Speaker Diarization with Speech Enhancement

學生姓名：劉品彥 鄭元富　　指導老師：洪維廷

摘要

新冠病毒衝擊下，全世界的生活模式都有極大的改變，人與人的互動由實體改為線上，使線上通訊軟體需求量暴增。以線上會議為例，會議成員無法在同一空間內面對面對談，需要分神去關注目前是誰在說話，增加會議難度與注意力的消耗，後續的會議紀錄也顯得相對困難，此時，語者自動分段標記(Speaker Diarization)就能用來解決這項問題。

語者自動分段標記是一項用對應於語者身份的類別來標記音頻或視頻紀錄的任務，簡而言之，是一項識別「誰在何時說話」的任務。本文中，我們使用 34 層的 ResNet 為主幹架構，接上經常應用於臉部辨識的 GhostVLAD 層，以聚合語者特徵，進行端到端的模型訓練，產生語者的聲紋辨識嵌入碼。之後，採用一種完全受監督的分段標記方法，無界交錯狀態回歸神經網路(UIS-RNN)，對一段指定的音頻做分段標記。為了減輕聲音環境中的任何干擾，執行音頻分段標記前，會使用語音增強(Speech Enhancement) 對音頻進行優化，消除與人聲不相關的雜音，增加語者分段標記任務的效率與準確率。

1.Speaker Diarization

語者自動分段標記已經成為語音研究中一個重要的專業領域。“Diarize”是指在日記中紀錄或保存事件，而“Speaker Diarization”就像是在這樣的日記中紀錄事件一樣，通過在多人發言的音頻數據上記錄特定發言者的標記來解決"誰在何時說話"的問題。自動分段標記的過程中，音頻數據將被劃分和聚類(Clustering)為具有相同語者標籤的語段組。因此，非語音、語音轉換或語者轉換等特別事件，會被自動檢測出來。一般來說，此過程不需要任何音頻內語者的相關資料，像是他們的真實身分或是參語數量，就能達到自動分段標記的功能。

傳統的語者分段標記算法可以大致分成兩個部分：語者分割(speaker segmentation)和聚類(clustering)[1]，根據這兩部份的順序差異，大多數先進的語者自動分段標記可分為：自下而上的與自上而下的方法[2]。自下而上的方法被稱之為聚合式階層分群法(AHC)。首先將整個語音紀錄切割成較小的片段，且其中的每片段最好只來自同一個說話者。這些片段通過一些距離指標，例如貝氏信息量準則(BIC)，來選擇較為相關的片段，疊代合併，直到滿足某個停止準則。反之，自上而下的方法連續地把語音片段分割到新的集群中，直到達到說話者的數量。一般來說，自下而上的方法遠比自上而下的方法更受歡迎。研究[3]表明，i-vector 作為一個強大的矢量，不僅包含語者差異信息，同時也存在信道差異信息，將其引入分段標記任務以增強語者的具體信息。此外，概率線性鑑別分析評分法(pLDA)被學習來區分兩個 i-vector 是否來自同一個人。而在[4]中指出，相較於 i-vector，使用神經網路嵌入的 d-vector 可以顯著提高分段標記的性能，這主要是由於神經網路可以用大量的數據集進行訓練，這樣的模型對不同使用場景中不同說話人口音和升學條件有足夠的魯棒性(robustness)。

除了上述的分段過程，一個實用的語者自動分段標記系統應該包含預處理階段，其中包括語音去噪、多聲道聲束成形和語音活動檢測，去除背景噪音、混響和其他干擾。在真實場景中，背景噪音、混響和其他干擾會極大幅度地損害整體的分段性能。因此，在整個任務過程中，累積的誤差將變得不受控制和無法追蹤。特別是在空間信息有限的單通道情況下，一個有效的語音去噪方法，作為前端的預處理，將發揮重要作用。在[5]中，證明了與傳統方法相比，基於深度學習的去噪方法在應對現實的噪音環境方面具有更強的潛力。複雜的聲學環境也會影響語音活動的檢測，這對於分段標記來說相當重要，通過良好的預處理，更好的語音質量和更準確的語音邊界定位可以確保任務的性能有一個更高的上限。

2.模組化自動分段標記系統

本節的目的是將語者分段標記任務模組化，對每個模塊的目的與功能做出相對應的解釋。如圖 1 所示，進行分段標記任務之前，會先將音頻進行語音增強，去除音頻內與人聲無關的環境雜音，執行分段標記時能更加專注於語者的分析與標記。語者自動分段標記會經過分割、聲紋嵌入碼與聚類分析三個階段，最後輸出標記完成的結果。

2.1 語音增強(Speech Enhancement)

在現實世界的聲學環境進行準確的語音處理，往往需要自動的語音增強。由於此課題對語音處理技術的重要性，人們已經提出許多方法來解決這一問題。以前大多數語音分離(增強)方法都是在混和信號的時頻中制定，用短傅立葉轉換(STFT)從波形中估計出來[6]的。雖然在時頻中處理仍然是最常用的語音方離方法，但這種方法有幾個缺點。第一，STFT 是一種通用的信號變換，對語音分離來說不

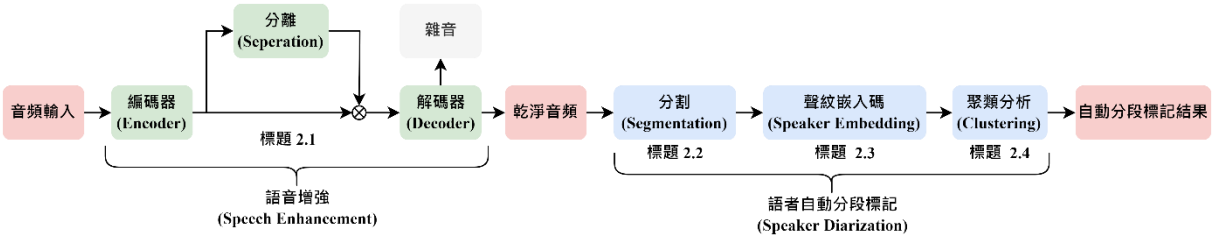

圖 1. 語者自動分段標記系統

一定是最佳的；第二，準確地重建乾淨來源的相位是一項難以解決的問題，對相位的錯誤估計會對重建音頻的準確性引入一個上限。即使在音頻的混和過程中，加入理想的乾淨音訊頻譜，這個音源的不完美重建導致的問題仍然非常明顯；第三，從時頻表示中成功分離需要對混和信號進行高分辨率的頻率分解，這需要一個長時間的窗口來計算 STFT，這一要求增加系統的最小延遲，從而限制其在即時、低延遲應用中的適應性，如線上語音通訊與線上會議等。因為這些議題是由在時頻域中制定分離問題而產生的，所以一個合理的方法是避免將聲音的大小和相位分開處理，直接在時域制定分離。我們將使用時域音頻分離網路(TasNet) [7]進行語音分離，在 TasNet 中，混和波形是用卷積編碼器與解碼器架構來建模，該架構由一個對其輸出有非負性約束的編碼器和一個用於將編碼器輸出反轉為聲音波形的線性解碼器組成。TasNet 取得了比以前各種時頻域系統更好的性能，顯示其有效性和潛力，但在最初的 TasNet 中使用深度長短期記憶(LSTM)網路作為分離模塊明顯限制其適用性。首先，在編碼器中選擇較小的核(kernel)會增加編碼器輸出的長度，使 LSTM 的訓練難以管理；第二，深層 LSTM 網路中的大量參數增加計算的成本，限制它對低資源、低功率平台的適應性。所以我們決定採用由[8]提出的完全卷積 TasNet(Conv-TasNet)，Conv-TasNet 使用堆疊的 1-D 卷積塊來代替深度 LSTM 網路進行分離步驟，卷積的使用對連續的幀或片段進行並行處理，從而加快分離過程，也能降低模型大小。Conv-TasNet 包括三個處理階段，如圖 1 語音增強部分所示：編碼器、分離和解碼器。首先，編碼器(Encoder)用於將混和波形的短片段轉換為其在特徵空間的相應表示式，這個表示式被用來估計每個時間步長的音源乘法函數(Mask)。通過使用解碼器(Decoder)對屏蔽的編碼器特徵進行轉換，重建來源波形。

Conv-TasNet 的分離(Seperation)模塊是受時間卷積網路(TCN)的啟發而成，TCN 被提出來作為各種序列建模任務中 RNN 的替代品，每一層都由一維擴張卷積塊的堆疊組成，對應的擴張因子以指數形式增加，以確保有足夠大的時間背景窗口來利用語音信號的長距離依賴性。而在堆疊一維卷積塊的設計過程中，應用了殘差(residual)路徑和跳過連結(skip-connection)路徑。殘差路徑做為下一個區塊的輸入，且所有區塊的跳過連接路徑被加起來，作為 TCN 的輸出。為了進一步減少參數的數量，深度可分離的卷積層(S-conv(.))被用來取代每個卷積塊中的標準卷積。深度可分離卷積已被證明在圖像處理任務[9][10]和神經機器翻譯任務[11]是有效的。

2.2 分割(Segmentation)

對語者自動分段標記任務而言，語音分割(speech segmentation)是將輸入的音頻切割成多個片段，獲得語者均勻片段的過程。因此，語者自動分段標記系統的輸出單位是通過分割過程確定的。一般情況下，用於分段標記的分割方法分為兩大類：通過偵測語者轉換點的分割與均勻分割。

通過偵測語者轉換點進行的分割是早期自動分段系統的黃金準則，其中語者轉換點是通過比較兩種假設來檢測的。H_0 假設左右兩個語音窗口都來自同一個語者，而 H_1 假設兩個語音窗口都來自不同語者。為了測試這兩個假設，基於度量(metric-based)的方式[12]與[13]被最廣泛地應用。在基於度量的方法中，語音特徵的分布被假定為遵循高斯分布 $N(\mu, \Sigma)$，平均值為 μ，協方差為 Σ，則兩個假設 H_0 和 H_1 可以表示如下：

$$\begin{aligned} H_0 &: \boldsymbol{x}_1 \cdots \boldsymbol{x}_N \sim N(\mu, \Sigma), \\ H_1 &: \boldsymbol{x}_1 \cdots \boldsymbol{x}_i \sim N(\mu_1, \Sigma_1), \\ &\quad \boldsymbol{x}_{i+1} \cdots \boldsymbol{x}_N \sim N(\mu_2, \Sigma_2) \end{aligned} \qquad (1)$$

其中 $(\boldsymbol{x}_i \mid i = 1, \cdots, N)$ 是一連串有利於假設驗證的語音特徵。為了量化兩個假設的相似性，人們提出了一系列基於度量的標準。如果使用偵測語者轉換點的方法進行語音分割，每個片段的長度是不一致的，因此，在 i-vector 出現和 DNN 的嵌入後，基於偵測語者轉換點的方法大多被替換成均勻分割，改善不同長度的分割在語者表示中產生額外變化，使語者表示的保真度下降的問題。

2.3 聲紋嵌入碼(Speaker Embedding)

語者代表(Speaker representation)在分段標記任務中起到至關重要的作用，用來衡量語音片段之間的相似度。在 i-vector 或 x-vector 等語者代表法出現之前，時常將基於高斯混和模型的通用背景模型(GMM-UBM)[14]應用於聲學特徵，它將空間分布的概率密度用多個高斯概率密度函數的加權來擬合，可以平滑的逼近任意形狀的概率密度函數，並且是一個易於處理的參數模型，具備對實際數據極

強的表徵力。但反過來說，GMM 規模越龐大，表徵力越強，其負面效應也會越明顯，參數規模也會等比例的膨脹，需要更多的數據來驅動 GMM 的參數訓練才能得到一個更加通用的 GMM 模型。UBM 由一個大型的 GMM(通常有 512 到 2048 個混合物)組成，經過訓練，可以代表獨立於說話人的聲學特徵分布。因此，一個 GMM-UBM 模型可以用以下參數來描述：混和物的權重、平均值和協方差矩陣。在 GMM-UBM 系統框架中，UBM 擬合出大量說話人的特徵分布，目標用戶的數據散落在 UBM 某些高斯分布的附近。其中自適應的過程就是將 UBM 的每個高斯分布向目標用戶數據偏移，而最大後驗概似估計值(MAP)就是解決這種問題的計算方法之一。

在傳統的 GMM-UBM 識別系統中，由於訓練環境和測試環境的失配問題，會導致系統性能不穩定。之後，聯合因子分析(JFA)[15][16]被提出來，通過對語者差異性和信道差異性分別建模，來彌補差異性問題。在 JFA 方法中，給定的 GMM 超矢量(supervector)被分解為與語者相關的超矢量，表示說話人之間的差異；信道相關的超矢量，表示同一說話人不同語音片段的差異。

在 JFA 方法之後，[17]的研究發現，JFA 模型中，信道因子也會攜帶部分語者的信息，在進行補償時，會損失一部分語者信息。因此，Dehak 等人提出了全局差異空間模型，將語者差異和信道差異作為一個整體進行建模。這種方法改善了 JFA 對訓練語料的要求，和計算複雜度高的問題，同時性能也與 JFA 相當，逐漸流行起來。給定語者的一段語音，與之相應的高斯均值超矢量可以定義如下：

$$M = m + T\omega \qquad (2)$$

其中，M 為語音的高斯均值超矢量，m 為 UBM 的高斯均值超矢量(與語者及信道無關)；T 為全局差異空間矩陣；ω為全局差異空間因子，被稱為 i-vector，也被認為是語者代表向量。相對於 JFA 而言，i-vector 的計算量大大降低，可以應對大規模數據，同時因為 i-vector 本身具有不錯的跨信道能力和 pLDA 信道補償法的引入，i-vector 的魯棒性也比 JFA 更好。通過使用 i-vector，語者代表的想法得到極大的響應，其中語者代表的向量可以描述每個說話人聲道的數字特徵。它不僅被用於語者識別研究，也被用於語者自動分段標籤的研究，並且與上面提到的基於度量法、GMM 與 JFA 相比，顯示出更優越的性能。

在本文中，我們使用由[18]提出的語者識別深度網路來產生聲紋嵌入碼。早期深度神經網路的語者識別系統採用了在視覺識別任務中獲得成功的均勻池化層(average pooling)或全連接層(fully connected layers)，將幀級(frame-level)層面的信息濃縮到話語級(utterance-level)層面的表示中。儘管這些方法的目的是將幀級信息聚集到一個單一的表示式中，而且這些表示式仍然允許反向傳播，但聚合(aggregation)不依賴於內容，導致神經網路不能考慮輸入信號的哪些部分包含最相關的信息。於是，作者建議將卷積神經網路和基於字典的 NetVLAD[19]層結合起來，前者是捕捉局部模式的方法，而後者可以通過鑑別性的訓練，將信息從任意大小的輸入彙總到一個固定大小的記述子(descriptor)中，從而使話語的最終表示不受無關信息的影響。

對於語者識別，理想的模型應該具有以下特性：(1)它應該能夠接受任意時間長度的輸入，並產生一個固定長度的話語級描述；(2)輸出的記述子應該是低維度的；(3)輸出的記述子應該具有鑑別力，例如能夠分辨出不同語者記述子之間的距離比相同語者記述子之間的距離還大。為了滿足上述特性，[18]使用一個修改過的 ResNet，以完全卷積的方式對輸入的二維頻譜進行編碼，然後用 GhostVLAD 層沿時間軸進行特徵聚合。此方法能產生一個固定長度的記述子，而且是可訓練的判別性聚類(每一個幀級記述子都將被分配到不同簇，而殘差被編碼為輸出特徵)。為了實現高效的驗證方法，作者進一步增加一個全連接層，用於降維。網路結構分為兩個部分，第一部分是特徵提取，從輸入頻譜中提取特徵。在此階段作者採用一個有 34 層的改良 ResNet，與標準的 ResNet 相比，減少了每個剩餘模塊中的通道數量，使其成為一個薄型的 RestNet。第二部分我們採用[20]內提出的 GhostVLAD 層，作者對 NetVLAD 的架構進行擴展，以包括幽靈群組。它使網路能夠學會忽略沒有信息量的語音片段，把它們分配到被丟棄的幽靈聚類中。因此，在匯集總幀級特徵時，語音片段的噪聲和不良部分在正常的 VLAD 集群會被降低權重，使網路在訓練過程中不會受雜訊影響。

2.4 聚類分析(Clustering)

根據上一節的介紹，應用聲紋嵌入碼執行語者代表任務後，需要以聚類分析將語者代表和模擬相似性方法對語音片段進行聚類。本文中，我們使用[4]提出的無界交錯狀態 RNN(UIS-RNN)，UIS-RNN 模型以 RNN 取代傳統的聚類模塊。在大多數的語者分段標記系統，仍有一個組件是無監督式的聚類模塊，其算法包括高斯混和模型、均值飄移(mean shift)[21]、k-均值聚類(k-mean)與譜分群(spectral clustering)。然而，RNN 為監督式學習，代表我們可以藉由增加訓練集來提升訓練效果，達到資料訓練的最佳化。UIS 的 U 代表無界(Unbounded)，不需要提前知道說話者的數量，可由模型的訓練得到；IS 代表(Interleaved-state)，不同的說話者有不同的 RNN 狀態，在相同的時間軸上交錯運行。圖 2 為 UIS-RNN 的模型生成過程，顏色代表說話者語音片段的標籤。在辨識之初，給定一個初始化的參數 h_0，經由第一個語音向量得到關於第一位說話者的隱藏層參數 h_1，輸出時同時給第一位說話者標記顏色，

當後面出現之前出現的說話者時，更新該說話者的隱藏層參數，如果出現新的說話者，重新訓練，得到新的隱藏層參數。當有一段新的語音輸入時，分別以當前說話者的代表隱藏層計算說話者變換的機率，以 y_7 預測為例，這時會出現三大種情況：(1)仍是黃色說話者說話；(2)變回藍色或紅色說話者說話；(3)一個全新的綠色說話者。模型分別計算這三種情況的機率值，取最大的值為新的結果。

給定一個說話者，我們將從嵌入碼模塊中取得一個觀察序列$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T)$，序列的每一個數值都是對應於原始語者說話片段的 d-vecotr。之後，我們會取得基準真相(ground truth)的語者標籤序列$\boldsymbol{Y} = (y_1, y_2, \cdots, y_T)$。以圖 2 為例，經過模型的生成過程後，會產生$\boldsymbol{Y}$ = (1,1,2,3,2,2)，代表有 6 個語音片段，且整段語音中有 3 位說話者。其中，模型計算三種情況的機率函數為：

$$p(\boldsymbol{x}_t, y_t, z_t | \boldsymbol{x}_{[t-1]}, y_{[t-1]}, z_{[t-1]}) \quad (3)$$
$$= p(\boldsymbol{x}_t | \boldsymbol{x}_{[t-1]}, y_{[t]}) \cdot p(y_t | z_t, y_{[t-1]}) \cdot p(z_t | z_{[t-1]})$$

Z 為說話者轉換的二元指標，如果說話者改變就為 1，沒有改變就為 0。機率函數分為三部分，第一部分是：

$$p(\boldsymbol{x}_t | \boldsymbol{x}_{[t-1]}, y_{[t]}) \quad (4)$$

基於 RNN 在不同說話者皆有不同狀態的特性，產生不同序列的機率，在 UIS-RNN 中，是使用 GRU 來記憶長期的序列相關性。第二部分是：

$$p(y_t | z_t, y_{[t-1]}) \quad (5)$$

此段函數表示換回之前曾出現過的說話者的機率，如果某一個說話者時常打斷他人，加入會話，代表他再次加入的機會較高，機率較大。第三部分則為說話者交換的機率。最後，使用 MAP 法則進行解碼，假設模型收到一個測試序列 Xtest，理想的情況下，模型要找到：

$$\mathbf{Y}^* = \arg_{\mathbf{Y}} \max \ln(p(\mathbf{X}^{\text{test}}, \mathbf{Y})) \quad (6)$$

由此結果判斷出最有可能的說話者。

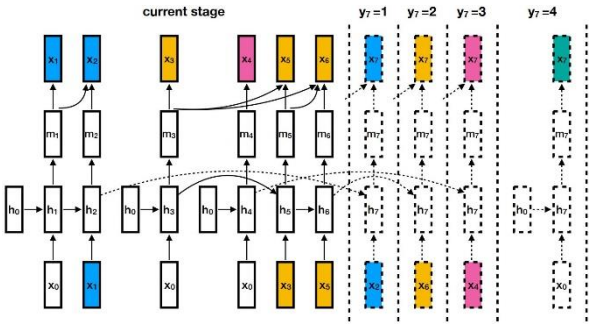

圖 2. UIS-RNN 的模型生成過程

3.實驗

3.1 資料集

我們使用 AISHELL-1[22]為中文的資料集，總共 178 小時，400 個說話者，其中包含訓練集 340 人，測試集 20 人，驗證集 40 人，每個人大概講三百多句話，每個說話者講的話都放在同個資料夾裡面。在英文方面，我們使用 VoxCeleb1 與 VoxCeleb2 為資料集。VoxCeleb1 包含 1251 個說話者，超過十萬個語音片段，VoxCeleb2 包含 6112 個說話者，超過一百萬個語音片段。

3.2 過程

首先，先訓練語者嵌入碼生成模型，採用薄型 ResNet34s 接上 GhostVLAD，其中包含十個 VLAD 聚類與兩個幽靈聚類，損失函數為採用 softmax 的 adam，特徵維度為 512。訓練完成後，使用語者嵌入碼生成模型產生中文與英文資料集內說話者的嵌入碼資料集，並用此資料集訓練分段標記模型。分段標記模型為 UIS-RNN，他的序列生成模型由一層 512 個具有 tanh 激活函數的 GRU 單元組成，接上兩層完全連接層，每個完全連接層有 512 個節點與 ReLU 激活函數。分段標記模型的實驗流水線為：(1)載入說話者嵌入碼資料集，(2)訓練模型，(3)測試模型，(4)輸出實驗結果。最後，我們使用 YouTube 上的新聞報導當作測試，並將分段標記的結果視覺化。

3.3 實驗結果

表格 1.UIS-RNN 訓練結果

資料集	Learning rate	Batch size	accuracy
AISHELL	0.001	16	**0.897**
VoxCeleb1			0.858
VoxCeleb2			0.409

圖 3. #iNEWS 最新普丁出招反制西方制裁! "不友善國家"購買天然氣須以盧布支付 德法怒批敲詐勒索 並為俄羅斯天然氣斷供做準備【國際局勢】20220401｜三立 iNEWS

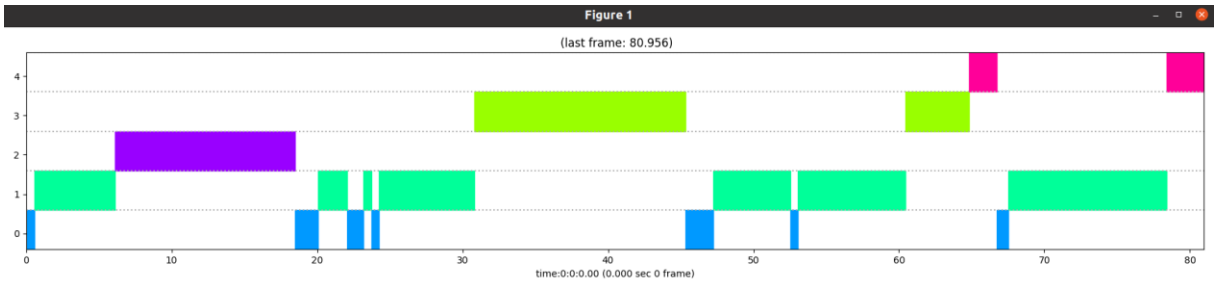

圖 4.依影片時間將自動分段結果視覺化

表格 2.模型測試結果之 EER

訓練資料集	測試資料集	gVLAD	EER
AISHELL	VoxCeleb1	2	0.297
VoxCeleb1	VoxCeleb1		**0.283**
VoxCeleb2	VoxCeleb1		0.420

```
========= 0 =========
0:00.64 ==> 0:00.564
0:18.496 ==> 0:20.72
0:22.72 ==> 0:23.168
0:23.732 ==> 0:24.232
0:45.317 ==> 0:47.233
0:52.553 ==> 0:53.53
1:06.716 ==> 1:07.536
========= 1 =========
0:00.564 ==> 0:06.128
0:20.72 ==> 0:22.72
0:23.168 ==> 0:23.732
0:24.232 ==> 0:30.808
0:47.233 ==> 0:52.553
0:53.53 ==> 1:00.437
1:07.536 ==> 1:18.420
========= 2 =========
0:06.128 ==> 0:18.496
========= 3 =========
0:30.808 ==> 0:45.317
1:00.437 ==> 1:04.801
========= 4 =========
1:04.801 ==> 1:06.716
1:18.420 ==> 1:20.956
```

圖 5.自動分段標籤

```
Finished diarization experiment
Config:
  sigma_alpha: 1.0
  sigma_beta: 1.0
  crp_alpha: 1.0
  learning rate: 0.001
  regularization: 1e-05
  batch size: 10

Performance:
  averaged accuracy: 0.822673
  accuracy_numbers for all testing sequences:
    0.779661
    0.866071
    0.786885
    0.641921
    0.905405
    0.802721
    0.857143
    0.747368
    0.851852
    0.882979
    0.711111
    0.758621
    0.682028
    0.888889
    0.794643
    0.602941
    0.708861
    0.716346
    0.786517
    0.835443
    0.880435
    0.731343
    0.831325
    0.913043
    0.762376
    0.862069
    0.758170
    0.816901
    0.859375
    0.711864
    0.800000
    0.884211
```

圖 6.模型訓練結果

4.結論

本文提出一個模組化的語者自動分段標記系統，能應用於各種語音音頻的分段標記。一開始藉由語音增強去除雜訊，嵌入碼生成模型產生不同語系和語者的嵌 入碼，再由分段標記模型標籤音頻，達成辨識誰在何時說話的任務。

語者自動分段標記系統已經在許多領域得到應用，包括會議記錄、對話分析、音頻索引和對話式人工智能系統。 然而，仍有許多可以更加完善的地方，像是配合語者嵌入碼，不只分辨出有幾個不同的說話者，還能辨識出說話者的身分與各種資料；不只對音頻進行分段，還能辨識出段落內說了什麼。相信在未來的語者分段標記技術能帶給人類更便利的生活，對於諸多創新技術會越來越成熟，甚至發展出更多相關應用。

5.參考文獻

[1] L. Sun, J. Du, C. Jiang, X. Zhang, S. He, B. Yin, and Chin-Hui Lee, “Speaker Diarization with Enhancing Speech for the First DIHARD Challenge,” University of Science and Technology of China, Hefei, Anhui, P. R. China, 2018.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker diarization: A review of recent research,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[3] S. Madikeri, I. Himawan, P. Motlicek, and M. Ferras, “Integrating online i-vector extractor with information bottleneck based speaker diarization system,” Idiap, Tech. Rep., 2015

[4] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C.Wang, “Fully supervised speaker diarization,” Google Inc., USA, Columbia University, USA, 2019.

[5] L. Sun, J. Du et al., “A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions,” in *ICASSP*, 2018.

[6] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[7] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[8] Y. Luo, N. Mesgarani, “Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation,” *arXiv preprint*, 2019

[9] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” *arXiv preprint*, 2016

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang,T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[11] L. Kaiser, A. N. Gomez, and F. Chollet, “Depthwise separable convolutions for neural machine translation,” *arXiv preprint arXiv:1706.03059*, 2017.

[12] S. Chen, P. Gopalakrishnan, et al., “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in: *Proceedings DARPA broadcast news transcription and understanding workshop*, volume 8, Virginia, USA, 1998, pp. 127–132.

[13] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, “Strategies for automatic segmentation of audio data,” in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, 2000, pp. 1423–1426.

[14] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital signal processing 10 (2000) 19–41.

[15] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, “Speaker and session variability in gmm-based speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing 15* (2007) 1448–1460.

[16] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, “A study of interspeaker variability in speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing 16* (2008) 980–988.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing 19* (2011).

[18] W. Xie, A. Nagrani, J.-S. Chung, and A. Zisserman, “Utterance-level aggregation for speaker recognition in the wild,” *arXiv preprint*, 2019

[19] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and ´ J. Sivic, “NetVLAD: CNN architecture for weakly supervised place recognition,” in Proc. CVPR, 2016

[20] Y. Zhong, R. Arandjelovic, and A. Zisserman, ´ “GhostVLAD for set-based face recognition,” in Asian Conference on Computer Vision, ACCV, 2018.

[21] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, “A study of the cosine distance-based mean shift for telephone speech diarization,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 217–227, 2014

[22] H. Bu, J. Du, X. Na, B. Wu, H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” *arXiv:1709.05522*, 2017