錄音資料中的語者切割與分群

Speaker Segmentation and Clustering for the Recorded Speech

蘇峻慶、王小川
Chun-Ching Su and Hsiao-Chuan Wang

國立清華大學電機工程學系
Department of Electrical Engineering, National Tsing Hua University

Email: g923990@oz.nthu.edu.tw hcwang@ee.nthu.edu.tw

摘要

本論文主要在探討錄音資料中語者切割與語者分群的問題，在語者切割方面，採用三個步驟，第一步是利用貝氏資訊準則約略找出語者轉換點大概的位置，第二步利用交叉偵測法作精確化，第三步再確認是否為轉換點，實驗上顯示此方法擁有運算量少及高準確率的優點。在語者分群方面，群集之語者模型採用高斯混合模型，音段與每一個群集模型作最大概似法估測，找出最靠近之群集，然後再利用一個門檻值判斷是要合併或是分離出新的群集。實驗結果顯示音段群中包含語者數愈多，其整體分群效能愈低。

關鍵詞：語者切割、語者分群、語者轉換點偵測、群集模型

一、 緒論

語言是人類溝通及傳達意念最自然的方法，語音訊號不只包含了說話者所要表達的意思，更是隱含了說話者的個人特徵，因此在一段語音信號中，我們不僅要能夠聽出其中所要表達的意思，更要知道這一段話究意是誰所講的。

近年來從有線或無線網路上以語音擷取資訊的應用增加，身份確認或說話人辯識變得更為重要，愈來愈多人投入自動語者辨識的研究領域。在多人說話的環境下，變成需要先對語音做分段，然後再辨認各個音段是誰在說話，因此就需事先作切割與分群。舉例來說，在一個重要會議場合的錄音，其內容包含若干人的談話，若想將這些語者的語音訊號分開，利用人工方法是既費時又不經濟，因此有必要發展出一套正確率高，速度又快的切割與分群方法。

過去已有許多語者切割的方法被提出[1][2]，而這些被提出的方法大致可分類為以解碼為基礎之切割法(Decoder-Guided Segmentation)、以模型為基礎之切割法(Model-based Segmentation)、以及以距離為基礎之切割法(Metric-Based Segmentation)。以上三種方法都有其優缺點，像以解碼為基礎之切割法，只能粗略地分類出語音、音樂、靜音等，並無法用來偵測出語者轉換點的位置。以模型為基礎之切割法，需要事先搜集相關語料建立相對應的模型，這並不符合實際。以距離為基礎之切割法，則需設定門檻值(Threshold Value)來決定語者轉換點的位置，因此缺少穩定性(Stability)和強健性(Robustness)。

語者分群是一個活躍多年的研究領域，大致上在作語者分群時有幾個基本的問題[3]：

1. 聚集(agglomeration)：對一群音段作語者分群時，其形成群集的方式有兩種，一種是凝聚，另一種是分裂。
2. 停止準則(stopping criteria)：在作語者分群時，通常是不曉得音段群裡包含多少個語者，因此需設立一個停止準則，當群集數達到此一停止準則，即停止再分新群。
3. 距離量測(distance measures)：利用一個距離量測的方法，用以決定所偵測的音段是屬於哪一群。

本文在語者切割方面，採用三個步驟，第一步是利用貝氏資訊準則約略找出語者轉換點大概的位置，第二步利用交叉偵測法作精確化，第三步再確認是否為轉換點，實驗上顯示此方法擁有運算量少及高準確率的優點。在語者分群方面，群集之語者模型採用高斯混合模型，音段與每個群集模型作最大概似法估測，找出最靠近之群集，然後再利用一門檻值判斷是要合併或是分離出新的群集。

本文內容安排如下：第二節詳細說明語者切割的基本技術及本論文所使用的方法，第三節說明本論文使用的語者分群方法，第四節是實驗設計及對實驗結果做討論，第五節為結論。

二、 語者切割

2.1 語者轉換點偵測(Speaker Change Detection)

語者轉換點偵測就是偵測說話者改變時的轉換點，最常被使用來偵測的方法，一為貝氏資訊準則(Bayesian Information Criterion, BIC)，另一為廣義概似比(Generalized Likelihood Ratio, GLR)，以下分別介紹這兩種方法。

(A) 貝氏資訊準則(Bayesian Information Criterion, BIC)[4]

假設 $M = M_1, M_2, M_3, ..., M_k$ 是所有的候選模型集合，k_j 是 M_j 這一個模型的參數數目，$X = X_1, X_2, X_3, ..., X_N$ 為一群資料集，根據定義，BIC 可寫成下式：

$$BIC(M_j) = \log L\left\langle X_1, X_2, ..., X_N \middle| M_j \right\rangle - \frac{1}{2}\lambda k_j \log N \quad (1)$$

其中 $L\left\langle X_1, X_2, ..., X_N \middle| M_j \right\rangle$ 為模型 M_j 和資料集 X 的最大概似值 (Maximum Likelihood)，λ 為損失權重，根據(1)式，就可從眾多模型中找出一個最佳的模型來描述資料集 X。

(B) 貝氏偵測法[1]

假設 $X = \{x_1, x_2, x_3, ..., x_N\}$ 代表一語音段的特徵向量，且只包含一個語者轉換點，如圖 1 所示。假設語者轉換點發生在 i 的時間點上，我們設定二個假說測試(Hypothesis Testing)，其定義如下：

$$H_0 : x_1, x_2, ..., x_N \sim N(\mu, \Sigma) \quad (2)$$

$$H_1 : x_1, x_2, ..., x_i \sim N(\mu_1, \Sigma_1); x_{i+1}, x_{i+2}, ..., x_N \sim N(\mu_2, \Sigma_2) \quad (3)$$

(2) 式表示全音段的特徵參數序列，呈高斯分布。(3) 式表示分成兩段音段的特徵參數序列，也是呈高斯分布。

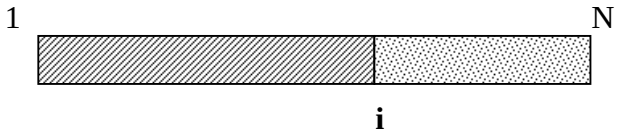


圖 1　長度為 N 並包含一個語者轉換點的語音段

將 H_0 與 H_1 兩模型作比較，比較的式子定義如下：

$$\Delta BIC = BIC(H_1) - BIC(H_0) \tag{4}$$

把(1)式、(2)式及(3)式代入上面的(4)式，可得到下列的結果：

$$\Delta BIC = R(i) - \lambda P \tag{5}$$

其中 λ 為一個加權值，$R(i)$ 為最大概似比(Maximum Likelihood Ratio)：

$$R(i) = N\log|\Sigma| - i\log|\Sigma_1| - (N-i)\log|\Sigma_2| \tag{6}$$

P 為懲罰值(penalty)：

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1))\log N \tag{7}$$

d 為特徵參數維度，N為特徵參數數量。

若該 i 點的 ΔBIC 值最大，而且為正值，我們認為此時間點為一語者轉換點。

$$\arg\max_i \Delta BIC(i) > 0 \tag{8}$$

(C) 廣義概似比(Generalized Likelihood Ratio, GLR)偵測法[5]

圖 2 所示為廣義概似比偵測法的流程圖，其演算和貝氏偵測法一樣，必須先定義兩個假說測試 H_0 與 H_1，不過貝氏偵測法是移動可變時間點 i 作語者轉換偵測，廣義概似比偵測法則以兩個固定長度的語音段作語者轉換點偵測，其測量距離的式子定義如下，

$$R = \frac{L(X, N(\mu,\Sigma))}{L(X_1, N(\mu_1,\Sigma_1))L(X_2, N(\mu_2,\Sigma_2))} \tag{9}$$

X_1 與 X_2 是相鄰的兩段語音參數序列，其連接的語音訊號序列就是 $X = X_1 \cup X_2$，呈高斯分佈，即 $X \sim N(\mu,\Sigma)$。X_1 與 X_2 也是呈高斯分佈，$X_1 \sim N(\mu_1,\Sigma_1)$，$X_2 \sim N(\mu_2,\Sigma_2)$。當 R 值愈小，代表兩個相鄰的語音段愈可能為不同說說者，反之，則愈可能為同一說話者。廣義概似比偵測法最大的缺點，即比較難去定義門檻值來判斷是同一說話者或不同說話者。

2-2 本論文使用之方法

A. 偵測單一語者轉換點

本論文偵測單一語者轉換點的方法，是當語音段進行特徵參數抽取後，會先將貝氏資訊準則應用到以距離為基礎之順序偵測法(Sequential Metric-based segmentation via BIC)[6]，找出語者轉換點大概的位置，然後再透過交叉偵測法[7]，將剛才所找出的語者轉換點作精確化，也就是

讓偵測到的轉換點離真實轉換點更近，最後再確認是否為轉換點，各功能方塊描述如圖 3 所示。

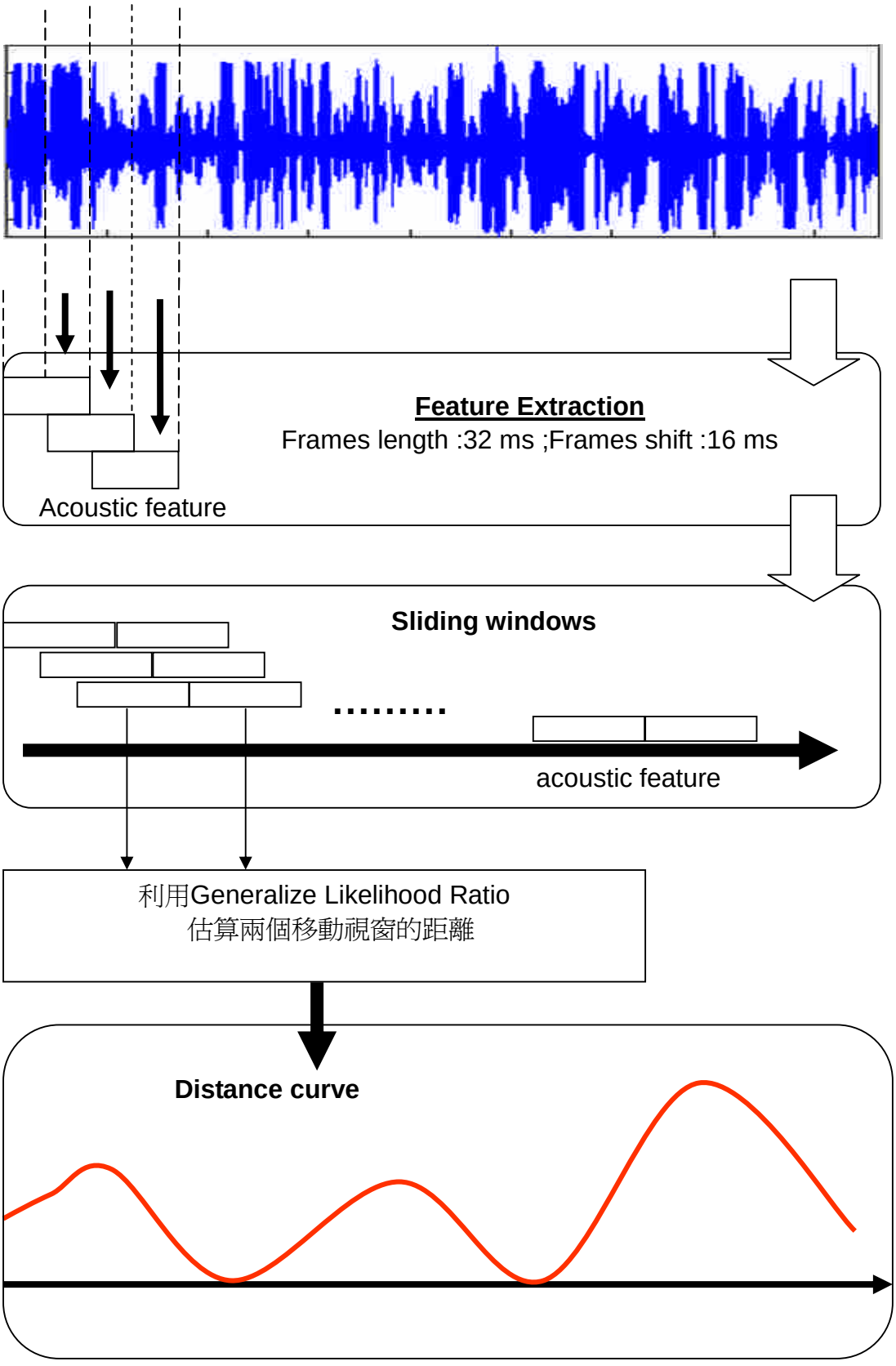


圖 2　廣義概似比偵測法之流程圖

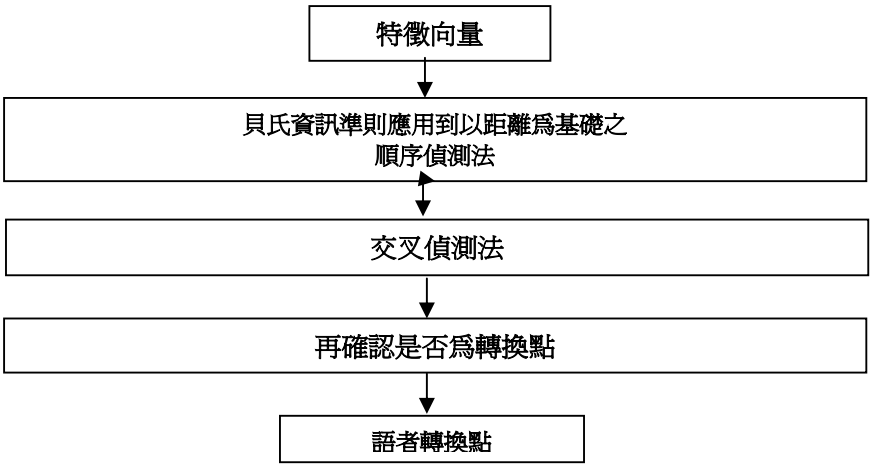


圖 3　偵測單一語者轉換點之架構流程圖

(1)貝氏資訊準則應用到以距離為基礎之順序偵測法(Sequential Metric-based segmentation via BIC)

前述的貝氏偵測法，是根據不同的時間點 i 建立兩個假說測試 H_0 與 H_1，然後計算其每個不同點 i 的 ΔBIC 值，最後由這些 ΔBIC 值決定語者轉換點。若我們將不同時間點 i 估計 ΔBIC 值的方式，改成廣義概似比固定長度的方式，來做 ΔBIC 值的估計，如圖 4 所示。

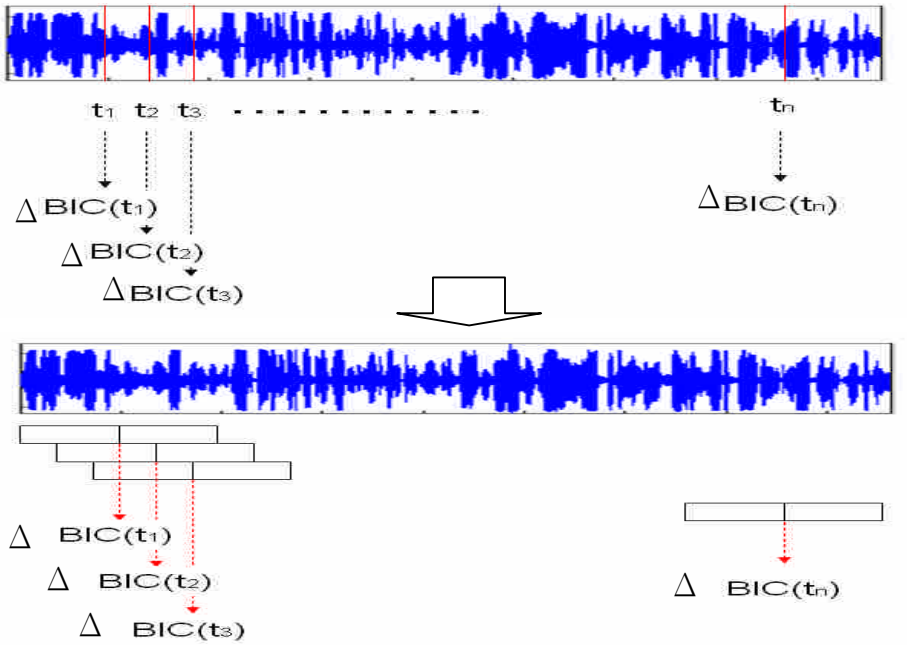


圖 4　計算方式由不同點改為固定長度之示意圖

計算方式改變後可得到下列的結果：

$$
\begin{aligned}
\Delta BIC &= BIC(H_1) - BIC(H_0) \\
&= \log pr\langle X \mid \mu x, \Sigma x\rangle + \log pr\langle Y \mid \mu y, \Sigma y\rangle \\
&\quad - \log pr\langle Z \mid \mu, \Sigma \rangle - P \\
&= \log \frac{pr\langle X \mid \mu x, \Sigma x\rangle pr\langle Y \mid \mu y, \Sigma y\rangle}{pr\langle Z \mid \mu, \Sigma \rangle} - P \\
&= GLR - P.
\end{aligned}
\tag{10}
$$

由(10)式可知，以這樣的方式估算ΔBIC值，其實就好像是計算GLR值，再多加個懲罰項P。這種方式就是貝氏資訊準則應用在以距離為基礎的偵測法(Metric-based segmentation via BIC)。

若再進一步的改成如圖5的方式來計算，則稱貝氏資訊準則應用在以距離為基礎的順序偵測法 (Sequential Metric-based segmentation via BIC)。

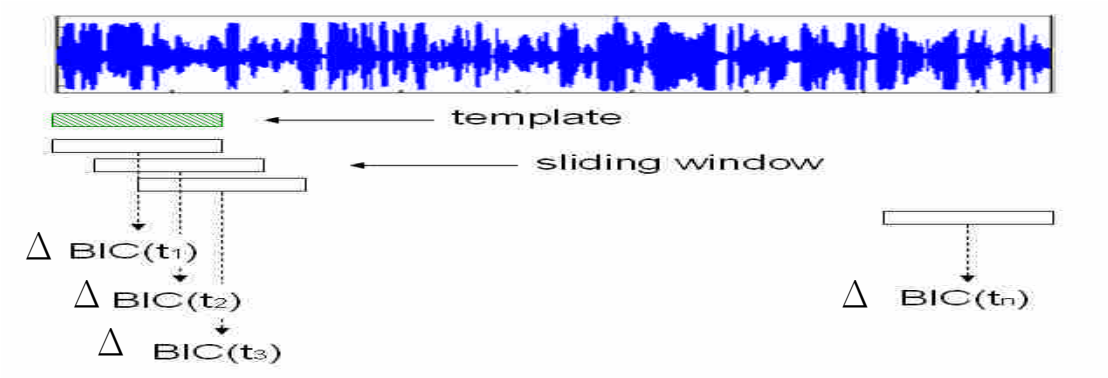


圖5 貝氏資訊準則應用在以距離為基礎的順序偵測法示意圖

首先我們取語音最開始時的短視窗(約2~3秒)作為樣式(template)，之後將此樣式和每個滑動視窗(長度和樣式相同)作ΔBIC的計算，可獲得ΔBIC的曲線，如圖6所示，

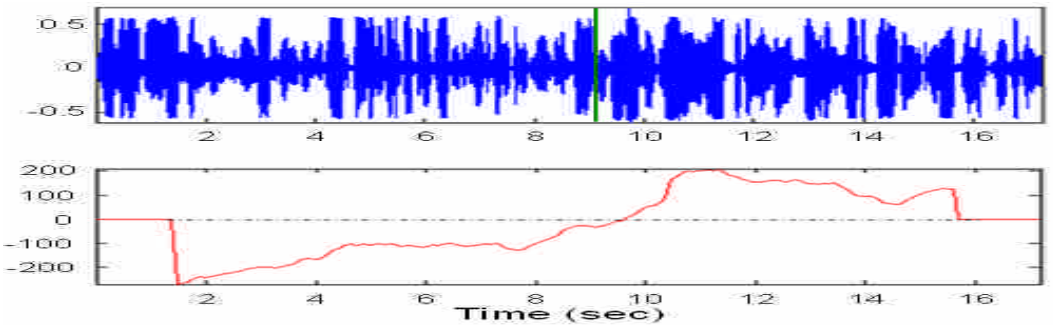


圖6 貝氏資訊準則應用在以距離為基礎的順序偵測法求出的ΔBIC曲線

由圖中觀察可發現，當滑動視窗在語者一的範圍內時，樣式和移動視窗均為語者一的聲音，所以ΔBIC值為負。當滑動視窗到達語者二的範圍內時，滑動視窗變為語者二的聲音，樣式還是語者一的聲音，所以ΔBIC值為正，這正是貝氏資訊準則的特性。在滑動視窗從語者一移到語者二時，ΔBIC值也由負變正，所以我們可以定義，在ΔBIC值為0時，其附近可能有轉換點存在。

(2)交叉偵測法

採用貝氏資訊準則應用在以距離為基礎的順序偵測法來偵測出語者轉換點後，在其轉換點向右延伸 0.5 秒處，往後抓取語者二的樣式，如圖 7 所示，其中向右延伸是為了確保抓取的樣式全包含語者二的語音訊號，而延伸長度選取 0.5 秒，是因為在實驗中發現，貝氏資訊準則應用在以距離為基礎的順序偵測法所偵測出的語者轉換點，大約在真實轉換點的左邊 0.5 秒到右邊 2 秒之間，所以只要向右延伸 0.5 秒就可確保樣式 2 包含語者二的語音訊號。

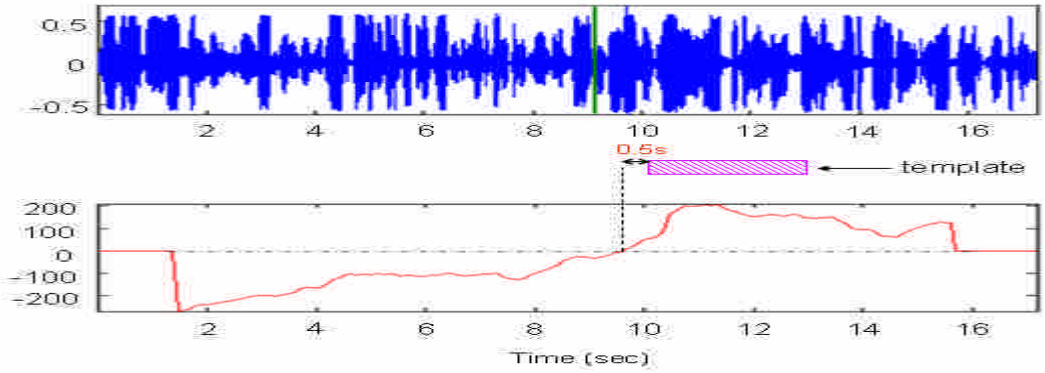


圖 7　尋找語者二的樣式

在找出語者二的樣式後，此樣式和每個滑動視窗作 ΔBIC 估算，可得一條 ΔBIC 曲線，如圖 8 中的藍色曲線，而這條曲線和原先貝氏資訊準則應用在以距離為基礎的順序偵測法所求之曲線的交叉處，即為語者轉換點的地方。會認為在曲線交叉的地方有語者轉換點存在的原因，是因為當滑動視窗移到真實轉換點時，會同時包含語者一和語者二的語音訊號，因此滑動視窗和語者一的樣式作 ΔBIC 計算以及和語者二的樣式作 ΔBIC 計算，其值會差不多。

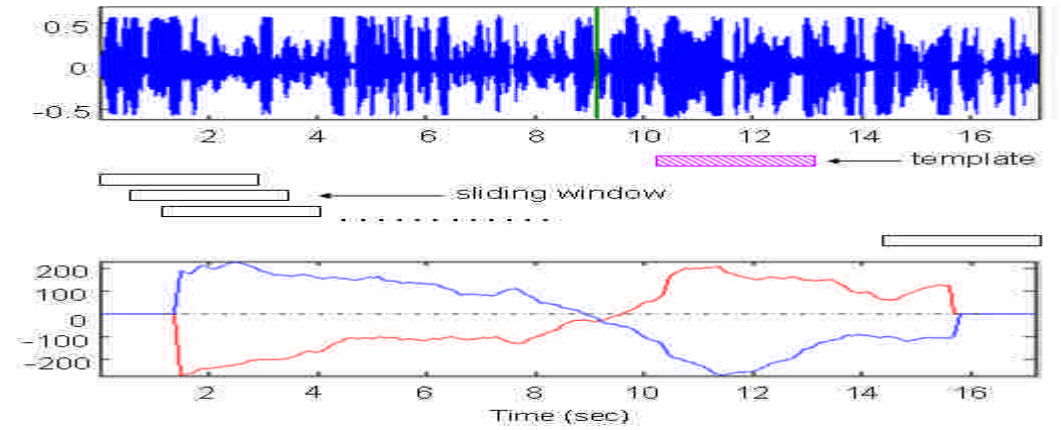


圖 8 交叉偵測法尋找語者轉換點

(3)再確認是否為轉換點

利用第二條曲線起始到交叉的這個區段(如圖 9 綠色框框部份)，將其區段中所有的 ΔBIC 值作符號函數運算後相加，若相加後值為正，則接受此點為語者轉換點，若為負，則拒絕此點為語者轉換點。

$$\begin{cases} \sum_{i=1}^{N} sign(\Delta BIC(i)) > 0 & accept \\ \sum_{i=1}^{N} sign(\Delta BIC(i)) < 0 & reject \end{cases} \tag{11}$$

其中 sign( )為符號函數。

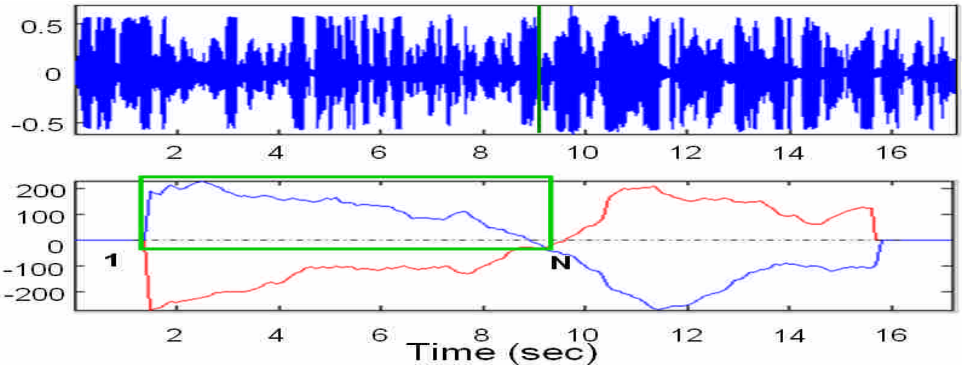


圖 9 作確認之區域圖例

利用(11)式作再確認的原因，是因為在觀察假警報錯誤的情況後，發現一些錯誤是第一條曲線有突起的狀況所造成的，導致本來應該找出語者二的樣本，結果找到語者一的樣本。一般而言，若正確找出語者二的樣本，其所求出的第二條曲線，會如圖 9 中藍色曲線所示一樣，前半段之 ΔBIC 值會大於 0，後半段之 ΔBIC 值會小於 0。若錯誤地找出語者一的樣本，就會如圖 10 中綠色曲線一樣，前半段之 ΔBIC 值小於 0，後半段之 ΔBIC 值大於 0。利用這樣的特性，第二條曲線起點到交叉點的區域，計算 ΔBIC 值是否大於 0，即可進一步確認此點有無偵測錯誤。

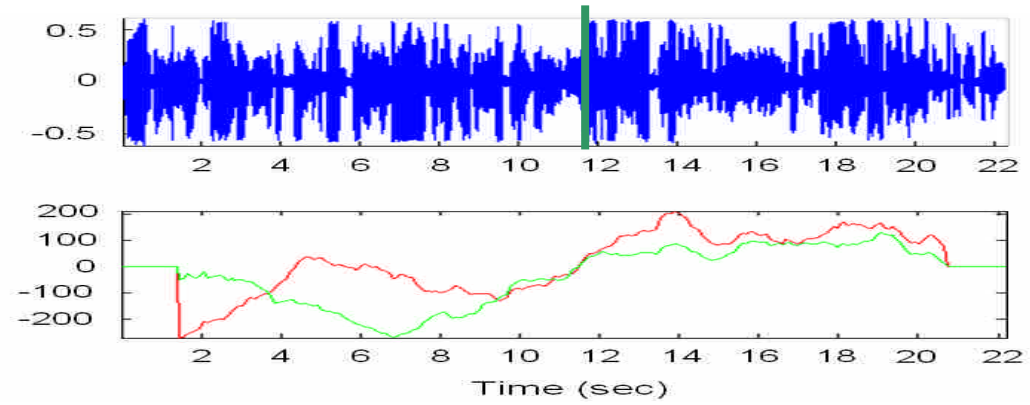


圖 10　錯誤偵測到語者轉換點之範例

B. 偵測多重語者轉換點

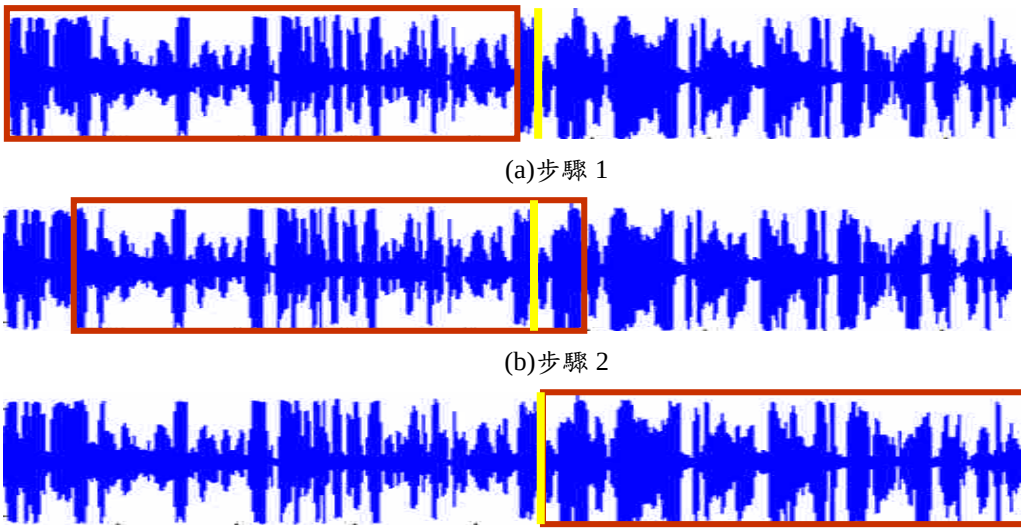
偵測單一語者轉換點的觀念，可以用來偵測多重語者轉換點。其步驟如下，

1.首先設一視窗(長度為 14 秒)，在視窗內作單一語者轉換點偵測。

2.若在上一步驟沒找到語者轉換點，則將視窗向右移動(向右移動 2 秒)，重新在視窗內作單一語者轉換點偵測。若還是沒找到轉換點，再將視窗向右移動，直到找到語者轉換點，或是直到語音結束。

3.若是找到語者轉換點，則記錄此轉換點，並將視窗的起始點設在此語者轉換點上，重新作步驟一及步驟二，直到找到下一個語者轉換點，或是直到語音結束。

圖 11 展示各個步驟之示意圖，圖中黃色線為人工標注之語者轉換點。



(a)步驟 1

(b)步驟 2

(c)步驟 3

圖 11　偵測多重語者轉換點之示意圖

三、 語者分群(speaker clustering)

語者分群系統

圖 12 為語者分群之系統架構，假設有音段 S1…Si…Sm 進入系統，其分群的步驟如下：

1. 開始時(i=1)，將音段 S1 作為第一個群集，訓練出高斯混合模型[8]。
2. 音段 Si+1 和每個群集訓練出的高斯混合模型作最大概似法計算。
3. 找出最靠近的群集。
4. 若最大概似值大於門檻值，則和最靠近的群集合併，重訓練高斯混合模型，若概似值小於門檻值，則創造新的群集，並訓練其高斯混合模型。
5. i=i+1，回到步驟 2，直到 i > m。

本論文所使用的最大概似法，有作了一點修改，就是多除了一個音段本身的長度 T，原因是因為每個音段長度都不同，導致每個音段算出的概似值無法比較，故除以音段長度 T 可以使其基準都一致，如此便可相互比較。

$$\hat{S} = \arg\max_{1\le k\le N} \frac{1}{T}\sum_{t=1}^{T} \log p\left\langle \vec{seg}_t \middle| \lambda_k \right\rangle \tag{12}$$

門檻值之選定，由實驗中得來，如圖 13 所示，藍色曲線為錯誤創新群曲線，紅色曲線為錯誤合併曲線，當門檻值設的越大，則該合併而沒合併的錯誤率越高，當門檻值設的越小，則該創新群而沒創的錯誤率越高。將門檻值選在兩錯誤曲線交叉的地方，實驗中觀察交叉的值，得知其門檻值為-41.3。

四、 實驗結果與討論

4.1 語者切割實驗

A. 實驗語料

此實驗之語料庫(Data Base)為坊間空中英語教室所轉錄出來，其取樣頻率為 44.1kHz，取樣點位元數為 16bits，由雙聲道轉為單聲道，總共有 210 個語者轉換點。以 512 取樣點為音框長度，

音框位移為 256 點，對每一個音框計算其 12 階梅爾刻度倒頻譜係數(MFCC)，做為 12 維語音特徵向量。同時也計算 MFCC 之差分值(Delta MFCC)及二次差分值(Delta Delta MFCC)，連同 MFCC 分別組成 24 維語音特徵向量，或 36 維語音特徵向量。

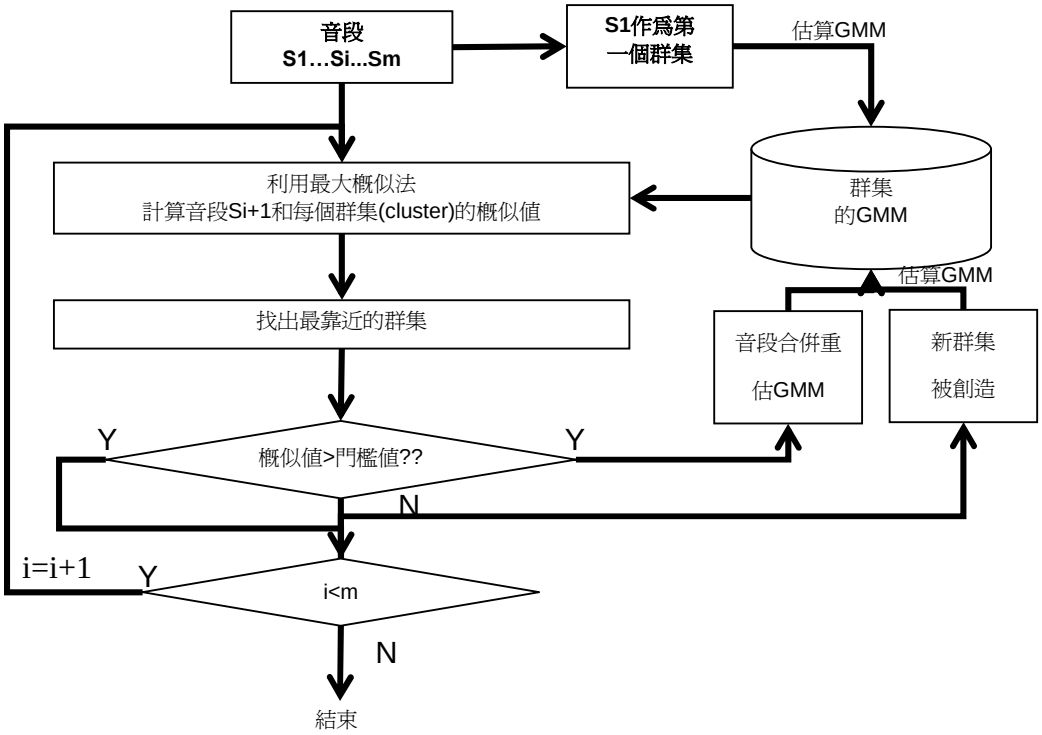


圖 12　語者分群之系統架構圖

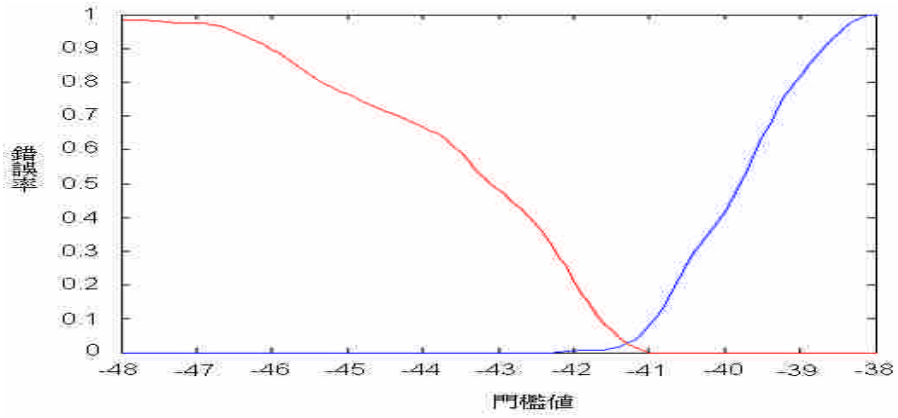


圖 13　合併與分新群之錯誤機率曲線

B. 評估方式[9]

首先定義語者切割會發生的兩種錯誤如下：

1.若在真實轉換點附近(左右 0.6 秒內)，沒偵測到轉換點，則稱這種錯誤為遺失偵測(Miss Detection)。

2.若在偵測出的轉換點附近，沒有真實轉換點存在，則稱此錯誤為假警報(False Alarm)。

利用上述的遺失偵測錯誤和假警報錯誤，可定義出召回率(Recall)和精確度(Precision)：

$$召回率(\mathrm{Re}\,call)=\frac{正確偵測的數目}{正確偵測的數目+遺失偵測} \tag{13}$$

$$精確度(\mathrm{Pr}\,ecision)=\frac{正確偵測的數目}{正確偵測的數目+假警報} \tag{14}$$

將召回率和精確度合併成為一個單一估測值，稱為 F-估測值 (F-measure)：

$$F\text{-評估值}=\frac{2\cdot 召回率\cdot 精確度}{召回率+精確度} \tag{15}$$

F 估測值愈大，代表偵測到的語者轉換點愈準確。

C. 貝氏資訊準則應用到以距離為基礎之順序偵測法與交叉偵測法所偵測到之語者轉換點落點比較

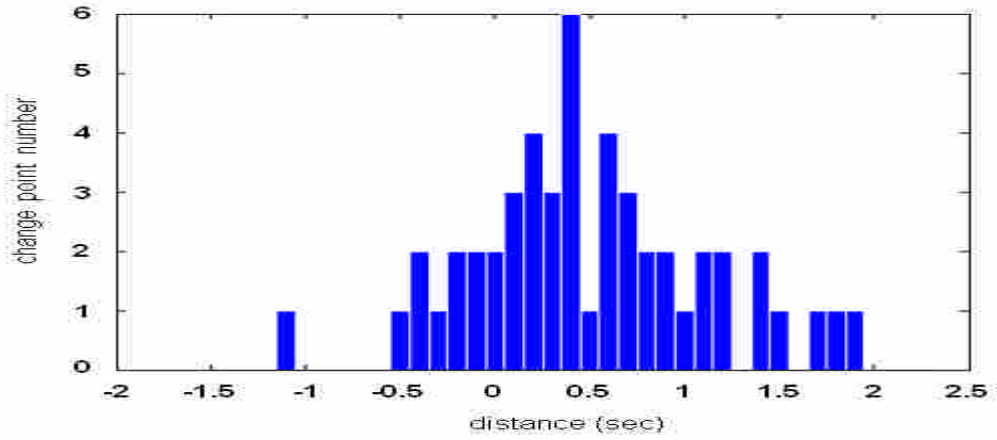


圖 14　貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測出的語者轉換點落點分布之直方圖

圖 14 為貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測出的轉換點落點分布情形，縱軸為轉換點數目，橫軸為偵測出之轉換點與真實轉換點間的距離，以秒為單位，由圖中發現偵測出的轉換點分布大約在真實轉換點左邊 0.5 秒至右邊 2 秒之間。

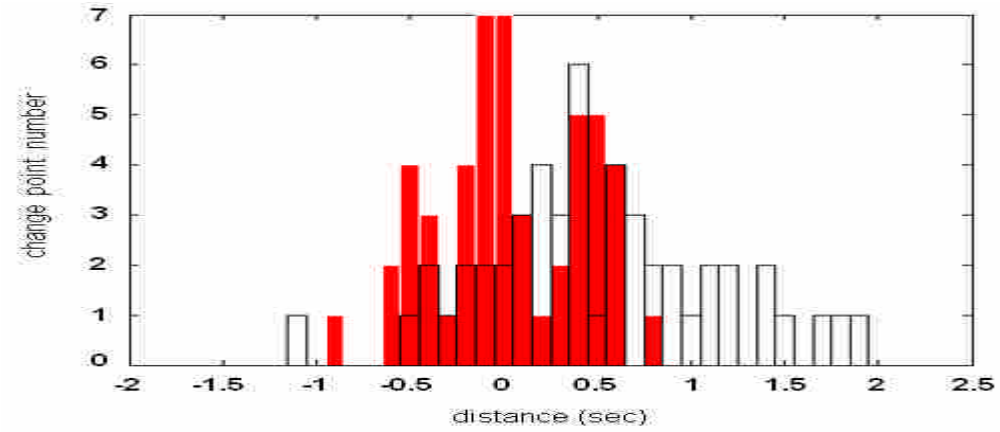


圖 15 貝氏資訊準則應用在以距離為基礎之順序偵測法與交叉偵測法所偵測出的語者轉換點分布之直方圖

圖 15 中紅色的部份為交叉偵測法所偵測出的語者轉換點落點分布直方圖，白色部份為貝氏資訊

準則應用在以距離為基礎之順序偵測法所偵測出的語者轉換點落點分布直方圖，比較兩者，可以發現交叉偵測法所偵測之語者轉換點落點範圍明顯較小，大約是在真實轉換點左邊 0.6 秒至右邊 0.6 秒之間，因此交叉偵測法可以有效將貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測出的語者轉換點作更精確化的處理。

D. 判斷式之影響

在偵測單一語者轉換點時，所用的判斷式對系統效能有何影響，在此針對這個問題作實驗。

	遺失偵測數	假警報數	F-估測
沒加判斷式	18	33	88.26%
加判斷式	22	16	90.81%

表 1　有無判斷式對切割效能的影響

從表 1 中觀察得知，加判斷式後遺失偵測數會有少許的增加，但假警報數卻大量的減少，所以利用此判斷式會犧牲少數對的偵測點來減少多數的假警報數。在整體效能方面，F-估測值從 88.26% 增加到 90.81%，因此加入判斷式對系統偵測是有幫助的。

E. 與貝氏偵測法及廣義概似比法之比較

演算法	花費時間(s)	遺失偵測數	假警報數	F-估測
GLR	66.96	25	48	84.47%
BIC	6714.8	17	29	88.9%
本方法	532.57	22	16	90.81%

表 2　與貝氏偵測法及廣義概似比法所需時間、假警報數、
遺失偵測數與 F-估測之比較

從表 2 中的實驗數據可以發現，廣義概似比法偵測所花費的時間相當短，但對於整體的偵測效能並不是很好。貝氏偵測法剛好和廣義概似比法相反，整體的偵測效能不錯，但偵測所花費的時間相當長。本論文方法，在整體效能方面均優於廣義概似比法和貝氏偵測法，在花費時間方面，雖然比廣義概似比法慢，但卻比貝氏偵測法快的多。

F. 不同特徵維度之影響

在這個實驗中採用不同的特徵維度，觀察本方法在不同特徵維度時，對於語者轉換點偵測是否有不同的影響。

	12 維	24 維	36 維
假警報數	16	18	13
遺失偵測數	22	28	35
F-估量	90.81%	88.78%	87.93%

表 3　不同的特徵維度對偵測效能之影響

由表 3 可看出，增加特徵參數維度並無法增加偵測的成效，尤其在遺失偵測數上，特徵維度越高，遺失偵測數也越多，因此增加特徵維度不但使運算量變大，且對於偵測的效果也沒幫助。

G. 不同取樣頻率之影響

這個實驗是對本論文所使用的方法，觀察其在不同取樣頻率下的效果。

	44.1kHz	32kHz	16kHz	8kHz
假警報數	16	14	39	23
遺失偵測數	22	25	23	50
F-估測	90.81%	90.45%	85.77%	81.42%

表 4　不同取樣頻率對偵測效能之影響

由表 4 可知，本方法在取樣頻率高時可得到較好的成效，在取樣頻率為 32kHz 與 44.1kHz 的情況下，其 F-估測值並不會差太多，而在取樣頻率為 16kHz 時，其 F-估測值已有明顯的下降，在取樣頻率為 8kHz 時，F-估測值降到 81.42%，非常不理想。

4.2 語者分群實驗

A. 實驗語料

本實驗採用三個測試檔案，如表 5 所示，取樣頻率為 44.1kHz，取樣點位元數為 16bits，由雙聲道轉為單聲道。

檔案	音段數	語者數
檔案一	63	3
檔案二	74	5
檔案三	88	7

表 5　三個測試檔案的音段數及語者數一覽表

B. 評估方式[10]

我們計算以下兩個數據，作為評量指標。

1.平均群集純度(Average Cluster Purity,　ACP)

$$p_m = \sum_{j=1}^{R} \frac{n_{mj}^{\ 2}}{n_m^{\ 2}} \qquad acp = \frac{1}{N}\sum_{m=1}^{M} p_m \cdot n_m \tag{16}$$

2.平均語者純度(Average Speaker Purity,　ASP)

$$p_j = \sum_{m=1}^{M} \frac{n_{mj}^{\ 2}}{n_j^{\ 2}} \qquad asp = \frac{1}{N}\sum_{j=1}^{R} p_j \cdot n_j \tag{17}$$

其中 M 為群集的數目，R 為語者的數目， N 為音段的數目。n_m 為第 m 個群集裡的音段數目。n_{mj} 為第 m 個群集裡由第 j 個語者所講的音段數目。n_j 為第 j 個語者所講的音段數目。

進一步的將上述的 acp 與 asp 作幾何平均數，可得到一個單一參數 K：

$$K = \sqrt{acp \cdot asp} \tag{18}$$

以上幾個評估參數的特性如下：

(1)平均群集純度愈高代表該群集包含人數愈接近一。

(2)平均語者純度愈高代表該語者被分配到的群數愈接近一。

(3)K 的值愈大，整體分群效能愈好。

C. 高斯混合數對分群之影響

分別對高斯混合數 2、4、8、16 及 32 作實驗，觀察高斯混合數對 K 值有什麼影響。由圖

16 中可觀察到，隨著高斯混合數的增加，其分群的效能也愈好，而當高斯混合數等於 16 時，*K* 值已達最好結果，繼續增加高斯混合數，*K* 值並不會再增加。

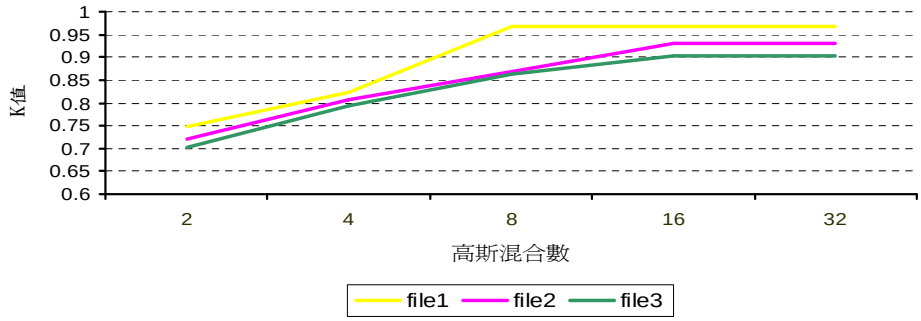


圖 16　高斯混合數對分群之影響

D. 各檔案分群實驗結果

實驗中得知在高斯混合數為 16 時，整體的分群效果已達最好，因此在這個實驗中，將高斯混合數設在 16，來觀察檔案一、檔案二、與檔案三的分群結果。

	實際語者數	群集數	平均群集純度	平均語者純度	K 值
檔案一	3	3	0.970	0.969	0.969
檔案二	5	7	0.974	0.887	0.929
檔案三	7	10	0.960	0.852	0.903

表 6　各檔案分群實驗結果

從表 6 中得知，在包含三個語者的檔案一中，作分群後正確分出三個群集，在包含五個語者的檔案二中，作分群後分出七個群集(多兩個群集)，在包含七個語者的檔案三中，作分群後分出十個群集(多三個群集)。由此可見，當檔案包含的語者數愈多，會錯誤多分出的群集也愈多，而這樣的原因也導致平均語者純度隨語者數的增加而下降。在整體分群效能方面，雖然平均群集純度不太受語者數多寡的影響，但由於平均語者純度的關係，使得 *K* 值也是隨語者數的增加而下降。

五、 結論

本論文探討錄音資料中之語者切割與分群，在語者切割方面，交叉偵測法的確是修正了貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測到的語者轉換點，也証明了利用判斷式作再確認的動作，能有效使假警報數下降。本方法與廣義概似比偵測法及貝氏資訊準則偵測法的比較，從實驗數據中發現，廣義概似比偵測法偵測轉換點花費的時間少，但偵測效能比較差，而貝氏資訊準則偵測法是偵測效能好，但偵測轉換點花費的時間相當長，本方法花費的時間雖比廣義概似比偵測法稍長，但比貝氏資訊準則偵測法卻短很多，且偵測效能為三者之冠，可說是同時擁有廣義概似比偵測法運算量少的優點及貝氏資訊準則偵測法高準確率的優點。在實驗中也發現，增加特徵維度對於整體切割效果並沒有幫助，反而使偵測的效能往下掉，而且偵測轉換點花費的時間

也增多。另外，在取樣率為 32kHz 及 44.1kHz 時，其結果差不多，都有不錯的偵測效能。

在語者分群部份，主要針對 3 個測試檔案做實驗，在實驗中可發現，增加高斯混合數對分群的結果是有幫助的，高斯混合數等於 16 時，其結果已達最好。當要分群的音段群中包含語者數愈多，其整體分群效能愈低。

致謝

本研究受國科會專題研究計畫補助，計畫編號 NSC-93-2213-E-007-019。

參考文獻

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *DARPA Speech Recognition Workshop*, 1998.

[2] 詹順凱,"在多語者環境下之語者分割與語言辨認研究", 電機工程研究所，國立清華大學，中華民國九十一年六月。

[3] Y. Moh, P. Nguyen, and J.-C. Junqua, "Towards domain independent Speaker clustering," *Proc. ICASSP 2003*, pp. I-85-88.

[4] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[5] J.F. Bonastre, P. Delacourt, C. Fredouille, "A Speaker Tracking System Based On Speaker Turn Detection For NIST Evaluation," *Proc. ICASSP 2000*, paper no. 1628.

[6] S. S. Cheng and H. M. Wang, "A sequential metric-based audio Segmentation method via the Bayesian Information Criterion," *Proc. Eurospeech 2003*, pp. 945-948.

[7] A. Adami, S. Kajarekar and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," *Proc. ICASSP 2002*, pp. IV-3908-3911.

[8] D. Reynolds and R. Rose, "Robust test-independent speaker identification using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol.3, No.1, 1995.

[9] J. Ajmera, I. McCowan, and H. Bourlard, "Robust Speaker Change Detection," *IEEE Signal Processing Letters*, pp. 649-651, Vol. 11, No. 8, pp. 649-651, August.2004

[10] I. Lapidot, "SOM as Likelihood Estimator for Speaker Clustering," *Proc. Eurospeech 2003*, pp. 3001-3004.