電視新聞語料場景的自動切割與分類

姜柏亘 [1]，呂仁園 [1]，楊博厚 [2 3]，謝鴻文 [2]

1. 長庚大學資訊工程研究所
2. 長庚大學電機工程研究所
3. 中央研究院資訊所

E-mail: rylyu@mail.cgu.edu.tw, TEL:886-3-2218800ext5967

摘要

在本篇論文中，我們提出場景自動切割與分類的演算法，我們將一小時新聞分成爲四種場景：新聞主播報導（Anchor Reporting）、現場採訪報導（Live Reporting）、氣象主播報導(Weather Anchor Reporting)與廣告（Commercials）。我們擷取了時域與頻域的特徵值用以描述場景的特性，並使用高斯混合模型(Gaussian Mixture Model)當作場景分類器。場景切割的策略有兩種：(1)每秒移動策略、(2)快速策略。每秒移動策略，是利用每次移動一秒，並觀察3秒的聲音去決定場景的轉換點，效能評估方面，其Deletion Rate爲5.56%，Insertion Ration爲5.56%。由於上述的方法耗費計算的時間較久，因此我們也開發了一套快速策略，其Deletion Rate爲2.27%，Insertion Ration爲5.4%。在場景分類方面，我們使用了MFCC、LSTER、HZCRR、SF與MFS去將經過真實轉換點切割出的一段段聲音去作分類，可以達到92.5%的平均正確率。

1. 簡介

隨著網際網路的蓬勃發展，越來越多的新聞資訊可以直接從網路下載。而新聞資訊裡富含語音、音樂、文字、顏色樣型及影像圖形。雖然人類可以快速的透過觀察來解釋這些內容的含意，但是透過電腦分析去瞭解其內容還是處於初步的階段。新聞資訊的檢索、分析應該也要像我們人類的頭腦一樣去處理，換句話說就是在作處理前應先透過電腦先分析及瞭解其內容。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，因此考慮到場景轉換的語音切割與分類方法便是需要且直觀地，而瞭解場景內容對於以內容爲基礎的新聞資料庫索引與檢索是相當重要的。近幾年越來越多的研究在這領域努力。

一般來說，研究場景的切割與分類可以使用Model-based segmentation及Metric-based segmentation，其中Model-based segmentation的方法是將不同的場景聲學群組(acoustic class)建立不同的模型，例如高斯混合模型或隱藏式馬可夫模型等。舉例來說，若我們要切割電視新聞的話，我們便可以爲棚內主播、外場記者、外場受訪者、氣象主播等建立個別的模型，之後測試的聲音透過Model Testing便可以依照既有模型去算出此分析音框的Maximum Likelihood，進而可決定轉換點。另一方面，Metric-based segmentation的方法是利用距離量測的概念，選擇某一相異度量測公式，計算相鄰兩個frame的相異度，並決定一個門檻值去決定轉換點，而常用的相異度量測公式有KL distance、Common Component GMM-based Divergence[2]、Delta BIC[1]等。

Hsin-min Wang [1]收集了公共電視新聞語料，並利用Bayesian Information Criterion定義一個相異度量測的方法去偵測環境或是語者的轉換點。Yih-Ru Wang [2]則是使用GMM來描述相異兩個

聲音片段的統計特性，利用共用的mixture component來減少估計混合權重的計算量，以估計出權重向量來代表聲音片段的特性，進而量測相鄰聲音片段間的相異度，決定可能的轉換點。Tong Zhang[8]則是使用四種特徵值：平均過零率、能量、基礎頻率與頻譜鋒的追蹤(spectral peak tracks)，並設計一套有規則的策略將電視audio訊號分成語音、音樂、環境聲音、含音樂背景的語音、含音樂背景的環境聲音與靜音等，正確率可達到90%以上。Lekha Chaisorn [9]使用多個特徵值與技術將影片分析成一個個shot與場景，在shot階段，配置一個選擇樹去分類shot到13種與先定義的類別其中一種。Zhu Liu利用12種音訊特徵值，並結合神經網路分類器（neural network classifier）[4]與隱藏式馬可夫模型(HMM)[3]將電視節目場景分爲廣告、籃球賽、足球賽、新聞及氣象報告。Lie Lu [6]使用梅爾倒頻譜參數、過零率、能量、亮度與頻寬（Brightness and bandwidth）、頻譜流量、頻帶週期性（Band periodicity）、噪音音框比率（Noise frame ratio）並結合支援向量機（Support vector machine）將聲音串流切割分類爲靜音、音樂、背景聲音、純語音、含有音樂的語音。

論文接下來的架構：第 2 節描述了特徵值擷取與分析，第 3 節研究整個系統的切割與分類演算法流程，第 4 節爲實驗的效能評估與分析，第 5 節爲結論與未來展望。

2. 特徵值擷取

2.1 梅爾倒頻譜參數(Mel-frequency cepstral coefficients)

$$Cn = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log S_K) \cos[n(k-0.5)\pi / K] \quad n = 1,2,3...L \tag{2.1}$$

2.2 短時距低能量比率(Low Short Time Energy Ratio)

我們使用能量的變化率當成我們的特徵向量的成分，而不是準確的短時距能量值。我們使用短時距低能量比率(LSTER)去表示短時距能量的變化率。

$$LSTER = \frac{1}{2} \sum_{n=0}^{N-1} [\mathrm{sgn}(0.5 avEng - STE(n) + 1)] \tag{2.2}$$

其中 n 代表音框索引，N 代表一秒內的音框總數，sgn[.]是符號函式，以及 Eng(n)代表在第 n 個音框的能量，avgEng 是一秒內的能量平均值。

LSTER 是一個很有效的特徵，特別是在區分語音與音樂。通常，在語音中有許多的靜音，所以測量 LSTER 的值會高於音樂。下圖 2.5 代表 LSTER 的機率分佈曲線：(a)代表語音 (b)代表音樂。

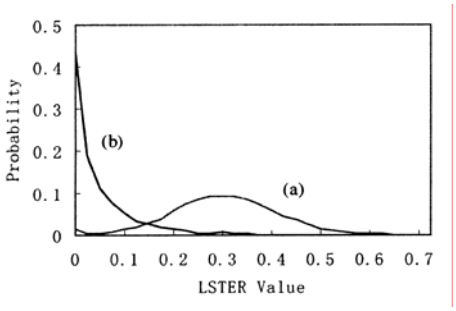

圖 2.1 LSTER 的機率分佈曲線

2.3 高過零率比率(High Zero Crossing Rate Ratio)

過零率在特徵化不同類型的音訊上被證明是非常有用的，他被使用在很多先前的語音與音

樂的分類演算法上。在我們的實驗中，我們發現過零率的的變化比原先的過零率的值更有辨識性，所以我們利用高過零率比率(HZCRR)當成演算法中的特徵值，如下定義：

$$HZCRR = \frac{1}{2}\sum_{n=0}^{N-1}\left[\mathrm{sgn}\left(ZCR(n) - 1.5avZCR\right) + 1\right] \qquad (2.3)$$

其中 n 代表音框索引，N 代表一秒內的音框總數，sgn[.]是符號函式，以及 ZCR(n)代表在第 n 個音框的過零率。通常語音訊號是由交替的 voice 聲音與 unvoice 聲音所組成，另一方面，音樂並沒有這種組成結構。因此，對於語音而言，它的過零率變化將會大於音樂。

下圖 2.7 代表 HZCRR 的機率分佈曲線：(a)代表語音 (b)代表音樂

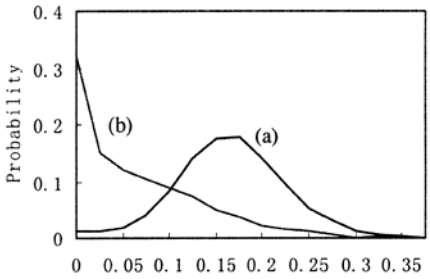

圖 2.2 HZCRR 的機率分佈曲線

2.4 頻譜流量(Spectrum Flux)

頻譜流量被定義成一秒內相鄰兩個音框的平均變化率的值，公式如下：

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}\left[\log\left(A(n,k)+\delta\right) - \log\left(A(n-1,k)+\delta\right)\right]^2 \qquad (2.4)$$

其中 A(n,k)是輸入信號第 n 個音框的離散傅利葉轉換(Discrete Fourier Transform)：

$$A(n,k) = \left|\sum_{m=-\infty}^{\infty} x(m)w(nL - me^{-j(2\pi/L)km})\right| \qquad (2.5)$$

x(m)是原始輸入的訊號，w(m)是窗函式，L 代表窗的長度，K 是 DFT 的階數，N 則是音框的總數，以及 δ 爲一個極小的數值避免計算時的溢位。

在我們實驗中，我們發現通常場外記者的 SF 值高於廣告，因爲場外的主要成分是爲語音或是環境聲，而廣告大部分都是由音樂組成。

下圖 2.3 代表 spectrum flux 特徵值的曲線：0~200 代表語音，200~350 代表音樂，350~450 代表環境聲音。

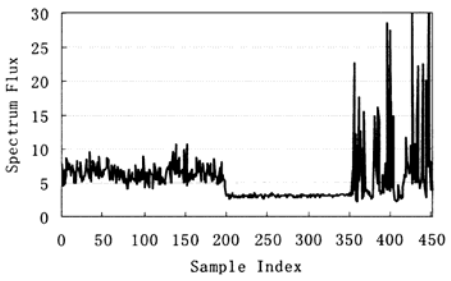

圖 2.3 spectrum flux 特徵值的曲線

2.5 梅爾頻率頻譜(Mel Frequency Spectrum)

在頻譜圖(spectragram)中，其實就可以明顯的看出語音與音樂的不同，但是若是直接取 FFT 完後的 512 或 1024 的值又太多，而且我們觀察頻譜時也不是全部觀察，而是去看它的密度比較深的地方。所以我們希望在取 MFCC 時，不要作最後的離散餘弦轉換，而是在做完 FFT 經過梅爾濾波組後的 26 個值當作一個特徵向量，稱之爲 MFS。

3.場景切割與分類

我們的系統架構同如圖3.1，首先我們會訓練出四種場景，輸入爲一小時的測試新聞，輸出爲四種場景的切割與分類。第一階段爲場景轉換點的偵測，第二階段爲場景的分類，我們會再接下來詳細介紹。

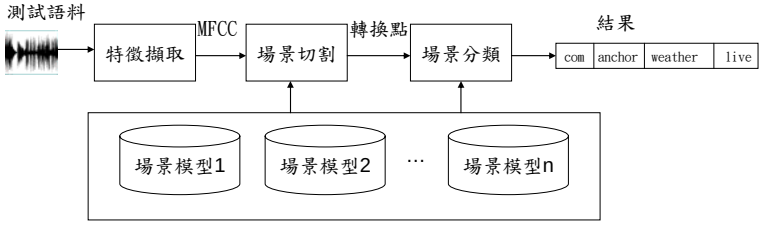

圖 3.1 系統架構圖

3.1 場景轉換點的偵測

在公共電視新聞語料中，我們可以發現研究中要切割與分類的場景，可以由新聞主播報導的場景將所有的場景類型切割出來。如圖3.2所示，我們將新聞主播報導場景的開始時間點與結束時間點找出來的話，這樣就可以順利的找到所有場景轉換點。

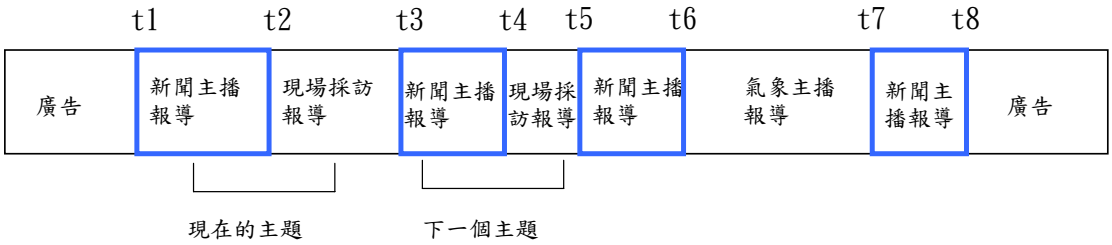

圖3.2 場景轉換時間點

t1代表第一個新聞主播報導的開始時間點，同時也是廣告場景轉換到新聞主播報導場景的轉換時間點，t2代表第一個新聞主播報導的結束時間點，同時也是新聞主播報導場景轉換到現場採訪報導場景的轉換時間點，t3、t4…以此類推。新聞主播報導場景的主要是由單一主播的語音構成，而攝影棚內的背景環境很安靜，並沒有背景環境聲音干擾。在許多語者辨識的系統中都是利用梅爾倒頻譜參數當成重要的特徵，將不同的語者區分出來，所以分類新聞主播報導方面，很適合用梅爾倒頻譜參數來描述主播的口腔組成，進而達到分類的效果。而語音訊號中富含了許多重要的因素讓我們來辨識新聞主播報導。

3.1.1 模型訓練

在研究中，如圖 3.3 所示，首先，從公共電視新聞語料中隨機選出三小時的測試語料，並把這三小時的語料，分出新聞主播報導、現場採訪報導、氣象主播報導、廣告等四種場景類型，

再對每一種場景透過特徵擷取子系統(Feature Extraction Subsystem)後，以特徵向量的形式儲存 39 維梅爾倒頻譜參數，之後利用模型訓練子系統(Model Training Subsystem)後，以模型高斯混合模型參數的型態儲存下來，訓練出四種不同的模型參數。

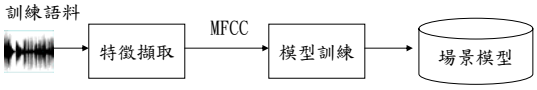

圖 3.3 場景模型訓練流程圖

3.1.2 模型測試

接著隨機選取出一小時的測試語料，經過特徵擷取子系統後，儲存 39 維的梅爾倒頻譜參數向量，利用模型測試子系統(Model Testing Subsystem)來對之前訓練出的模型參數找出最大事後機率（Maximum A Posteriori，MAP）的高斯混合模型，以辨識出場景的種類，如圖 3.4 所示：

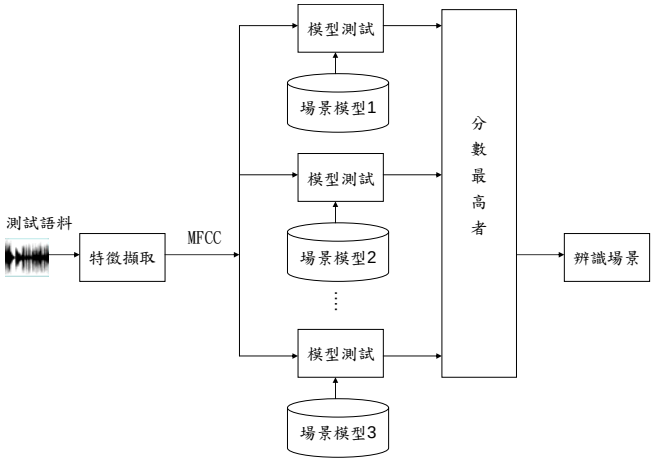

圖 3.4 模型測試流程圖

測試過程中，我們並不是拿整整一小時的新聞所擷取出的 MFCC 去跟每一個模型作比對，而是一段段的聲音片段去觀察與辨識場景，其中觀察的聲音片段爲 3 秒，並移動 1 秒計算一次 MAP，如圖 3.5 所示：

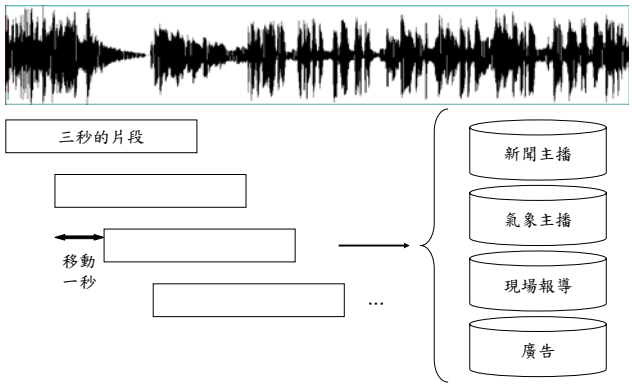

圖 3.5 一小時新聞測試的示意圖

換句話說，就是將測試的語音片段(3 秒)分別與新聞主播報導、氣象主播報導、現場採訪報導及廣告場景的高斯混合模型參數計算出可能的機率，之後選出機率最大者，然後我們判定此測

試語音的片段屬於此參數模型所對應的場景。由於我們是每移動一秒去計算是屬於哪個場景，所以當下一秒計算出的場景與現在不同時，我們便標示此時間點爲一個場景的轉換時間點。

3.1.3 後處理分析

另一方面，在每秒計算出的場景中，偶爾會出現 1 到 7 秒的錯誤判斷，我們稱之爲一個 error。對於此現象，我們利用 Median Filter 將其同化成相鄰的場景，根據實際聽取場景維持的秒數，平均一個場景的片段在 8 秒以上，所以我們針對 1 到 7 秒的 error 進行同化，以提升場景轉換時間點的正確率，另一方面透過實驗，我們也發現當同化到 6 秒以上正確率並不會再提升，反而有時還會下降，是因爲會同化到正確的轉換時間點，因此我們最多同化到 5 秒。

由以上 3 個步驟的步驟我們可以找出場景間轉換的時間點。

3.2 場景分類

在 3.1 節中，我們找出了所有場景轉換時間點，透過這些轉換時間點，便可以將一小時的新聞語料切割出一段段的場景片段，但是 MFCC 並不能正確的描述廣告的特性，因爲廣告與現場採訪報導最大的不同在於廣告含有音樂成分，因此在這部分，我們將非主播部分去擷取它的 LSTER、HZCRR、SF 與 MFS。將這一段段的非主播場景片段當作測試語料，並透過 GMM 分類器將這一整段的場景去作分類，並計算出是屬於現場採訪報導還是廣告場景。

3.3 加速策略

若是利用一秒一秒的去判斷屬於哪個場景，這樣一小時的新聞總共需要判斷 3600~3800 次，這樣所需的計算時間大致上要 10 分鐘，所以這部分我們開發了一套新的加速策略，希望可以將判斷切割與分類場景的計算時間縮短。發現當轉換點發生時，大致上都會有一段 0.2 到 1.0 秒的 silence，而在這 silence 的左右兩部分的特徵值分佈也不盡相同，如圖所示：

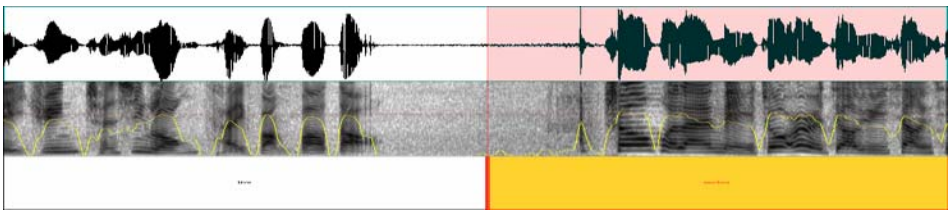

圖 3.6 現場採訪報導<->新聞主播報導

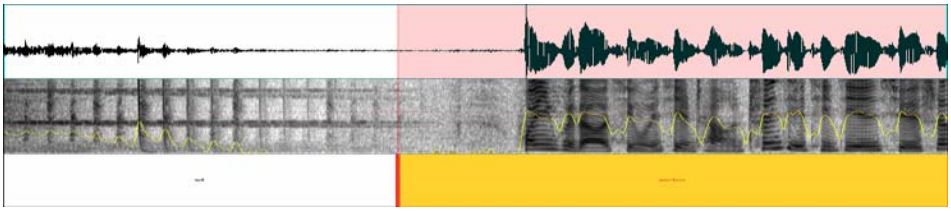

圖 3.7 廣告<->新聞主播報導

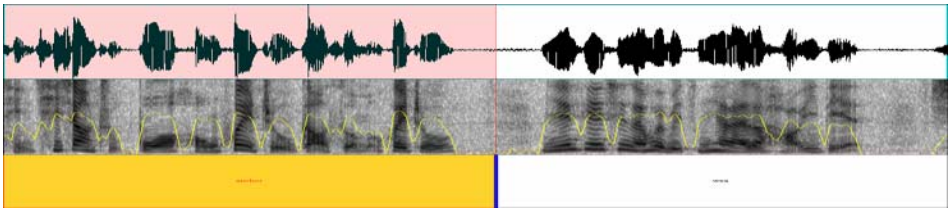

圖 3.8 新聞主播報導<->氣象主播報導

因此我們設計了一套快速判斷場景演算法去判斷這四大場景的轉換點，計算出的暫時轉換點(temp change point)大約有 600 到 700 個，而真實轉換點是這些暫時轉換點的子集合，相較於之前方法的 3600 個到 3800 個，我們大概可以省下大約 1/2 的時間，下面爲快速策略的演算法：

Step1：計算一小時新聞的能量

Step2：if(能量維持一段 0.1 秒長的 silence)

此時取這一段 silence 開始與結尾的 1/2 的時間點當成一個暫時轉換點。

Step3：將暫時轉換點的左右各 3 秒，總共 6 秒的聲音送進場景辨識器辨識。若左右兩段的場景不同，則此暫時轉換點爲真實轉換點。若左右兩段場景相同，則刪除此暫時轉換點。

Step4：合併出兩個真實轉換點間的場景類型。

圖 3.9 中，綠色曲線代表 step1 計算出的能量，紅色箭頭代表此時的靜音長在 0.1 秒以上，此時就設定藍色直線爲暫時轉換點，若暫時轉換點的左右場景不同則設爲真實轉換點，相同則刪除。

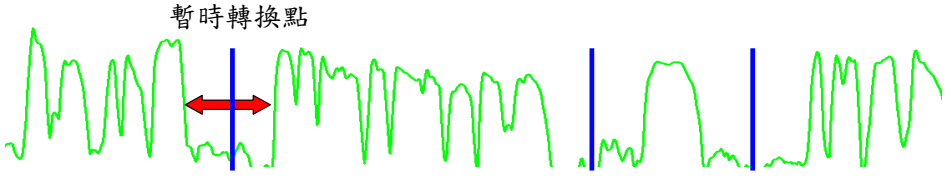

圖 3.9 快速策略

接下來，我們再利用另一個圖來解釋演算法，圖 3.10 代表一小時的新聞，其中 t1，t2，t3…代表在 Step2 之後計算出的暫時轉換點。在 t1 的左邊爲廣告，右邊爲新聞主播報導，所以我們變標示此點爲真實轉換點(實線)。另一方面，由於在 t3 的左邊爲現場採訪報導，同時右邊也是現場採訪報導，因此我們會刪除此暫時轉換點(虛線)，以下以此類推。因此 t1、t2、t4、t5 與 t7 爲真實轉換點。同時我們也將不同的場景類型分類了出來。

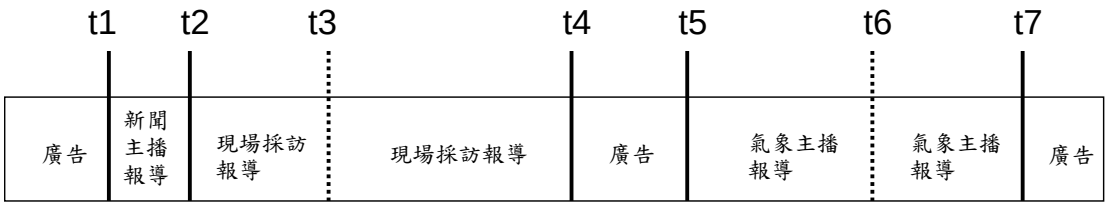

圖 3.10 快速策略

4. 實驗設計與實驗結果

4.1 公共電視新聞語料簡介

本論文所使用的語音資料庫爲公共電視新聞語料庫(Public Television Service News Database, PTSND)，是由中研院王新民教授以及助理研究團隊所整理規劃的中文電視新聞語料，收集了西元2001~2003年共220小時的新聞wave檔；其錄音的參數爲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，因考量檔案傳輸及讀取速度的問題，所以每個DAT都經由人爲處理成16kHz 16-bit單聲道的WAV檔。接下來我們簡述一下PTSND語料庫的一些統計特性，如表4.1所示；首先若我們以語者類別來區分的話，因爲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爲一類，稱之爲現場採訪報導，而氣象主播因爲其背景大多爲音樂，因此獨立出來統計；此外，新聞主播報導無背景聲音，故自成一類。

表4.1 PTSND 基本統計特性

Scene types	Percentage(in time)
新聞主播報導	17.68%
氣象主播報導	15.12%
現場採訪報導	59.20%
廣告	8.00%

4.2 系統效能評估

實驗的效能評估是利用插入率（insertion rate）以及刪除率（deletion rate）來評估我們的方法。

如圖 4.1 所示，Reference Boundaries 是指經由人工標示出的正確場景轉換點，Testing Boundaries 是指電腦經過我們設計出的策略後，計算出的場景轉換點。Insertion 代表電腦有計算出來的轉換點，但是人工並沒有標示此轉換點。Deletion 則是相反情況，及人工有標示此轉換點，但是電腦沒計算出此點。Matching 是指電腦計算出的轉換點跟人工標示的正確轉換點差距在 2 秒內。

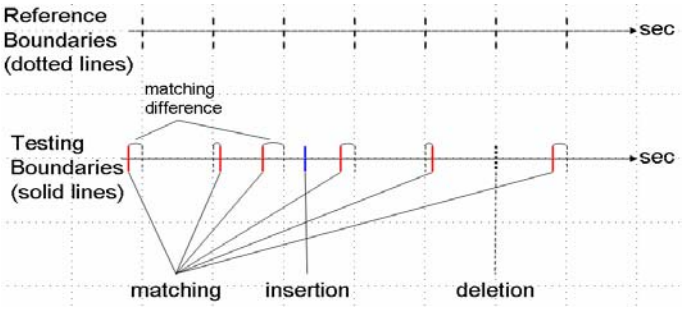

圖 4.1 效能評估

而它們之間的關係如下式所示：

$$N_{matching} + N_{insertion} = N_{Testing} \quad ,$$

$$N_{matching} + N_{deletion} = N_{\text{Re}\,ference}\ ,$$

其中$N_{matching}$ 代表Matching轉換點的數量， $N_{insertion}$ 代表Insertion轉換點的數量， $N_{deletion}$ 代表Deletion轉換點的數量。 $N_{\text{Re}\,ference}$ 和 $N_{Testing}$ 各代表人工標示與電腦計算的轉換點數量。接著我們定義了插入率(insertion rate)以及刪除率(deletion rate)的公式：

$$Insertion\_Rate = \frac{N_{insertion}}{N_{Testing}} \times 100\%$$

$$Deletion\_Rate = \frac{N_{insertion}}{N_{\text{Re}\,ference}} \times 100\%$$

4.3 實驗參數設定與結果

4.3.1 不同特徵向量維度對系統效能的影響

首先我們先評估不同特徵向量維度對轉換時間點辨識率的影響，如表所示：我們可以發現當特徵向量維度提高時，Insertion Rate 與 Deletion Rate 都會下降，這代表越高的特徵向量維度去描述場景的機率分佈與特性會越好。當特徵向量維度增加時，計算量會大幅升高，而導致耗費計算很多的時間，所以兩者是一種 trade off。最後，我們也發現當 mixture 增加時，Insertion Rate 與 Deletion Rate 會漸漸降低，但是 mixture 到了 64 有些不降反升，這是由於特徵值的分佈用了太多高斯分佈去描述了。

表 4.3 不同特徵向量維度的效能

mixture	condition / Evaluation	MFCC(13 dims)	MFCC+delta (26 dims)	MFCC+delta+ Deltadelta (39 dims)
4	Insertion Rate(%)	39.58	23.81	25
	Deletion Rate(%)	19.44	11.11	16.67
8	Insertion Rate(%)	25	20	20
	Deletion Rate(%)	16.67	11.11	11.11
16	Insertion Rate(%)	27.27	13.16	8.33
	Deletion Rate(%)	11.11	8.33	8.33
32	Insertion Rate(%)	26.19	15	8.33
	Deletion Rate(%)	13.89	5.56	8.33
64	Insertion Rate(%)	23.80	17.5	5.56
	Deletion Rate(%)	11.11	8.33	5.56

4.3.2 不同觀察片段時間長對系統效能的影響

接下來我們評估觀察片段時間長對轉換時間點辨識率的影響，如表所示，當觀察片段增加到 4 秒的時候，使由於可能包含到下一場景的特徵，導致影響這個時候的判斷。而觀察片段爲 2 秒時，判別場景的特徵不夠充分，所以 Insertion Rate 與 Deletion Rate 也會上升一些。因此我們設定觀察片段時間長 3 秒爲我們研究中的評估。

表 4.4 不同觀察片段時間長的效能比較

mixture	condition Evaluation	2秒	3秒	4秒
4	Insertion Rate(%)	23.68	25	22.5
	Deletion Rate(%)	19.44	16.67	13.89
8	Insertion Rate(%)	21.05	20	23.8
	Deletion Rate(%)	16.67	11.11	11.11
16	Insertion Rate(%)	11.11	8.33	18.42
	Deletion Rate(%)	11.11	8.33	13.88
32	Insertion Rate(%)	15.78	8.33	13.16
	Deletion Rate(%)	11.11	8.33	8.33
64	Insertion Rate(%)	8.33	5.56	8.33
	Deletion Rate(%)	8.33	5.56	8.33

4.3.4 每秒移動策略與快速策略速度比較

在研究中，主要影響計算時間的地方在模型測試子系統，因此我們針對兩種不同的策略，將它們的模型測試的時間作比較。如表 4.6，我們發現快速策略比每秒移動策略快了大約兩倍時間，主要是因爲每秒移動策略總共要測試模型 3600~3800 次，但是經由快速策略後的測試模型只要 600~700 次，但是快速策略要測試左右 3 秒的聲音各一次，判斷是屬於哪種場景，所以大致上也要測到 1200~1400 次。所以，快速策略比每秒移動的速度省了 1/2 以上。

表 4.6 每秒移動與快速策略的比較

method feature	每秒移動	快速策略
MFCC	153 sec	79 sec
MFCC+delta	230 sec	120 sec
MFCC+ delta+ delta_delta	294 sec	163 sec

4.3.5 每秒移動與快速策略的效能比較

我們發現快速策略的 Insertion Rate 與 Deletion Rate 比每秒移動的策略低。快速策略幾乎所有的轉換點都會去判斷，但是若轉換點出現在不是 silence 並維持 0.1 秒以上的話，快速策略就無法找出來。另一方面，由於每秒移動策略是每移動一秒就去判斷場景，這樣對於偵測場景轉換來說太細微了，也就是說一點點的差異就容易被誤判爲錯誤的場景，而這部分的錯誤無法利用同化場景的方法更正。而快速策略是利用場景轉換間會存在一段 silence，這是一個關鍵，而這種方法可以偵測出幾乎全部的場景轉換點，也因此快速策略的效能表現都比每秒移動策略佳。

表 4.7 每秒移動與快速策略比較

feature	condition Evaluation	每秒移動	快速策略
MFCC	Insertion Rate(%)	27.27	8.33
	Deletion Rate(%)	16.67	8.33
MFCC+delta	Insertion Rate(%)	20	5.40
	Deletion Rate(%)	11.11	2.77
MFCC+ delta+ delta_delta	Insertion Rate(%)	8.33	5.40
	Deletion Rate(%)	8.33	2.77

4.3.6 維持靜音長對快速策略效能影響

實驗中，靜音維持 0.1 秒，可以找出 862 個暫時轉換點，靜音維持 0.2 秒時可以找出 598 個暫時轉換點，0.3 秒可以找出 335 個。其中維持 0.1 秒可以找出較多的暫時轉換點，所以真實轉換點不容易遺漏，但也因此要判斷更多次場景的轉換。而有幾個轉換點，其之間的轉換就是小於 0.2 秒，因此若設爲維持 0.2 秒以上的話，此類的暫時轉換點就無法找出。

表 4.8 維持靜音長對快速策略效能比較

mixture	condition Evaluation	0.1秒	0.2秒	0.3秒
MFCC	Insertion Rate(%)	8.33	13.51	6.45
	Deletion Rate(%)	8.33	11.11	19.4
MFCC+delta	Insertion Rate(%)	5.40	11.11	6.66
	Deletion Rate(%)	2.77	11.11	22.22
MFCC+ delta+ delta_delta	Insertion Rate(%)	5.40	11.11	9.67
	Deletion Rate(%)	2.77	11.11	22.22

4.3. 分類結果的評估

我們評估場景分類的正確率，分類正確率的定義如下：

$$\text{分類正確率} = \frac{\text{正確辨別的測試檔案數}}{\text{所有的測試檔案數}} \times 100\%$$

每一種測試類型都包含 20 個測試檔案，總共 80 個檔案。接著將每個測試檔案透過場景分類器去作分類，例如第一行，代表新聞主播共 20 個檔案被分類到新聞主播的比率爲 100%。另一方面，我們發現在廣告的部份還是容易出錯，是由於廣告的組成相當複雜，其中主要成分是語音的測試檔案很容易出錯，在這種檔案類型中，音樂的聲音幾乎都被語音覆蓋掉，因此會容易辨識成現場採訪報導，我們用人耳去聽的確不是很清楚的可以辨別出其差異性。另外一方面，就分類結果而

言，我們的平均正確率可達 92.5%。

表 4.9 分類結果評估

分類類型 測試類型	新聞主播報導	現場採訪報導	氣象主播報導	廣告
新聞主播報導	100%	0%	0%	0%
現場採訪報導	5%	90%	5%	0%
氣象主播報導	0%	0%	100%	0%
廣告	0%	20%	0%	80%

5. 結論與未來展望

首先我們要感謝中研院王新民教授，他們研究團隊所開發的公共電視新聞語料(PTSD)，提供我們有關新聞中很多詳細的資訊，例如語者資訊、場景的類型、新聞的內容、轉換的時間點以及如何錄製新聞語料的步驟等等，讓我們能快速的瞭解與進一步分析新聞語料的內容。我們也利用了這份語料當成我們效能評估的依據。

在本論文中，我們分析各種場景時域與頻域的特性，使用高斯混合模型來做場景的切割與分類的分類器。在場景切割方面，我們透過新聞主播報導的開始與結束時間點去決定場景的轉換點，使用每秒移動策略，其 Deletion Rate 爲 5.56%，Insertion Ration 爲 5.56%。在場景分類方面，我們使用了 MFCC、LER、HZCRR、SF 與 MFS 去將經過真實轉換點切割出的一段段聲音去作分類，可以達到 92.5%的平均正確率。由於上述的方法耗費計算的時間較久，因此我們也開發了一套快速策略，計算時間節省了大約 1/2，其 Deletion Rate 爲 2.27%，Insertion Ration 爲 5.4%。本論文與其他研究不同在於我們利用知識爲基礎的想法、當人類碰到此問題實是如何解決的，如何去判斷場景的不同與場景轉換時會發生那些現象，將這些解決的想法轉成知識，進一步將這知識設計一套演算法而把場景的切割與分類自動化，而傳統處理場景轉換點偵測是利用相異度的判斷，但是這種方法會容易受到環境的影響（如環境聲、音樂聲、噪音等等）而導致誤判，所以有較高的 Deletion Rate 與 Insertion Rate。最後，我們期望未來可以利用我們分類出來的四種場景去擷取出更多的音訊類型，並研究更多頻域與時域的有效特徵值，另一方面，結合視覺特徵及文字上的資訊，開發更快速正確的方法，將場景的自動切割與分類辨識率提升。

6. 參考文獻

[1] Hsin-min Wang, Shi-sian Cheng, and Yong-cheng Chen, “The SoVideo broadcast news retrieval system for Mandarin Chinese.” International Conference on Spoken Language Processing 2004

[2] Yih-Ru Wang, Chi-Han Huang, “Speaker-and-environment change detection in broadcast news using the common component GMM-based divergence measure.”, International Conference on Spoken Language Processing 2004, pp1069-1072.

[3] Jincheng Huang, Zhu Liu, Yao Wang, ”Joint scene classification and segmentation based on hidden

Markov model", Multimedia, IEEE Transactions on Volume 7, Issue 3, June 2005 Page(s):538 - 550

[4] Zhu Liu, Yao Wang, Tsuhan Chen, "Audio feature extraction and analysis for scene segmentation and classification." Journal of VLSI Signal Processing Systems 1998, Vol.20, pp61 – 79.

[5]Lie Lu, Hong-Jiang Zhang, Hao Jiang, "Content analysis for audio classification and segmentation." IEEE Trans. on Speech and Audio Processing 2002, Vol.10, No.7, pp.504-516.

[6] Lie Lu, Hong-Jiang Zhang, Stan Li, "Content-based audio classification and segmentation by using support vector machines."ACM Multimedia Systems Journal 2003, pp. 482-492.

[7] Lie Lu, Hao Jiang, Hong-Jiang Zhang, "A robust audio classification and segmentation method." ACM International Conference on Multimedia 2001, pp203-211.

[8]Tong Zhang, C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification." IEEE Transactions on Speech and Audio Processing 2001, Vol.9, No.4.

[9]Lekha Chaisorn and Tat-Seng Chua, "The Segmentation and Classification of Story Boundaries in News Video" , Proceeding of 6th IFIP working conference on Visual Database Systems VDB6 2002, Australia 2002

[10]TingYao Wu, Lie Lu, Hong-Jiang Zhan, "UBM-based real-time speaker segmentation for broadcasting news."Speech and Signal Processing (ICASSP) 2003, Vol. II, pp. 193-196.

[11]Ting-Yao Wu, Lie Lu, Ke Chen, Hong-Jiang Zhang, "Universal background models for real-time speaker change detection."Proc. of the 9th International Conference on Multi-Media Modeling 2003, pp.135-149.