國　立　交　通　大　學

電信工程學系碩士班

碩士論文

對於公共電視新聞語料之語者與環境轉換偵測

Speaker and environment change detection in PTSND broadcasting news

研究生：黃祺翰

指導教授：王逸如 博士

中華民國九十三年六月

對於公共電視新聞語料之語者與環境轉換偵測

Speaker and environment change detection in PTSND broadcasting news

研究生：黃祺翰　　Student：Chi-Han Huang

指導教授：王逸如 博士　　Advisor：Dr. Yih-Ru Wang

國立交通大學

電信工程學系碩士班

碩士論文

A Thesis
Submitted to Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Electrical Engineering
June 2004
Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

對於公共電視新聞語料之語者與環境轉換偵測

研究生：黃祺翰　　指導教授：王逸如 博士

國立交通大學電信工程學系碩士班

中文摘要

在本論文中，我們提出非監督式的語者及環境轉換偵測，其中使用 GMM 來描述相異兩個聲音片段之統計特性，利用共用的 mixture component(Common Component GMM，CCGMM)來減少估計混合權重時之計算量，以估計出之權重向量來代表聲音片段之特性，進而量測相鄰聲音片段間的相異度，決定出可能之轉換點。我們以公共電視新聞語料(PTSND)來測試我們提出之相異度量測方法，並且與傳統的貝氏資訊法則(BIC)之相異度量測方法比較，有著較佳的轉換點之辨識率，其 MDR 與 FAR 分別為 18.8 %及 16.8 %。

Speaker and environment change detection in PTSND broadcasting news

Student: Chi-Han Huang Advisor: Dr. Yih-Ru Wang

Department of Communication Engineering

National Chiao Tung University

Abstract

In this thesis, a GMM with common mixture components, referred to as the common component GMM(CCGMM), is proposed to be the diversity measure for the speaker and environment change detection in Public Television Service Database (PTSND). The use of GMM is to increase the accuracy of audio signal modeling while the use of common mixture components is to solve the complexity problem of parameter estimation and divergence measure evaluation. Experimental results on the PTSND showed that it outperformed a conventional BIC-based method. A MDR of 18.8% with 16.8% FAR is achieved.

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

根據語者識別(speaker identity)或者環境狀況(environmental condition)乃至於故事邊界(story boundary)來自動地切割聲音片段的技術，在最近幾年已經投入越來越多的努力，例如卡內基美隆大學(Carnegie Mellon University)的 Informedia Digital Video Library project【1】及中研院的 SoVideo 中文影音新聞檢索系統(The SoVideo Mandarin Chinese Broadcasting News Retrieval System)【2】。舉例來說，假設我們要搜尋某一主題的新聞片段，我們必須把有關這個主題的整個聲音片段及文字資訊列舉出來，然而傳統的語音辨認系統並無法藉由文字資訊來切割出主題式的片段，因此考慮到語者或環境狀況轉換的語音切割方法便是需要且直觀地。

對於一個成功的影音檢索系統而言，兩項重要的先決條件是必須的：令人滿意的語音辨識率及精確地對於影音串流做主題式的分段。在本論文中，我們只專注於聲音串流之分段，而對語音辨識而言，可參考在同一個電視新聞語料庫下【3】之實驗方法及結果。

此外，達成主題式聲音串流之分段的方法大概都是分為兩個階段：切割和分類(clustering)；首先先把聲音串流切割成相關性的小片段，例如相同語者的片段或相同環境聲的片段，之後便對切割出來的小片段利用某些方法，如 BIC(Bayesian Information Criterion)【2】等，合併成一個故事片段(story segment)。

1.2 文獻概述

傳統上根據語者識別及環境狀況的切割方法大致可分為 feature-analysis segmentation、model-based segmentation 及 metric-based segmentation 三種，簡介如下：

1. Feature-analysis segmentation【4】：

   此方法是利用某些特徵參數，例如 HZCRR(High Zero Crossing Rate Ratio)、LSTER(Low Short-Time Energy Ratio)、SF(Spectrum Flux)等，對於某些聲音類型如音樂、靜音、噪音有不錯的鑑別度，對聲音串流做有系統的分類，進而達成聲音的切割。

2. Model-based segmentation：

   此方法是替不同的聲學群組(acoustic class)建立不同的模型，例如 GMM 或 HMM 等；舉例來說若我們要切割電視新聞的話，我們便可以為棚內主播、外場記者、外場受訪者等建立個別的模型，之後測試的聲音串流經由 sliding window 便可以依照既有模型去算出此分析音框的似然機率(likelihood probability)，進而可決定轉換點。

3. Metric-based segmentation：

   此方法是利用距離量測的概念，選擇某一相異度量測公式，例如 KL distance【5】、KL2 distance 等，測量相鄰兩個分析音框間的相異度，在經過某一 decision rule 決定轉換點。

基本上每種切割的方法都有其缺點及限制，例如 feature-analysis segmentation 只能粗略地分類聲音串流成語音、音樂、背景雜音、靜音等，

並不能再往上合併成語者片段【4】；不管是 model-based segmentation 或者是 metric-based segmentation 都需要訂定某一臨界值來決定聲音片段為哪一個群組，這使得此兩種方法缺乏穩定性(stability)及強健性(robustness)；此外，model-based segmentation 不能處理未曾出現過的群組，只能分辨出已經訓練過模型的群組。

因此混合、修正許多不同切割方法或者提出創新的切割方法便是提升效能的不二法門；Microsoft Research Asia 便針對廣播新聞之語者轉換偵測便做了相當程度的努力【4,6】，他們利用 divergence shape distance【7】作為其距離量測工具，切割出較短的聲音片段，之後利用 GMM 語者模型合併成一個較長的語者片段，而且利用合併的結果更新語者模型，使合併結果更能反映語者之轉換；IBM T. J. Watson Research Center 利用 BIC (Bayesian Information Criterion)來做轉換點的偵測(change point detection)，得到了環境轉換點，然後再依照偵測出之轉換點利用 BIC 執行階層式分群(hierarchical clustering)的動作，就此獲得語者轉換點【8】。

此外，電視新聞之切割與分類除了可以利用聲音來偵測轉換點，還可以融合影像處理的技術，例如人臉辨識、black frame 等，以其得到更精確的語者及故事切割。例如卡內基美隆大學(Carnegie Mellon University)主持了一項有關數位影音圖書館(Digital Video Library)之計畫【1】--Informedia Digital Video Library Project--便是融合了聲音切割及影像處理的技術，對於電視新聞做主題式的切割，達成資訊擷取(Information Retrieval)之目的；IBM T. J. Watson Research Center 對於電視新聞之聲音及影像抽取不同的特徵參數【9】，例如 motion、face、music/speech types、prosody 及 text information，利用 ME (Maximum Entropy) 統計模型來融合這些不同形式的特徵參數，進而可以對電視新聞片段分類，達到語者及故事切割之目的。

1.3 章節簡介

本論文總共分為四章，摘要如下：

1. 在第一章中我們簡介了語音切割傳統上的作法，以及國內外相關研究所用之方法。
2. 在第二章中我們詳細地描述了 CCGMM 的由來，以及推導出其相異度量測(divergence measure)之公式；此外我們也對一些問題之解決方法，以及相異度測量公式之修正做了詳盡的說明。
3. 在第三章中我們先簡介實驗的語料庫--公共電視新聞語料(Public Television Service News Database)之基本統計特性及其標註內容(transcription)的方式，之後我們對我們提出的相異度量測方法作效能上的分析。
4. 在第四章中我們總結我們提出的距離量測公式之優劣，並且點出未來之展望。

第二章 基本及相關應用原理

2.1 廣播新聞語料轉換點偵測之緣起

對於廣播電視新聞之切割而言，通常一個非監督式(unsupervised)語者及環境狀況之轉換點偵測為第一個階段，然後再基於第一階段所得到的候選轉換點來達成語者或故事切割與追蹤(tracking)。但是就語者與環境狀況之轉換點偵測而言，要達成令人滿意的結果是相當困難的，因為在電視新聞語料中包含了許多不同的聲音來源，如不同的語者、不同的說話方式、背景人聲及非語音的聲音等，在在都使得語者及環境狀況之轉換點偵測變得十分困難；此外，不同語者的錄音環境，如棚內主播、外場記者及外場受訪者等，也會增加其轉換點偵測之困難度。

常被用來解決此問題的方法有三種，如第一章所提到的，而最常使用的方法就是描述出候選轉換點左右兩個聲音片段之聲學上的統計特性，進而利用某種相異度量測(divergence measure)得到此兩個聲音片段之相異度(divergence)。例如 Kullback Leibler(KL) distance、symmetric Kullback Leibler(KL2) distance【5】、divergence shape distance【6】及 Bayesian Information Criterion(BIC)【2,8,10】等相異度量測方法。這些相異度量測方法都有著下列假設：在 2～3 秒聲音片段中的特徵向量(feature vector)之機率分佈為一高斯分佈；直覺上，此假設似乎不是非常合理，但是更精準的特徵向量模型，如混合高斯模型(Gaussian Mixture Model，GMM)，會造成估計模型參數及相異度量測時計算複雜度的增加。

在本論文中，我們提出有著相同 mixture component 的 GMM--我們稱

之為 Common Component GMM(CCGMM)--來描述聲音片段之統計特性，進而利用 CCGMM 之參數來做相異度量測，以達成對於電視新聞語料語者與環境狀況轉換之偵測。我們使用混合高斯模型來精確地描述聲音片段之統計特性，而利用相同的 mixture component 來降低在參數估計及相異度量測時的計算複雜度。

2.2 混合高斯模型之定義

混合高斯模型(Gaussian Mixture Model，GMM)是由不同權重的高斯混合機率密度函數所組成之機率密度函數,其形式如下列所示：

$$p(\mathbf{x}|\Lambda)=\sum_{k=1}^{M} w_k b_k(\mathbf{x}) \tag{2.1}$$

其中 $\mathbf{x}$ 是 D 維之隨機向量，$b_k(\mathbf{x}), k=1,\cdots,M$ 為不同權重之混合機率密度函數，$w_k, k=1,\cdots,M$ 為其相對應之混合權重(mixture weight)，且混合權重有 $\sum_{k=1}^{M} w_k=1$ 之限制。每個混合機率密度為 D 維的高斯函數，有著下列之形式：

$$b_k(\mathbf{x})=\frac{1}{(2\pi)^{D/2}\left|\mathbf{C}_k\right|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{C}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\} \tag{2.2}$$

其中 $\boldsymbol{\mu}_k$ 與 $\mathbf{C}_k$ 分別為高斯混合機率密度之平均向量(mean vector)及共變異矩陣(covariance matrix)。

圖 二-1 為一 GMM 之例，(a)~(d)為四個不同的平均值與變異數之 mixture component，亦即 $b_k(\mathbf{x}), k=1,2,3,4$，(e)為混合高斯模型 $p(\mathbf{x}|\Lambda)$，其四個相對應之混合權重，$w_k, k=1,2,3,4$，分別為 0.05、0.5、0.1 與 0.35。

一個完整的混合高斯模型主要是由所有混合高斯函數之平均向量、共變異矩陣及其相對應之混合權重等參數所描述，而這些參數可以用下列記號來表示

$$\Lambda=\left\{w_k,\boldsymbol{\mu}_k,\mathbf{C}_k\right\}, k=1,\cdots,M$$

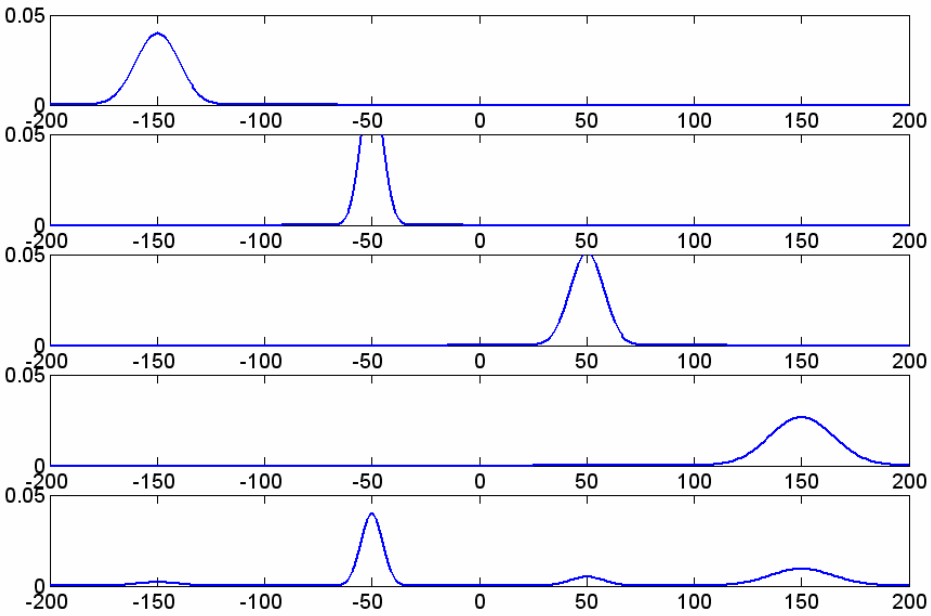

圖 二-1 4 個 mixture component 之 GMM 範例

2.3 混合高斯模型之描述

為什麼在本論文中我們要用 GMM 來描述某一個群組(class)的特性呢？最主要的動機為 GMM 描述任意的機率分佈可以有相當不錯的結果；而在本論文中，某個聲音片段中特徵向量之分佈可視為隨機分佈，因此必須找到一個可以完整描述其分佈之模型，再找出某一種可以測量兩個特徵向量分佈之相異度，使相異度測量可以更精確，所以我們使用 GMM 描述聲音片段之特徵向量的動機因此產生。

圖 二-2 說明了 GMM 對於某一聲音片段中特徵向量之分佈的能力【11】；圖 二-2(a)為某一聲音片段中某一維倒頻譜參數(cepstral coefficient)的 pdf，圖 二-2(b)則是用 unimodal Gaussian model 來描述此聲音片段，圖二-2(c)為使用 10 個 mixture component 的 GMM 之 pdf，圖 二-2(d)則為利用 10 個 codeword 之 VQ 所做的結果；從圖中我們可以明顯地看出 GMM 對於任意分佈的機率密度函數皆可以描述得很不錯，而這也是為什麼我們要使用 GMM 來描述某一個聲音片段的統計特性。

我們藉由 GMM 之混合權重、平均向量及共變異矩陣來代表每個聲音片段的特性，概括來說，平均向量可以代表聲音片段在特徵空間(feature space)的絕對位置，共變異矩陣描述了此聲音片段中特徵向量的分散程度，而混合權重可視為描述此聲音片段細節的參數，也正因上列三個參數造就了 GMM 的優點：描述任意聲音片段其特徵向量分佈的能力。

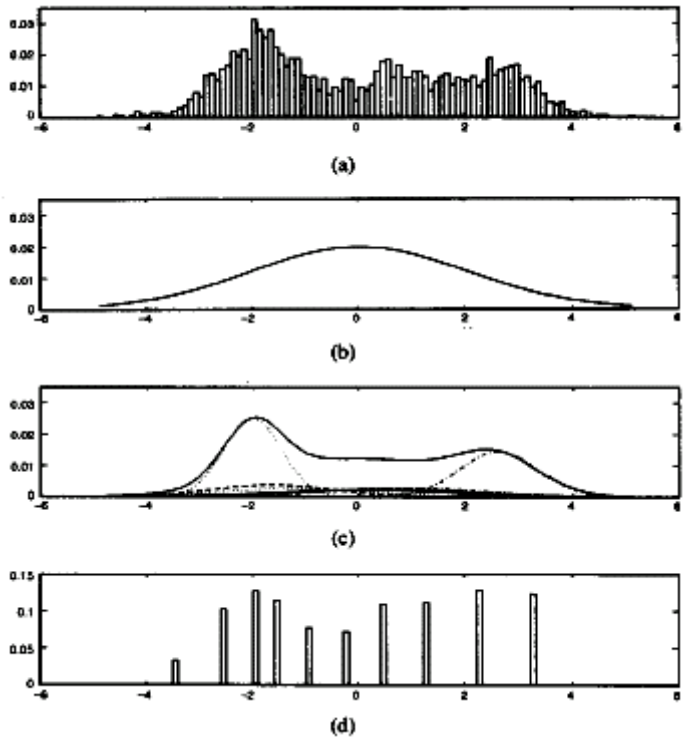

圖 二-2 GMM 與 VQ 對於特徵向量的描述能力之比較

2.4 利用最大似然機率參數估計法估計 GMM 之模型參數

我們提出的基於 CCGMM 之相異度量測方法中，首先必須要有 GMM 之模型參數，也就是共用的平均向量與共變異矩陣。所以我們利用訓練語料之聲音串流，估計代表此語料庫特徵向量分佈特性之 GMM 模型參數 Λ。

有許多方法可以估計混合高斯模型的參數；到目前為止，最常見的方法是最大似然機率估計法(Maximum Likelihood Estimation, MLE)。使用最大似然機率估計法其目的在於：對於一群給定的訓練語料，找到使得 GMM 的似然函數(likelihood function)最大的一組模型參數 Λ；也就是說，對於一個有 T 組獨立之訓練特徵向量 $X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_T\}$，其 GMM 的似然機率函數可以寫成如(2.3)式之形式：

$$p(X\mid\Lambda)=\prod_{t=1}^{T}p(\mathbf{x}_t\mid\Lambda) \qquad (2.3)$$

不幸地，某個訓練特徵向量屬於哪一群組(class)對於我們而言是未知的，所以我們借用 EM 演算法中對於 ML 的遞迴關係式來估計 Λ。

EM 演算法的基本概念如下：給定一初始模型參數 $\overline{\Lambda}$，欲找到新的模型參數 Λ 使得 $p(X\mid\Lambda)\geq p(X\mid\overline{\Lambda})$，更新初始模型參數後，上述步驟重複至某個收斂臨界值；相同的概念我們也用來估計 HMM 的模型參數，也就是著名的 Baum-Welch 演算法。在每一次重複的步驟中，下列的更新公式可以保證我們 GMM 模型參數的似然機率值單調遞增(monotonic increase)：

首先我們定義第 i 群的事後機率(a posteriori prob.)如下：

$$p(i \mid \mathbf{x}_t, \Lambda) = \frac{w_i b_i(\mathbf{x}_t)}{\sum_{k=1}^{M} w_k b_k(\mathbf{x}_t)} \qquad (2.4)$$

則經過每次的疊代(iteration)後，GMM 模型參數之更新公式如下所示，假設共變異矩陣為對角矩陣，也就是 $c_{kij} = \begin{cases} \sigma_{kd}^2 & i = j = d \\ 0 & \text{others} \end{cases}$，

混合高斯權重：

$$w_k = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \mathbf{x}_t, \Lambda) \qquad (2.5)$$

混合高斯平均向量：

$$\boldsymbol{\mu}_k = \frac{\sum_{t=1}^{T} p(i \mid \mathbf{x}_t, \Lambda) \mathbf{x}_t}{\sum_{t=1}^{T} p(i \mid \mathbf{x}_t, \Lambda)} \qquad (2.6)$$

混合高斯變異數：

$$\sigma_{kd}^2 = \frac{\sum_{t=1}^{T} p(i \mid \mathbf{x}_t, \Lambda) x_{td}^2}{\sum_{t=1}^{T} p(i \mid \mathbf{x}_t, \Lambda)} - \mu_{kd}^2 \qquad (2.7)$$

其中 *k*, *t*, *d* 分別為 mixture component、音框(frame)及維度的指標(index)。在本論文中，似然機率值之相對變化小於 0.1%或者疊代的次數超過 50 次，我們便認為 EM 演算法已經收斂。

2.5 利用 GMM 描述聲音片段特性時相異度之定義

非監督式的語音切割一般用高斯機率密度函數來描述聲音片段中特徵向量之分佈，再量測相鄰兩個聲音片段特徵向量分佈之相異度，但是高斯分佈的假設缺乏合理性，因此我們便想要以 GMM 來精確描述聲音片段；然而為了降低計算相異度時之運算量，每個聲音片段共用 mixture components，亦即本論文提出的基於 CCGMM 之相異度量測，接下來我們推導其相異度量測如下：

根據消息理論，兩個機率分佈 $p_{right}(\mathbf{x})$ 與 $p_{left}(\mathbf{x})$ 間的相異度可用(2.8)式來量測【7】，

$$D=\int\left[p_{right}(\mathbf{x})-p_{left}(\mathbf{x})\right]\ln\frac{p_{right}(\mathbf{x})}{p_{left}(\mathbf{x})}d\mathbf{x} \tag{2.8}$$

在本篇論文中，$p_{right}(\mathbf{x})$ 與 $p_{left}(\mathbf{x})$ 所代表的意義為某一時間點的左右兩個聲音片段中特徵向量分佈的機率密度函數。

先前我們提到了混合權重可代表此段聲音之細節，我們認為此項參數比較能代表聲音片段的特性，因此假設所有相鄰聲音片段共用一組 GMM 之平均向量及共變異矩陣，也就是本論文所提出之 CCGMM。我們可以利用 CCGMM 之混合權重來代表聲音片段的特性，亦即

$$\begin{aligned} p_{right}(\mathbf{x})&=\sum_{k=1}^{M}w_{right,k}b_k(\mathbf{x})\\ p_{left}(\mathbf{x})&=\sum_{k=1}^{M}w_{left,k}b_k(\mathbf{x}) \end{aligned} \tag{2.9}$$

其中 $b_k(\mathbf{x})=\frac{1}{(2\pi)^{1/2}\left|\mathbf{C}_k\right|^{D/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{C}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\},k=1,\cdots,M$ ，為由訓練語料

訓練出來之 CCGMM 之平均向量和共變異矩陣所組成的 mixture component。

我們可以把這種方式想像成一種將特徵向量從較高維度的倒頻譜空間(cepstral space)轉換到某一較低維度之混合權重空間(mixture-weight space)，每一個 mixture component 我們可視為在此空間之基底(basis)，而此空間轉換的過程便是對特徵向量做基底展開。在本篇論文中我們稱之為特徵向量轉換(feature transformation)，如圖 二-3 所示，其中 X 代表某一聲音片段之特徵向量的集合，經過特徵向量轉換後，我們可以用一組混合權重向量來代表此聲音片段之特性；然而由於空間已由倒頻譜空間轉換到了混合權重空間，因此我們必須找到適用於混合權重空間之距離量測公式。

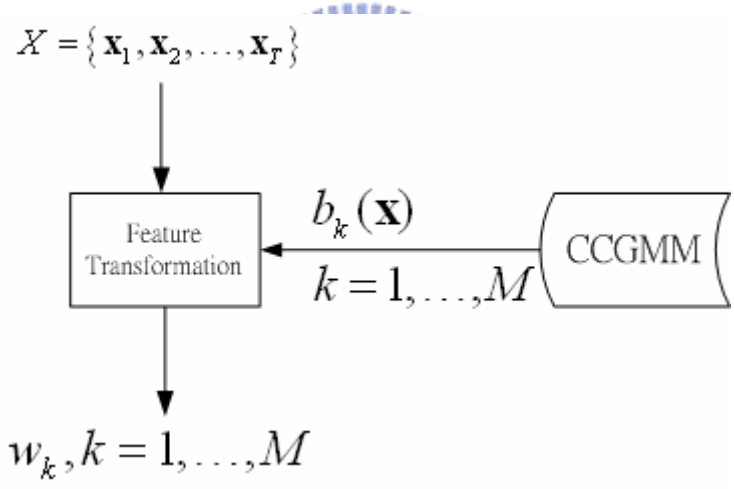

圖 二-3 特徵向量轉換之示意圖

首先我們把(2.9)式代入(2.8)式中，可得

$$D = \int_{RX_k} \left\{ \sum_{k=1}^{M} w_{right,k} b_k(\mathbf{x}) - \sum_{k=1}^{M} w_{left,k} b_k(\mathbf{x}) \right\} \ln \frac{\sum_{k=1}^{M} w_{right,k} b_k(\mathbf{x})}{\sum_{k=1}^{M} w_{left,k} b_k(\mathbf{x})} d\mathbf{x} \qquad (2.10)$$

因為積分為線性運算，所以把積分與 summation 互換之後如下式所示，

$$D=\sum_{k=1}^{M}\int_{RX_k}\left[w_{right,k}-w_{left,k}\right]b_k(\mathbf{x})\ln\frac{\sum_{k=1}^{M}w_{right,k}b_k(\mathbf{x})}{\sum_{k=1}^{M}w_{left,k}b_k(\mathbf{x})}d\mathbf{x} \qquad (2.11)$$

在此我們為了簡化相異度量測時之計算複雜度，因而假設在 $RX_k, k=1,\ldots,M$ 中，混合高斯分佈可近似為單一高斯分佈，亦即

$$\sum_{k=1}^{M}w_{right,k}b_k(\mathbf{x})\approx w_{right,k}b_k(\mathbf{x}) \quad \forall\mathbf{x}\in RX_k \qquad (2.12)$$

因此相異度量測公式(2.11)式在(2.12)式的假設下，可近似為下列形式：

$$\begin{aligned}D&\approx\sum_{k=1}^{M}\int_{RX_k}(w_{right,k}-w_{left,k})b_k(\mathbf{x})\ln\frac{w_{right,k}b_k(\mathbf{x})}{w_{left,k}b_k(\mathbf{x})}d\mathbf{x}\\&\approx\sum_{k=1}^{M}(w_{right,k}-w_{left,k})\ln\frac{w_{right,k}}{w_{left,k}}\int_{RX_k}b_k(\mathbf{x})d\mathbf{x}\end{aligned}$$

因為我們已經假設在 RX_k 中為單一高斯分佈，因此 $\int_{RX_k}b_k(\mathbf{x})d\mathbf{x}\approx 1$，所以兩個聲音片段間的相異度便可以化簡成下列形式：

$$D\approx\sum_{k=1}^{M}(w_{right,k}-w_{left,k})\ln\frac{w_{right,k}}{w_{left,k}} \qquad (2.13)$$

與(2.8)式比較，兩者有相同的形式；我們可以把(2.8)式視為兩個連續機率分佈間的相異度，而(2.13)式則為兩個離散機率分佈間的相異度。

在本論文中，為了計算出相鄰兩個聲音片段間的相異度，我們採用 GMM 之平均向量與共變異矩陣所組成的 mixture component 當作每個聲音片段的基底，也就是本論文提出之 CCGMM，接下來再利用(2.5)式估計出相鄰聲音片段的混合權重，亦即 $w_{right,k}, k=1,\ldots,M$ 與 $w_{left,k}, k=1,\ldots,M$ ，最後藉由(2.13)式計算出相鄰聲音片段之相異度。顯然地，當我們要估計出聲音片段的混合權重時，也就是當我們要做特徵向量轉換時，我們可以假設在 EM 演算法的每次疊代中，平均向量和共變異矩陣為定值，單純地只更新每個聲音片段的混合權重，其更新公式之形式與(2.5)式相同。

為了加快當我們利用(2.5)式更新混合權重時的收斂速度，小心地設定其初始值是必要的，因此每個混合權重的初始值可以如下式所示：

$$w_k = \frac{\left(\sum_{t=1}^{T}\sum_{I_{\mathbf{x}_t}=k}^{M} 1\right)+1}{\left(\sum_{t=1}^{T} 1\right)+M}, \quad k=1,\cdots,M \tag{2.14}$$

其中 $I_{\mathbf{x}_t} = \arg\max_k b_k(\mathbf{x}_t), k=1,\ldots,M$ ，其中 $\mathbf{x}_t$ 為時間指標 t 的特徵向量。

2.6 利用信號相異度做轉換點偵測之系統架構

2.6.1 快速 CCGMM 權重參數之抽取

在本論文中我們以 3 秒的分析音框(analytic window)中特徵向量之分佈來代表轉換點左右的聲音片段特性，而這似乎不合理且缺乏彈性，因為我們不能保證 3 秒是否足以代表整個聲音片段之特性，所以可變的分析音框長度是必需的。然而在本實驗中混合權重之更新公式，亦即(2.5)式，是最耗費計算量的步驟，因此我們希望在整個實驗過程中，加大或縮小分析音框長度時不要再重新估計混合權重向量。基於此理由，我們先以 50 frame 長的分析音框估計權重向量，得到混合權重序列(mixture weights series)，$\mathbf{w}_{sub}(i), i=1,\ldots$，之後再以(2.15)式組合成任意長度 $N\times 50$ frame 的分析音框之混合權重向量，$\mathbf{w}$，

$$\mathbf{w}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{sub}(i) \qquad (2.15)$$

而此步驟在本論文中我們稱之為混合權重近似(weight approximation)。

圖 二-4 為(2.15)式之示意圖；在圖 二-4 中，我們先估計出以分析音框長度為 50 frame 來估計出混合權重向量之序列，$\mathbf{w}_{sub}(i)$，之後再利用(2.15)式來近似分析音框長度為 300 frame 的混合權重向量，$\mathbf{w}$；由此可看出一旦我們估計出整段聲音串流的 50-frame 長之混合權重向量，我們便可在任意時間點(以 0.5 sec 為單位)求出在其時間點，左右任意分析音框長度之聲音片段的相異度。

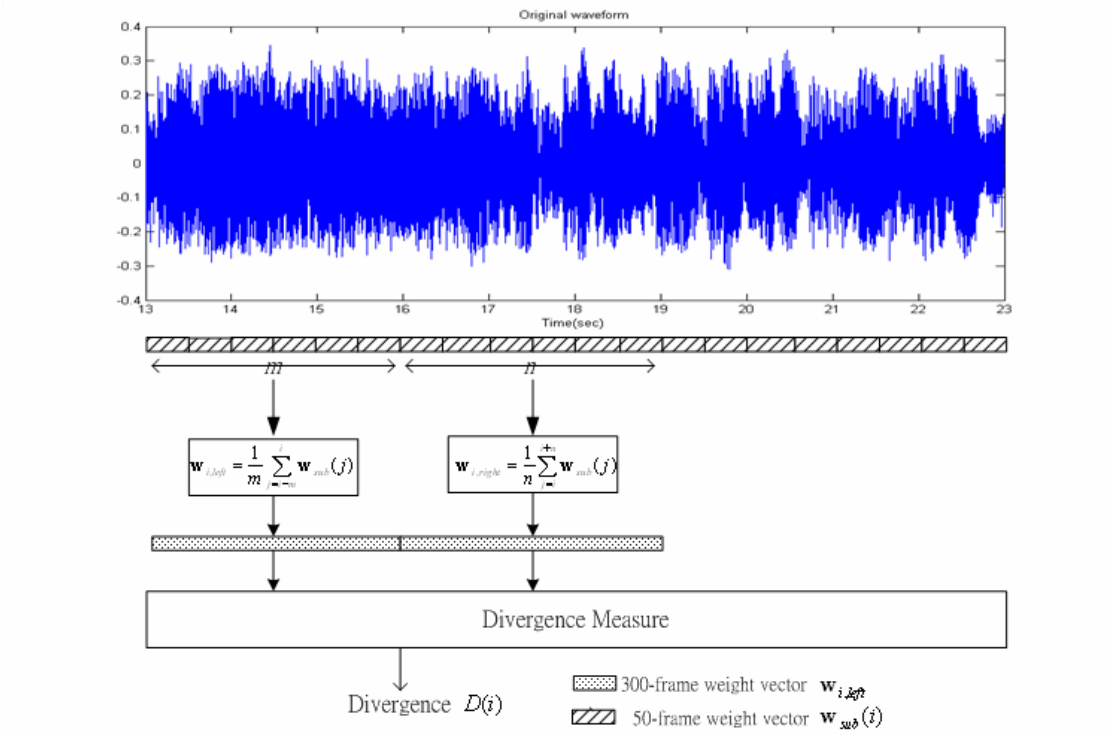

圖 二-4 相異度量測之示意圖

2.6.2 相異度量測之加權

此外，若我們只用(2.13)式所計算出來的相異度做 decision，發現不能用簡單的 decision rule 來找出候選轉換點，這是 metric-based segmentation 的缺點之一；如同 Microsoft Research Asia 在【6】中為了過濾出準確且數目合理的候選轉換點，所提出的 decision rules 便相當的複雜。

為了解決上述問題，我們觀察相異度序列 $D(i)$ 之起伏，如圖 二-5 所示，發現若在相鄰之分析音框中包含了真實轉換點，也就是在真實轉換點附近時，其相異度 $D(i)$ 便會相對較高，因為相鄰分析音框之特性差異已經逐漸地增加；而當真實轉換點落在相鄰兩個分析音框之交接的邊界時，相

異度序列便會出現峰值，因此我們可以藉助匹配率波器的概念來加強相異度序列(divergence series)中峰值的大小，增加候選轉換點和非轉換點間振幅的差距，也就是說對原本的相異度序列與三角波做迴旋積分(convolution)；因此經過此步驟後的相異度序列，$D'(i)$，可用(2.16)式表示：

$$D'(i)=\sum_{j=i-W}^{i+W}\left(1-\left(\frac{i-j}{W}\right)^{\alpha}\right)D(j) \qquad (2.16)$$

其中 α 為 1 時，表示 $D(i)$ 與一個三角波做迴旋積分，而 W 為此三角波的寬度；在本實驗中為了考慮在(2.13)式中自然對數之比值也會反應出聲音片段間的相異度，因此我們把 α 設為 2。

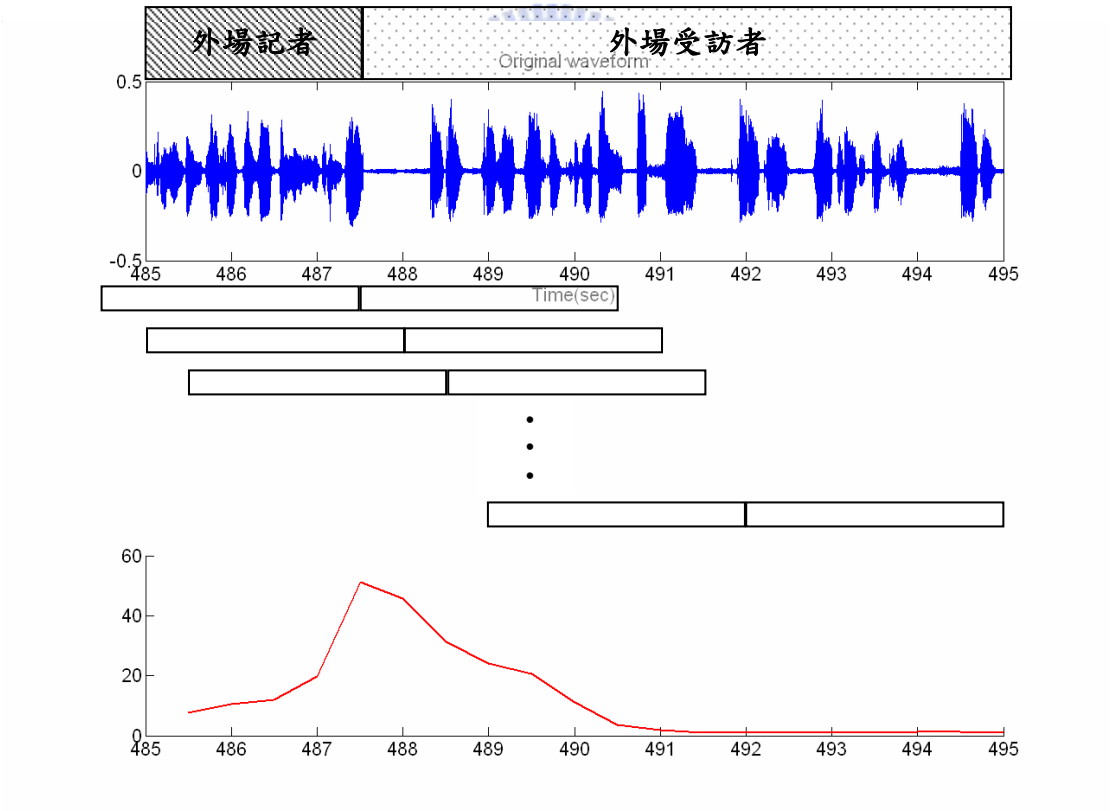

圖 二-5 相異度序列之範例

接下來我們以圖例來說明 $D'(i)$ 與 $D(i)$ 之差異；圖 二-6(b)為原始的聲音波形，其中包含了廣告、純音樂及主播聲音片段，並且標註了背景環境狀況，如圖 二-6(a)所示；而圖 二-6(c)呈現了相對應時間之 $D'(i)$ 與 $D(i)$；我們可以很明顯地看出轉換點都落在 $D'(i)$ 的峰值，而且在非轉換點時，$D'(i)$ 始終相對較低，這也隱含了我們在做 decision 時，可以使用簡單的 threshold-based decision rules。

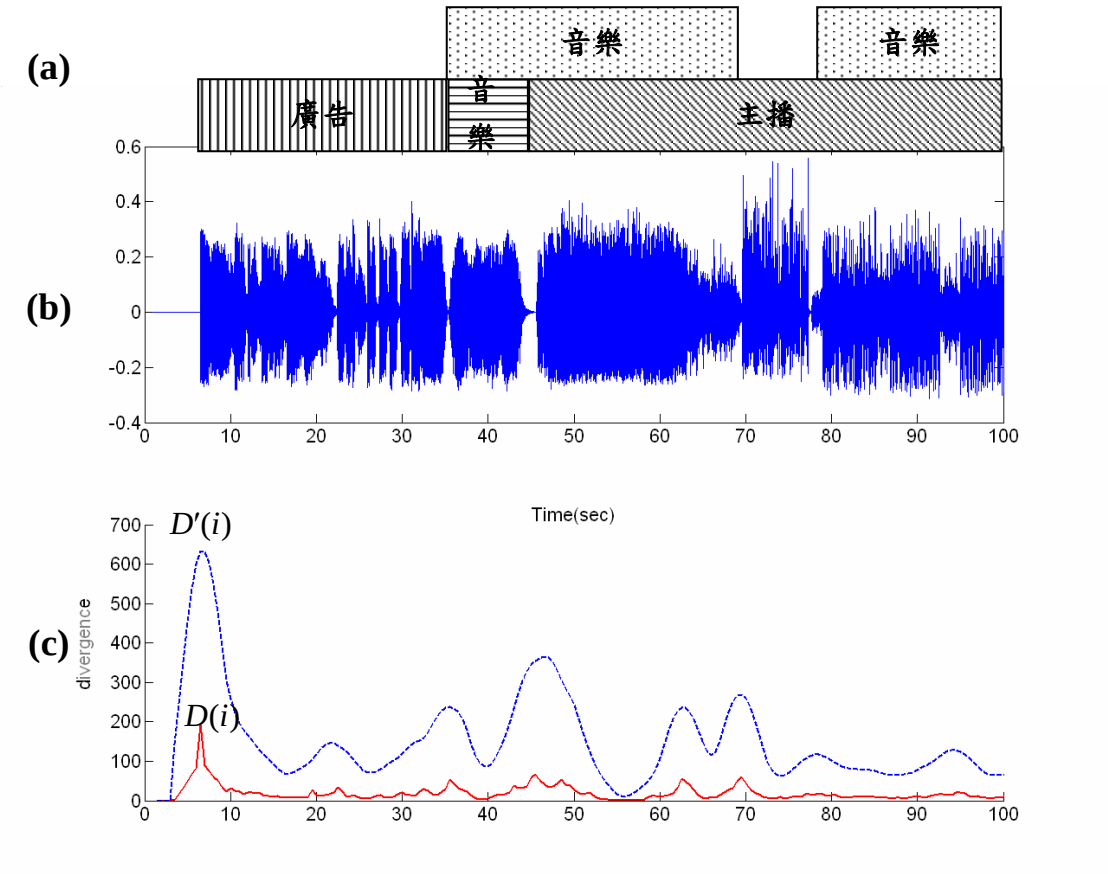

圖 二-6 $D(i)$ 與 $D'(i)$ 之比較

2.6.3 屬於靜音的 mixture component 之移除

此外，在我們做相異度量測時，發現若相鄰聲音片段間之靜音在數量上不平衡，會影響到相異度量測之結果，因而我們可以把屬於靜音的 mixture component 在相異度量測的過程中忽略，進而減輕此現象。假設

$\forall s \in \mathbf{S}$ 為屬於靜音的 mixture component，則經過忽略靜音 mixture component 後的權重向量變成了下列結果：

$$\mathbf{w}_{new} = \frac{1}{\sum_{k \notin s} w_k}\left[w_k, k = 1,\ldots,M; \text{and } i \notin s\right] \tag{2.17}$$

2.6.4 使用 Global covariance matrix

最後，假設每個 mixture component 有各自之共變異矩陣，因而可能會發生如圖 二-7(a)之情境，這會使得在 $RX_k, k = 1,\ldots,M$ 中不再可以近似為單一高斯分佈，所以為了使得(2.12)式之假設更為合理，我們假設在 $RX_k, k = 1,\ldots,M$ 中有相同的共變異矩陣，如圖 二-7(b)所示，亦即每個 mixture component 有相同的共變異矩陣；如此，我們在(2.13)式做相異度量測時，以 discrete convergence 相似之誤差量會變為較小。因此我們在估計 CCGMM 之模型參數時，共變異矩陣之更新公式如下所示：

$$\sigma_d^2 = \sum_{k=1}^{M}\left\{\frac{\sum_{t=1}^{T} p(i \mid \mathbf{x}_t, \lambda) x_{td}^2}{\sum_{t=1}^{T} p(i \mid \mathbf{x}_t, \lambda)} - \mu_{kd}^2\right\} \tag{2.18}$$

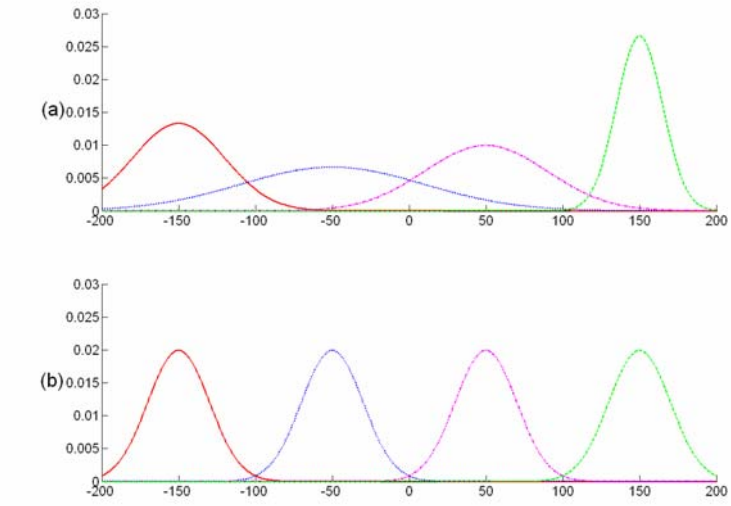

圖 二-7 使用 global covariance 之示意圖

2.6.5 系統架構簡述

接下來我們簡述一下整個實驗的流程如圖 二-8 所示；首先我們先拿一天的節目(i.e.一小時)來訓練出 GMM 之模型參數Λ，其中當我們更新模型參數時，每個 mixture component 共用同一個共變異矩陣，也就是利用(2.18)式之更新公式，然後保留其平均向量及共變異矩陣，$\{\boldsymbol{\mu}_k, \mathbf{C}_k; k=1,\ldots,M\}$，亦即估計出電視新聞語料在特徵空間之基底，$b_k(\mathbf{x}), k=1,\ldots,M$。一旦我們有了基底後，我們接下來便以分析音框長度為50 frame 的特徵向量做基底展開，得到 50-frame 權重向量序列，$\mathbf{w}_{sub}(j), j=1,\ldots$，之後我們就只對此序列做處理。在經過(2.15)式的混合權重近似之步驟後，我們便可以得到在某一時間點左右任意分析音框長度的聲音片段之權重向量，之後利用(2.13)式求出在此時間點相鄰的兩個聲音片段之相異度；在此處要注意的是我們已經移除了靜音的mixture component。

因為利用(2.13)式求出的結果對於聲音片段間的差異過於敏感，要過濾出有效且數目合理的轉換點並不容易，所以我們利用匹配濾波器的概念對於$D(i)$做處理，來增加候選轉換點與非轉換點間相異度的差距，如(2.16)式所示，使得當我們要決定候選轉換點時，可以直接定一個臨界值決定；以下是本實驗中候選轉換點的決定規則：

$$\begin{aligned} &D'(i) > D'(i \pm 1), \\ &D'(i) > D'(i \pm 2), \\ &D'(i) > TH \end{aligned} \qquad (2.19)$$

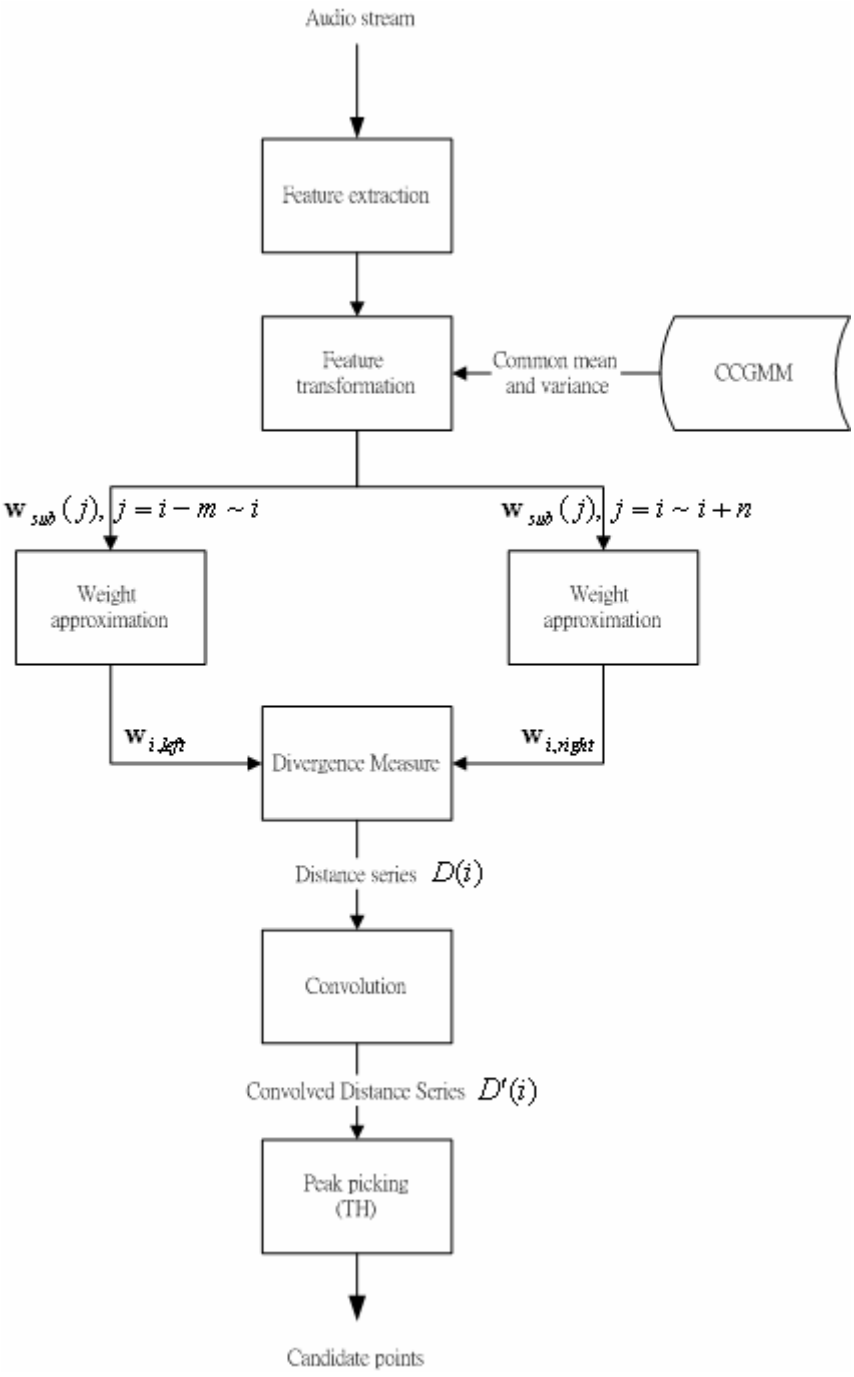
Audio stream
Feature extraction
Feature transformation
Common mean and variance
CCGMM
$\mathbf{w}_{sub}(j), j = i - m \sim i$
$\mathbf{w}_{sub}(j), j = i \sim i + n$
Weight approximation
Weight approximation
$\mathbf{w}_{i,left}$
$\mathbf{w}_{i,right}$
Divergence Measure
Distance series $D(i)$
Convolution
Convolved Distance Series $D'(i)$
Peak picking (TH)
Candidate points

圖 二-8 基本系統架構圖

圖 二-9 為套用(2.19)式後所得到的結果，其中我們設定 TH=100；在圖中我們可以看到與圖 二-6 相似的結果。我們可以看出在此段聲音中語者類別依序為受訪者 1、外場記者、受訪者 2、受訪者 3、外場記者、受訪者 4、外場記者、受訪者 5，而在圖 二-9 中我們發現受訪者 2 與受訪者 3 間的真實轉換點無法偵測出來，經過實際聽過音檔後發現兩者有著相同的背景聲且兩者皆為男性語者，因此較難偵測出來，但是從圖中我們可看出 $D'(i)$ 在其附近仍有峰值，只是其高度沒有大於 TH，也就是說 $D'(i)$ 仍然有反映出其相異度的變化。

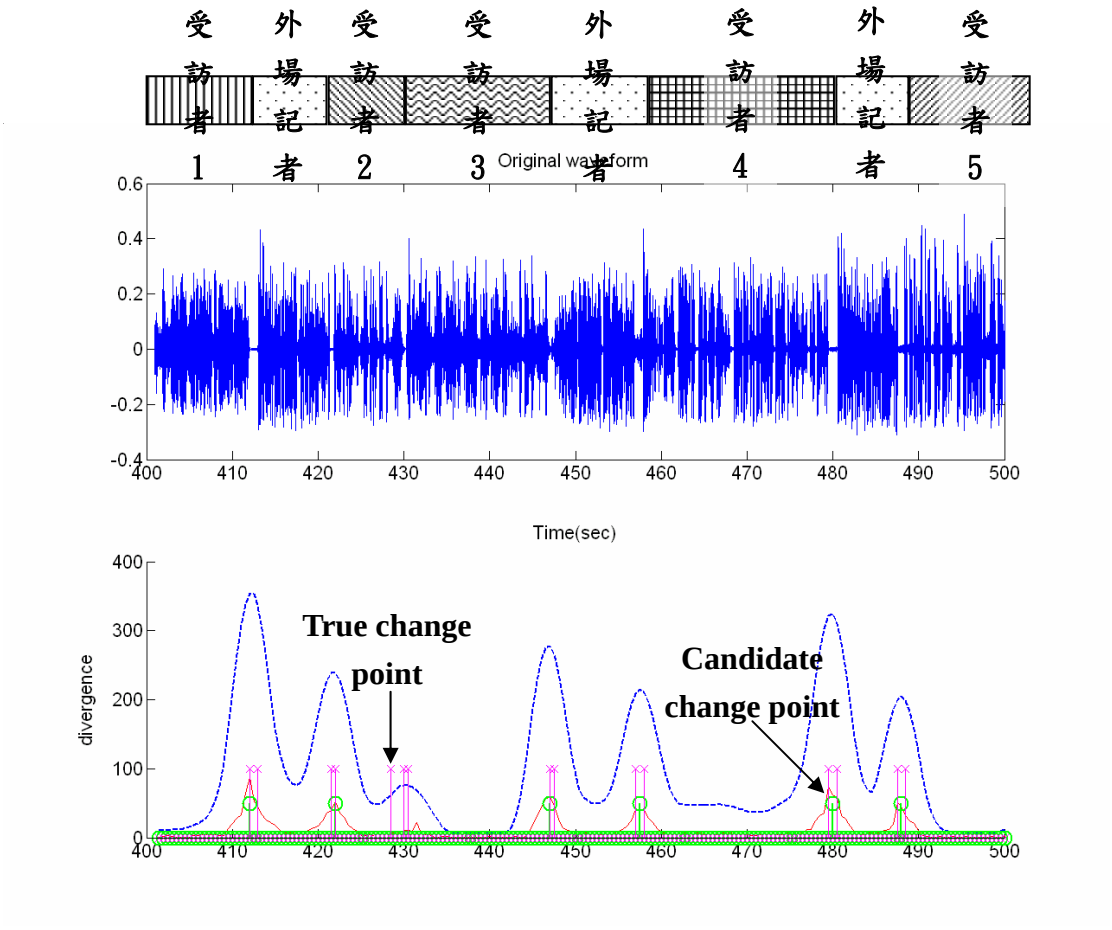

圖 二-9 轉換點偵測範例

第三章 實驗結果及討論

3.1 PTSND 電視新聞語料庫之簡介

2001 年八月，王新民教授所率領的團隊開始了一個語料收集的計畫，其目的為分三年收集 220 小時的中文電視新聞語料，名為公共電視新聞語料庫(Public Television Service News Database, PTSND)【2】；其錄音的參數為 44.1kHz 的取樣率，16-bit 的解析度，而每段節目長約 60 分鐘，由數位錄音機(DAT recorder)直接由公視新聞的主控台所錄製而成，且每個 DAT 都經由人為處理成 16kHz 16-bit 單聲道的 WAV 檔。此電視新聞語料庫都已經以 SGML(Standard General Makeup Language)的語法標註了語音內容(transcription)、背景環境(background condition)、語者轉換之邊界(speaker turn boundaries)、故事邊界(story boundaries)等標籤，而這些標籤都包含了時間戳記，圖 三-1 為其內容標註之例子。

PTSND 若以語者來區分段落的話，可歸類成下列四種：棚內主播(studio anchor)、氣象主播(weather anchor)、外場記者(field reporter)、外場受訪者(interviewee)及 no speaker，而前三種段落都以人工標註內容；此外，no speaker 的聲音段落，例如廣告、純音樂以及主播段落中的氣象主播等，都只標註整個段落之起迄時間戳記*，並沒有標註內容。

在圖 三-1 之聲音波形下的內容標註分四層，由上到下依序為聲學背景層(acoustic background layer)、故事層(story layer)、語者層(speaker turn layer)及標註層(orthographic transcription layer)；

*在 PTSND 中，只有幾天的節目之氣象主播段落有詳細的標註內容，其它都只標註整段氣象之起迄時間。

其中要注意的是最上層（聲學背景層）是與其它三層獨立的，也就是說聲學背景層的起迄時間並不一定要與其它三層有關，這是源起於在外場訪問的新聞中，常會有不可預期的背景聲，如汽車聲、人聲、機器聲等，因此在標註此現象時，自然是獨立於其它三層之外。

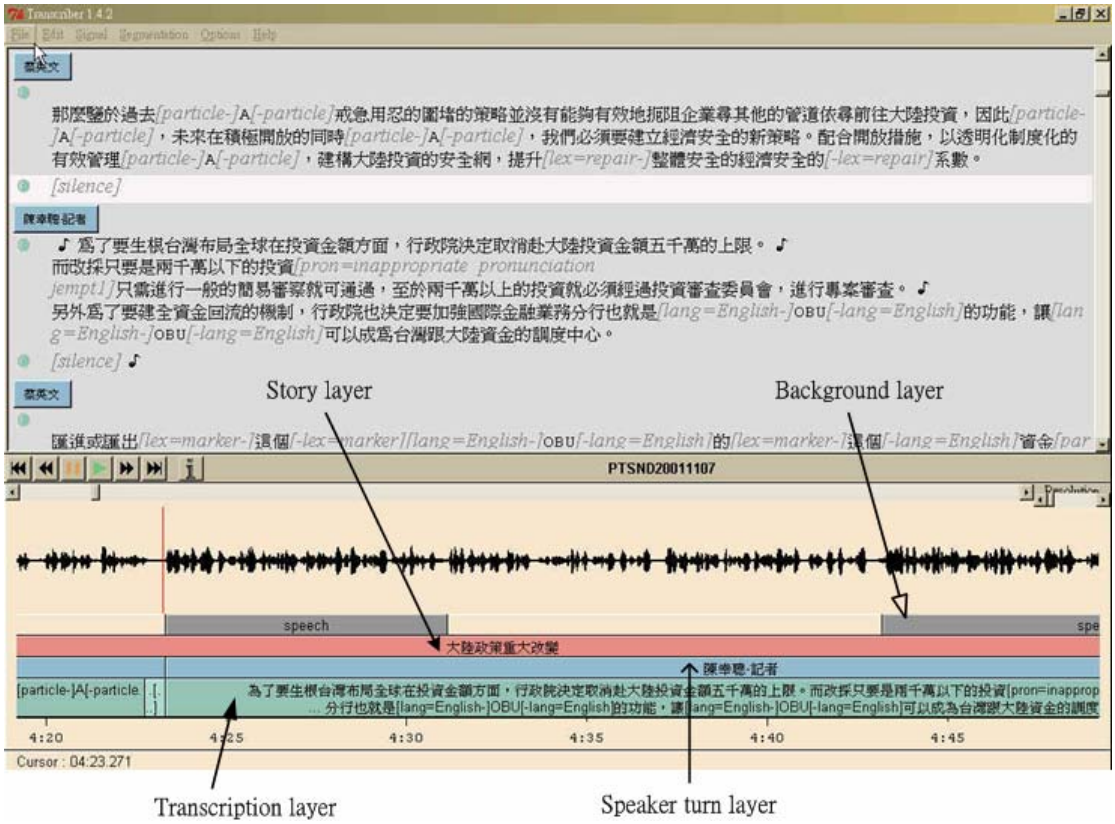

圖 三-1 PTSND 內容標註之範例

接下來我們簡述一下 PTSND 語料庫的一些統計特性，如表 三-1 所示；首先若我們以語者類別**來區分的話，因為外場記者及受訪者有相似的背景聲音，所以我們把兩者合併為一類，而氣象主播因為其背景大多為音樂，因此獨立出來統計；此外，棚內主播無背景聲音，故自成一類。我

**在表 三-1 中我們只對有語者的聲音片段分為表中的三類，而沒有語者的聲音片段，如廣告、純音樂等皆忽略不計。

們可以看出外場記者及受訪者之段落佔了此語料庫近四分之三的比例，表

示我們要處理的聲音串流有相當大的比例可能有背景聲。

若我們以聲音訊號的狀況來對 PTSND 分類，可分為純語音、有背景聲***的語音、純背景聲、廣告、氣象播報及靜音，若我們把純背景聲與有參雜背景聲的語音歸成一類的話，其比例佔了 61.4 %（34.5+8.9+10.0+8.0），這也暗示著**我們要偵測的轉換點大都在有背景之聲音片段中**，這是電視語料切割的困難處之一。

表 三-1 PTSND 基本統計特性

Speaker types	Percentage (in time)		Signal condition	Percentage (in time)
Studio anchor	15.1 %		Speech only	36.0 %
Weather anchor	10.1 %		Speech with background	34.5 %
Field reporter and interviewee	74.8 %		Background only	8.9 %
			Advertisement	10.0 %
			Weather report	8.0 %
			Silence	2.6 %

附註：此表是以 PTSND 第一年中 4 個小時的新聞所統計出來

***背景聲包括背景人聲、音樂、機器聲等，而有背景聲的片段並不包括氣象播報及廣告。

3.2 效能評估參數之定義

首先我們先定義兩個實驗中會發生的兩種典型的錯誤—FA(False Alarm)及 MD(Miss Detection)，如下：

1. 在候選轉換點的左右各 1.5 秒間若沒有真實轉換點出現的話，我們便稱此候選轉換點為 FA。
2. 在真實轉換點的左右各 1.5 秒間沒有候選轉換點的出現的話，此真實轉換點我們稱之為 MD。

依照上面的定義，我們可定義出 FAR(False Alarm Rate)及 MDR(Miss Detection Rate)這兩種效能評估參數，如下：

$$\text{FAR} = \frac{\#\text{ of FA}}{\text{total }\#\text{ of candidate change points}} \tag{3.1}$$

$$\text{MDR} = \frac{\#\text{ of MD}}{\text{total }\#\text{ of true change points}} \tag{3.2}$$

此外，我們依照(3.1)及(3.2)兩式分別定義了 PRC(Precision Rate)=1-FAR 及 RCL(Recall Rate)=1-MDR 兩項參數。

接下來我們來看一下常見的 FA 與 MD 發生的情況，如圖 三-2 所示；在圖中我們可以看到大部分的真實轉換點都是成對出現，這是因為在內容標註的過程中，語者間的轉換常伴隨著短停頓(short pause)，然而標註員在標註時也真實反應此現象，因此單獨出現的真實轉換點相對少見。然而此現象會引起下列的狀況：若語者間的短停頓過長，這會使得候選轉換點只能對應到其中一點，因此就會產生了 MD，例如在圖 三-2(b)所標示出來的兩個 MD 皆是因為此現象所產生；此外，在圖 三-2(a)中的三個 MD 雖然沒被偵測出來，但是基本上 $D'(i)$ 仍然有反映出在這三個真實轉換點之語者轉換；在圖 三-2(b)中也呈現了在 2.6 節提到的靜音不平衡所造成的 FA。

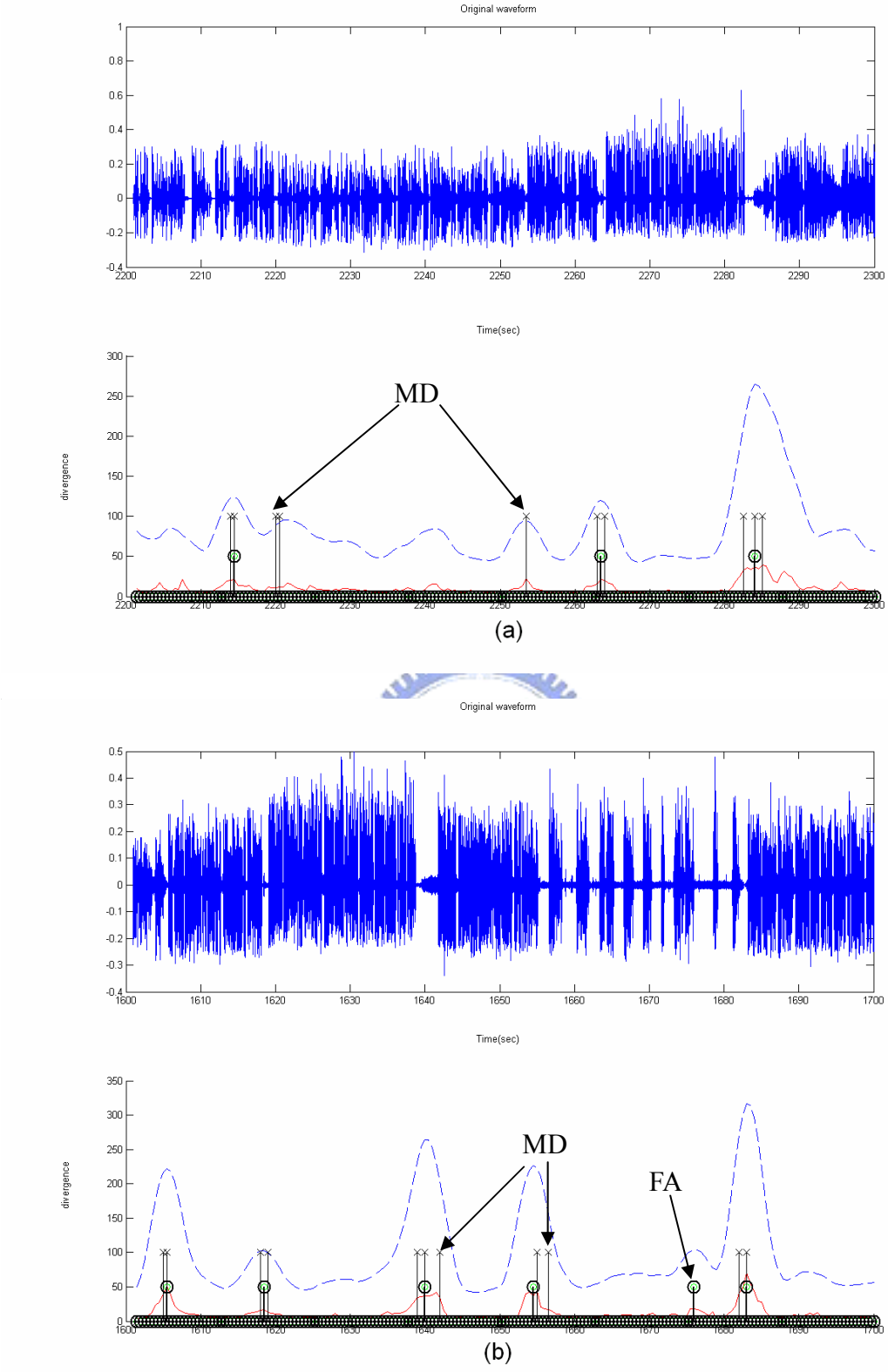

圖 三-2 FA 與 MD 之範例

對於大多數切割的演算法而言，系統操作在不同的操作點，例如不同的決定臨界值(decision threshold)，便有不同對的(FAR，MDR)；而我們可以藉由對於不同操作點的(FAR，MDR)做圖，藉以表現出此切割演算法的特性，進而找到最佳的操作點，這就類似在通訊系統的 ROC 圖(Receiver Operator Characteristic plot)，因而我們可以藉此來比較出出演算法間的優劣；有時我們希望能有單一參數來評估切割演算法的效能，因此有人便提出了 F-measure【12】【13】這個結合 PRC 及 RCL 的參數，其定義如下：

$$\text{F-measure} = \frac{2 \times \text{PRC} \times \text{RCL}}{\text{PRC+RCL}} \tag{3.3}$$

此評估參數有著 decreasing marginal effectiveness 的特性，我們以圖 三-3 之例來說明如下：

假設 PRC 大於 RCL 之情況下，我們欲犧牲 ε 的 PRC 來換取 ε 的 RCL 增加量，卻不想犧牲另一 ε 的 PRC 來換取 RCL 之 ε 增加，也就是說

$$\begin{aligned} &\text{F}(\text{RCL}+\varepsilon, \text{PRC}-\varepsilon) > \text{F}(\text{RCL}, \text{PRC}) \\ &\text{but} \\ &\text{F}(\text{RCL}+\varepsilon, \text{PRC}) > \text{F}(\text{RCL}+2\varepsilon, \text{PRC}-\varepsilon) \end{aligned} \tag{3.4}$$

從(3.4)式我們可以得到下列結論：區間[RCL,RCL+ε]對於 F-measure 的影響超過區間[PRC,PRC-ε]，然而區間[RCL+ε,RCL+2ε]比前兩者之影響力更小。因此對於 RCL 而言，其 marginal effectiveness 是 decreasing；同理，在 RCL 大於 PRC 的情況下，PRC 亦然。因此對於 RCL 與 PRC 之和為定值的條件下，(RCL,PRC)越接近中線有較佳的 F-measure。對於 PRC 與 RCL 相等的情況而言，F-measure 與 RCL 及 PRC 相等，因此(RCL,PRC)越接近座標(1,1)時 F-measure 越高。換句話說，**對於 ROC 圖而言，其操作點越往中線且越接近座標原點，F-measure 越高**。基於上述特性，F-measure 在我們接下來的實驗中因此成為主要的效能評估參數。

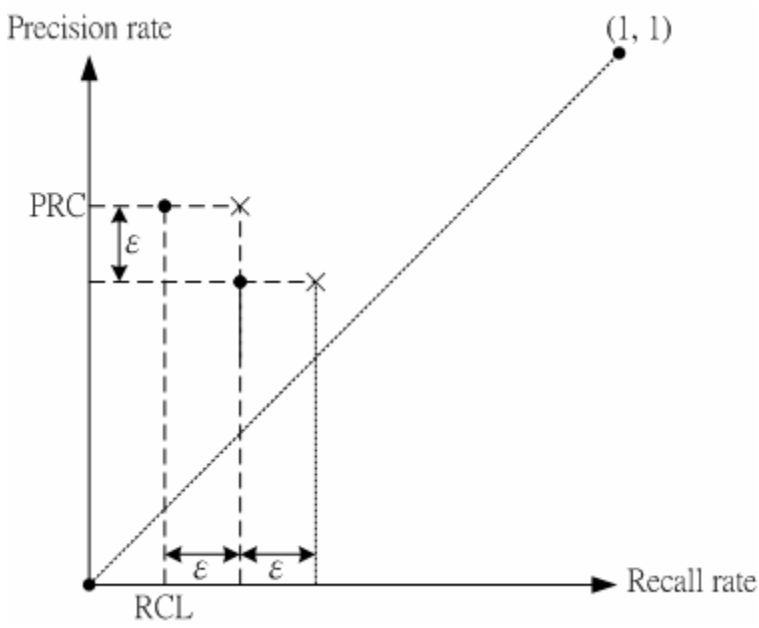

圖 三-3 F-measure 之特性說明

3.3 實驗參數設定及其結果

在我們的實驗中，所有的聲音訊號皆先通過 $1-0.97z^{-1}$的預強調濾波器，再對聲音訊號音框化成 30 ms大小的片段，其中音框與音框間的重複區間為 20 ms，之後我們從每個音框中求取出 **12 維**的梅爾刻度倒頻譜參數(MFCC)當作一個特徵向量。接下來我們選取了**一個小時**的節目來訓練出有 **256 個mixture component**的混合高斯模型，對**四個小時**的測試語料取此256 個mixture component做為CCGMM之mixture component，然後每隔 **0.5 秒**對於此時間點左右相鄰之分析音框做相異度量測，以期找出候選轉換點。

在我們實驗的過程中，發現到在節目中的廣告只有標註整個段落的時間戳記，而沒有對每一個廣告段落做標記的動作；此外，氣象播報只有少數天數的節目有標註內容，其它則只有標註整段氣象的起迄時間。為了改正此缺失，我們以人工對於廣告及氣象播報做了初步的分段，標註出每個小段落的起迄時間，以期得到更正確的真實轉換點。經由初步的人為統計真實轉換點之特性，如表 三-2 所示，我們發現在第二欄中最常發生之情況有兩種，一是棚內主播與外場記者或者外場記者與受訪者間的轉換，這種情況對我們來說是容易偵測的；另一為背景聲突然消失或出現，這種情況就不易處理。在第三欄中也可分為兩種情況，背景聲及語者皆不相同之轉換為其一，這相對來說較易處理，然而對於有著相同背景聲之情況下，背景聲之大小對於偵測的難易度便扮演了相當重要的角色。從表 三-2 我們可以呼應 3.1 節之結果：我們要偵測之轉換點大多與背景聲有關，這也是電視新聞語料之轉換點偵測的困難處。

表 三-2 真實轉換點之特性統計

Types of true change point[註]	Percentage (in numbers)
Speech ↔ Speech	20.1%
Speech ↔ Speech+bg	33.9%
Speech+bg ↔ Speech+bg	31.1%
Speech ↔ Music	4.6%
Speech+bg ↔ Music	9.3%
Music ↔ Music	1.0%

[註]

1. X↔Y 表示此轉換點為 X 轉換至 Y 或者是 Y 轉換到 X。
2. Speech 表示純語音片段， Music 表示純音樂之片段，Speech+bg 表示有背景聲之語音片段，然而背景聲包括音樂、人聲、機器聲及其它等噪音。
3. 廣告片段則屬於 Speech+bg。
4. 氣象播報在有前景語音時屬於 Speech+bg，在純背景音樂片段時則屬於 Music。

3.3.1 不同 α 值對於系統效能之影響

首先我們評估在(2.16)式中，α=1 與 α=2 對於轉換點之辨識率的影響，如圖 三-4 所示；在圖中我們可以瞭解雖然對於相異度序列做三角波的迴旋積分之結果已經相當不錯了，但是若我們設定 α 為 2 的話，可以得到小幅度的效能改善，而這也證明了自然對數裡的比值會對 $D'(i)$ 造成些微的影響，因此我們在接下來的所有實驗中，都使用 α 為 2 之 $D'(i)$。

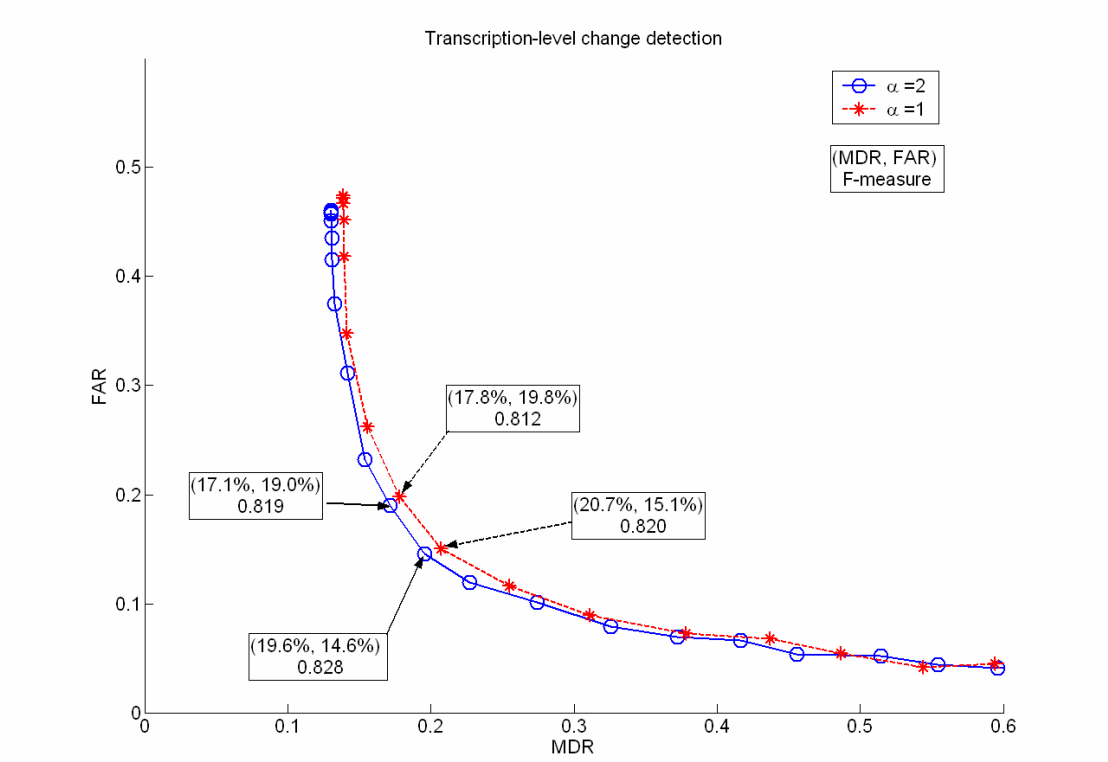

圖 三-4 在(2.16)式中不同 α 值對於系統效能之影響

3.3.2 共用共變異矩陣之效能改進

接下來我們為了使得(2.12)式之近似值更為合理，因此我們使每個 mixture component 共用同一個共變異矩陣，也就是當我們在更新共變異矩

陣時使用(2.18)式，其結果如圖 三-5 所示。

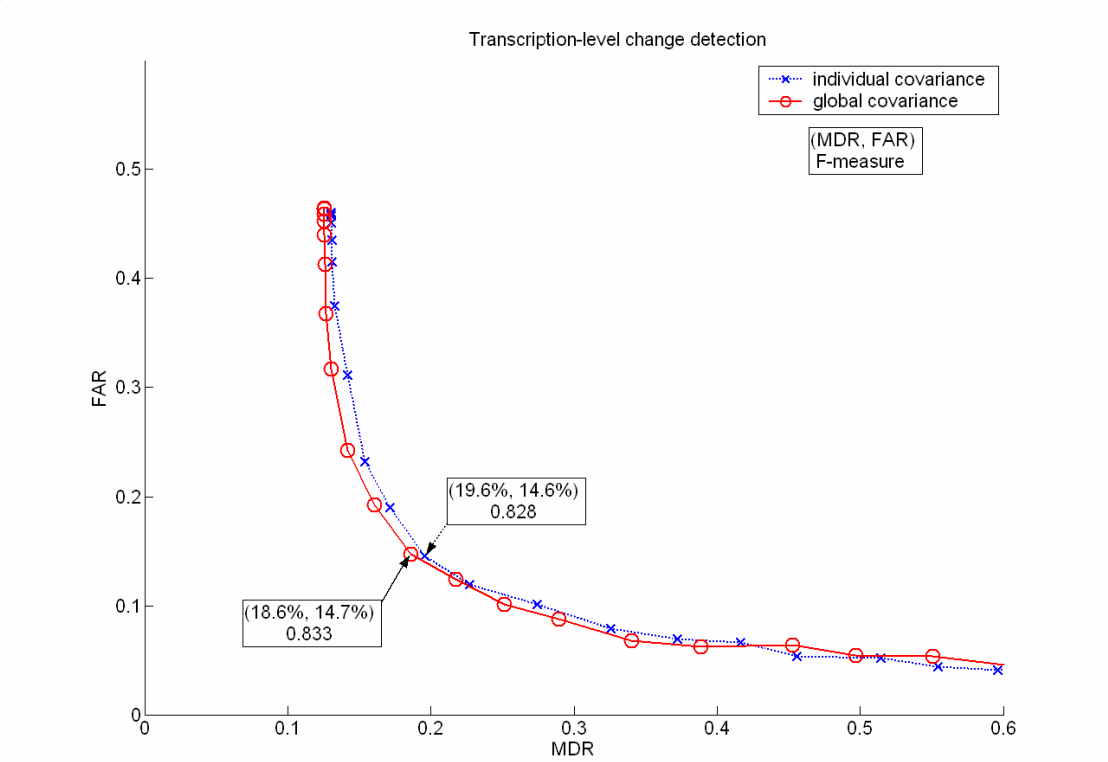

圖 三-5 mixture components 共用共變異矩陣之效能改進

3.3.3 不同 mixture 數對於系統效能之影響

直觀而言，對於 CCGMM 之 mixture component 數目越多的話，描述聲音片段中的特徵向量分佈之精確度越高，然而 mixture 數越高會使得估計參數時之計算複雜度越高，因此如何選擇 mixture component 的數目便需要 trade off。我們接下來便評估其影響如圖 三-6 所示；從圖中我們觀察到隨著 mixture 數為 64、128 與 256 時，F-measure 正比於 mixture 數，分別為 0.813、0.820 及 0.828，但是其數值都在令人滿意的範圍內，這隱含著我們可以只用較少的 mixture 數便可以得到不錯的結果。

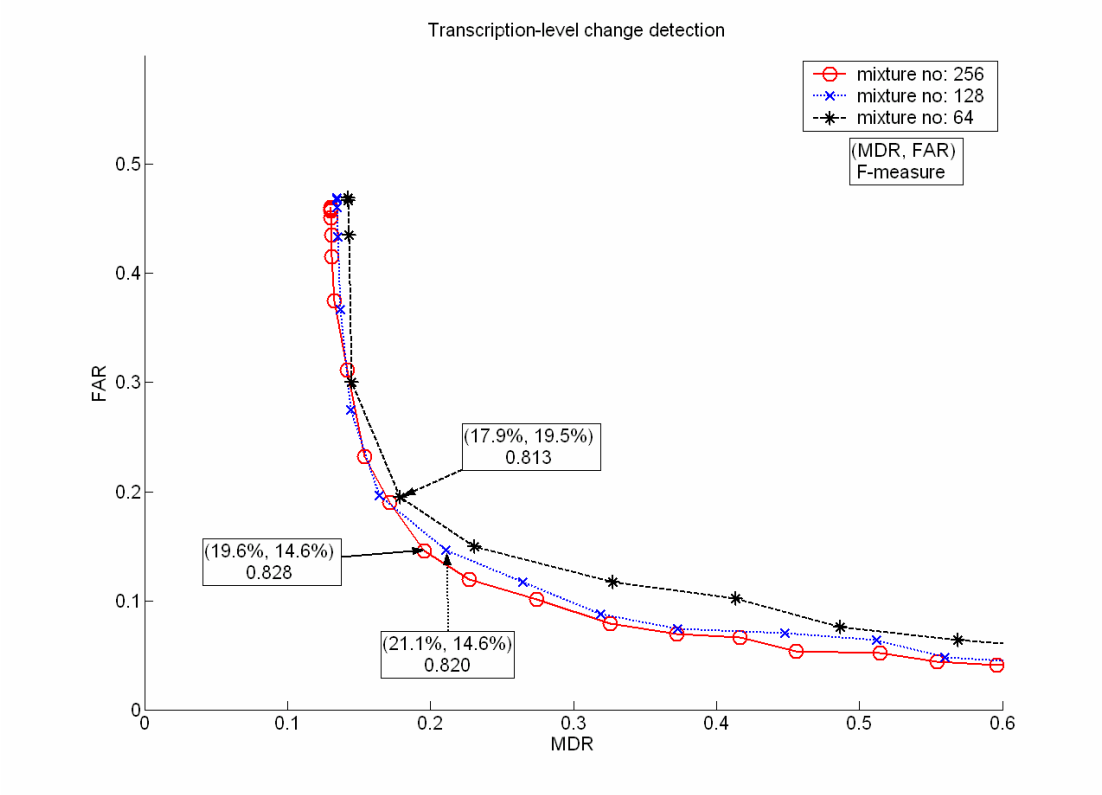

圖 三-6 不同 mixture 數之影響

3.3.4 混合權重近似對於系統效能之影響

再來我們評估混合權重近似對於切割效能的影響，如圖 三-7 所示。在圖 三-7 之(a)(b)(c)分別為 256、128 及 64 個 mixture component 的 CCGMM 有無做混合權重近似之 ROC 圖，其中此圖為針對標註層所做的轉換點偵測之結果。從圖 三-7(a)中我們可以明顯地看出，對於 CCGMM 的 mixture component 數目為 256 時，混合權重近似使得 F-measure 由 0.828 下降至 0.820，並沒有造成系統效能明顯地減低；然而在 3.3.3 節中我們發現 mixture 數為 128 時，其 F-measure 與本實驗中 mixture 數為 256 且使用混合權重近似時相同，也就是說我們利用增加的 mixture 數來換取混合權重近似，這雖然會增加參數估計時之計算量，但是由於我們在初始化混合權重向量時使用了(2.14)式，因此可以減輕計算量的問題。

在本論文中，我們先以 256-mixture CCGMM 求出所有的 50-frame 權重向量，再利用(2.15)式近似出 300-frame 權重向量，之後便可以輕易地計算出相對應之相異度序列。

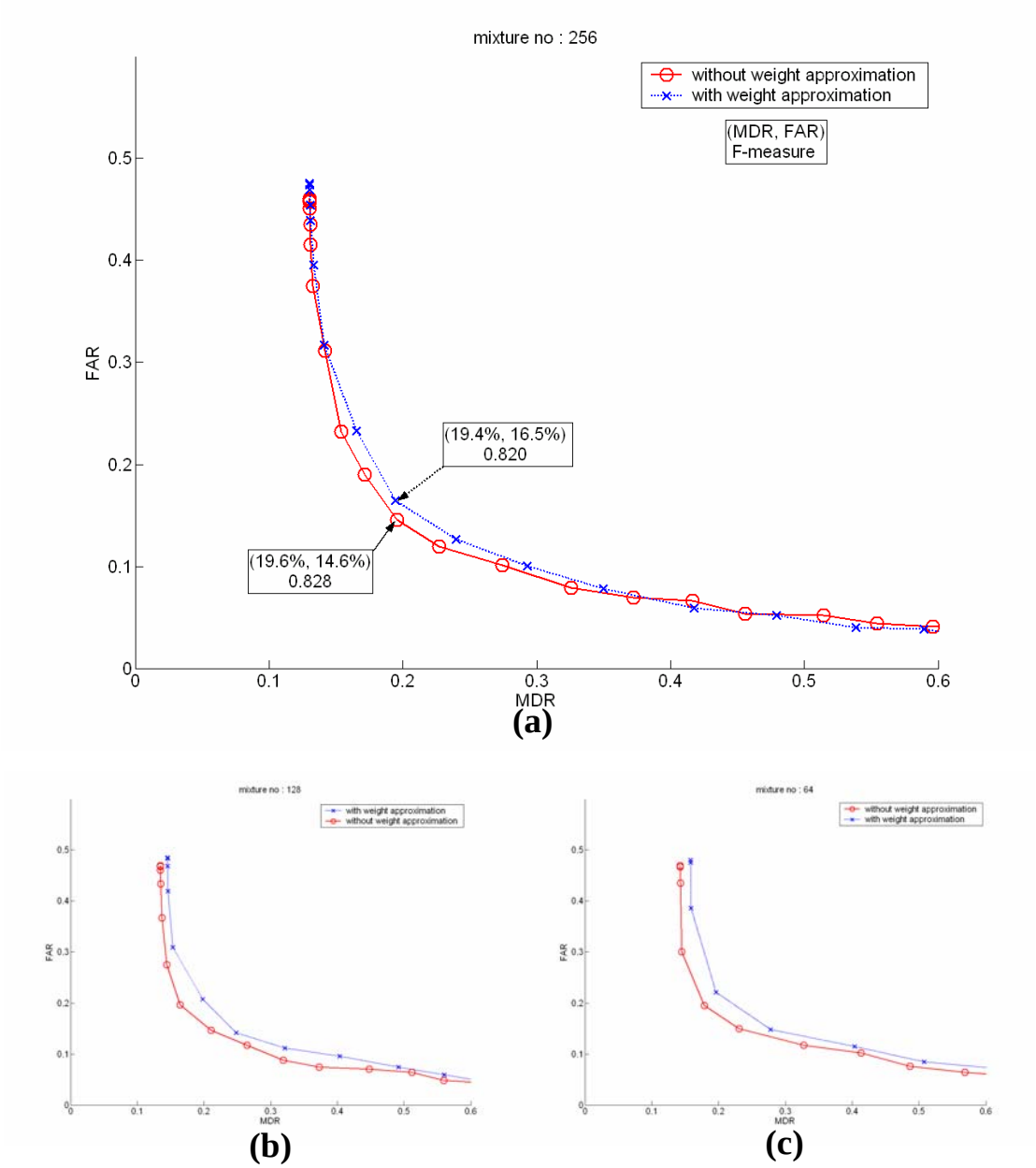

圖 三-7 不同 mixture 數對於混合權重近似之影響

3.3.5 移除與靜音相關之 mixture component

在 2.6 節中我們提到了相鄰片段間的靜音不平衡會導致相異度序列 $D(i)$ 的升高，造成一些不必要的 FA 產生。我們嘗試解決此現象，因此當在計算相異度時，利用 (2.17) 式來忽略與靜音有相關性的 mixture component，其結果如圖 三-8 所示。從圖 三-8 中我們可以看出移除靜音之 mixture component 對於轉換點的辨識率並沒有明顯的改進，這可以分為兩方面解釋：一是靜音不平衡的現象出現頻率雖然不算小，但是常發生在語者轉換之邊界，因此其影響並無想像中大；另一為我們無法明確地指出屬於靜音的 mixture component 為何，因此移除屬於靜音的 mixture component 之影響不明顯，然而在本論文中我們移除了 9 個與靜音有相關性之 mixture component。

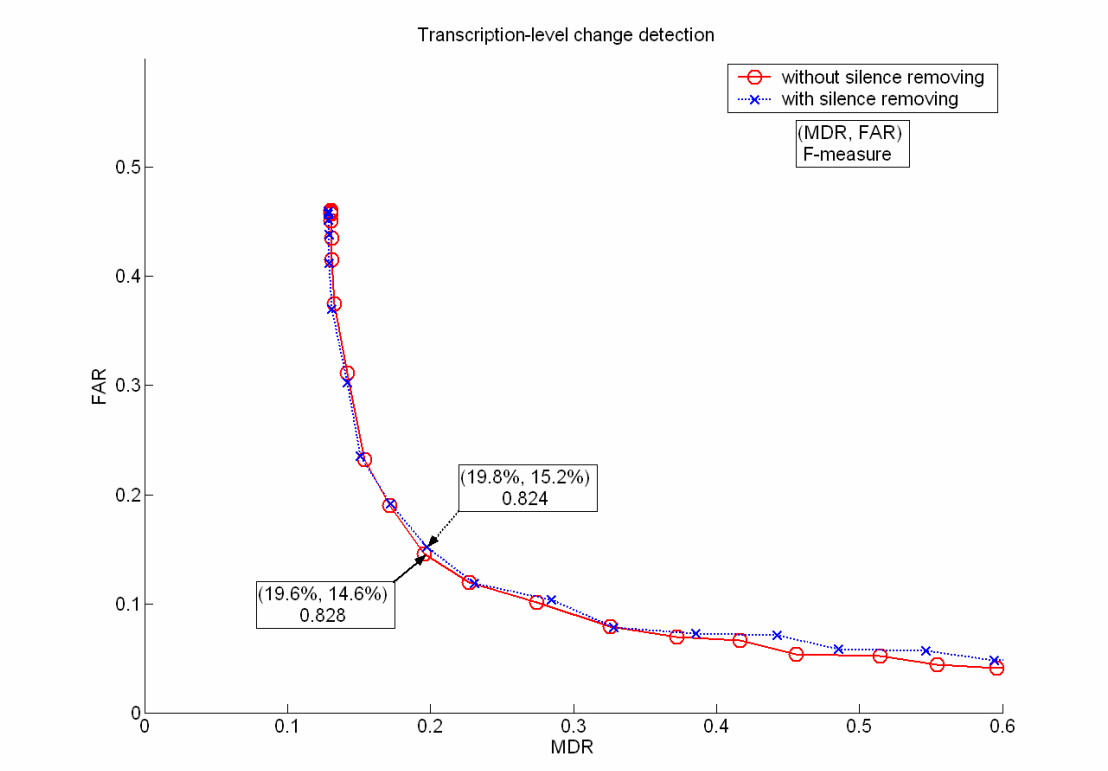

圖 三-8 忽略靜音之 mixture component 之影響

3.3.6 標註層與加上聲學背景之標註層之比較

最後來我們分別對於標註層與標註層加上聲學背景層做轉換點偵測，其結果如圖 三-9 所示；從圖 三-9 中我們可以看出，若考慮背景聲的情況下，MDR 會上升約 4~5 個百分點，這是相當顯而易見的，因為通常背景聲音的時間戳記都與標註層的時間戳記一致，只有少數的情況下背景聲音會突然地消失或加入，而在這種情況下聲音訊號的訊噪比(SNR)大部分都會相對較低，聲音的轉換不明顯，自然而然地造成了 MD 的產生。

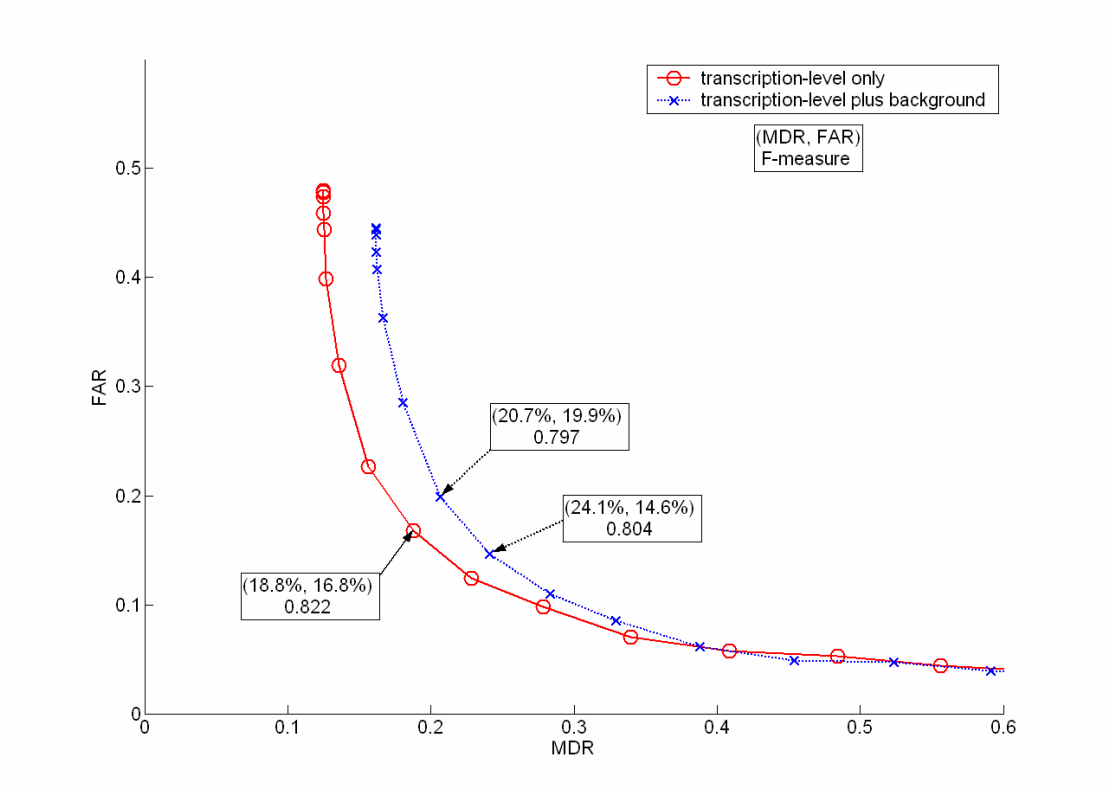

圖 三-9 聲學背景層對於 ROC 圖之影響

3.3.7 候選轉換點之錯誤分析

為了瞭解在本實驗中 FA 發生的主要原因，因此我們對於測試語料之 4 個小時的結果，以人工的方式小心地檢查所有發生 FA 的時間點是否真有人耳可察覺到的轉換點，並對其做歸類在表 三-3。檢查後我們發現，**所有發生 FA 的時間點都有明顯的聲學特性之轉換**，而且我們發現語料中之廣告片段有著下列問題：

1. 廣告中的音樂特性變化太快，例如常有某一種樂器突然加入，造成 FA 的產生。
2. 在廣告中前景的語音並不是持續出現，這也是造成 FA 產生之原因。

因為在整個 PTSND 新聞語料中，廣告出現了近 10%的時間，而在所有發生的 FA 中，存在於廣告片段的比率近 24%，也就是說在廣告中發生 FA 之可能性相較於其它聲音片段高，影響程度不可不謂重大。此外，除了廣告對於 FAR 的影響外，每天節目中的氣象播報也有一定程度的影響，因為當氣象主播在播報各地氣象時，背景音樂聲相當明顯，因此我們的相異度量測公式便會反映出音樂的變化，造成不必要的 FA；幸運的是氣象播報的背景音樂相對於廣告之背景音樂柔和，變化也沒有那麼大且頻繁，因此對於 FAR 的影響比較輕微。

此外，從表 三-3 中可知，FA 最有可能發生在有背景聲（包括音樂、人聲及其它），這是由於電視新聞語料不可避免地有外場的採訪，就 PTSND 來說就有近四分之三的語料為外場新聞，而外場新聞常伴隨著背景聲，所以電視新聞語料切割之困難處由此而來；我們在表 三-3 中還看到 FA 發生在純音樂片段佔了五分之一強，這並不令人驚訝，因為音樂在聲音的特性上本來就變化相當大，就我們的觀察來看，最常發生的現象就是某一樂器聲的突然加入或消失，造成其聲音特性上的劇烈改變。

表 三-3 FA 之統計特性

Types of speech conditions	Percentage of all false alarms
Pure speech	9 %
Speech with music/others	40 %
weather reports	5 %
advertisements	24 %
Pure music	22 %

至於對於 MD 而言，最常發生之情況有下列三種型態：

Type Ⅰ：在發生 MD 附近左右 5 秒間，沒有候選轉換點的出現。

Type Ⅱ：成對或成群的真實轉換點沒有全部被候選轉換點對應到，或者單獨出現但是左右 5 秒間有候選轉換點。

Type Ⅲ：MD 發生在有相同背景音樂的語音片段。

我們把上述情況整理如表 三-4 所示；對於 Type Ⅱ的 MD 而言，我們已經偵測出其聲音片段之變化，只是由於標註方法之緣故以致於候選轉換點無法完全對應至真實轉換點；然而對於 Type Ⅲ的 MD 來說，由於其發生在有著相同背景音樂之聲音片段，而且通常 SNR 值都較低，因此很困難來偵測出真實轉換點。

表 三-4 MD 之統計特性

Types of MD	Percentage
Type Ⅰ	54.1%
Type Ⅱ	27.2%
Type Ⅲ	18.7%

3.4 利用合併的方式實現語者轉換偵測

一旦偵測出標註層之候選轉換點，許多人都是利用合併(merging)的方式來決定出語者轉換點，接下來我們採用合併的方式來決定出語者轉換點，並且與基於臨界值之語者轉換偵測之結果做比較。

首先我們列出合併的規則如下：

1. 對於所有的標註層之候選轉換點依照其時間順序依序測試是否可以合併。
2. 若在第 i 個候選轉換點左右兩個聲音片段的相異度大於某個臨界值 TH2 的話，此兩個聲音片段片段即可以合併成一個。
3. 一旦某個候選轉換點左右兩個聲音片段可以合併的話，其它的聲音片段便不可以與此兩個聲音片段合併。
4. 執行上述兩個步驟直到沒有相鄰的兩個聲音片段可以再合併。

為了更明白地解釋我們的作法，我們以圖 三-10 來說明我們整個合併的過程。首先我們使用 3.3 節所切割出來的結果，得到了標註層之候選轉換點(i.e.圖 三-10 中 peak picking 的輸出)，之後我們依照時間順序選擇某一候選轉換點之左右聲音片段(從現在之候選轉換點至下一個及前一個候選轉換點)做相異度量測，以 TH2 來決定此兩段聲音是否可以合併，直到所有的候選轉換點皆不可以再合併了。

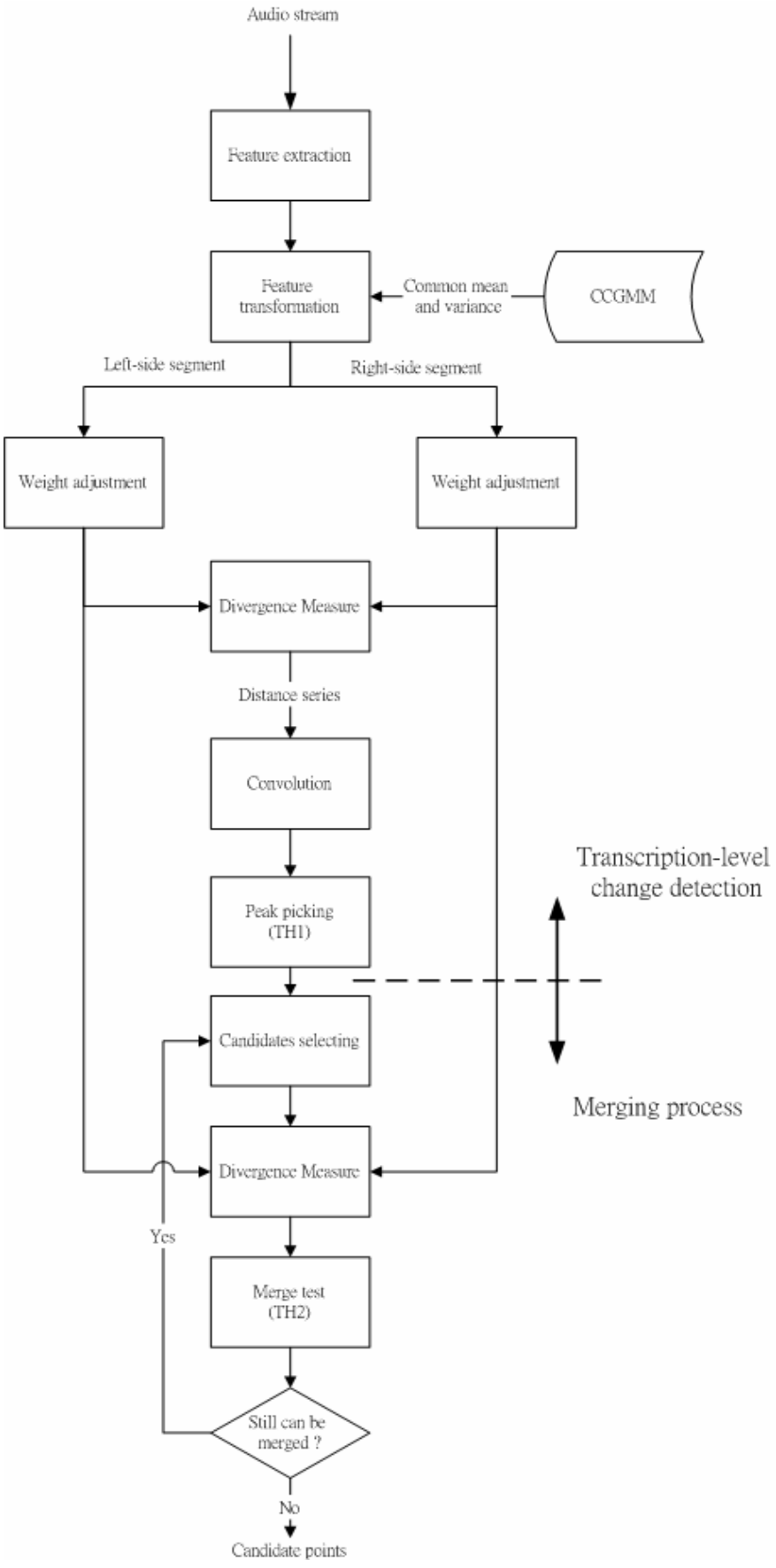

圖 三-10 加入合併過程之系統流程圖

以圖 三-10 為例，在第一階段(1st phase)的合併過程中，依序測試每個候選轉換點是否可以合併，也就是測試$D([t_{i-1},t_i],[t_i,t_{i+1}])$是否大於 TH2，圖中的第 1、2、5 個候選轉換點皆不可合併，而第 3 個候選轉換點可以合併，因此在第二回合中第一階段中的[t2,t3]、[t3,t4]這個兩個聲音片段便合併為一個聲音片段，然而因為第一階段中[t3,t4]的聲音片段已經合併了，所以在第 4 個候選轉換點時便不做合併的測試。

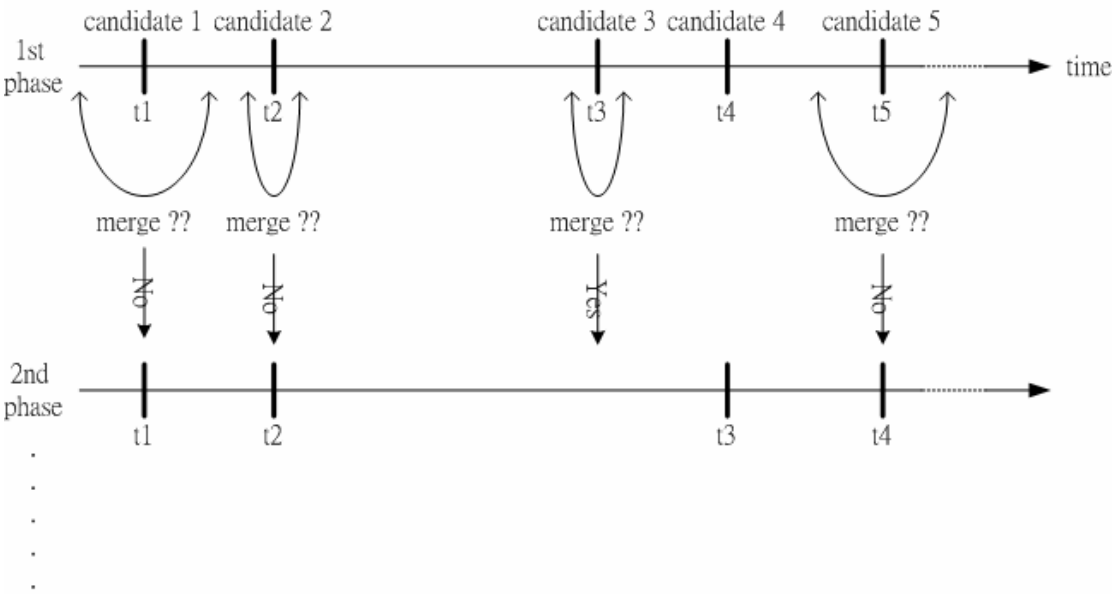

圖 三-11 標註層的候選轉換點合併之示意圖

從圖 三-10 中我們可以看到混合權重近似的好處，因為在合併的過程中，相鄰兩個分析音框的長度不再固定為 3 秒，而是依照標註層之相鄰候選轉換點之時間長度，這使得在合併的每階段中我們使用混合權重近似，因而一旦要做合併測試時，只要套用(2.15)式便可以得到任意分析音框長度之權重向量，然後便可以利用(2.13)式計算出兩個任意長度的聲音片段之相異度，且計算量大幅減少。圖 三-10 為利用合併的方法實現語者轉換偵測之系統架構圖，與圖 2-8 之架構比較的話多了合併的步驟，而切割出來的結果如圖 三-12 所示。

在圖 三-12 中，我們比較兩種不同方法做語者轉換偵測，一是單純利用 threshold-based decision rule 來決定候選轉換點，如圖 2-6 所示；另一為如圖 三-10 中使用合併方法來合併出語者片段。從圖 三-12 我們可以看出兩者的最佳操作點之效能差不多，這也隱含了當我們要做語者轉換偵測時，只要單純地加大臨界值便可以決定出不錯的候選轉換點，而不需要傳統語者轉換偵測所常用的合併步驟。

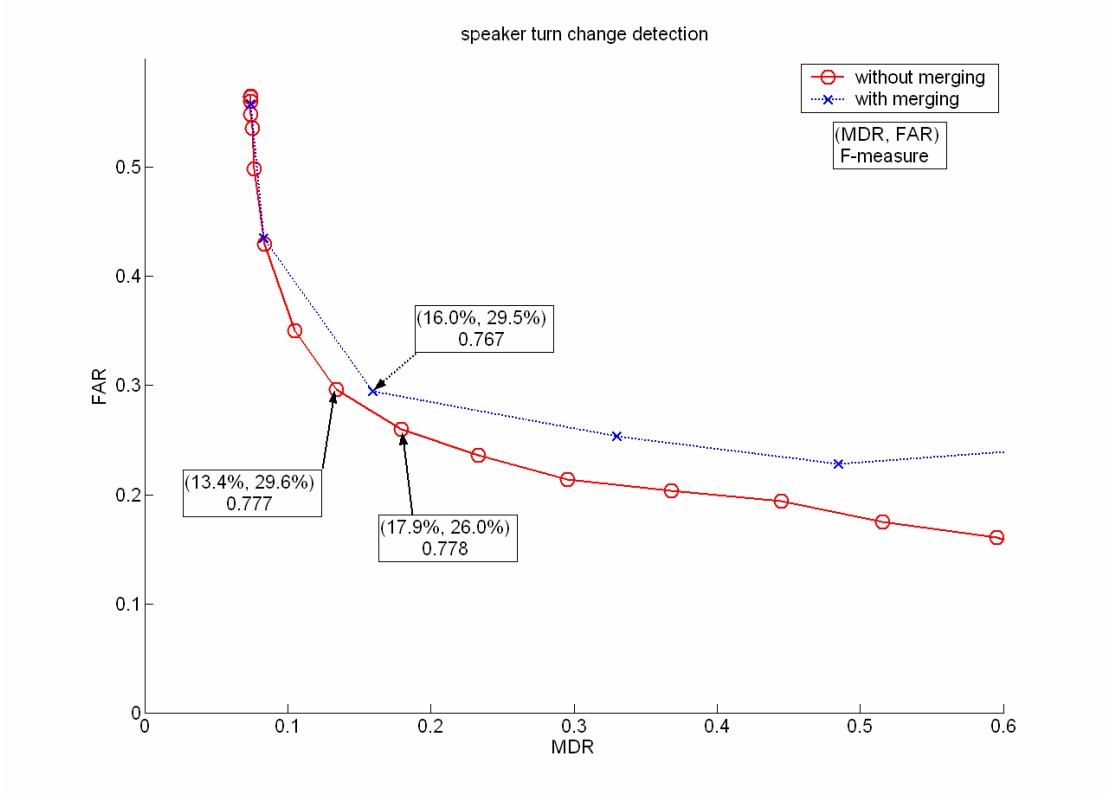

圖 三-12 語者轉換之 ROC 圖

上圖之結果乍看之下有點令人疑惑，因為合併步驟中的相異度量測我們分析音框的長度不再是固定 3 秒，是以相鄰候選轉換點間的聲音片段來做相異度量測，因此應該可以得到較佳的辨認率，但是結果卻並非如此，主要可以由下列之觀點來解釋：

假設兩組待量測之資料 $X=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$ 及 $Y=\{\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_n\}$ ，根據相異度量測之特性【14】

$$D\left(\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\},\{\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_n\}\right)\le\frac{1}{n}\sum_{i=1}^{n}D(\mathbf{x}_i,\mathbf{y}_i) \tag{3.5}$$

where the equality holds $\mathbf{x}_i,\mathbf{y}_i$ are i.i.d.

我們可以發現使用較多的特徵向量來描述一個聲音片段來求取其聲音的轉換點並不會得到任何的好處，但是其效能也不會變差很多。

3.5 利用基於BIC相異度量測之轉換點偵測的結果

在此節中我們先簡介一下前人所提出的BIC相異度量測，並以PTSND作為測試語料庫評估其效能，進而能與本論文所提出之基於CCGMM相異度量測做比較，簡介如下【8】：

假設 $\mathbf{x}=\{x_t;t=1,\ldots,N\}$ 為我們要建立模型的資料，在此即為聲音片段之特徵向量，$\{M_i,i=1,\ldots,Z\}$ 為其可能的模型，因此BIC可定義成下列形式：

$$BIC(M_i)=\log L(\mathbf{x},M_i)-\lambda\frac{1}{2}\#(M_i)\times\log(N) \tag{3.6}$$

其中 $L(\mathbf{x},M_i)$ 為 $\mathbf{x}$ 對於模型 M_i 之似然機率，λ 為 penalty weight，$\#(M_i)$ 為模型 M_i 之參數個數。

接下來我們假設 $\mathbf{x}$ 有兩種可能模型，一是在此聲音片段中沒有候選轉換點，也就是說在 $\mathbf{x}$ 中所有特徵向量有相同高斯分佈--M_0；另一則是在此聲音片段中存在候選轉換點，亦即 $\mathbf{x}$ 中特徵向量之高斯分佈有兩種，以 M_1 來描述此模型。簡言之，

$$\begin{aligned}&M_0:x_1,x_2,\ldots,x_N\sim N(\mu,\Sigma)\\&M_1:x_{i-N_L+1},\ldots,x_i\sim N(\mu_L,\Sigma_L);x_{i+1},\ldots,x_{i+N_R}\sim N(\mu_R,\Sigma_R)\end{aligned} \tag{3.7}$$

其中 $N=N_L+N_R$；在本論文中我們設定 $N_L=N_R=\frac{1}{2}N=300$。

然後我們將某一聲音片段對此兩種模型做比較，定義為 *delta*BIC，如下所示：

$$deltaBIC=BIC(M_1)-BIC(M_0) \tag{3.8}$$

也就是說當 *delta*BIC 值為正時，此聲音片段之特徵向量的分佈傾向於兩個不同的高斯分佈，亦即在此聲音片段有聲音轉換的發生；反之，若 *delta*BIC 值為負值，此聲音片段便傾向於不存在轉換點。因此我們把(3.7)式之假設與(3.6)式之定義代入(3.8)式中，我們可以得到下列結果【8】【10】：

$$deltaBIC = \frac{N}{2}\log|\Sigma| - \frac{N_L}{2}\log|\Sigma_L| - \frac{N_R}{2}\log|\Sigma_R| - \frac{\lambda}{2}\left(D + \frac{D(D+1)}{2}\right)\times\log(N) \tag{3.9}$$

其中 D 為特徵向量之維度；藉由 *delta*BIC 曲線，如圖 三-13 所示，我們可以瞭解聲音串流是否可能存在轉換點，進而達成轉換點偵測之目的；在本論文中我們設定 penalty weight λ 為 2.5。

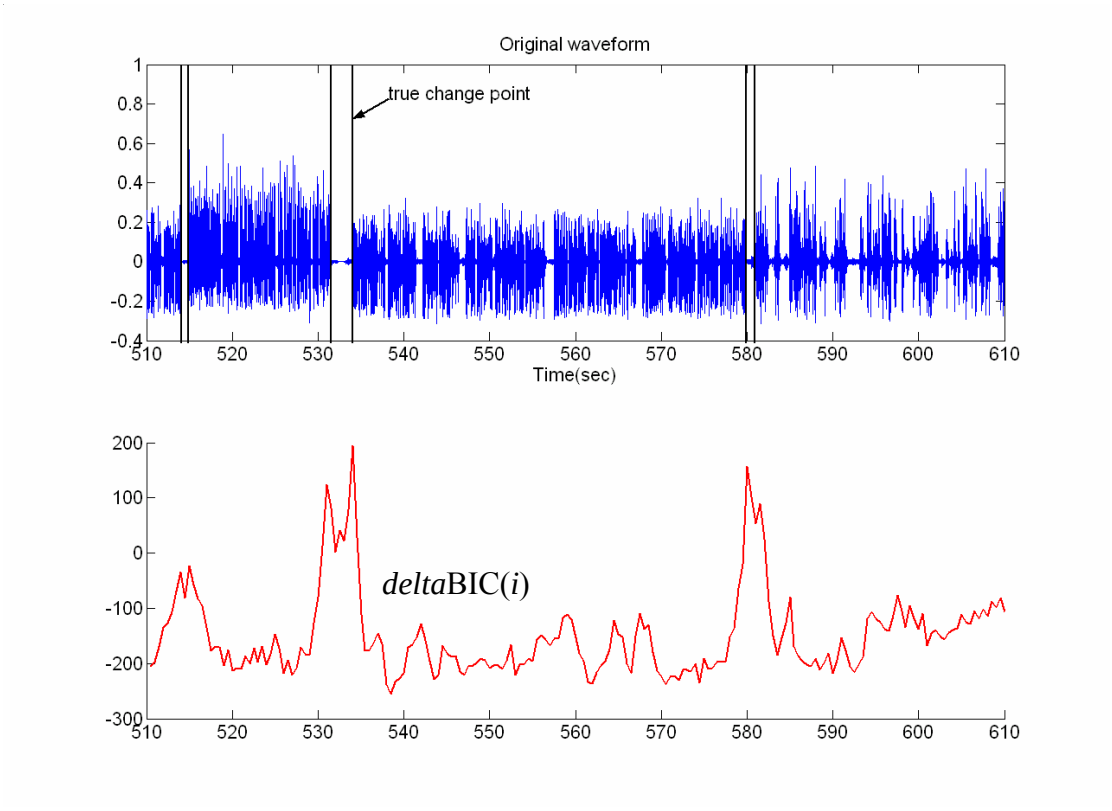

圖 三-13 *delta*BIC 曲線之範例

最後我們列出候選轉換點之 detection rule 如下：

$$\begin{aligned} &deltaBIC(i) > deltaBIC(i \pm j),\ j = 1 \sim 5 \\ &deltaBIC(i) > 0 \end{aligned} \tag{3.10}$$

在加入上式之 detection rule 後，便可以計算出其 MDR 及 FAR 在標註層之轉換點偵測分別為 24.3%與 18.5%。

3.6 與基於 BIC 相異度量測之轉換點偵測的結果比較

在【2】中，中研院王新民教授建構了一套 broadcast news retrieval system，SoVideo，其中他們利用 *delta*BIC 相異度量測方法，來實現故事轉換、語者轉換和標註層轉換的聲音切割；因為我們與王新民教授之實驗結果與我們之實驗都是針對 PTSND 作聲音轉換切割，所以我們可以拿其結果來作為我們新提出的相異度量測方法之對照組，其結果摘要如表 三-5；此外我們也利用前一節所提出的方法，對於 PTSND 電視新聞語料庫做了相同的轉換點偵測，其結果也表列於表 三-5 中。

在表 三-5 中，第一欄為本論文提出之基於 CCGMM 相異度量測的結果，第二欄為中研院王新民教授使用 *delta*BIC 相異度量測方法對於同一電視新聞語料庫所做出的結果【2】，第三欄為我們使用前一節所提到之 *delta*BIC 相異度量測之結果，其中**第二、三欄的差別在於因為我們對於廣告及氣象片段之起迄時間標註做了些修正**，因此為了公平起見我們重新評估基於 BIC 相異度量測之辨識率；我們可以看出無論是標註層抑或語者轉換層的切割結果，我們提出的方法都相對較好，對於標註層轉換偵測，在大約相同的 FAR 下【2】中的 MDR 之結果有了近 30%的錯誤下降率，想當然爾 F-measure 也都比【2】之結果高；比較第一、三欄，本論文提出的方法不管是 FAR 及 MDR 在最佳操作點，其平均錯誤下降率分別約為 12%與 20%，當然 F-measure 都比 *delta*BIC 相異度量測之 F-measure 好，這也表示**我們所提出之基於 CCGMM 之相異度量測方法對於聲音變化及轉換有令人滿意的鑑別度。**

表 三-5 與 *delta*BIC 相異度量測之比較

Test condition \ Method		CCGMM-based Divergence Measure	*delta*BIC 【2】	*delta*BIC $\lambda = 2.5$
Transcription-level	MDR	18.8 %	26.97 %	24.3%
	FAR	16.8 %	15.63 %	18.5%
	F-measure	0.822	0.783	0.785
Both transcription -level and background	MDR	24.1 %	32.30 %	29.3%
	FAR	14.6 %	14.49 %	17.1%
	F-measure	0.804	0.756	0.763
Speaker turn level	MDR	17.9 %	22.33 %	X
	FAR	26.0 %	40.17 %	X
	F-measure	0.778	0.676	X

第四章 結論與未來展望

4.1 結論

在本論文中，我們使用了基於 CCGMM 相異度量測方法來做語者及環境狀況轉換之偵測，並且與基於 BIC 之相異度量測方法比較其轉換點辨認率，其結果分列如下：

1. 本論文提出之相異度量測方法對於標註層之轉換點偵測可以得到令人滿意的結果，其 MDR 為 19.6%，FAR 為 14.6%，而 F1-measure 高達 0.828，相較於傳統的基於 BIC 之相異度量測方法有著更好的辨識率。
2. 因為我們利用 GMM 來描述聲音片段之統計特性，而共用 mixture component 來大幅減少其計算量，因此在不增加太多的計算量下，基於 CCGMM 之相異度量測有顯著的效能增加。
3. 對於語者轉換偵測而言，利用本論文所提出的相異度量測方法可以單純地利用 threshold-based decision rule 來決定出語者之轉換點。
4. 在我們觀察 $D'(i)$ 曲線且對照實際的聲音波形後，發現因為基於 CCGMM 之相異度量測方法對於聲音特性變化有著相當好的鑑別度，幾乎在真實轉換點附近 $D'(i)$ 都有峰值，只是由於電視新聞語料之特性，使得常有不可預期之 FA 發生。

4.2 未來展望

對於語者及環境轉換之偵測而言，其最終目標不外乎切割出有著相同主題的聲音片段，以達到資料檢索之目的，因此有著令人滿意之標註層及語者層之轉換點辨識率便是我們的首要目標，而我們提出的相異度量測方法可以為其解決方案之一；然而下一步的目標便是如何利用已偵測出來之轉換點來合併及歸類成主題式的聲音片段。

對於故事層之轉換點偵測而言，如同我們在第一章所提到的，已經有許多人投入了相當程度的努力，例如利用影像處理、語音辨識、Dynamic Programming 等技術來對於影像或聲音之內容分類，已經獲得了不錯的結果；然而這些技術都牽涉到了許多不同的層次，而且在不同的地方新聞播報的次序也不盡相同，因此前述之故事轉換偵測方法有著地域性的限制，所以如何找到簡單且通用的偵測方法便是我們未來努力的目標。

參考文獻

【1】Hauptmann, A.G., and Witbrock, M.J., “Story Segmentation and Detection of Commercials in Broadcast News Video,” ADL-98 Advances in Digital Libraries, Santa Barbara, CA, April 22-24, 1998.

【2】Hsin-Min Wang, Shi-Sian Cheng and Yong-Cheng Chen, “The SoVideo Mandarin Chinese News Retrieval System”, *Int. Journal of Speech Technology*, Vol. 7, pp189-202, 2004

【3】陳俊良， “中文廣播新聞語音辨識之研究”，*國立交通大學碩士論文*，民國九十三年六月。

【4】Lie Lu, Hao Jiang, and HongJiang Zhang, “A robust audio classification and segmentation method,” *Tech. Rep., Microsoft Research*, 2001.

【5】M. Siegler, U. Jain, B. Ray and R. Stern, “Automatic segmentation, classification and clustering of broadcast news audio”, *Proceedings of the Speech Recognition Workshop*, pp 97-99, 1997.

【6】TingYao Wu, Lei Lu, Ke Chen, and HongJiang Zhang, “UBM-based Real-time Speaker Segmentation For Broadcasting News,” *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2003*, Vol. II, pp. 193-196, Hong Kong, April 4-10, 2003

【7】Joseph P. Campbell, “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, pp. 1436-1462, Sept 1997.

【8】S.S. Chen and P.S. Gopalakrishnan, “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion,” *DARPA Broadcast News Transcription and Understanding Workshop*, 127-132, 1998.

【9】Winston Hsu, Shih-Fu Chang, Chih-Wei Huang, Lyndon Kennedy, Ching-Yung Lin, and Giridharan Iyengar, “Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation,” *TRECVID 2003 Workshop*.

【10】Trischler, A. & Gopinath, R.A., “Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion” *Proceedings of EuroSpeech99*, pp679-682.

【11】Douglas A. Reynolds, “Robust text-independent speaker identification using gaussian mixture speaker models,” *IEEE Transactions on speech and audio processing*, 3(1), Jan. 1995.

【12】T. Kemp, M. Schmidt, M. Westphal, A. Waibel, “Strategies for automatic segmentation of audio data,” *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

【13】C. J. van Rijsbergen, *Information Retrieval*, London, Butterworth, 1979.

【14】J. T. Tou, R. C. Gonzalez, *Pattern Recognition Principles*, R. Kalaba, Ed. Reading, MA: Addison-Wesley, 1974.