotated mandarin conversational(ramc) speech dataset, 2022.

[104] R. Yin, H. Bredin, and C. Barras. Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks. In Interspeech 2017, Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)., Stockholm, Sweden, Aug. 2017. ISCA.

[105] R. Yin, H. Bredin, and C. Barras. Neural speech turn segmentation and affinity propagation for speaker diarization. In Annual Conference of the International Speech Communication Association, Hyderabad, India, Sept. 2018.

[106] Z. Yin, J. Tian, X. Hu, X. Xu, and Y. Xiang. Large-scale learning on overlapped speech detection: New benchmark and new general system, 2023.

[107] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation, 2017.

[108] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge, 2022.

[109] N. Zeghidour and D. Grangier. Wavesplit: End-to-end speech separation by speaker clustering, 2020.

[110] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang. Fully supervised speaker diarization, 2019.

[111] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition, 2022.

[112] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Strohman, B. Ramabhadran, T. Sainath, P. Moreno, C.-C. Chiu, J. Schalkwyk, F. Beaufays, and Y. Wu. Google usm: Scaling automatic speech recognition beyond 100 languages, 2023.

[113] Özgür Çetin and E. Shriberg. Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition. In Proc. Interspeech 2006, pages paper 1915–Mon2A2O.6, 2006.