MPEG-2 AAC電影音效的內涵式自動分段

蕭聖峰、劉志俊

中華大學資訊工程系

sefe.hsiao@msa.hinet.net、ccliu@chu.edu.tw

摘要

從 90 年代至今，電腦軟、硬體相對地成長到一個前所未有的高度，而隨著許多超大型資料庫的建立，資料的擷取、檢索和分類，其重要性已經成為不可忽視的研究議題。在數位多媒體方面，藉由 ISO 所規範的 MPEG 標準，各種的儲存型式及功能，讓數位影音得以不斷地演化而趨於完善。就音訊格式的演進而言，從單音到立體聲，進而到多聲道音訊格式。以 MPEG-2 AAC 格式為例，其為多聲道的音訊格式，是 ISO 用以取代立體聲音訊格式以及奠定 MPEG-4 音訊核心的標準。

在本文中，將以電影音訊 MPEG-2 AAC 為背景，對此音訊格式進行分析，藉由音效的分析方法和內涵描述工具，對於電影的內容加以識別。在 MPEG-7 音效特徵描述子的基礎下，MPEG-2 AAC 的音效自動化分段技術，將對電影音效做自動化的內涵分析。

關鍵字：MPEG-2 AAC、MPEG-4、MPEG-7、音效分段

一、導論

基於數位多媒體的快速發展，人們對於數位多媒體的審視標準逐漸提高，對於影像及原音的重現不斷地苛求，而對應的 ISO 標準在資料格式的容量和品質上不斷地提升，以期在儲存和應用的範疇上，趨於完美。但是不斷產生的數位多媒體資料，使得多媒體資料庫的容量達到一個驚人的地步，龐大的資料庫難以有效率地檢索，而各種多媒體資料的特殊內涵，也不易查詢以及做出合適的索引。所以對於多媒體資料的內涵特徵，給予合適的定義及描述顯得關鍵而重要。

2001 年 ISO 制定了 MPEG-7 的正式規範[28]，用以定義及描述數位多媒體資料的基本結構和相關特性。在 MPEG-7 的定義下，其並非經由資料內涵描述的方法對多媒體資料編碼或解碼，而是對於資料內涵實現具體的描述以及有效率地提取其相關的特徵值。藉由 MPEG-7 描述多媒體資料的標準特徵值，各種形式的特徵值過濾器，得以對多媒體資料的內涵加以識別，進而使用在許多的多媒體應用上，例如，資料索引的建立、資料內容的識別、資料類型的分類等等[7][18][21]。因此，本文將在 MPEG-7 音效特徵值組的基礎下，針對電影音訊 MPEG-2 AAC 做自動化分段處理。

就近代商業電影而言，觀眾的審視角度不再侷限於故事的情節，更多的是受到聲光特效的撼動，進而決定主觀的評價，如同法國社會學家尚‧布希亞(Jean Baudrillard, 1929- 2007)所闡釋，在大量模擬效果的影響下支配著人們接受過度的真實[25]。而音效在一部電影中，是一個重要而不可或缺的元素，二維的影像畫面佐以適當的音效，對於人類來說，呈現出來的是多層次的臨場感覺，例如動態音效的傳導所造成的移動感、低頻音效的厚實感和高頻音效的銳利感等等。不同類型的電影，其特殊音效的類型各異，我們將藉由 MPEG-2 AAC 音效內涵分析方法，以特徵值擷取電影音效，加以歸納分類，從而可自動對電影內容分段及索引，並應用在多媒體資料庫的檢索以及多媒體資料的內涵分析。

在音效分段的相關研究中，有許多研究提出以影像及音訊兩者分別對電影鏡頭做偵測，以取得分段資料[2][3][4][5]，再將音訊內容分割為語音(Speech)、音樂(Music)和音效(Sound Effect)三類[15][30][31]。由於音訊內涵的特徵差異，音效可分為感知和聽覺的特徵值，以便將不同類型的音效加以自動分類[13][14][16]。音效類型和其他音訊類型的音效特徵值變異量存在著明顯的差別，如頻譜通量 SF(Spractrum Flux)和框架雜訊比 NFR (Noise Frame Ratio)等等。在聲紋和語音識別的相關研究，基於人耳的聽覺感知，以音訊頻率域或時間域的各種特徵係數，從而辨識分段，如梅爾倒頻譜係數 MFCC(Mel Frequency Cepstral Coefficients)和線性預測係數(Linear Predictive Cepstral Coefficients)等[8][19]。而 Saunders 等人則以能量函數(Energy function)、均方根(RMS)和越零率(Zero-Crossing Rate，ZCR)等等方法分段音訊的種類[17][26][35]。近年 Panagiotakis 和 Tziritas 提出均方根配合越零率的分段系統，其實驗結果，正確分段為 97%，分類部份為 95%[9]。

本文的自動音效分段研究，將以多組不同的 MPEG-7 特徵值進行音訊分析，基於特徵值的差異，我們能對音訊自動分段，藉以分辨出各式的音效類型。本文起始介紹電影音訊及音效分段相關研究，在第二節為 MPEG-2 AAC 多聲道音訊壓縮標準簡介，第三節架構 MPEG-2 AAC 電影音效分段之系統，第四節敘述電影音效自動分段的方法，第

五節說明實驗與分析結果，第六節總結研究及未來工作方向。

二、MPEG-2 AAC 音訊壓縮標準簡介

聲音紀錄自愛迪生 1877 年發明留聲機起，至 1977 年日本出現 PCM(Pulse Code Modulation)脈波編碼調變技術，經歷百年才由類比音訊跨入數位音訊的新世代，ISO/IEC 13818-7 (Advanced Audio Coding)簡稱 MPEG-2 AAC，在 1994 年 3 月由 Fraunhofer IIS、AT&T、Dolby Labs、Sony、Hanover University 和 NEC 等等協力參與，於 1997 年 JTC1/SC29/WG11 訂定，Bosi 和 Brandenburg 等人提出[27]。彈性而有效率的壓縮方式奠定多聲道音訊世代的新標準。

由於電影院或劇場的特殊環境限制，大量的聽眾使得 ITU-R 5.1 聲道的配置[22]難以符合劇院的需求，人們向多聲道應用尋求解答。藉由目前數位媒體的迅速發展，使得電影音訊已不再局限於影片膠卷上的齒孔間距，在 DVD 的大量使用下，解決了多聲道的音訊格式的儲存空間。多聲道的環場音效豐富了電影的品質，也提供了人們對於多聲道音訊格式的選擇，如 MPEG-2 AAC、Dolby AC-3[6]、DTS[23]和其它延伸的格式等等。

其中，MPEG-2 AAC 綜合了 MP3、MPEG-2 和 Dolby AC-3 的優點，提升了壓縮的效率及編碼的彈性，更大幅地壓縮了資料傳輸比，每聲道只用 64kbit/s 的資料比即可符合 ITU-R BS. 1116 的音質標準[24]，獲得優於舊有音訊格式的音質，並讓音訊在低頻寬的狀態下，能有較佳的品質。MPEG-2 AAC 提供單聲道至 48 個主聲道的選擇，並支援 16 個低頻音效聲道、16 個配音或多語言聲道和 16 個內嵌資料流程，使其在多聲道環繞立體聲有充裕的應用範圍。再者，由於 MPEG-4 格式廣泛的應用於網際網路、DAB、DVB、劇院系統及手持通訊裝置上，使其音訊核心 MPEG-2 AAC 在日常音訊應用上漸漸取代 MP3 音訊格式，這許多的優點及改良都顯示出 MPEG-2 AAC 在下一個音訊世代的核心價值。

三、MPEG-2 AAC 電影音效分段系統

對於 MPEG-2 AAC 的電影音效分段系統，以下將從幾個步驟加以分解並敘述：

步驟一. 以商業電影為背景資料，分離電影的數位視訊及音訊，儲存音訊為 MPEG-2 AAC 檔案格式，做為音訊樣本。

步驟二. 在電影音訊上，以多組 MPEG-7 音效特徵值公式為基礎，經由程式計算音效切片內相關的音效特徵值。

步驟三. 依據音效斷點偵測法做自動音效分段分析，取得音訊樣本的斷點，並依斷點位置做電影音訊的音效分段。

步驟四. 從取得的正確音效分段，建立電影音效資料庫的摘要及索引。

鑒於多聲道環場音效的普及度以及目前儲存媒體的容量限制，本文所提 MPEG-2 AAC 檔案格式尚未在商業電影音訊上大量使用，但在部分為第二區域碼(日本、歐洲和南非等地)的 DVD，已可用做電影音訊資料的取樣來源，自動音效分段的主要關鍵技術如後文敘述。電影音效分段系統架構如圖 1. 所示。

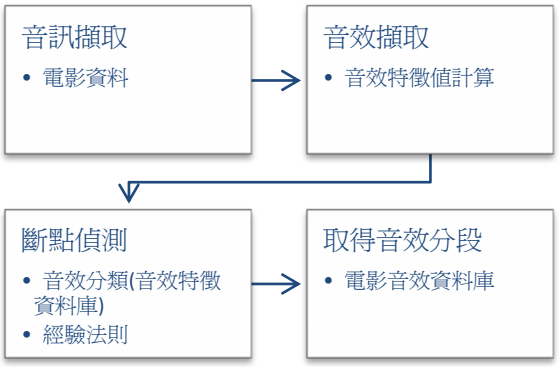

圖 1. 電影音效分段系統架構圖

(一)MPEG-2 AAC 音效特徵值擷取

從 MPEG-2 AAC 的壓縮檔案格式下，我們將在解碼過程中取得各個聲道的音訊內涵特徵值，用以 MPEG-7 音效描述子的計算。MPEG-2 AAC 定義了三種不同的格式，為主要格式(Main profile)、低複雜度格式(LC，Low Complexity profile)和可變取樣率格式(SSR，Scalable Sampling Rate profile)。「主要格式」包含了除增益控制(Gain Control)外所有的其他功能，提供品質最好的音訊。「低複雜度格式」使用限制的時域雜訊重整(TNS，Temporal Noise Shaping)，沒有預測(Prediction)、前置處理(Pre-processing)及增益控制的功能，雖降低了少許的音質，但卻大量地減少編碼和解碼的複雜度。「可變取樣率格式」使用限制的時域雜訊重整和頻寬，沒有預測的功能，經由一個增益控制來執行預先處理的部分。其格式能提供變動頻率的訊號。

本文所使用的 MPEG-2 AAC 內涵特徵值擷取於 MPEG-2 AAC 解碼程序中的濾波器模組(Filter Bank)輸出，如圖 2. 所示。在 MP3(ISO/IEC 11172-3)使用的是混合型濾波器組，而 MPEG-2 AAC 使用的則是 MDCT(Modified Discrete Cosine Transform)濾波器組。MPEG-2 AAC 的濾波器組被設計成允許視窗改變大小，用來適應輸入信號的狀態。視窗的大小隨著編碼器及解碼器同時改變，讓濾波器組

能有效率地分辨變化多端的輸入訊號，而分離其頻譜成分。其較長的轉換視窗長度，可變換的視窗型態，及可變更轉換區塊的長度，使得 MDCT 優於使用預先編碼法的濾波器組，並且提供濾波器組更好的頻率選擇性。雖然量化和編碼都是在頻率域裡執行完，解碼濾波器組的功能是利用 IMDCT(Inverse Modified Discrete Cosine Transform)，將解碼器輸入端的頻譜值，轉換成時間域的輸出值。對每個聲道而言，經由 IMDCT，N/2 個的時間-頻率值 Xik 被轉換到 N 個的時間域值 Xin 裡。IMDCT 的表示法如下：

$$x_{in} = \frac{2}{N}\sum_{k=0}^{N/2-1} X_{ik}\cos[\frac{2\pi}{N}(n+n_0)(k+\frac{1}{2})], n = 0,...,N-1 \quad (1)$$

公式(1)中，n 為樣本指標，N 為轉換視窗長度，i 為區指標，而 $n_0 = (N/2+1)/2$ 。

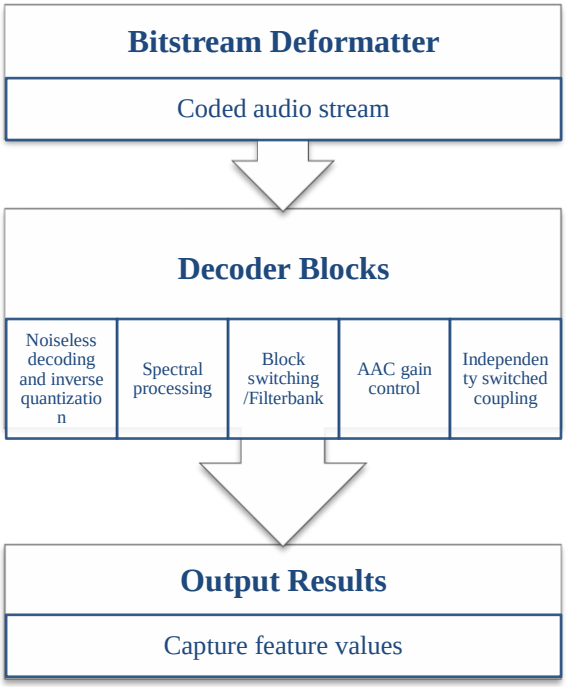

圖 2. MPEG-2 AAC 解碼流程圖

MPEG-2 AAC 的資料流(Bit Stream)同樣以框架(Frame)為單位，每一個框架為一個包含 1024 個 MDCT 係數的音訊區塊(Block)，或八個包含 128 個 MDCT 係數的音訊區塊，因此每一個框架包含 1024 個 MDCT 係數。對於不同內涵特徵值分析所需的音訊資料可能不同，因此必須要選取適當的框架數量來調整適當的資料量。一般 DVD-9 取樣頻率為 48kHz，所以 MPEG-2 AAC 的檔案格式，每一秒包含有 46.87 個框架（48000/1024＝46.87）。一般取這些 MDCT 係數的平方來計算該頻帶的能量。

我們所擷取的 MPEG-2 AAC 音訊內涵特徵值，是從 MPEG-2 AAC 解碼過程中擷取出 1024 個 MDCT 係數。然後經由擷取出的內涵特徵值作正規化(Normalization)後[29]，再將正規化的音效特徵值傳送到音效特徵值計算模組，做各種自動分段與分類的音效特徵值計算及內涵分析。在程式實作實驗的部份，以修改 FAAD(Free Advanced Audio Decoder)發展的解碼程式，從所輸入的 AAC 檔案，在其解碼過程中（如圖 2.），分析適當的內涵特徵值輸出位置及擷取數值，圖 3. 為 MPEG-2 AAC 音效內涵特徵值擷取程式的實作畫面。

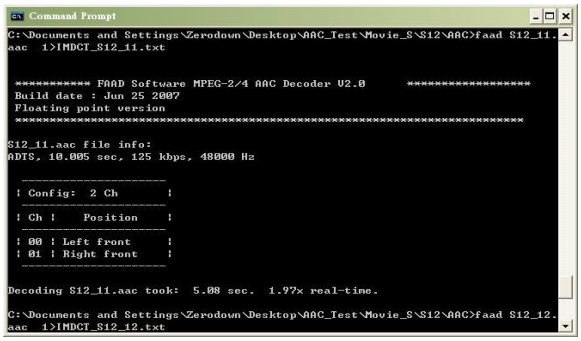

圖 3. MPEG-2 AAC 音效內涵特徵值擷取程式畫面

(二)分段用途之音效特徵值計算

本文實驗中，在分段前先將原始資料做前置處理的音訊切片，再從音訊切片取固定長度的分析樣本，以利於特徵值的擷取與分析。基於各種音效特徵值的差異性，所能偵測的音效種類各異。因此我們採用多種特徵值來做分析，以期能盡量涵蓋所有音訊類型的特徵。實驗之特徵值主要根據傳統音訊分類常用的特徵值及 MPEG7 音訊描述子，作為環場音效特徵[1]，最後形成由多維所構成的特徵向量。以下介紹本篇論文所採用的主要音訊特徵值。

甲. 平均能量(AveRMS)

在 MPEG-2 AAC，音訊經由 MDCT 轉換後，每個框架具有 1024 個 MDCT 值，其各自代表音訊資料在不同頻率下的能量表現，將框架中 MDCT 係數平方加總則表示單一框架的總能量，MPEG-2 AAC 的檔案格式每秒鐘大約含有 46.87 個框架，因此平均能量(Average Root Mean Square)特徵值公式如下所示：

$$RMS = \sqrt{\frac{\sum_0^{1023}(M[i]^2)}{1024}}, 0 \le i \le 1023 \quad (2)$$

$$AveRMS = \frac{\sum_0^{f-1}(RMS)}{f} \quad (3)$$

公式(2)中，M[i]表示在第 i 個框架中的 MDCT 值，公式(3)的 f 為音效之框架數。就差異性而言，普通在音樂中的平均能量會比在語音中低，利用片段能量比的差異，可加權於分段斷點的判斷。

乙. 靜音比例(AveSR)

在 MDCT 轉換後，當框架中的 MDCT 係數小於最大能量的 0.05 倍時，作其加總，並除以整個框架的能量，最後統計靜音比例。

(4) $$SR[k]=\frac{\sum M[i][k]\le 0.05\times \max M}{\sum_0^{1023} M[i][k]}$$

(5) $$AveSR=\frac{\sum_0^{k-1} SR[k]}{K}$$

公式(4)中，maxM 為具有最大能量的 MDCT 係數。一般在語音中出現靜音的比例會比音樂高，因此使用平均靜音比，可藉以有效的分辨音樂和語音。

丙. 4 赫茲調變能量(4ME)

一般而言，語音在一秒鐘有四個音節的能量語調高峰。藉由語音所特有的性質，在計算 MPEG-2 AAC 的 4ME 特徵值時，將正規化後的音訊能量，依照自我相關計算(Autocorrelation)找出該音訊與 4Hz Cosine 波形的相似度，如公式(6)。

(6) $$4HZ=Max\left(\sum_{i=0}^{18}\sin\left(\frac{8j}{19}\pi+\frac{8i}{19}\pi\right)\times RMS[i]\right),0\le j\le 18$$

丁. MPEG-7 音訊描述子

從 MPEG-7 的音訊描述子擷取音訊特徵，可分為頻譜特性、能量特性、頻率特性和頻率能量特徵四類，以下說明我們使用 MPEG-7 計算 MPEG-2 AAC 特徵值之公式。

- 頻譜特性

總平均頻譜通量：計算 MPEG-2 AAC 音效檔中 1024 個平均頻譜通量的總平均值。如公式(8)。

(7) $$SF_f[i]=[\log(|M_f[i]|+\delta)-\log(|M_{f-1}[i]|+\delta)]^2$$

(8) $$AveSF_f=\frac{\sum_{i=0}^{1023} SF_f[i]}{1024},0\le i\le 1023$$

(9) $$AveFrameSF=\frac{\sum_0^{f-2} AveSF_f}{f-1}$$

平均頻譜質量中心：頻譜的平衡點，可用來區分有聲語音和無聲語音。如公式(10)。

(10) $$SC=\frac{\sum_{i=0}^{N/2-1} f(k)M[i]}{512},0\le i\le 511$$

(11) $$AveSC=\frac{\sum_0^{f-1} SC}{f}$$

平均頻譜偏斜：以 85%的能量頻譜分布狀況，量測頻譜形狀偏移的不對稱性。其目的用來分辨無聲語音中有聲語音的部份，主要因為一般的無聲語音在頻譜的高頻部份會有很高比例的能量，而音樂則分佈在低頻部分。如公式(12)。

(12) $$\sum_{i=0}^{K_{roll}}|M[i]|\ge 0.85\sum_{i=0}^{N/2-1}|M[i]|$$

(13) $$AveSpectralRolloff=\frac{\sum_0^{f-1} K_{roll}}{f}$$

公式(12)中，Kroll 為符合此公式的最小值。

平均非零頻譜通量：所有大於零的頻譜通量平均值。如公式(14)。

(14) $$AveNZSF=\frac{\sum_0^{1023} SF[i]}{Count\left(\sum_0^{1023} SF[i]>0\right)},0\le i\le 1023$$

- 能量特性

平均能量：即平均音量。如公式(15)。

(15) $$AvePow=\frac{\sum_0^{f-1}\sum M[i]}{f\times 1024},0\le i\le 1023$$

平均框架低能量比。如公式(17)。

(16) $$LE=\frac{\sum M[i]\le 0.3\times AvePower}{\sum M[i]},0\le i\le 1023$$

(17) $$AveLowEnergy=\frac{\sum_0^{f-1} LE}{f}$$

平均框架中能量比。如公式(19)。

(18) $$ME=\frac{0.3\times AvePower\le\sum M[i]\le 0.7\times AvePower}{\sum M[i]},0\le i\le 1023$$

(19) $$AveMidEnergy=\frac{\sum_0^{f-1} ME}{f}$$

平均框架高能量比。如公式(21)。

(20) $$HE=\frac{\sum M[i]\ge 0.7\times AvePower}{\sum M[i]},0\le i\le 1023$$

(21) $$AveHigEnergy=\frac{\sum_0^{f-1} HE}{f}$$

平均負能量。如公式(22)。

(22) $$AveNegPow=\frac{\sum_0^{f-1} Count\left(\sum M[i]<0\right)}{f\times 1024},0\le i\le 1023$$

- 頻率特性

平均框架頻率，即平均框架音高。如公式(24)。

(23) $$frameFeq=\frac{\sum line\times 46.87}{Count\left(\sum M[i]>0\right)},0\le i\le 1023$$

(24) $$Avefeq=\frac{\sum_0^{f-1} frameFeq}{f}$$

- 頻率能量特徵

平均框架最大能量之頻率。如公式(25)。

(25) $$AveMaxPowFeq=\frac{\sum_0^{f-1}(f\max MDCT\times 46.87)}{f}$$

平均框架低頻能量比：低於 200Hz 頻率所佔之能量比。如公式(27)。

(26) $$low=\frac{\sum_0^{4} M[i]}{\sum_0^{1023} M[i]}$$

(27) $$AveLowFeqPow = \frac{\sum_{0}^{f-1} low}{f}$$

平均框架中低頻能量比：200Hz 到 500Hz 頻率所佔之能量比。如公式(29)。

(28) $$midlow = \frac{\sum_{5}^{11} M[i]}{\sum_{0}^{1023} M[i]}$$

(29) $$AveMidLowFeqPow = \frac{\sum_{0}^{f-1} midlow}{f}$$

平均框架中頻能量比：500Hz 到 1kHz 頻率所佔之能量比。如公式(31)。

(30) $$mid = \frac{\sum_{12}^{21} M[i]}{\sum_{0}^{1023} M[i]}$$

(31) $$AveMidFeqPow = \frac{\sum_{0}^{f-1} mid}{f}$$

平均框架中高頻能量比：1kHz 到 2kHz 頻率所佔之能量比。如公式(33)。

(32) $$midhig = \frac{\sum_{22}^{43} M[i]}{\sum_{0}^{1023} M[i]}$$

(33) $$AveMidHigFeqPow = \frac{\sum_{0}^{f-1} midhig}{f}$$

平均框架高頻能量比：高於 2kHz 頻率所佔之能量比。如公式(35)。

(34) $$hig = \frac{\sum_{44}^{1023} M[i]}{\sum_{0}^{1023} M[i]}$$

(35) $$AveHigFeqPow = \frac{\sum_{0}^{f-1} hig}{f}$$

四、電影音效自動分段的方法

傳統的音訊分段方法利用特徵值所產生的差異性(Differential)來偵測音效的斷點，取音訊內的各類音效類型作為分段的依據。所以我們將電影音訊內容分段成語音、音樂、音效和靜音四類，並以 GMM、BPN 和 SVM 等分類器對音訊樣本的音效切片內所計算出的相關音效特徵值，進行訓練並建立分類模型，利用這些分類模型，標記音效切片作為音效類型的識別。在標記不同的相鄰片段即為音訊斷點的位置(如圖 4.)，藉由音訊切片標記的差異來判別不同的音效類型。在識別出不同的音效分段後，經音效特徵值資料庫比對並加以分類。

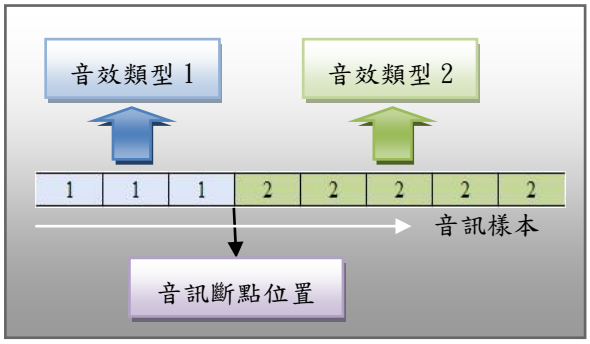

圖 4. 音效切片標記分段圖

(一)分類器斷點偵測法

甲. 高斯混合模型 (GMM)

直至目前，有許多機率函數模型用於語音及語者識別，如通用背景模型(Universal Background Model，UBM)、隱藏式馬可夫模型(Hidden Markov Model，HMM)和高斯混合模型(Gaussian Mixture Model，GMM)[32]等等，其中又以高斯混合模型的效果優於其他方式。對於音頻訊號在維度空間的特徵向量表示如圖 5.，而以混合加權的機率密度函數則稱為高斯混合密度函數或高斯混合模型。

一個高斯混合模型具有三種參數，即混合加權值(w_i)、平均向量值(μ_i)和共變異矩陣(Σ_i)，如下所示。

(36) $$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1...M$$

其中，λ 代表一段音訊，M 為高斯分佈的特徵個數。而對於一個 D 維空間的特徵向量 X，其混合密度用於機率函數的定義如下：

(37) $$p(x \mid \lambda) = \sum_{i=1}^{M} w_i p_i(x)$$

而密度的線性加權 $p_i(x)$，則為：

(38) $$p_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'(\Sigma_i)^{-1}(x-\mu_i)\right\}$$

其中的混合權重 w_i，須滿足 $\sum_{i=1}^{M} w_i = 1$。

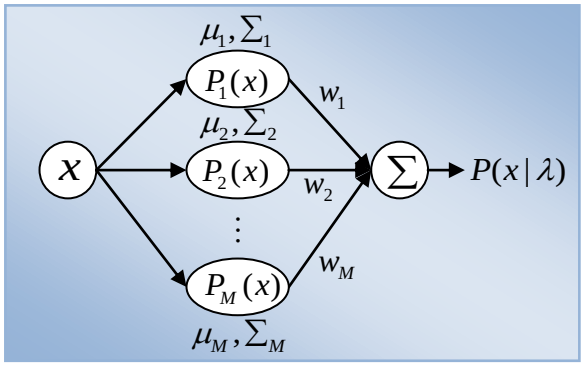

圖 5. 高斯混合模型架構

乙. 倒傳遞類神經網路 (BPN)

類神經網路理論源於 50 年代，科學家模仿人類大腦組織及運作，提出「感知機」(Perceptron)的神經元模型，而感知機通常作為分類器(Classifier)使用。近年許多學者針對不同的問題，提出許多的類神經網路模型，每一種的演算法並不相同，常見的網路有：倒傳遞網路(Back-propagation Network)、霍普菲爾網路(Hopfield Network)、半徑式函數網路(Radial Basis Function Network)，其中應用最廣泛地是倒傳遞類神經網路(Back-propagation Network，BPN)[20][33]。

典型的倒傳遞類神經網路有三層架構(如圖 6.)，第一層為輸入層，第二層為隱藏層，第三層為輸出層。每一層為多節點組成，且每一層之節點與相鄰層的每一節點相互連結，形成網路架構。倒傳遞類神經網路具有一層至多層的隱藏層，使網路利用平滑可微分轉換函數表示輸入與輸出單元間的映射關係，並可利用最陡坡降法（The steepest gradient descent method）將誤差函數最小化，使網路導出修正的加權值，進而最佳化[11][12]。倒傳遞類神經網路的資料運算，是由輸入層向隱藏層傳遞，由隱藏層運算可得隱藏層第 j 個節點的輸出值(如公式 39)，再傳遞至輸出層可得到輸出層第 k 個節點的輸出值(如公式 40)。

$$h_j = g\left(\sum_{i=1}^{m} w_{ji}x_i + \theta_{wj}\right), j = 1\cdots n \quad (39)$$

$$y_k = g\left(\sum_{j=1}^{n} w_{kj}h_j + \theta_{wk}\right), k = 1\cdots o \quad (40)$$

公式(39)及公式(40)中，w_{ji} 與 w_{kj} 為連結輸入層、隱藏層和輸出層的加權值，x_i 為輸入層第 i 個節點輸入值，θ_{wj} 與 θ_{wk} 為轉換函數 g 的門檻值或閥值(bias)具有偏移的效果，m、n 和 o 為各層的節點個數，轉換函數 g 可為線性或非線性函數，倒傳遞類神經網路的回想速度快，學習率高，本文使用此模型來進行分類實驗。

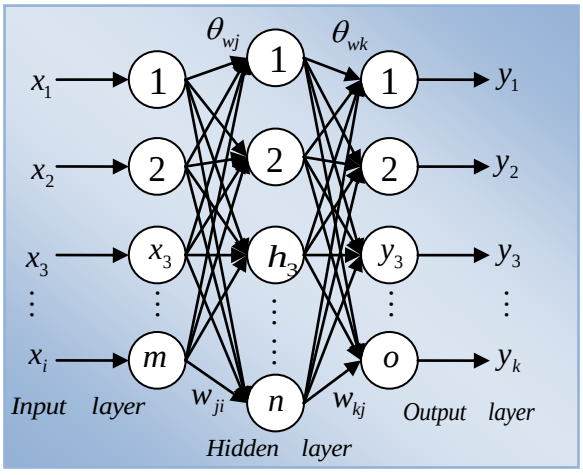

圖 6. 倒傳遞類神經網路架構

圖 6. 中，輸入層的節點數量為所需的音效特徵空間的維度，輸出層的節點數量為分類音效類型，隱藏層節點則可依照需求設置。

丙. 支援向量機 (SVM)

支援向量機(Support Vector Machine，SVM)是 Vapnik 等人以統計學習法則中的結構風險最小化(Structural Risk Minimization，SRM)[34][36]為基礎所發展的機器學習演算法，支援向量機可運算兩個至多個不同類別的線性樣本空間之最佳分割超平面(Optimal Separate Hyperplane)，以取得樣本分類，對於線性不可分割的非線性分類問題，其可將低維度的樣本向量轉換到更高維度的特徵空間中進行線性分割(圖 7.)。以一組二元訓練樣本集合 S 為例，如公式(41)。

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\},$$
$$x_i \in R^n, y_i \in \{+1, -1\}, i = 1\cdots m \quad (41)$$

其中 m 為樣本個數，n 為樣本維度，存在一超平面(Hyperplane)可分割此二類樣本如公式(42)：

$$f(x) = sign(w\cdot x + b) \quad (42)$$

$$f_D(x_i) = w\cdot x_i + b \begin{cases} \geq +1, & if \quad y_i = +1, \\ \leq -1, & if \quad y_i = -1, \end{cases} i = 1,\ldots,m \quad (43)$$

公式(43)中，$w \in R^n$ 且 $b \in R$；如果存在決策函數(Decision Function)的超平面參數(w,b)，使得 $\forall x_i$ 滿足公式(43)，則此集合 S 為線性可分割。如果集合 S 無法被線性分割，則可透過非線性核心函數(Non-Linear Kernal Function)來對映更高維度的特徵空間進行分割，如公式(44)。本文也採用此分類器進行音效分類與分段實驗。

$$x_i \cdot x_j \rightarrow \phi(x_i)\cdot\phi(x_j) = K(x_i, x_j) \quad (44)$$

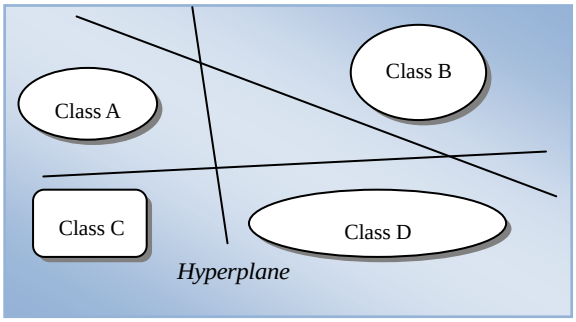

圖 7. 支援向量機分類圖

(二)循序偵測斷點法

循序偵測斷點法相似於鏡頭切換的判斷方法[10]，是以單一框架在音訊切片上循序移動的方式，比較兩個相鄰框架的各種特徵值差異，並藉由臨界值的設定來判斷音效斷點的位置。但是當實際音效斷點的位置不在兩音效交界時，則從音訊所包含的音效類型判定其斷點位置。

在判斷音效特徵向量的差異性上，本文使用歐基理得距離來計算，但由於各種特徵值分辨音效的能力各異，因此在計算出特徵值間的歐基理得距離之後，我們再乘上一加權值及加總，並進行正規化，藉以比較其差異性。

$$D_t = \sqrt{\sum_{i=1}^{19} w_i (f_{ti} - f_{(t+1)i})^2} \quad (45)$$

$$Df_t = \frac{1}{n+1} \left| \sum_{i=k-n}^{k} D_{ti} - \sum_{i=k}^{k+n} D_{ti} \right| \quad (46)$$

公式(45)中，*fti* 為第 *t* 個框架的第 *i* 個特徵向量，*wi* 為該項加權值。在公式(46)，*n* 為音效切片單位的框架數，*Dft* 為時間 *k* 前後框架數 *n* 的總平均值差，當在時間 *k* 的 *Dft* 值大於臨界值，則表示在該時間點 *k* 的特徵向量發生躍變，並判別為音訊斷點。(如圖 8.)

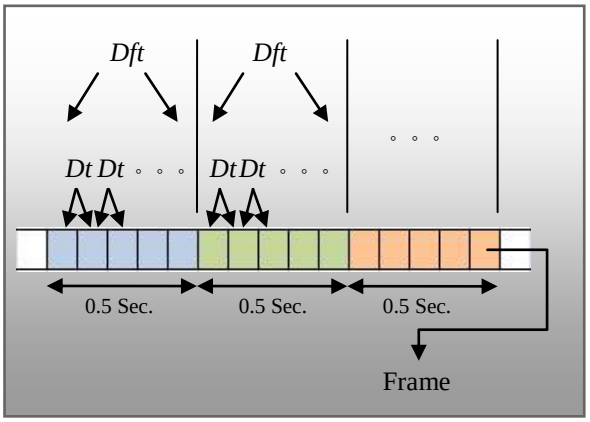

圖 8. 循序偵測斷點圖示

五、實驗

(一)實驗環境

本文所使用的電影音效分類樣本，主要是取自多部 DVD 格式的商業電影。在電腦硬體上，採用 Intel Pentium 4 3.0G CPU 處理器，記憶體為 DDR2 533-RAM 總容量 1G。作業系統為 Microsoft Windows XP Professional SP2。在系統開發上，使用 Borland C++ Builder 6 語言來撰寫特徵值擷取程式與分段系統。

(二)實驗樣本與評估方式

本論文的實驗樣本取自近年商業電影作品，如表 1.所列，其中類型包含科幻片、動作片以及劇情片。取每一音訊樣本為 10 秒，音效切片以 0.5 秒為一個單位，則電影音效取樣統計表，如表 2.所示。在實驗中有許多因素會影響到音效分段的正確性，包含了各組不同音效特徵值臨界值常數的選定和該特徵值的權重，以及分段系統對音訊資料取樣框架大小。為了呈現實驗結果的優劣狀況，我們採用準確率(Precision Rate，公式 47)及回覆率(Recall Rate，公式 48)作為斷點偵測評估的判斷方式。

表 1. 取樣電影列表

取樣電影列表		
動作片	**科幻片**	**劇情片**
血鑽石	決戰異世界	香水
生死格鬥	紫光任務	頂尖對決
墨攻	魔幻至尊	死亡筆記本 2
滿城盡帶黃金甲	毫髮人的冒險	當幸福來敲門

表 2. 電影音效取樣統計表

電影類型	樣本數量	切片數量	人工斷點數量	系統斷點數量
動作片	40	800	40	51
科幻片	40	800	40	46
劇情片	40	800	40	52
合計	120	2400	120	149

$$\text{Precision Rate} = \frac{\text{Number of Relevent Retrieved System}}{\text{Number of Retrieved System}} \quad (47)$$

$$\text{Recall Rate} = \frac{\text{Number of Relevent Retrieved System}}{\text{Number of Retrieved Manual}} \quad (48)$$

公式(47)為系統所辨識且符合人工判別的斷點數目被除以系統所識別的斷點數目；公式(48)為系統所辨識且符合人工判別的斷點數目被除以人工所識別的斷點數目。

(三)實驗結果

甲. 循序偵測斷點法

本節主要對循序偵測斷點法進行實驗，分段系統依設定的音效切片長度，取 0.5 秒內所含的框架數，然後計算每個框架內涵的特徵值，以框架與框架之間的特徵值差異為 *Dt*，求得 0.5 秒內特徵向量的總平均值差為 *Dft*，經比對臨界值可得音效斷點的位置，此為系統分析斷點的位置。而人工斷點的位置則經由非特定人選三人，對同一樣本，以人耳去判定不同音效類別之間的斷點位置並記錄。完成斷點分段後，斷點前後的分段內容，可藉由音效特徵資料庫的比對，從而分辨音效的類型，進而建立電影音效資料庫的摘要及索引。

表 3.為各類型音效變換分段評估結果，針對各個不同的音訊樣本作分類的混淆矩陣(Confusion matrix)表示，表中各行表示斷點後的音訊類型，各列表示斷點前的音訊類型。經由此架構做的分段系統，回覆率為 74.5%，而準確率為 68.25%。圖 9.為電影『生死格鬥』中，某一片段音效的系統斷點分析結果，圖 10.為其人工斷點位置及系統分析斷

點位置示意圖。

表 3. 各類型音效變換分段評估結果

Recall Rate	Music	Speech	Sound	Silence
Music	81%	30%	77%	77%
Speech	60%	63%	100%	100%
Sound	30%	60%	81%	43%
Silence	100%	53%	77%	73%
Precision Rate	Music	Speech	Sound	Silence
Music	75%	30%	60%	60%
Speech	43%	60%	100%	100%
Sound	27%	60%	73%	35%
Silence	77%	37%	63%	65%

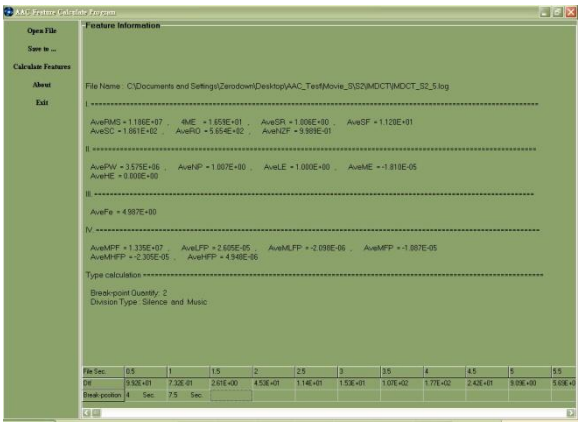

圖 9. 系統斷點分析結果

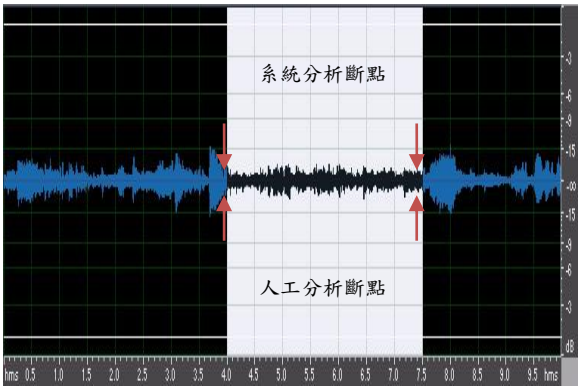

圖 10. 人工及系統分析斷點位置示意圖

乙. 分類器偵測斷點法

在相同的 MPEG-2 AAC 音效檔樣本下，我們透過類神經網路(BPN)、高斯混合模型(GMM)、向量支援機(SVM)為分類器，來對音效分類系統作訓練和測試。從實驗結果顯示，這三種分類器的正確率均未優於循序偵測斷點法。表 4. 為分類器實驗結果的整體統計。

表 4. 分類器正確率結果

Classification / Sorter	Music	Speech	Sound	Silence
GMM	58%	64%	47%	49%
BPN	67%	63%	42%	57%
SVM	70%	65%	55%	61%

六、結論

在本篇論文，我們提出一種針對 MPEG-2 AAC 音效格式的電影音效自動分段技術。我們利用擷取自 MPEG-2 AAC 解壓縮過程中的 MDCT 係數來計算出 MPEG-7 規範的音效描述子，例如：靜音比、4ME、頻譜及能量特性等等，用以分辨混合語音、靜音、音樂、音效四種音訊類型的音訊斷點。且利用經驗法則以及 GMM、BPN、SVM 等分類器，來偵測不同音訊斷點的發生位置，以自動識別電影音訊中的音效發生位置，作為進一步分析音訊內涵的前置技術。

在 MPEG 所規範的音訊格式中，我們未來亦將針對廣播、網際網路和手持裝置等日常應用的音訊技術進行相關的內涵分析研究。

誌謝

本篇論文得以完成，首先要感謝 劉志俊教授，劉老師對於學生研究方向的支持及指導，讓本篇論文的疑難之處得以迎刃而解。此外，系上諸位教授以及外系教授的指導和協助，不論在學業或是研究上都給予學生莫大的幫助。感謝我親愛的女友、同學以及實驗室的學弟們，在我的求學生涯中，不論在人生的議題上，或是研究瓶頸的諮詢上，都給予我相當多的建議及鼓勵。最後，我要向父母親獻上最誠摯的感激，並將榮耀歸於父母。

七、參考文獻

[1] 吳智偉、劉志俊,“支援 MPEG-7 之電影 AC-3 環場音效內涵描述工具,” 二○○五數位生活與網際網路科技研討會, 2005.

[2] 范世鎮、劉志俊,“利用特寫鏡頭偵測與主角辨識技術來自動建立電影摘要,” 第二屆數位典藏技術研討會, 2003.

[3] 陳信修、劉志俊,“一種利用特寫鏡頭對數位電影資料進行自動化摘要合成之技術,”第一屆數位典藏技術研討會, 2002.

[4] 葉億真、劉志俊,“音效資料的內涵式分類及其在電影資料庫的應用,” 第二屆數位典藏技術研討會, 2003.

[5] 鄭煒平、劉志俊,“網際網路電影資料庫之音效自動分段索引系統,” 第六屆網際網路應用與發展研討會, 2005.

[6] ATSC A/52, Digital Audio Compression (AC-3)Standard, United States Advanced Television Systems Committee.

[7] B.S. Manjunath, P. Salembier, and T. Sikora, Eds.," Introduction to MPEG-7:Multimedia Content Description Interface," John Wiley & Sons, 2002.

[8] B. Logan, "Mel frequency cepstral coefficients for music modeling," In Proc. Int. Symp. Music Information Retrieval (ISMIR), 2000.

[9] C. Panagiotakis, G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," IEEE Transactions on Multimedia, vol.7, no. 1, Feb. 2005.

[10] Changsheng Xu, Namunu C. Maddage, and Xi Shao, "Automatic music classification and summarization," IEEE Transactions on Multimedia, vol. 13, no. 3, pp. 441- 450, May. 2005.

[11] Dongbing Gu and Huosheng Hu, "Wavelet neural network based predictive control for mobile robots," In IEEE International Conference, Systems、 Man and Cybernetics, Oct. 2000.

[12] Duncan J.S. and Birkholzer T., "Edge reinforcement using parametrized relaxation labeling," In IEEE Computer Society Conference ,Computer Vision and Pattern Recognition, 1989. Proceedings CVPR '89., June 1989.

[13] Erling Wold et al, "Content-based classification, search, and retrieval of audio", IEEE Multimedia, pp.27-36, Fall 1996.

[14] F. Pachet and D. Cazaly, "A classification of musical genre," In Proc. RIAO Content-Based Multimedia Information Access Conf., Paris, France, Mar. 2000.

[15] Guojun Lu and Templar Hankinson, "A Teehnique towards Automatic Audio Classification and Retrieval," In Proceedings of ICSP, Australia, 1998.

[16] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, July 2002.

[17] G. Tzanetakis and P. Cook, "A framework for audio analysis based on classification and temporal segmentation," In Proc. 25th Euromicro Conf. Workshop on Music Technology and Audio Processing, 1999.

[18] Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora, "MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval", U.K.:Wiley, 2006.

[19] Hong Kook Kim et al, "On approximating line spectral frequencies to LPC cepstral coefficients," IEEE Transactions on Speech and Audio Processing, AT&T Bell Labs., Florham Park, NJ, USA, March 2000.

[20] Hecht-Nielsen, R., "Theory of the Back Propagation Neural Network," Proceeding of International Joint Conference on Neural Networks, IEEE, Vol. 1, pp. 593-605, 1989.

[21] ISO/IEC 15938-4:/FPDAM, "Information Technology — Multimedia Content Description Interface — Part 4: Audio," ISO/IEC, 2002.

[22] ITU-R BS. 775-1, "Multichannel Stereophonic Sound System With and Without Accompanying Picture," International Telecommunication Union, Geneva, Switzerland, 1992-1994.

[23] IEC CDV 61937-5: Digital audio - Interface for non-linear pcm encoded audio bitstreams applying IEC 60958 - Part 5: Non-linear PCM bitstreams according to the DTS (Digital Theater Systems) format(s) [IEC 100/974/CDV]

[24] ITU-R Recommendation BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," Geneva, Switzerland, 1994.

[25] Jean Baudrillard, "Simulacres et simulations", Paris, Galilée, 1981.

[26] J. Saunders, "Real-time discrimination of broadcast speech/music," In Proc. IEEE ICASSP, 1996.

[27] M. Bosi et al., "ISO/IEC MPEG-2 advanced audio coding," J. Audio Eng. Soc., vol. 45, no. 10, pp. 791–811, Oct. 1997.

[28] MPEG Requirements Group, "Information technology - Multimedia Content Description Interface - Part2 ： Description Definition Language," ISO/IEC JTC1/SC29/WG11 N4002, Singapore, Mar. 2001.

[29] Manian V., Vasquez R. and Katiyar, P.,"Texture classification using logical operators," IEEE Transactions on Image Processing, Oct. 2000.

[30] N. V. Pate1 and I. K. Sethi, "Auido Characterization for Video Indexing", SPIE Vol. 2670, pp. 370-384.

[31] Roben Gonzalez and Kathy Melih, "Content based retrieval of audio", Proceedings of Australian Telecommunication Networks & Applications Conference, pp. 357-362, Melbourne, 3-6 December.

[32] Reynolds D. A. and Rose R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. Speech Audio Process. 3 (1995), 72–83.

[33] Sung A.H. and Jun Lin, "Performance comparison of neural network models for engineering problems," In IEEE International Conference, Systems、Man and Cybernetics, 1997. 'Computational Cybernetics and

Simulation', Oct. 1997.

[34] Shawe-Tayloe J., Bartlett P.L., Williamson R.C. and Anthony M., “Structural Risk Minimization.,” IEEE Transactions on Information Theory, 1998.

[35] T. Zhang and J. Kuo, “Audio content analysis for on-line audiovisual data segmentation and classification,” IEEE Trans. Speech Audio Process., vol. 9, no. 3, pp. 441–457, May. 2001.

[36] Vapnik V., “Statistical Learning Theory,” Springer, N.Y., 1998.