The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 119-135

基於 i-vector 與 PLDA 並使用 GMM-HMM 強制對位之自動語者分段標記系統

Speaker Diarization based on I-vector PLDA Scoring and using GMM-HMM Forced Alignment

張乘若 Cheng-Jo Ray Chang[1]

李鴻欣 Hung-Shin Lee[2]

王新民 Hsin-Min Wang[2]

張智星 Jyh-Shing Roger Jang[1]

1 國立台灣大學資訊工程學系 Department of Computer Science and Information Engineering, National Taiwan University

2 中央研究院資訊科學研究所 Institute of Information Science, Academia Sinica

摘要

近年來，i-vector 搭配 PLDA（Probability Linear Discriminant Analysis）的系統已經在自動語者分段標記（Speaker Diarization）的研究上獲得了很好的結果。不過，由於 i-vector 需要由較長的音訊片段抽取出來才具有較佳的語者特性，所以較無法有效地處理時間極短的語句區段。為此，本論文提出一個新的自動語者分段標記框架：先由 K 平均（K-means）演算法得到初步的自動語者分段標記結果，並據此建立初步語者模型，再配合利用 GMM-HMM（Gaussian Mixture Models-Hidden Markov Models）進行強制對位（Forced Alignment）以及語者分群（Speaker Clustering）來進行自動語者分段標記。從實驗上我們可以發現，雖然單獨利用 GMM-HMM 語者分群並未比使用 GMM-HMM 強制對位所得到的召回率（Recall）以及精準率（Precision）來得好，但是利用 GMM-HMM 語者分群的結果再重新進行 GMM-HMM 強制對位卻可以得到較好的召回率以及精準率，故由 GMM-HMM 語者分群以得到更細小的語者說話區段對自動語者分段標記的問題是有幫助的。此外，這篇論文也探討針對不同時間長度的音訊片段對自動語者分段標記的影響。

關鍵字：自動語者分段標記，I-vector，PLDA，GMM-HMM，強制對位，語者分群

一、緒論

隨著時代不斷的演進，人們在處理語音的技術也愈來愈成熟。就拿語者辨識（Speaker Recognition）的領域來講，從當初使用藉著高斯混合模型（Gaussian Mixture Models, GMM）[1] 來建立廣義背景模型（Universal Background Models, UBM），及至聯合因素分析（Joint Factor Analysis, JFA），到目前最廣為流行的 i-vector [2][3]，在建立特定的語者模型上，其準確性已經有相當幅度的提升。然而，有時候我們不需要知道每一句對話是出自哪一位語者，因為在某些情境中，只有某一位語者是最重要的，而其他人的聲音相對上並沒有那麼關鍵。例如，在追蹤嫌疑犯的犯罪錄音中，我們只需要關注嫌疑犯的聲音，而其餘在錄音中出現的人聲就沒有辨識其身分的必要，只需要給予他們語者識別（Speaker Identity）即可。一般而言，我們會將這類只需把不同語者以語者識別的方式標記下來的問題統稱為自動語者分段標記（Automatic Speaker Diarization）的問題，而這種問題又可被稱為「Who Spoke When」，也就是要將一段錄音資料中的語者區分出來，並一一標示他們的身分識別以及時間戳記（Time Stamp）[4]。在本篇論文中，我們會將身分識別以及時間戳記統稱為語者區段（Speaker Region）。廣義來說，自動語者分段標記的問題主要會分為兩種類型，一種是會議錄音（Conference Recordings），另一種則為廣播新聞（Broadcast News）[5]。這兩種情境最大的差別在於，廣播新聞可以是預先演練過的，所以實際的錄音情境可能是許多語者一個接著一個討論議題；相對地，會議錄音中參與者的發言具有較高自發性，所以語者跟語者的對話可能在時間上會有重疊，在會議錄音當中也有可能會出現拍手，笑聲等情況出現，而在這篇論文所探討的情境是介於這兩種類型之間的電話語音，主要針對客服與客戶的電話錄音，因此在一般情況下只有客服與客戶兩位語者。

一般在處理自動語者分段標記的問題會涉及三個步驟：1）將錄音資料切割成許多音訊片段，我們希望在每一個音訊片段內只包含一位語者的聲音；2）對切割好的音訊片段進行語者分群，這是自動語者分段標記的問題中最為關鍵的步驟。傳統的自動語者分段標記是處理未知語者數目的語音紀錄，所以在針對語者分群的問題中，最首要的問題是「究竟有幾位語者」。最廣為人知的方法是，我們先假設一個足夠多的語者數

目，對這些語者建立簡單的語者模型，接著根據它們彼此之間的異同，嘗試去合併兩兩語者的模型，直到找到最佳的語者數目；3）對於每一個群集都給予一個語者識別，並記錄語者區間，最後再與正確標記（Ground Truth）比對。

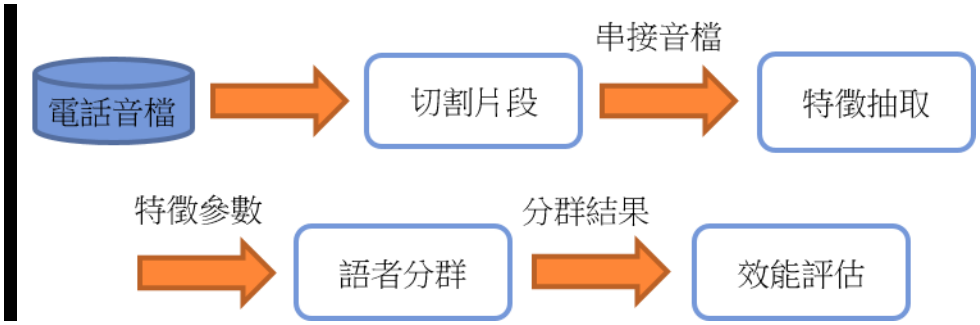
因為我們會將錄音記錄切割成許多只包含單一語者的音訊片段，所以在進行語者分群的步驟時，可以視為對每個音訊片段進行一連串的語者辨識。在這個想法之下，我們只需判斷兩個音訊片段是否出自於同一位語者 [6]。目前在語者辨識的研究中，i-vector 技術已經相當成熟且被廣泛使用，它能將不同長度的語音轉換成一個具有相同維度的向量且可以保留其中的語者資訊，並將音訊的通道噪音（Channel Noise）濾除。因此，在自動語者分段標記中，我們的主要任務是對所有音訊片段抽取其 i-vector，並計算所有 i-vector 彼此之間的 PLDA（Probability Linear Discriminant Analysis）分數– 也就是二者的聯合機率（Joint Probability）[7]，而得到一個自相似矩陣（Self-similarity Matrix），矩陣內的數值為兩個不同 i-vector 彼此間的 PLDA 分數。也就是說，PLDA 是用來計算這兩個 i-vector 是否來自同一位語者的評量標準，值愈高表示為同一位語者的可能性愈大，值愈低則表示這兩個音訊片段的語者為不同人的機率較小。在本篇論文中，我們會探討如何使用 i-vector 搭配 PLDA 來解決雙語者之自動語者分段標記問題，並且探討得到語者分群後的語者區域透過強制對位，以及再進行 GMM-HMM 語者分群是否有助於提升自動語者分段標記的召回率以及精準率。

以下為本論文的結構說明：我們將在第二節中介紹雙語者之自動語者分段標記的系統架構；在第三節中介紹我們使用的資料集，以及評估自動語者分段標記的標準；第四節將說明實驗結果與數據分析；在第五節我們將為本文做結論。

二、雙語者之自動語者分段標記系統架構

雙語者之自動語者分段標記的系統流程如圖一所示，主要分為四個部分: 第一部分負責切割片段（Segmentation），將電話語音切割成許多只包含一位語者的說話片段，之後將這些語者片段串接成一個串接音檔；第二部分為特徵參數的抽取（ Feature

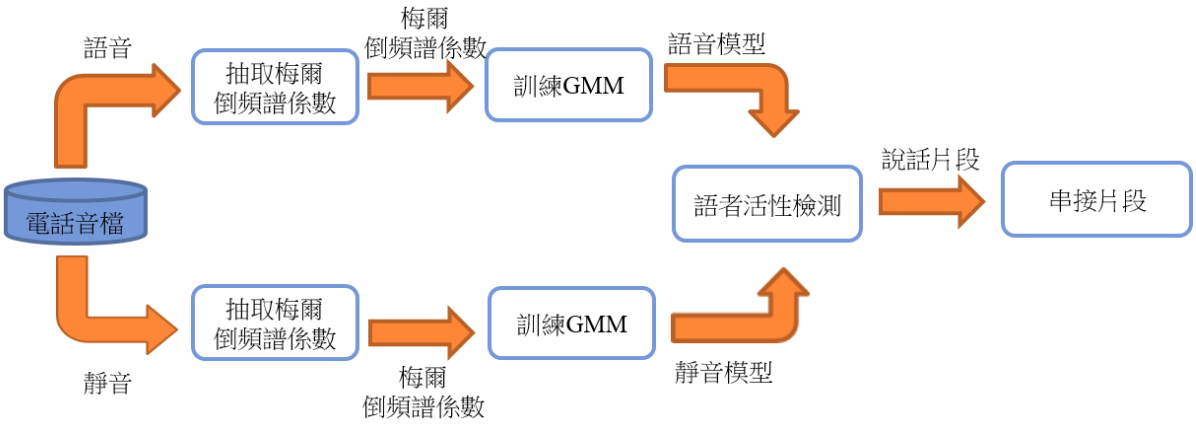
Extraction）；第三部分則是進行語者分群（Speaker Clustering）；第四部分則為系統效能的評估方式。這四個部分將會在本章的各個小節中一一介紹。其中特徵參數的部分我們是採用以下兩個特徵，分別是梅爾倒頻譜係數（Mel-Frequency Cepstrum Coefficients, MFCC），以及 i-vector。最後在進行語者分群時，我們提出了一個系統效能的評估機制，希望藉由此評估機制來判斷分群後的結果與正確標記的差距。



圖一、雙語者之自動語者分段標記系統流程。

（一）切割片段

為了從電話語音中得到只包含一位語者的說話片段，我們預先訓練一組靜音模型（Silence Model）以及語音模型（Speech Model）。首先從 50 句電話語音中針對靜音以及語音的片段抽取 13 維梅爾倒頻譜係數，用來訓練具有 32 個成分（Component）的 GMM。有了這兩個模型之後，就可以針對每個測試音檔透過此高斯混合模型進行語音偵測（Voice Activity Detection, VAD）而得到許多只包含一位語者的說話片段，最後再將這些片段串接成一個新的音檔，目的是為了之後在進行自動語者分段標記時可以將時間極短的段落（介於 0 至 1 秒）標示出來。

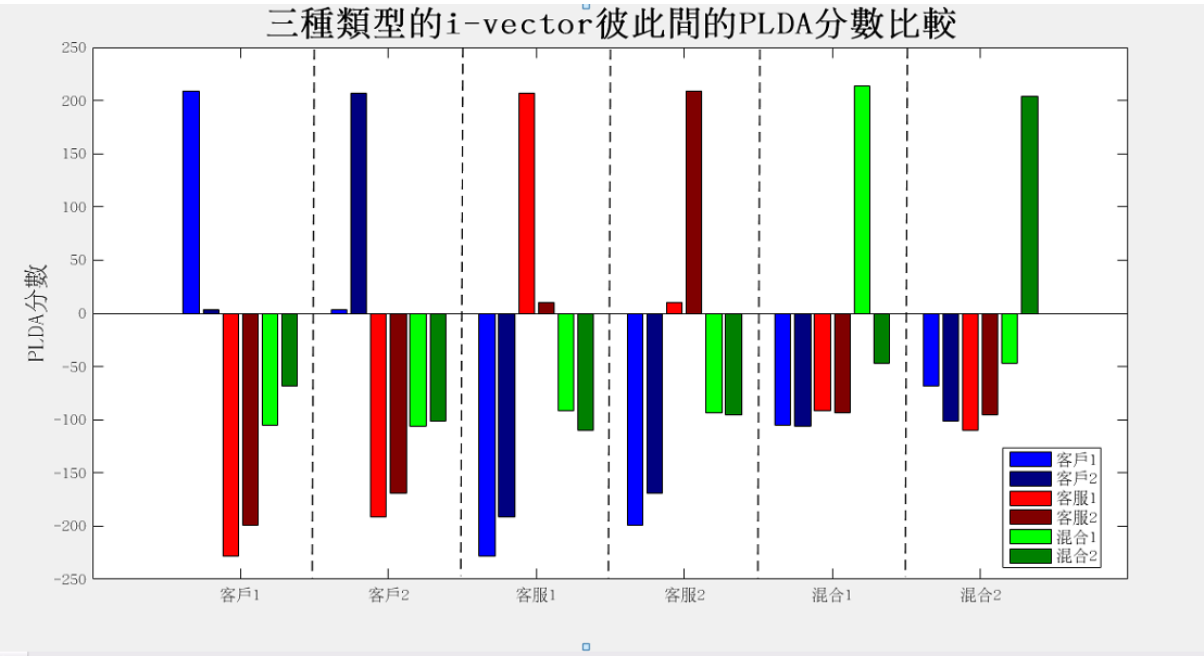


圖二、切割片段的流程。

（二）特徵抽取

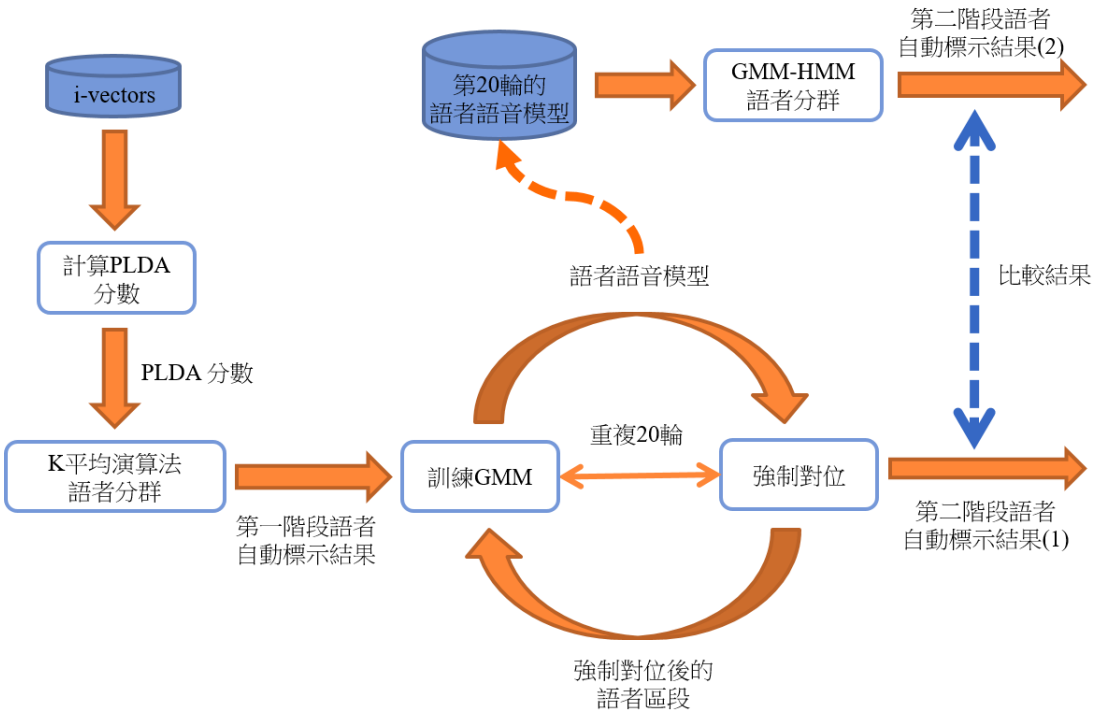
圖三、特徵抽取的流程。

接著，我們將上一小節得到的串接音檔切割成許多相同長度並且部分重疊的音訊片段 [8]，並對這些音訊片段抽取 13 維梅爾倒頻譜係數之後再分別求出其 i-vector。在此特別強調的是，此處的音訊片段內並非只包含一位語者，因為它經由語音的串接而來，所以裡面可能不只包含一位語者，分別為下面三種可能：1）音訊片段內只有客服的聲音；2）音訊片段內只有客戶的聲音；3）音訊片段內同時含有客戶與客服的聲音，以下簡稱為混合。，由圖四中我們任意取出三種語者（客戶、客服、混合）之各兩段音訊片段所抽取出的 i-vector，計算其彼此之間的 PLDA 分數，可以觀察出客戶對客戶或客服對客服的 i-vector 彼此間的 PLDA 分數是相對較高的，不過混合對混合的 i-vector 彼此間的 PLDA 分數卻沒有這樣的關係。因此根據不同客服與客戶聲音的混合程度，在 i-vector 的表示上可視為兩個不同的語者。此外，為了不讓語者的聲音變化太大，並且增加系統在處理自動語者分段標記上的精細度，我們嘗試使用可重疊的音訊片段，而且音訊片段的長度與重疊時間是可調整的，在第四節的實驗我會描述不同的長度與重疊時間對我們的系統會造成甚麼樣的影響。



圖四、不同語者的 i-vector 間的 PLDA 比較。

（三）語者分群



圖四、語者分群的流程

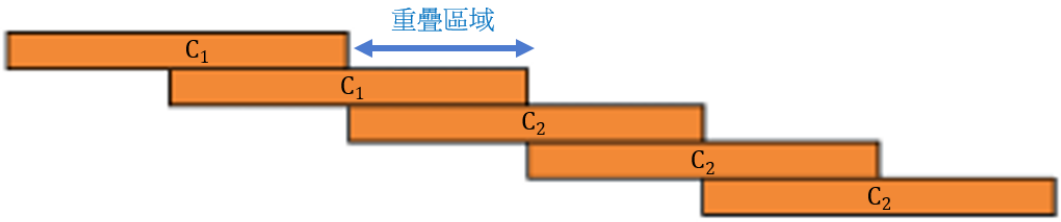
對所有音訊片段所抽取出來的 i-vector 計算其兩兩之間的 PLDA 分數而得到一個自

相似矩陣。在這個矩陣中，我們可以查詢到所有兩兩不同的 i-vector 之間的 PLDA 分數，並利用此矩陣，藉著 K 平均演算法來對語者分群，而得到第一階段的自動語者分段標記結果，其步驟如下：1）先隨機選取兩個 i-vector 當作兩群群心，我們將這兩群稱為 C_1、C_2，群心稱為 Q_1、Q_2；2）查詢剩餘的 i-vector 對 Q_1、Q_2 之 PLDA 分數，並且比較其大小，如果和 Q_1 者較高，則被分配到 C_1，反之則分配到 C_2；3）重新定義 C_1 和 C_2 的群心，目標為找一個和群內所有 i-vector 最相似的一個 i-vector，計算的方式如下式

$$\bar{Q}_i = \underset{j}{\operatorname{argmax}} \sum_{k \in C_i, k \neq j} PLDA(k, j) \quad , \forall j \in C_i \quad , \forall \mathrm{i} \in \{1, 2\}$$

其中，PLDA（k, j）表示查詢 k 和 j 這兩個 i-vector 的 PLDA 分數；4）重複步驟 2）和 3）直到 K 平均演算法收斂為止。

值得一提的是，此處對音訊片段進行語者分群，會遇到如圖六的問題：如果前一個音訊片段被分配到 C_1，而後一個音訊片段被分配到 C_2，那我們如何去決定重疊部分的類別？



圖六、重疊區域的語者分群問題

該如何決定重疊部分的類別勢必會對實驗結果產生影響，因此我們用一個簡單並且直覺的方式來解決這個問題：假設前一個 i-vector 叫做 i，落於 C_1 群內；後一個 i-vector 叫做 j，落於 C_2 群內；重疊部分寫作 $S_{i \cap j}$，想法就是找出較高的 PLDA 分數，表示與哪一群就越像，依照這樣的想法來決定重疊部分的分群。

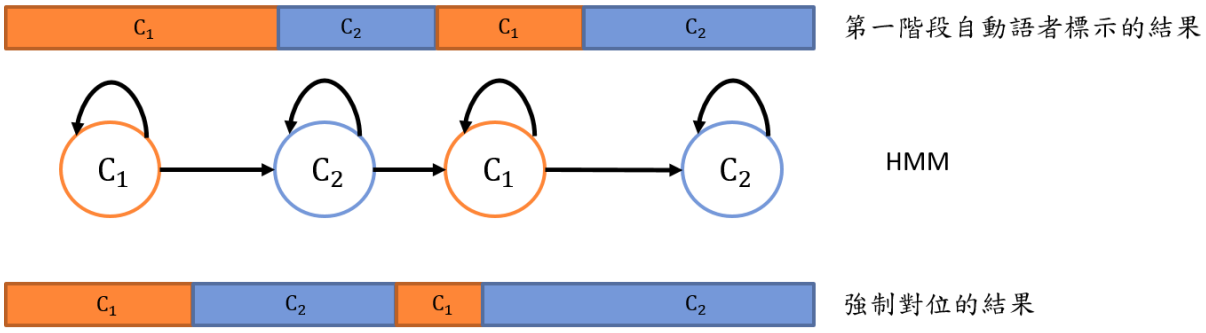
$$\mathrm{P}_1 = \mathrm{PLDA}(\mathrm{i}, \mathrm{Q}_1) \qquad (1)$$

$$\mathrm{P}_2 = \mathrm{PLDA}(\mathrm{j}, \mathrm{Q}_2) \qquad (2)$$

$$S_{i \cap j} \in \mathrm{C_k} \text{ , if } \underset{\mathrm{k}}{\operatorname{argmax}} \ \mathrm{P_k} \ , \ \forall \mathrm{k} \in \{1, 2\}$$

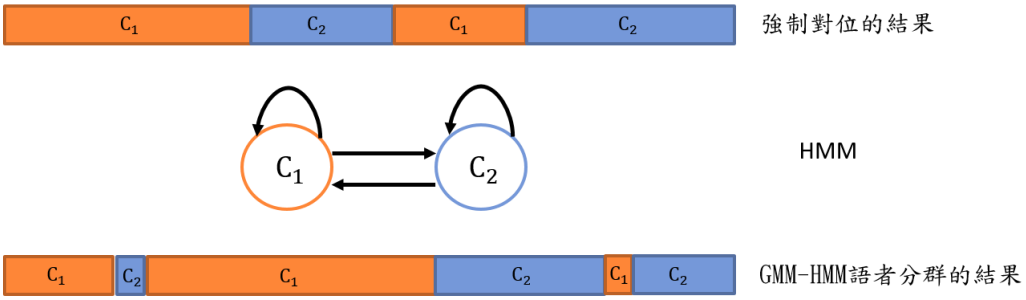
由 K 平均演算法我們得到了第一階段自動語者分段標記的結果。然而，以音訊片

段為分群的單位而得到的結果始終還是太過鬆散。一般而言，我們在處理語音的問題都是以音框（Frame）為單位，通常一個音框的時間為 32 微秒，相對於我們的音訊片段以秒為單位實在差距太大。因此，我們從第一階段的自動語者分段標記結果中得到了兩位語者所有的語者片段，針對這些語者片段抽取其 13 維的梅爾倒頻譜係數來訓練兩個語者模型，接著利用 GMM-HMM 去重新調整所有語者區段的範圍，這樣的動作稱做強制對位。圖七是強制對位的說明圖，針對第一階段自動語者標示的結果所訓練的語者模型，利用 GMM-HMM 對整個串接音檔做強制對位，其強制對位的結果並不會改變語者區段的數量，而是改變它們的相對範圍。第四節的實驗也會說明，強制對位的語者自動分段結果在進行 20 次之前就會達到收斂。



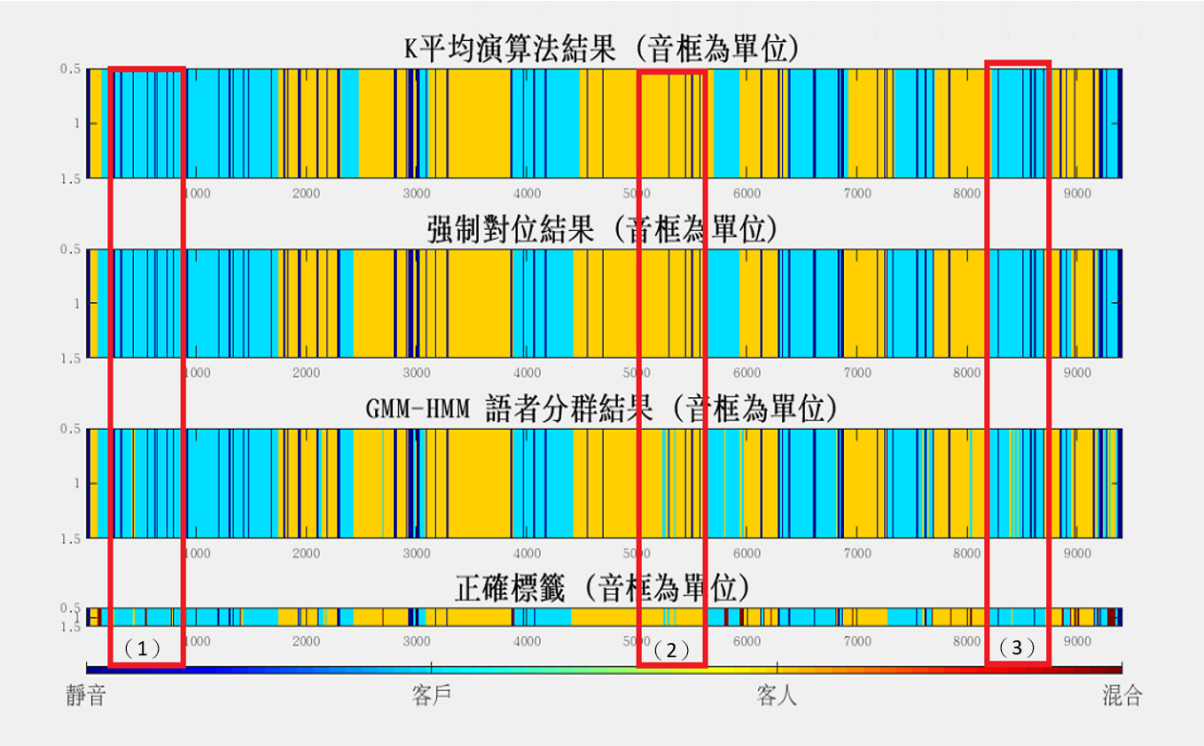
圖七、使用 GMM-HMM 進行強制對位示意圖

在確認了時間較長的語者區段之後，我們將面對在自動語者分段標記上的難題，就是將極短時間的語者區段標示出來。經由強制對位，我們取得了更為準確的語者區段，利用它們來訓練新的語者模型，之後再用 GMM-HMM 進行語者分群，概念如圖八所示，由第 20 輪的強制對位結果訓練出新的語者模型，接著對整的串接音檔做 GMM-HMM 路徑解碼（Decode）。和強制對位不同的是，利用 GMM-HMM 進行語者分群不只會改變語者區段的範圍，也會改變語者區段的數量。

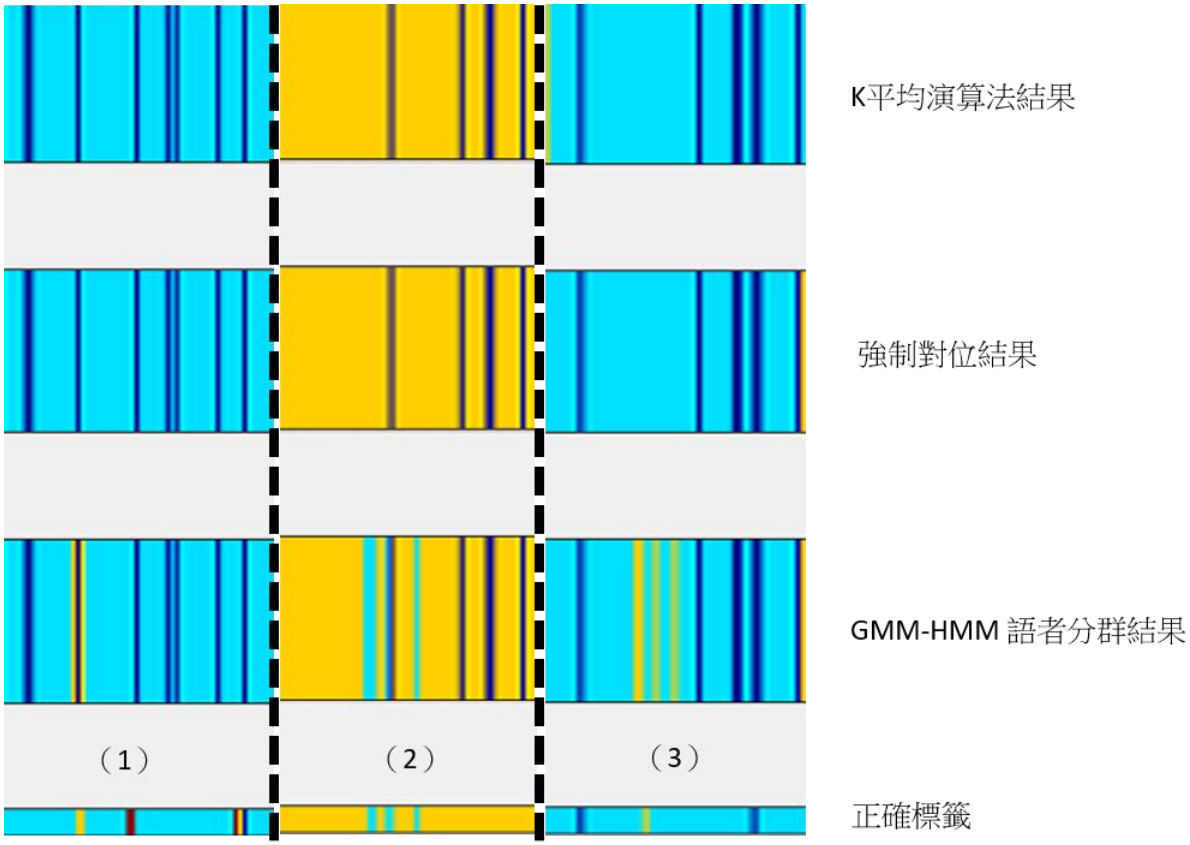


圖八、使用 GMM-HMM 進行語者分群

最後，圖九描述了 K 平均演算法、強制對位、利用 GMM-HMM 進行語者分群的結果比較，其中較明顯觀察到變化的我用紅色方框標記，並且放大顯示於圖十當中。由圖十的（1）我們可以發現，在淺藍色標記的客戶說話區段中尚存在著細小黃色區段，也就是客戶說話區段。因此利用 GMM-HMM 進行語者分群可以將電話錄音內細小的片段找出來。



圖九、K 平均演算法、強制對位、GMM-HMM 語者分群的結果比較。



圖十、放大顯示強制對位與 GMM-HMM 語者分群的差異圖。

除了第二階段的自動語者分段標示結果之外，為了想確認找出細小語者區段的結果再透過重新對位會不會有更好的效果，在第四章也會將實驗的結果展示出來。

三、資料庫與實驗評估

我們使用的資料庫是由中國信託（China Trust）提供之 100 段客服和客戶的電話語音，每段電話語音的取樣率為 16,000 Hz，且都只包含兩位語者，並由客服先開始對話，平均長度為 4 分 57 秒。其中，資料庫並未提供每段語音之客服和客戶的語者資料。

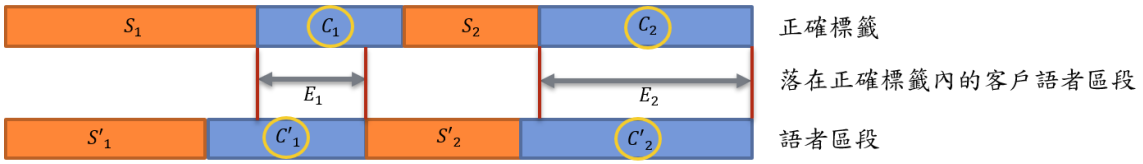
為了抽取梅爾倒頻譜係數與 i-vector，每段語音會先降低取樣率為 8,000 Hz。其中，每一音框的長度為 32 ms，而音框位移的長度則為 10 ms。因為上下文資訊（Contextual Information）較無關於語者特性，梅爾倒頻譜係數只去靜態的 13 維部分，而不考慮動態的差異。而在抽取 i-vector 以及計算 PLDA 分數的部分，我們以 NIST SRE 2004、2005、2006 並 Switchboard II-Phase 1-3 以及 Switchboard Cellular Part 1-2 來訓練通用背景模型、全變異（Total Variability）模型以及 PLDA 模型。其中，通用背景模型的成分

數為 2048，而 i-vector 的維度則為 600。

我們使用召回率（Recall）以及精準率（Precision）來作為評估自動語者分段標記的標準，其中我們的召回率和精準率是定義在客戶的語者片段。由於我們使用的資料庫都會由客服先開始說話，所以我們在做完自動語者分段標記之後取第二位語者做為客戶，針對其語者區段來計算召回率以及精準率。召回率表示落在正確標記內的客戶語者區段音框總數和正確標記內客戶音框總數的比例；精準率則表示落在正確標記內的客戶語者區段音框總數和客戶語者區段音框總數的比例。圖十一簡單給予一個計算召回率以及精準率的範例，其中 C_1、C_2 表示正確標記內客戶音框總數，C'_1、C'_2 表示客戶語者區段音框總數，而 E_1、E_2 代表落在正確標記內的客戶音框總數，則召回率以及精準率的計算如下：

$$\text{召回率(Recall)} = \frac{E_1 + E_2}{C_1 + C_2} \qquad (3)$$

$$\text{精準率(Precision)} = \frac{E_1 + E_2}{C'_1 + C'_2} \qquad (4)$$



圖十一、召回率、精準率的計算示意圖。

最後我們會由召回率以及精準率求得 F-Score，來做為挑選比較各種音訊片段的長度以及重疊時間的評估方式，其計算的公式如（5）。

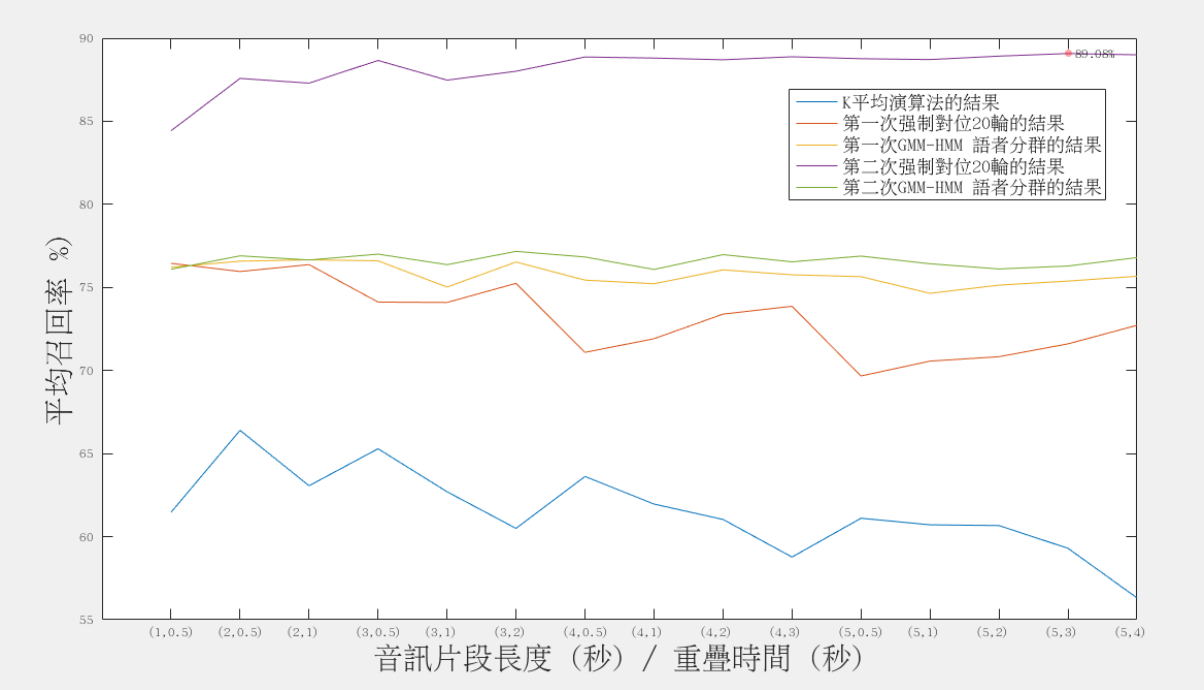
$$\text{F-Score} = \frac{2 * Recall * Precision}{Recall + Precision} \qquad (5)$$

四、實驗結果

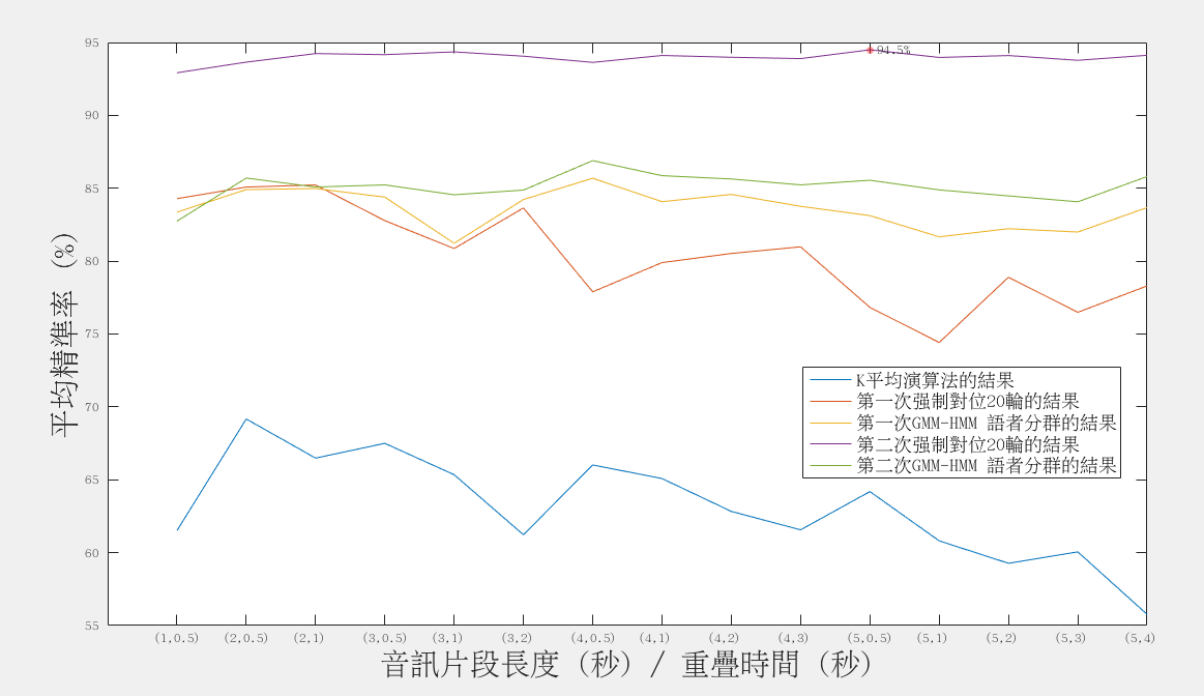
我們針對不同曲音訊片段的長度以及重疊時間做語者的自動分段標記的實驗，企圖找出最好的組合。除了比較 K 平均演算法得到的結果、20 輪強制對位的結果、GMM-HMM 語者分群的結果之外，我還想知道利用 GMM-HMM 進行語者分群得到細

小的語者片段再進行行強制對位以及語者分群是否有助於自動語者分段標記的效能，因此還會另外比較這兩個結果：第二次 20 輪強制對位的結果，和第二次 GMM-HMM 語者分群的結果。

由圖十二還有圖十三可以看得出來 K 平均演算法不論在召回率以及準確率都會得到最差的結果，因為它再進行語者分群的單位是以秒為單位進行的，相對於其他方式以音框做為分群單位太過粗略，不過也是音為建立於 i-vector 以及 PLDA 的機制下，它對接下來進行的強制對位有一個良好的分群基礎，使得語者區段的範圍進行微調之後可以使召回率以及精準率有大幅度的提升。最後在這五個比較方法中，我們由第二次強次對位得到最好的召回率（89.08 %）和精準率（94.55 %）。

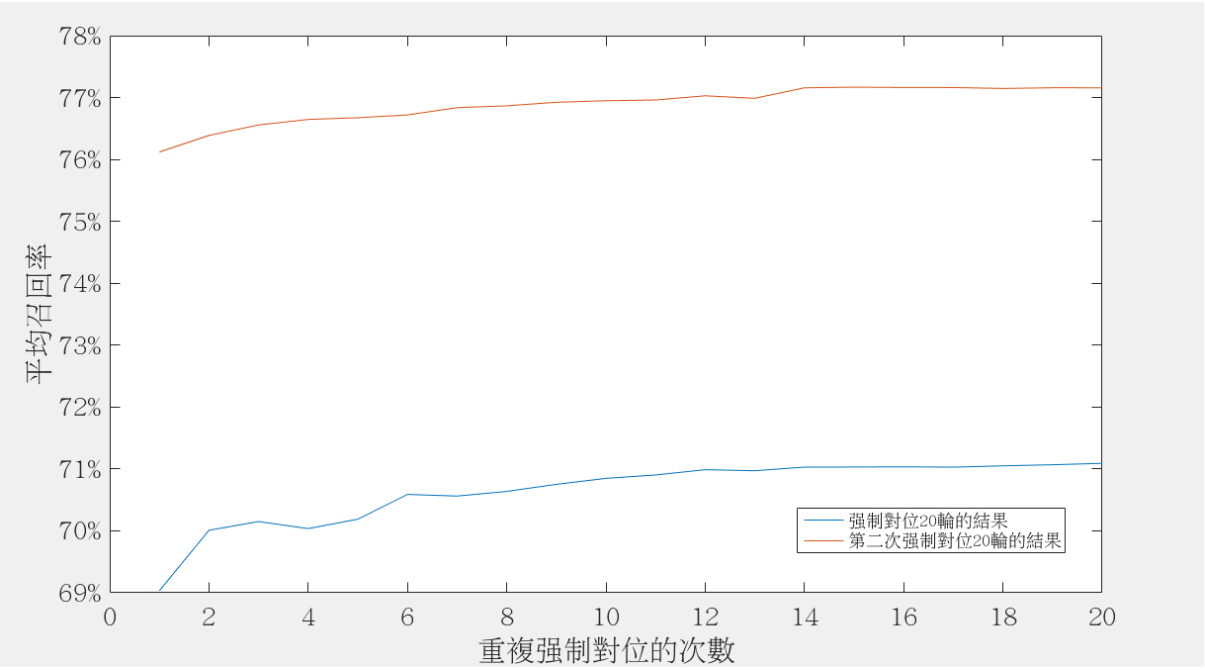


圖十二、平均召回率的比較。

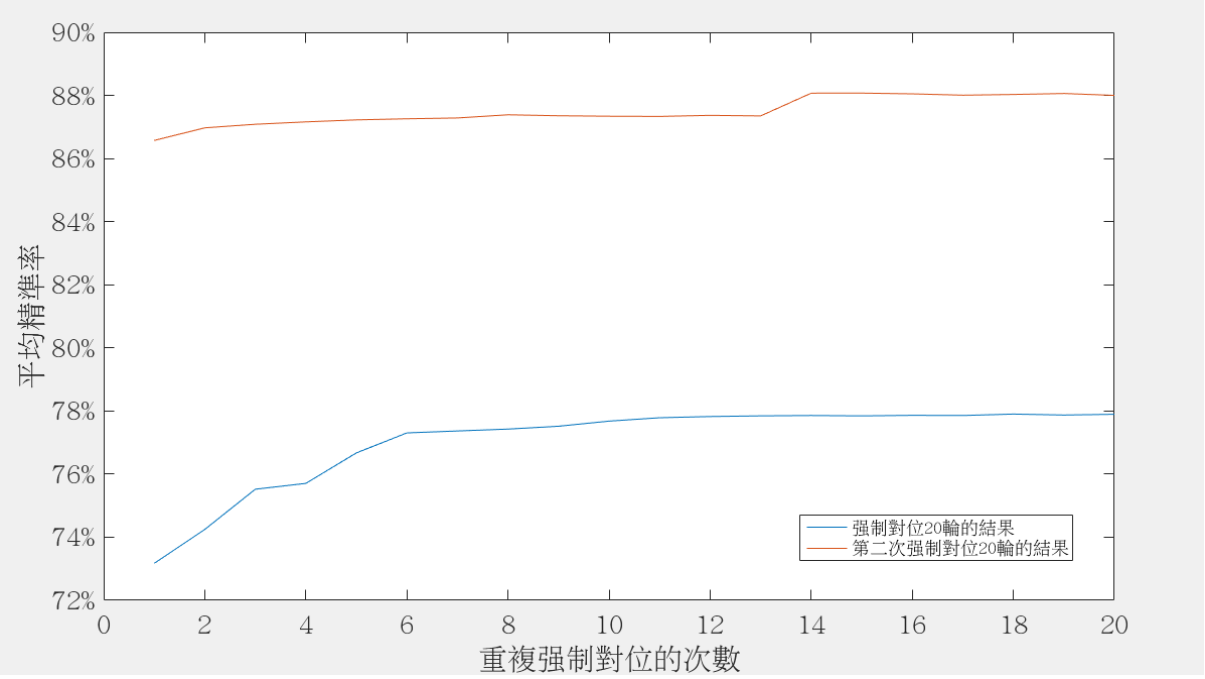


圖十三、平均精準率的比較。

此外由圖十四、圖十五可以觀察到，不論是第一次或者第二次強制對位在進行第 20 輪之前召回率以及精準率都會達到飽和，並且由圖十二和圖十三觀察得知，第一次用 GMM-HMM 進行語者分群的召回率以及精準率都比第一次強制對位還要好，所以利用第一次 GMM-HMM 的語者分群所得到的語者區段可以訓練出更好的語者模型，進而提升第二次強制對位的召回率以及精準率。不過第二次 GMM-HMM 語者分群卻比第二次強制對位得到的結果還要差，我們可以合理推測經過第二次強制對位後就不需要再進行第二次 GMM-HMM 語者分群來得到更細碎的語者區段。

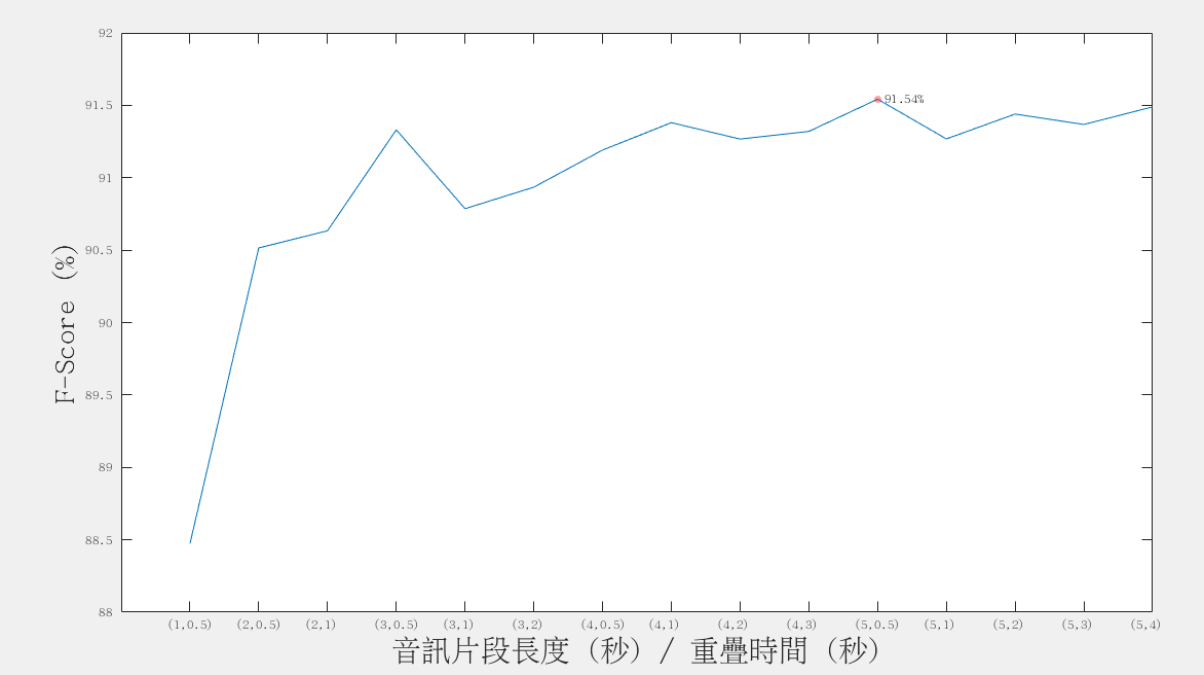


圖十四、音訊片段 4 秒，重疊時間 0.5 秒的強制對位和第二次強制對位的召回率曲線。

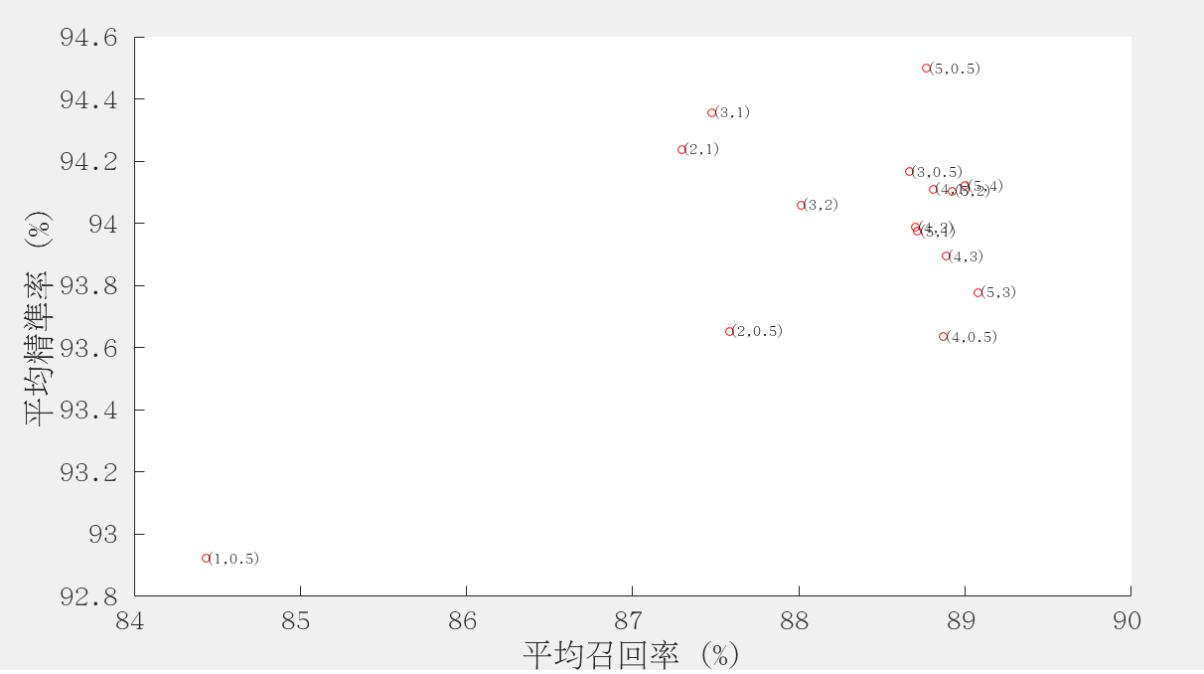


圖十五、音訊片段 4 秒，重疊時間 0.5 秒的強制對位和第二次強制對位的精準率曲線。

圖十五我們針對第二次強制對位下去評估每個音訊片段的長度與重疊時間的配對對自動語者分段標記的影響，我們發現在長度為 5 秒，重疊秒數為 0.5 秒的時候可以得到最好的 F-Score（91.54 %），不過整來來說，除了長度為 1 秒，重疊秒數為 0.5 秒的配對外，其他的配對組合對自動語者分段標記的影響並沒有很大的差別。圖十六描述了第二次強制對位 20 輪的召回率與精準率的關係。



圖十五、第二次強制對位 20 輪的 F-Score。



圖十六、第二次強制對位 20 輪的平均召回率與平均精準率的對應圖。

五、結論

我們由實驗得知強制對位之後進行的 GMM-HMM 語者分群得到較細小的語者區段有助於訓練出更好的語者模型，使第二次強制對位能得到更好的結果。值得觀察的是，第二次使用 GMM-HMM 進行語者分群的結果並沒有比第二次強制對位的結果還要來的好，所以我們認為再繼續進行 GMM-HMM 語者分群與和強制對位這樣的循環對自動語者分段標記不會再有顯著的進步。

我們也發現在多組音訊片段的長度與重疊時間的組合對自動語者分段標記的結果並沒有太大的差別，不過由於 i-vector 的抽取時間與音訊片段的數量呈正相關，所以建議可以使用時間較長，並且重疊時間較短的音訊片段來進行自動語者分段標記。

六、參考文獻

[1] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted gaussian mixture models,” *Dig. Sig. Proc.*, 2000.

[2] Najim Dehak, Patrick Kenny, R´eda Dehak, Pierre Dumouchel,and Pierre Ouellet,

"Front-End Factor Analysisfor Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[3] H.-S. Lee et al. , "Clustering-based i-vector formulation for speaker recognition," in *Proc. Interspeech*, 2014.

[4] S. Tranter and D. Reynolds, "An overview of automatic speaker diarisation systems," *IEEE Transactions on Audio, Speech, and Language Processing*, Sept. 2006.

[5] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio Speech Lang. Process*., vol. 20, no. 2, pp. 356–370, Feb. 2012.

[6] G. Sell and D. Garcia-Romero, "Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[7] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of Interspeech*, 2011.

[8] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and Jim Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *Proceedings of Interspeech,* 2011.