國立政治大學資訊科學系

Department of Computer Science
National Chengchi University

碩士論文

Master's Thesis

串流式音訊分類於智慧家庭之應用

Streaming Audio Classification for Smart Home Environments

研 究 生：溫景堯

指導教授：廖文宏

中華民國九十九年七月

July 2010

串流式音訊分類於智慧家庭之應用

Streaming Audio Classification for Smart Home Environments

研 究 生：溫景堯　　Student：Jing-Yao Wen

指導教授：廖文宏　　Advisor：Wen-Hung Liao

國立政治大學

資訊科學系

碩士論文

A Thesis

submitted to Department of Computer Science

National Chengchi University

in partial fulfillment of the Requirements

for the degree of

Master

in

Computer Science

中華民國九十九年七月

July 2010

致謝

謹以此論文獻予雙親與姊妹。兩年於政大學習的日子，得以完成這篇論文。不僅於求學過程中的支持，還有這二十多年來無盡的關心與照顧，讓我毫無後顧之憂的完成這篇研究，願能將此成就和喜悅與你們分享。

感謝指導教授 廖文宏老師，於這兩年間自老師所得不僅止於專業學識，學習如何完成研究，從問題的定義與切入，至思考及其研究方法，由此過程中學習與成長，學習所得都是一生受用不盡的財富。此外，感謝兩位口試委員 李蔡彥老師與 楊傳凱老師，給予我許多論文上的寶貴意見，得以讓此論文更為充實、完整。

感謝 VIPL 的每一位成員，感謝靈威、岡隆、榮聖三位學長給予太多太多的幫忙與指導，在你們身上所學遠遠超過了這份文憑的價值。感謝仁和、梃榮、慧文、立強、柏穎、正和、建堡、政明、浩瑋在這段求學生涯中的一路相伴，走過那些辛苦的白晝與黑夜。感謝政大資科的同學們，一起度過這段努力、汗水與歡笑交錯的日子。感謝高中與大學的同學們，給予學業與生活中的許多幫助。

最後，感謝恩儷在這些日子的陪伴，支持我走過最艱辛的時候，鼓勵我追求更高的目標，陪伴我完成這個人生階段。

景堯 於 2010 年秋

串流式音訊分類於智慧家庭之應用

摘要

聽覺與視覺同為人類最重要的感官。計算式聽覺場景分析(Computation Auditory Scene Analysis, CASA)透過聽覺心理學中對於人耳特性與心理感知的關連性，定義了一個可能的方向，讓電腦聽覺更為貼近人類感知。本研究目的在於應用聽覺心理學之原則，以影像處理與圖型辨識技術，設計音訊增益、切割、描述等對應之處理，透過相似度計算方式實現智慧家庭之環境中的即時音訊分類。

本研究分為三部分，第一部分為音訊處理，將環境中的聲音轉換成電腦可處理與強化之訊號；第二部分透過 CASA 原則設計影像處理，以冀於影像上達成音訊處理之結果，並以影像特徵加以描述音訊事件；第三部分定義影像特徵之距離，以 K 個最近鄰點(K-Nearest Neighbor, KNN)技術針對智慧家庭環境常見之音訊事件，實現即時辨識與分類。實驗結果顯示本論文所提出的音訊分類方法有著不錯的效果，對八種家庭環境常見的聲音辨識正確率可達 80-90%，而在雜訊或其他聲音干擾的情況下，辨識結果也維持在 70%左右。

關鍵字：計算式聽覺場景分析、串流式音訊分類

Streaming Audio Classification for Smart Home Environments

Abstract

Human receive sounds such as language and music through audition. Therefore, audition and vision are viewed as the two most important aspects of human perception. Computational auditory scene analysis (CASA) defined a possible direction to close the gap between computerized audition and human perception using the correlation between features of ears and mental perception in psychology of hearing. In this research, we develop and integrate methods for real-time streaming audio classification based on the principles of psychology of hearing as well as techniques in pattern recognition.

There are three major parts in this research. The first is audio processing, translating sounds into information that can be enhanced by computers; the second part uses the principles of CASA to design a framework for audio signal description and event detection by means of computer vision and image processing techniques; the third part defines the distance of image feature vectors and uses K-Nearest Neighbor (KNN) classifier to accomplish audio recognition and classification in real-time. Experimental results show that the proposed approach is quite effective, achieving an overall recognition rate of 80-90% for 8 types of audio input. The performance degrades only slightly in the presence of noise and other interferences.

Key words: computational auditory scene analysis, streaming audio classification.

目錄

圖目錄

表目錄

第一章

緒論

1.1 研究背景

聽覺於人類感官中與視覺同為最重要的認知感官，透過聽覺接收發生的訊息、音樂旋律、語言傳達等眾多資訊，隨著電腦硬體與音訊處理的快速發展，電腦聽覺也開始了長足的進步，讓電腦可以如同人類使用聽覺對週遭的環境與事件加以分析處理。

計算式聽覺場景分析(Computational Auditory Scene Analysis, CASA)一直以來都是聲音處理的一大重要課題，自 Bregman 於 1990 年提出 Auditory Scene Analysis[1]的架構以來，與此相關之研究持續不斷，但大多數之研究均注重於一般人類語音之辨識與分析，而隨著近年數位化影音多媒體資料大量增長，數位內容的分析與了解成為刻不容緩的熱門研究課題，再加上電腦硬體設備的進步，讓電腦理解聲音其中隱藏的內涵不再是個遙不可及的夢想。而居家安全、老年照護、智慧家庭等應用領域的萌芽，將計算聽覺場景分析技術應用於相關方面的研究亦開始受到重視。

所謂計算聽覺場景分析[1][3]的目標便是將單聲道的聲音(此聲音可能為由單一音源或是同時由多個音源(source)所產生)，經由電腦取樣轉換成電腦可讀的格式，透過我們對聲音於物理與心理上特性的了解，加以定義電腦對聲音訊號的處理[4][5]，包括音訊波形的轉換(transformation)與處理、增益與降噪、音訊特徵擷取、機器學習與分類等方式，分離不同音源所產生之聲音並加以擷取音訊事件，用於更進一步的計算與處理。

本實驗室於[6]中所提出階層式分類法(如圖 1-1)將聲音區分為人聲與非人聲，並將非人聲的部分以環境音(environment sound)加以概括，本論文將著重於此環境音的部分加以分析研究，過去的研究著重於非即時之人聲中的語音和非語音的部分，而本研究將重點放在即時發生之環境聲，將環境聲更細分為雜訊、環境聲與音訊事件，針對家庭環境可能發生之情境雜訊與環境聲，對於常見於家庭之音訊事件加以分類。

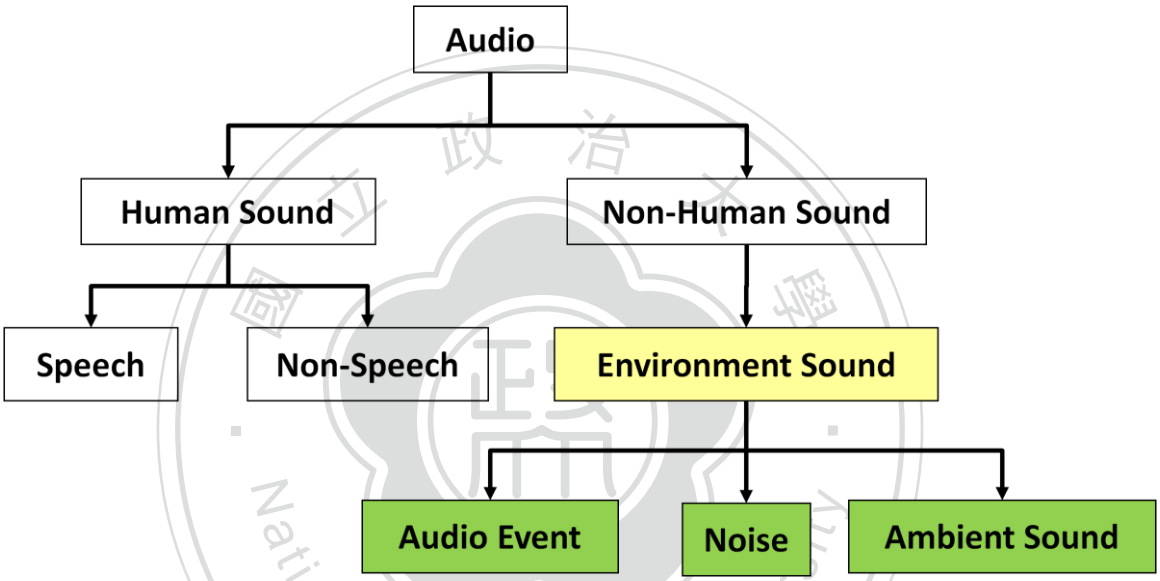


圖 1-1. 階層式分類

本論文提出的音訊處理架構，最大的特點是以影像分析與圖型識別的技術為基礎，進行各類音訊的解析。音訊處理的議題乍看之下與電腦視覺(Computer Vision)全然無關，但當我們將一維的聲音訊號波形轉換成二維的甚至三維的頻譜影像後，聲音訊息便以影像呈現另一不同面貌[7]，如圖 1-2 所示。本研究將於影像上分析音訊事件。將聲音訊號經過短時傅利葉轉換(Short-Time Fourier Transform, STFT) [8]得其時間-頻率頻譜圖，藉由我們對音訊事件特性的了解，定義其特性在時間-頻率頻譜圖上的影像特徵，利用影像處理的方式完成增益音訊事件訊號強度、降噪、音訊事件偵測等處理，並以影像特徵(image feature)描述偵測所得音訊事件，例如事件發生之頻率、事件內涵呈現紋理(texture)，最後將擷取所得特徵利用相似度搜尋加以分類。

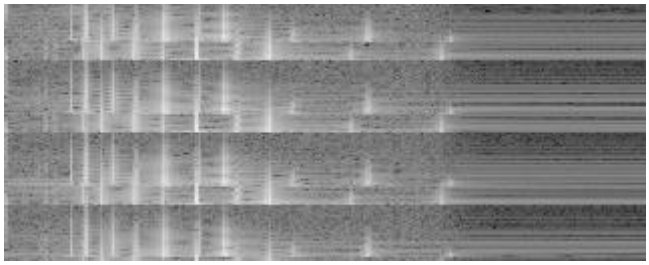


圖 1-2. 時間-頻率頻譜圖(縱軸為時間、橫軸為頻率)

1.2 研究目的

本研究主要包含三個階段：一是偵測平日居家環境中常見之聲音，例如電話鈴聲、門鈴聲、水壺汽笛、鬧鐘等；二是將所偵測到之音訊事件轉換成時間-頻率頻譜圖(time-frequency spectrogram)，利用影像處理(image processing)技術對音訊事件(auditory event)擷取其特徵(feature)用以描述;三則透過相似度搜尋(similarity search)方式對音訊事件特徵加以分析並完成其分類(classification)，透過以上方式實現串流式音訊分類於智慧家庭。

我們將於第二章介紹關於計算式聽覺場景分析、時間-頻率頻譜圖分析的相關研究文獻，第三章提出本論文的研究方法，第四章說明實驗結果，並於第五章闡述本研究之結論與未來的研究規劃。

第二章

相關研究

近來由於技術的大幅進步及網路頻寬的增加，多媒體的發展也越來越蓬勃，不論是視訊、音樂、影像等資訊，在網路上都能廣泛地流通與利用，也因此多媒體資料的處理及資料庫的管理，逐漸成為人們重視的課題。

計算聽覺場景分析的目標為分析聲音環境並加以辨識不同聲音事件[1]，然而這些聲音可能是個別單一出現，也有可能是同時出現。如此複雜的聲音背景，便增添了處理的困難度。處理這類聲音時，如果使用上述的方法，便需要先把音源分開，再把聲音檔分割成某些長度的聲音片段。在分割的過程中，還牽涉到區分語音、音樂或其他聲音的相關技術。

聲音分割的做法一般是透過信號的聲學分析[9] [10]，並找出聲音的轉變點，這裡的轉變點指的是聲音特徵向量突然改變的地方。利用轉變點把聲音信號分成若干區段，這些區段就可以當作個別的聲音來處理。於[9]中，Hu 等學者提出將音強變化視為偵測起點，透過不同程度的訊號平滑化(smoothing)，降低音量波動(intensity fluctuation)並強化音訊事件的起始點(onset)，最後以迭代(iterative)方式整合 Multiscale 的事件起始點以偵測音訊事件。S. H. Srinivasan 提出將音訊事件以區塊(blob)加以描述的概念，基於聽覺場景分析的原則，無關的音訊事件其起始點同時發生的機率低、同一音訊事件其變化程度為趨緩以及音訊事件因其和諧性所呈現樣式(pattern)，將頻率波峰(spectral peak)的連續延展定義為岸線(strand)，而區塊便是擁有同起始點的岸線其集合，以音訊事件中的頻率

內容、和諧性、能量變化、頻率變化四個主要特徵描述音訊區塊，最後計算其相似度加以分類。

時間-頻率頻譜圖的分析與觀察是常見於聲音或語音辨識的方法之一，Yan Ke 等人於[7]中將過去我們所習慣的觀察以影像處理的方式加以分析音訊事件於頻譜圖上的呈現，首先將音訊資料轉換成頻譜圖，計算頻譜圖上的區域描述子(local descriptor)，利用 RANSAC 校正每個音訊事件候選(candidate)，最後計算每個音訊事件候選間的相關度用以校正事件的配對是否正確。

Valerie Pierson 等人提出對於時間-頻率頻譜圖邊界偵測(boundary detection)的技術[10]，於研究中，在直角座標系(cartesian coordinates)上以區域正切角度(local tangent angle)與半徑(radius)偵測邊界，並對四個主要的邊界偵測技術加以比較其校能與速度。

於[12]中 Ruohua Zhou 等人提出基於共振器時間-頻率頻譜圖的音樂事件起始點偵測，於研究中將音強變化劇烈的起始點定義為硬性起始點(hard onset)，而變化緩和者則定義為軟性起始點(soft onset)，首先將聲音訊號轉換成時間-頻率頻譜圖，利用基於能量變化演算法(energy-based algorithm)可有效的處理硬性起始點偵測，而未能有效偵測的軟性起始點則輔以基於音調變化演算法(pitch-based algorithm)改善其偵測，利用不同轉換方式增強不同時間-頻率頻譜圖上所呈現訊號強度，例如以和諧性音源法則(harmonic grouping principle)強化音調、利用低通濾波器(low-pass filter)降噪，再以聽覺場景分析原則加以整合不同頻譜圖上偵測所得事件。

音訊分類大致可分為以下兩個部分[13][14]。第一部分是透過樣本訓練來分類。選擇一些表達某類特性的聲音樣本來訓練系統，建立這類聲音的模型。系統對每一個樣本找出其特徵向量，並計算這些訓練樣本的平均向量和共變異矩陣用以建構出表達這類聲

音的模型。第二部分是利用聽覺特徵進行檢索[15][16]。基於人類聽覺感知的特性，如基頻、振幅、音高等，找出特徵向量並用來區分不同聲音。[15]中 Zhu Liu 等人將音訊特徵階層式分類為三類，低階的基礎聲學特徵(acoustic feature)，又以時域與頻率域將其加以區分，中階的特定場景(scene)特徵，例如將場景限定於球場，以及高階的場景前設特徵(prior)。於[16]中 Silvia Allegro 等人提出將基礎聲學特徵更細分為三類，振幅變化、頻率特性及和諧性(harmonicity)。

於本研究中，我們將過去的研究所提出之概念加以整合，透過聽覺心理學的原則與人類感受器官的特性，試圖建立聽覺與視覺之間的轉換關係，應用電腦視覺與影像分析的方法，將原本的音訊分類問題轉化為圖型辨識的範疇。

圖 2-1 展示本研究所提出以影像處理技術為基礎的串流式音訊分類系統架構。透過麥克風接收測試環境中之聲音，經過取樣(sampling)與量化(quantization)，對此數位訊號做短時傅利葉轉換(Short-Time Fourier Transform, STFT)，取得其二維時間-頻率頻譜圖。利用雙向濾波器(bilateral filter)對此時間-頻率頻譜圖增強其音訊事件之結構，以助益於音訊起始點偵測(audio onset detection)，實現時間-頻率頻譜圖之影像分割(segmentation)，取得音訊事件影像(auditory event image)。利用閾值計算與設定(thresholding)的方法取得二值化影像(binary Image)做為輸入，經由區塊偵測(blob detection)演算法將音訊事件影像中所偵測之區塊視為音訊區塊(auditory blob)，擷取其區域二元化圖型(local binary patterns)[17]特徵用以描述此音訊區塊中所呈現紋理與輪廓，利用此編碼分布直方圖作為特徵向量，透過預先定義的距離(distance metric)計算以實現音訊分類。

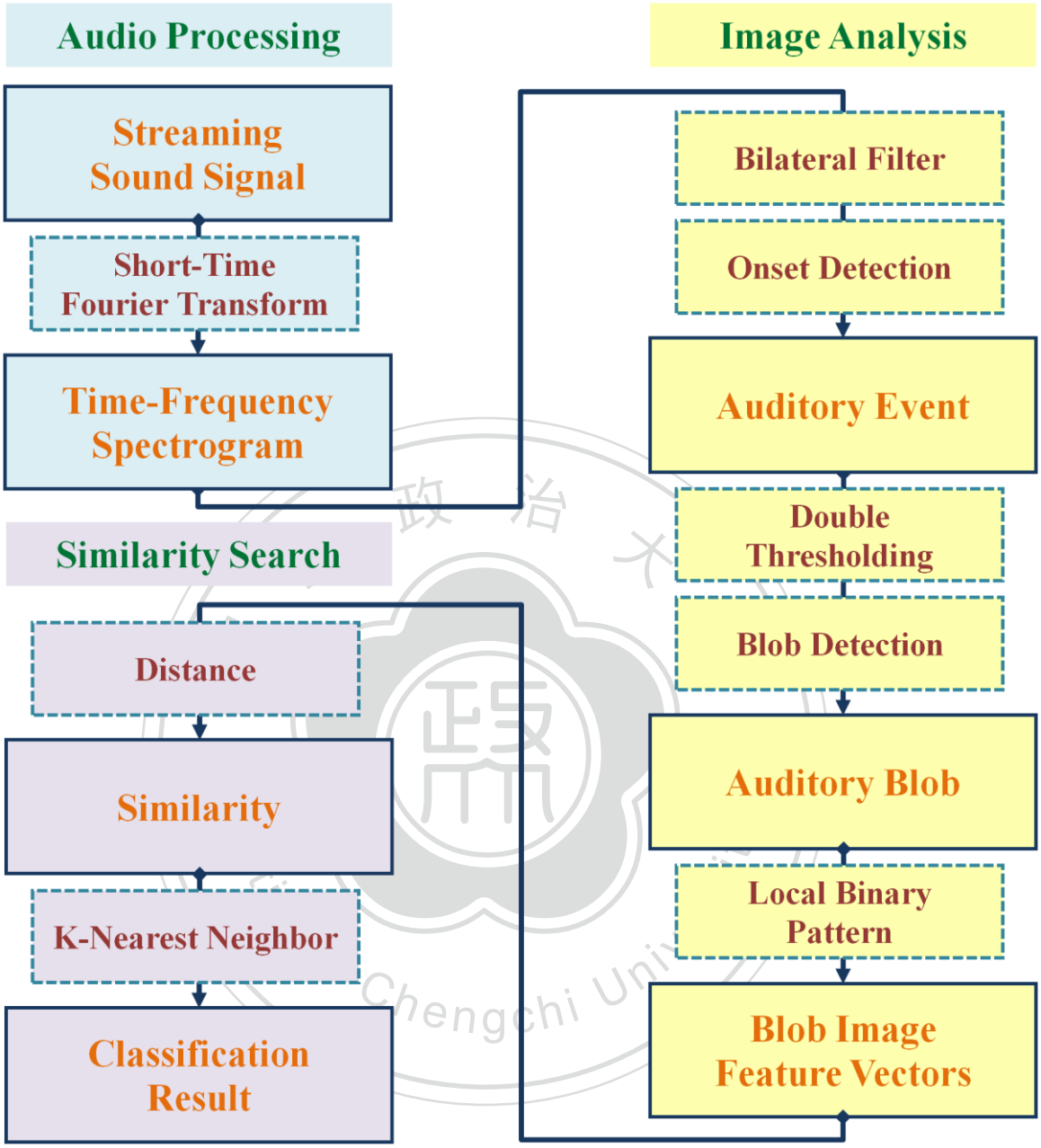

Audio Processing
Streaming Sound Signal
Short-Time Fourier Transform
Time-Frequency Spectrogram
Similarity Search
Distance
Similarity
K-Nearest Neighbor
Classification Result
Image Analysis
Bilateral Filter
Onset Detection
Auditory Event
Double Thresholding
Blob Detection
Auditory Blob
Local Binary Pattern
Blob Image Feature Vectors

圖 2-1. 系統架構圖

第三章

研究方法

本章節將介紹我們於本次研究中所提出基於影像處理之音訊分類流程，包含音訊的輸入、轉換、影像處理、事件偵測、特徵描述以及分類的方法。

3.1 音訊處理 (Audio Processing)

聲音訊號(Audio Signals)簡稱音訊，泛指由人耳聽到的各種聲音的訊號。當發音體產生震動對空氣產生壓縮與伸張的效果，形成聲波，當此聲波傳遞到人耳，耳膜會感覺到一伸一壓的壓力訊號，內耳神經再將此訊號傳遞到大腦，並由大腦解析與判讀，來分辨此訊號的意義。當聲波經過量化與轉換，變成電腦可讀之格式，對此數位訊號加以處理與分析

於本研究中，我們將居家環境中的音訊透過收音裝置，取得其數位訊號，基於聽覺心理學的原則，如聲音對於心理與物理之影響、音訊感知等訊息對此訊號加以定義，並以短時傅利葉轉換將其呈現於時間-頻率頻譜圖之上，以供下一階段處理之使用。

3.1.1 聽覺心理學 (Psychoacoustic)

因為人類耳朵中的聲音受器，讓我們對於聲音的感受產生某些特性。基底膜不同位置含有聽神經受器，不同頻率聲波又會使基底膜不同位置達到最大振幅，因此形成基底

膜上對不同頻率具不同敏感程度的聽神經分布。而因為基底膜的質地因不同位置而有不同變化，我們對不同頻率的音強感受性也有所不同。

雖然視覺是對光波做反應，而聽覺是對空氣分子所產生的壓力波有所反應，各有不同的受器與不同的大腦皮質對應區。但這兩個感官功能仍有許多相似的運作機制存在：

1. 透過傅利葉轉換於頻率域分析訊息

   視覺系統是將空間域訊息轉成空間-頻率訊息處理；聽覺系統則是將時域訊息轉換成時間-頻率的訊息處理。兩者都各有頻率敏感細胞(frequency tuning cells)或管道(channels)。這使得傅利葉轉換可以同時用在視覺與聽覺，並獲得類似的結果。

2. 選擇性的偏好反應(response selectivity)

   兩種感受器官各有對於不同改變之反應的細胞，讓我們對於不同的變化產生選擇性的反應。

3. 左右訊息交叉傳送到大腦的運作方式

   視覺在視交叉之後,是左視野訊息傳到右腦,右視野訊息傳到左腦,幾乎是 50%對稱交叉。然而聽覺只有部份左右訊息交換。

基於視覺與聽覺間的相似特性與聽覺心理學之原則，於本研究將嘗試將音訊透過短時傅利葉轉換成時間-頻率頻譜圖，對此影像中各種音訊所呈現之表徵加以分析討論。

3.1.2 音訊輸入 (Input Signal)

聲音代表了空氣的密度隨時間的變化，基本上是一個連續的函數，但是若要將此訊號儲存在電腦裡，就必須先將此訊號數位化。一般而言，當我們將聲音儲存到電腦時，有下列幾個參數需要考慮：

1. 取樣頻率 (sampling rate)

   每秒鐘所取得的聲音資料點數，以 Hertz(Hz)為單位。點數越高，聲音品質越好，但是資料量越大。

2. 取樣解析度(bit resolution)

   每個聲音資料點所用的位元數。

3. 聲道

   一般只分單聲道(mono)或立體聲(stereo)，立體音即是雙聲道。

3.1.3 短時傅利葉轉換 (Short-Time Fourier Transform)

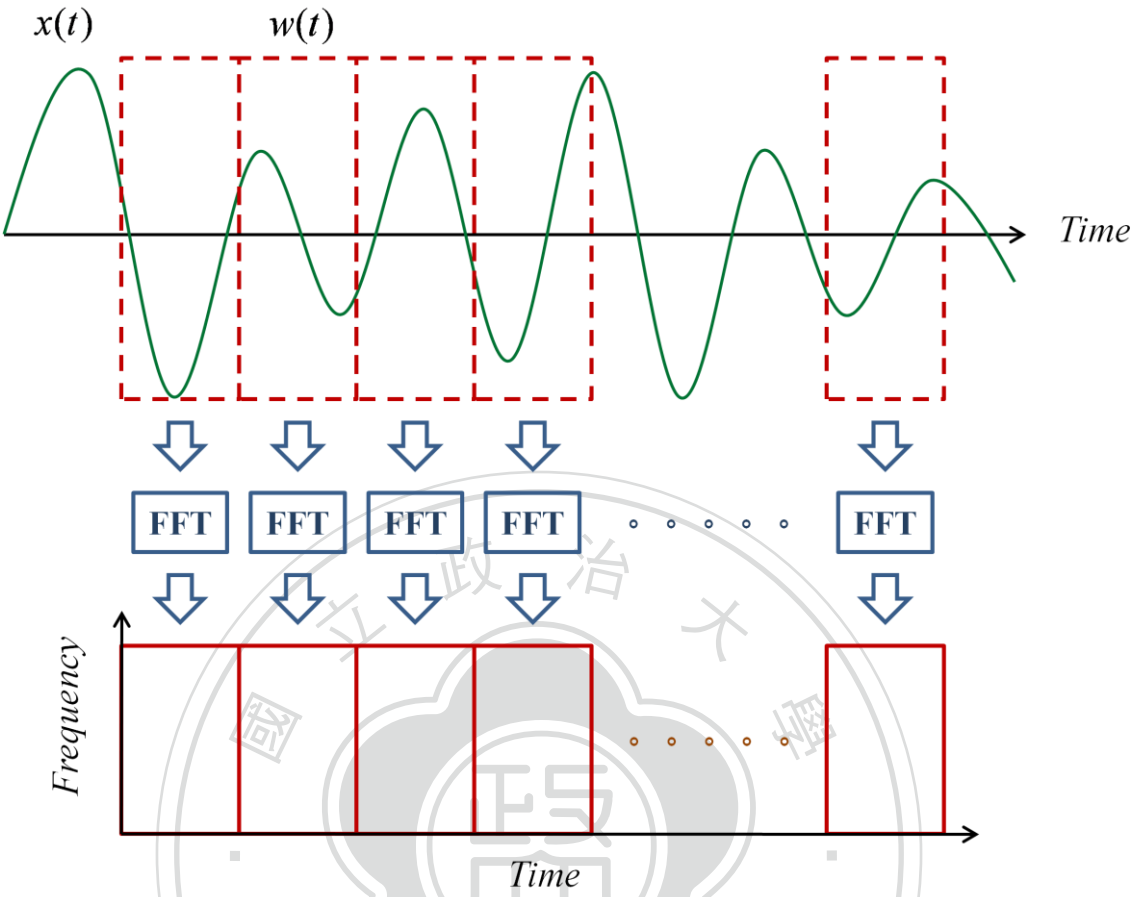
於實驗中量測所得訊號大多是以時間作為基礎的時域(time-domain)訊號，觀察時域訊號可以了解訊號隨時間變化的程度，但如欲得知此訊號在頻率域(frequency-domain)中之訊息，便可經由離散傅利葉轉換(discrete Fourier transform)後得知，算式如下：

$$X(k)=\sum_{n=0}^{N-1} x(n)e^{-i\frac{2\pi k}{N}n} \tag{3.1}$$

其中 $x(n)$ 為待分析之訊號，N 為分析訊號之長度。

傅利葉轉換雖具有良好的頻率域分析能力，但失去時域上的解析能力。Dennis Gabor 為了修正這個缺點，於 1946 年提出了視窗(window)的概念[18]，也就是將訊號切割成多個訊窗，先將訊號跟時間軸上不斷平移的窗型函數(window function)相乘，藉由窗型函數的平移，取出特定時間的訊號再做傅利葉轉換以找出其頻率分布，算式如下與圖 3-1：

$$S(\omega,\tau) = \int x(t)w(t-\tau)e^{-i\omega t}dt \tag{3.2}$$



$x(t)$：為待分析訊號

$w(t)$：為窗型函數

圖 3-1. 短時傅利葉轉換示意圖

由上述可知，窗型函數 $w(t)$ 的視窗長度(window length)選取決定了頻譜圖的解析度，較長的視窗長度可以得到較高的頻率域解析度，但時域的解析度就變差；反之選擇較短的視窗長度，雖有較高的時域解析度，卻犧牲了頻率域上的解析度，而解析度的高低並不會隨著時間或是頻率上的改變而有所變化。於解析度較低之頻譜圖上，系統容易忽略較小的音訊事件，或是造成數個小事件之合併；反之，於解析度較高之頻譜圖上，系統對於事件偵測過於敏感，造成將大事件過度切割為數個小事件。

3.1.4 時間-頻率頻譜圖分析

頻率為時間訊號的重要特徵，傳統用傅利葉分析來了解一段時間內頻譜的分布 ，但在某些情況下，我們更有興趣的是頻率隨時間變化的情形，分析各種不同頻率隨時間變化的情形稱為時頻分析。時頻分析相較於頻譜分析多了頻率對時間的解析。傅利葉轉換提供了從時域到頻率域的轉換，能進一步提供二維或三維的時頻分布圖形，進而在時間-頻率平面上表現的信號中，了解各種分量的時間變化與頻譜間關聯的特性。

於本研究中採用

- 取樣頻率：22.05 KHz
- 取樣解析度：16 bits
- STFT Window Size：1024 samples
- 聲道：單聲道(左聲道)

基於取樣理論(sampling theory)中，為避免訊號疊假(aliasing)所造成頻率域上頻譜的重疊，取樣頻率需大於訊號最大頻寬的 2 倍，並因人耳特性與實驗器材限制，故將取樣頻率設定為 22050 Hz。故本實驗所設定之時間-頻率頻譜圖解析度為 320*128，呈現 0～11025 Hz、為時約 6 秒之頻率分布影像。以門鈴聲-2 為例，如圖 3-2 所示，縱軸為時間，橫軸為頻率，灰階像素值為音訊強度，像素值越大表強度越大；反之。表 3-1 為本研究中所用常見於家庭之音訊分類。

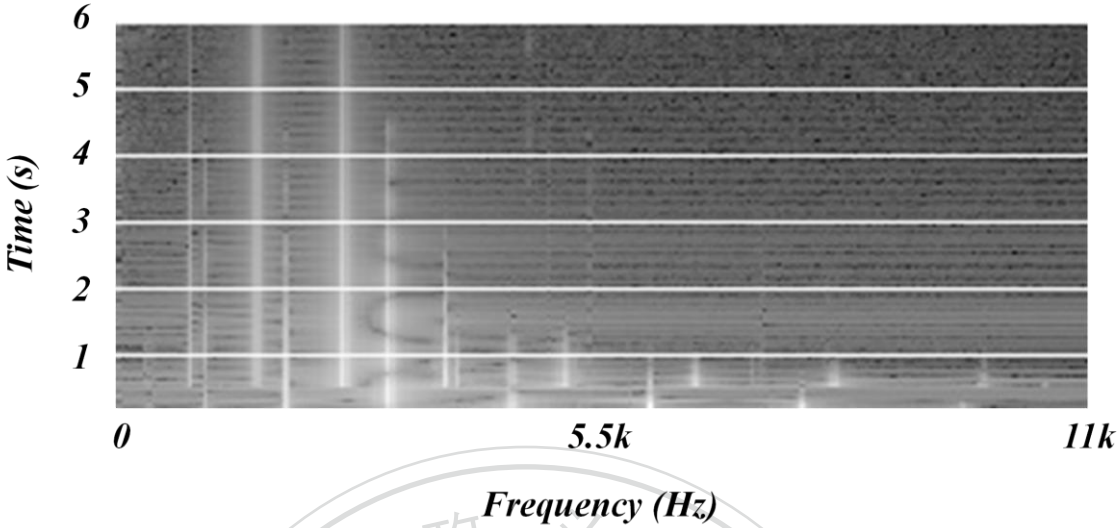
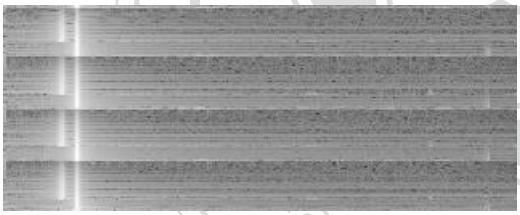
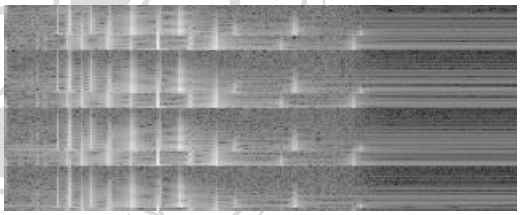
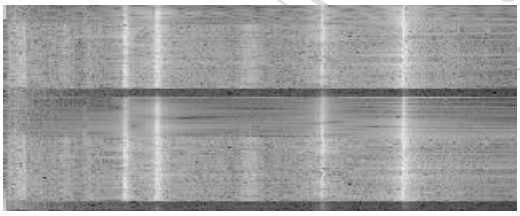
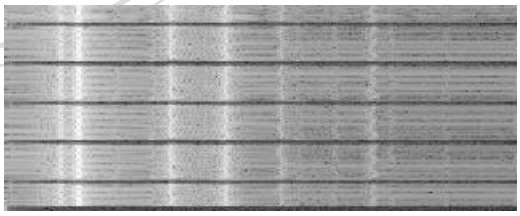
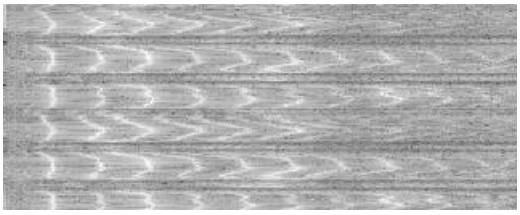
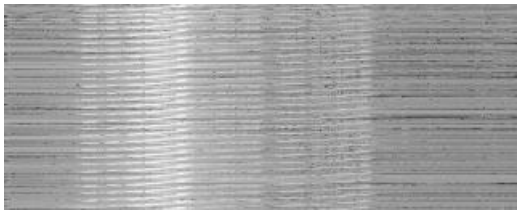
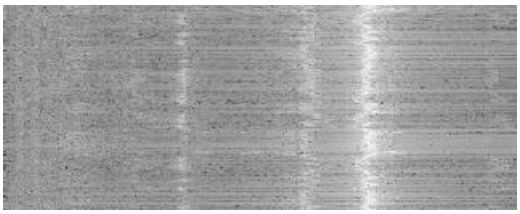
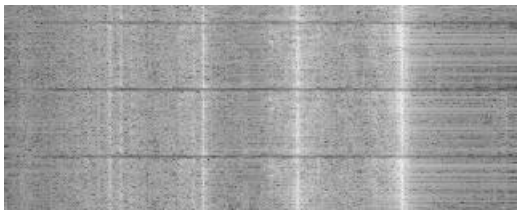


圖 3-2. 門鈴聲-2 (時間-頻率頻譜圖)

表 3-1. 常見於家庭之音訊分類

Class 1. 門鈴聲-1	Class 2. 門鈴聲-2
	
Class 3. 電話鈴聲-1	Class 4. 電話鈴聲-2
	
Class 5. 嬰兒哭聲	Class 6. 汽車警報聲
	

Class 7. 水壺汽笛聲	Class 8. 火災警報聲
	

相關研究中可見許多基於計算聽覺場景分析原則所提出的音訊描述模型，本研究主要利用其中的音訊起始點與音訊區塊輔助影像偵測音訊事件。

3.1.5 音訊起始點 (Auditory Onset)

Bello 等學者於[19]中指出，當一個音訊事件發生時，通常會產生一個較為突然的能量變化，造成能量曲線的突增(attack)，而後隨著能量的消減而造成曲線遞減(decay)，在這段時間可稱為瞬態(transient)。而瞬態發生的起始點，將之定義為音訊起始點(audio onset)，其概念如圖 3-3 所示：

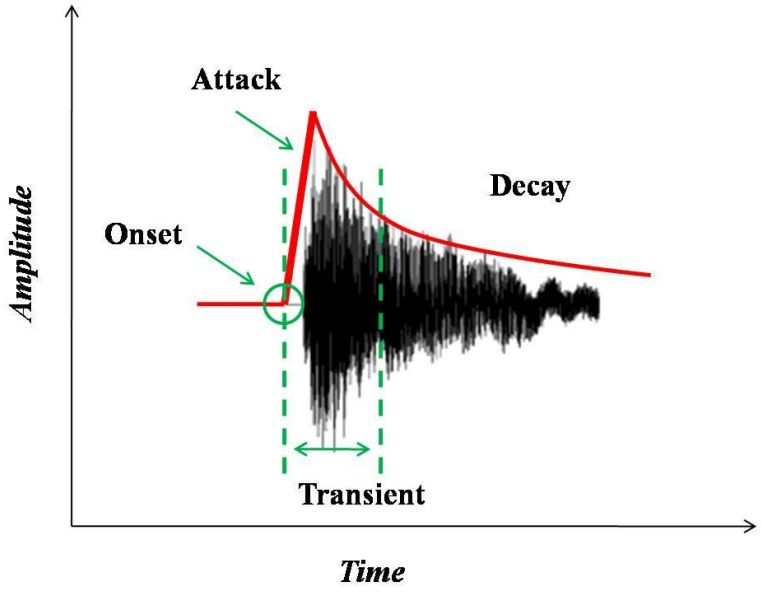


圖 3-3. 音訊起始點 (Audio Onset)

透過音訊起始點偵測，可以找出目前環境中是否有音訊事件發生，並針對此音訊事件以音訊區塊概念加以描述。

3.1.6 音訊區塊 (Auditory Blobs)

於文獻中提出將音訊事件以音訊區塊的概念建立其模型。首先定義頻譜圖上波峰的連續延展為岸線，在此實驗中限制一個訊框中僅含一個岸線。基於聽覺場景分析原則將擁有相同起始點之岸線集合定義為區塊，而此區塊視為一個音訊事件，如下圖所示。

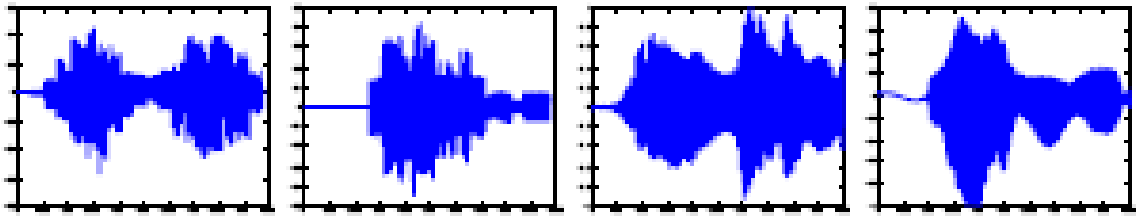


圖 3-4. 音訊區塊(Auditory Blobs)範例

文獻中提出利用以下四種特徵對此音訊區塊加以描述:

1. Frequency Content: 紀錄每個訊框的索引(index)與頻率分布之直方圖。
2. Harmonicity: 於不同頻率上呈現某種程度上相同的樣式。
3. Energy Dynamic: 此區塊中所含所有岸線之能量總和。
4. Frequency Dynamic: 此區塊中所含所有岸線之權重能量總合。

透過以上特徵加以定義之音訊事件，於本研究中則以影像特徵嘗試對此音訊事件於時間-頻率頻譜圖所呈現之紋理加以描述，我們利用影像偵測中的區塊偵測(blob detection)，對時間-頻率頻譜圖上的所呈現之能量變化偵測音訊區塊，並於多個頻率上

找尋具有相同起始點之影像區塊，將其視為一完整音訊事件，並以下述之影像特徵加以描述。

3.2 影像分析 (Image Analysis)

人們對於數位影像處理方法的興趣來自於兩個主要的應用領域：改善影像資訊供人理解之用，以及處理影像資料供機器自動感知所需的儲存、傳輸與表示。定義我們所稱影像處理之領域範圍，利用數位影像處理方法對此影像加以處理，藉以取得所需資訊。

在本研究中，我們將透過數位影像處理技術對時間-頻率頻譜圖加以處理，以取得其中蘊含訊息。透過生理與物理之聽覺感知原理定義濾波器(filter)，針對音訊事件於時間-頻率頻譜圖上所呈現之特徵，強化所需部分並捨去不必要之區塊，如此有助於音訊之起始點偵測，並針對此音訊區段中的事件，透過區塊偵測演算法偵測所需區塊並加以標記，再以影像描述子(descriptor) 描述此影像，讓電腦理解此區塊所呈現訊息。

3.2.1 雙向濾波器 (Bilateral Filter)

雙向濾波器是一種非線性濾波器(non-linear filter)[20]，可有效的將雜訊平滑化，又可將重要的邊緣保留下來。起源可由 1995 年 Aurich 等人所提出非線性高斯濾波器之概念說起[21]，而後由 Tomasi 等人給予正式之名稱[22]。之後雙向濾波器之應用急速的發展，至今十分廣泛的實現於各種影像處理應用中，例如降噪(denoising)、紋理編輯與增強(texture editing and relighting)、色調管理(tone management)等。

而雙向濾波器之所以如此被廣泛的使用，有以下幾個主要原因：

1. 算式簡單，每個像素以其周圍鄰點之權重平均加以計算。因其簡單的概念可以讓人更為直覺的將此技術滿足於特定之應用需求，並加以實現。
2. 此計算結果僅需取決於兩個主要參數，且這兩個參數不會在迭代中產生累積效應，使得參數可以更為簡單的設定。
3. 因為其計算的機制，即便在高解析度影像的處理中也可透過圖形裝置的輔助實現即時計算。

於本研究中，我們對於時間頻率頻譜圖使用雙向濾波器做為端起始點偵測中波動削減之用，以冀濾去因紋理與細小雜訊所造成之訊號波動，保留主要音訊事件結構以利於起始點偵測。

3.2.1.1 高斯濾波器 (Gaussian Filter)

模糊化(Blurring)是影像平滑化中最簡單的一種，透過計算周圍鄰點的權重加總的方法。高斯濾波器是透過計算周圍鄰點空間中的位置為其權重，權重取決於高斯分布(Gaussian Distribution)的σ參數，依高斯分布隨空間距離越遠而遞減。其算式如下所示，其中I表一灰階影像、$\mathbf{I}_{\mathrm{p}}$表位置p上之強度(intensity)、S表空間域上可能的像素位置、$\|p-q\|$表位置p、q兩點之歐氏空間距離。

$$GC[I]_p = \sum_{q \in S} G_\sigma(\|p-q\|) I_q \tag{3.3}$$

其中$G_\sigma(x) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2}{2\sigma^2})$。

其中強度的變化僅取決於兩點間距離而非它們的強度，故因強度差異所形成的邊緣

會因此特性而失去。如圖 3-5、圖 3-7 所示實驗結果，隨著 kernel size 與 σ 的增加，其模糊化效果上升，而原強度分布也因模糊化效果而呈現單閾值分布或是雙閾值分布，如圖 3-6、圖 3-8 所示。雖然高斯濾波器有著優異的模糊化效果，但其運算方式造成時間與頻率上的資訊被視為相同而模糊化，也失去音訊事件之輪廓，如此造成在之後的音訊區塊偵測錯誤，故我們改採以可保留其邊緣的平滑化技術，稱之為雙向濾波器，並對其定義加以改寫，對音訊事件的區塊邊緣與其時間頻率訊息可以更有意義的被保留。

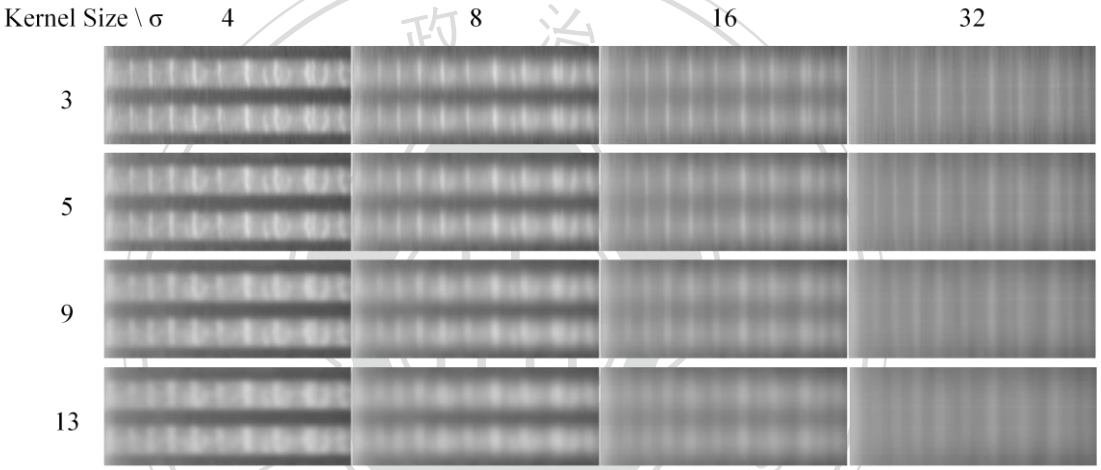


圖 3-5. 高斯濾波器(火災警報聲)

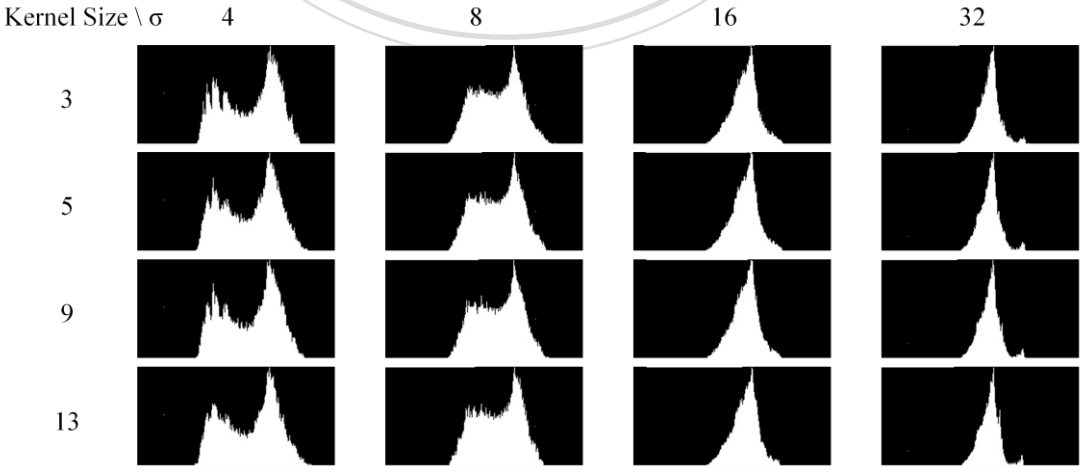


圖 3-6. 高斯濾波器-強度分布直方圖(火災警報器)

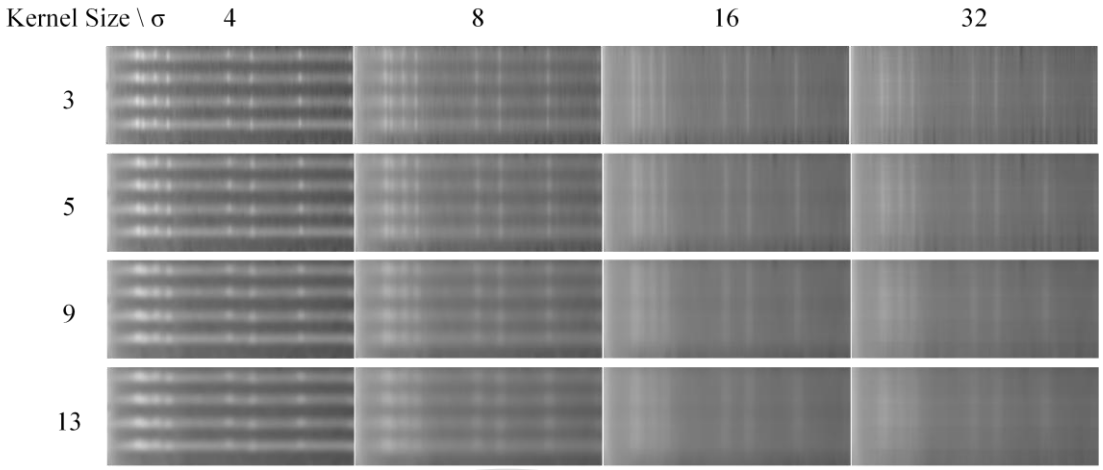


圖 3-7. 高斯濾波器(門鈴聲)

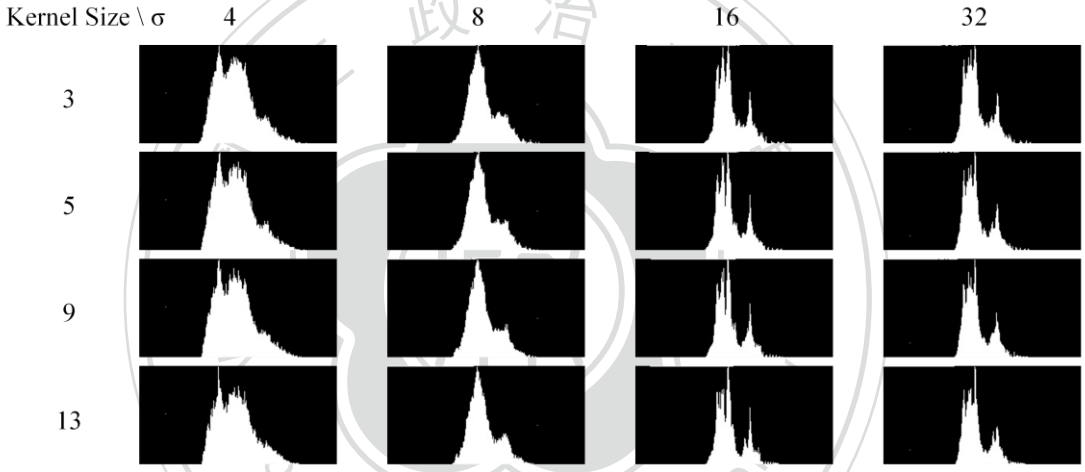


圖 3-8. 高斯濾波器-強度分布直方圖(門鈴聲)

3.2.1.2 雙向濾波器 (Bilateral Filter)

為使因強度差異所造成的邊緣能在平滑化的過程中保留下來，而加上了考慮周圍鄰點之像素值所產生的權重。而雙向濾波器的基礎概念建立在平滑化時不僅考慮兩點距離，同時也將兩點間的強度差異加入考量。將影像分割成大範圍的結構特徵與小範圍的紋理特徵處理，如下圖所示。

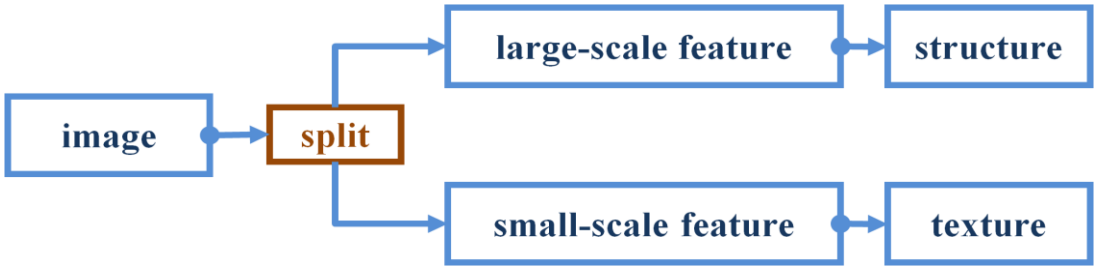


圖 3-9. Bilateral Filter 概念圖

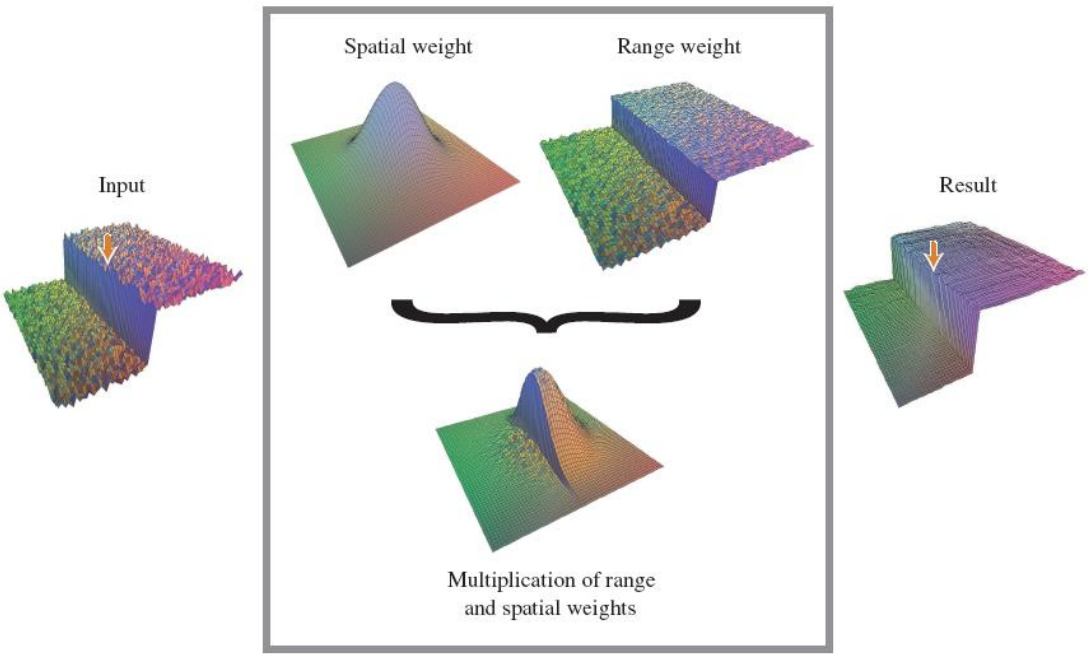


圖 3-10. Bilateral Filter 平滑化示意圖[23]

其算式如下

$$BF[I]_p = \frac{1}{W_p}\sum_{q\in S} G_{\sigma_s}(\|p-q\|)G_{\sigma_r}(|I_p - I_q|)I_q \tag{3.4}$$

其中 $W_p = \sum_{q\in S} G_{\sigma_s}(\|p-q\|)G_{\sigma_r}(|I_p - I_q|)$

$G_{\sigma_s}(\|p-q\|)$ 為 Gaussian spatial weighting

$G_{\sigma_r}(|I_p - I_q|)$ 為 Gaussian range weighting

參數 σ_s 控制 G_{σ_s} 以降低距離過遠的點所造成之影響，增加空間參數(spatial parameter) σ_s 會將結構上的特徵平滑化。而 σ_r 則控制 G_{σ_r} 以降低兩點間強度過大時所造成的影響，當範圍參數(range parameter) σ_r 增加時，因為 G_{σ_r} 逐漸平坦，雙向濾波器所產生效果會逐漸接近高斯濾波器。而圖 3-11、圖 3-12 呈現兩個參數間關係所造成之結果差異。由實驗結果可觀察其現象，因 σ_s 的增加造成結構上的平滑化，而 σ_r 的增加則可有效產生消除雜訊的結果，有助於取得對於判讀音訊事件所需資訊。

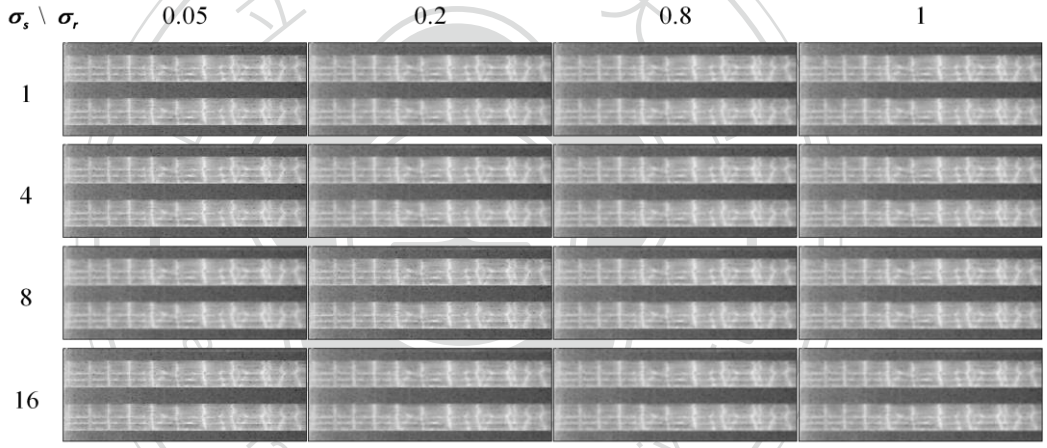


圖 3-11. 雙向濾波器(火災警報聲) - σ_r 與 σ_s 影響所得結果

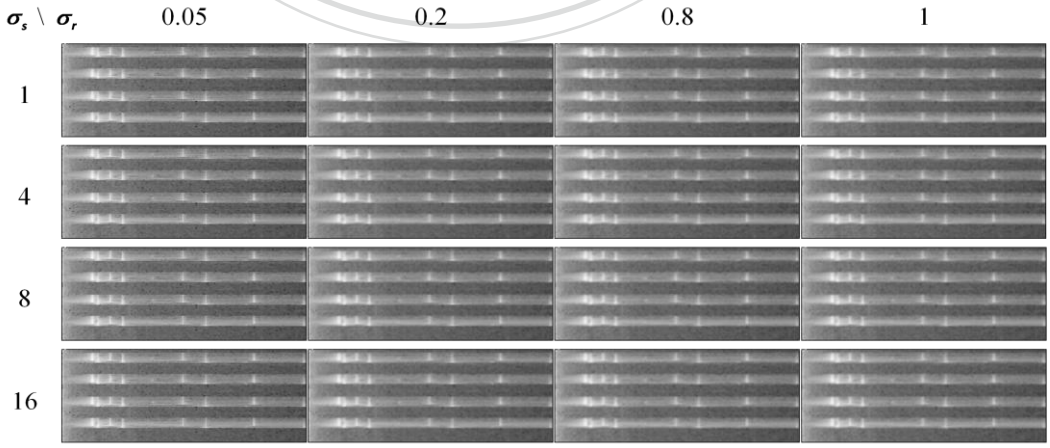


圖 3-12. 雙向濾波器(門鈴聲) -σ_r 與 σ_s 影響所得結果

同時，透過雙向濾波器的可迭代特性，可以讓區域更為平坦化，其影響與空間、範圍參數的設定不同，雖然一個較大的範圍參數σ_r可產生平滑化效果，有效的將紋理特徵濾去，但同時也讓邊緣更趨於平坦化。而迭代可以讓紋理的部分更為平坦，仍將結構特徵保留下來，於圖 3-13、圖 3-14 顯示其結果與其分布直方圖，由此實驗結果可見，迭代方法可產生與σ_r、σ_s 設定不同的效果，雖然增加σ_r可以將雜訊去除，但也因為σ_r的過度增加造成邊緣的平滑化，而迭代方法則可將較大的邊緣結構保留下來，對細微的邊緣結構加以平滑化，藉以找出此時間-頻率頻譜圖中音訊事件。

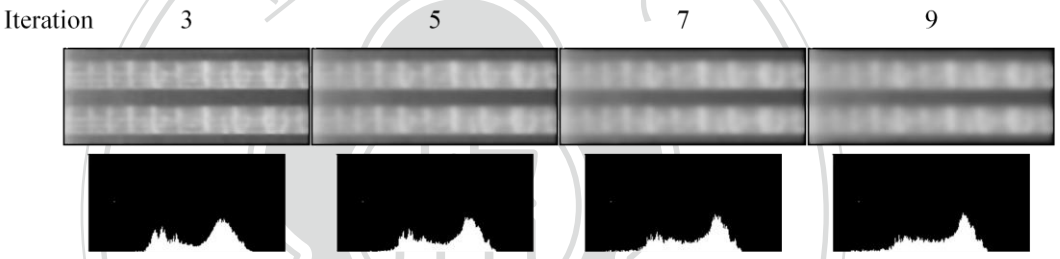


圖 3-13. 雙向濾波器(火災警報聲) -迭代次數之影響

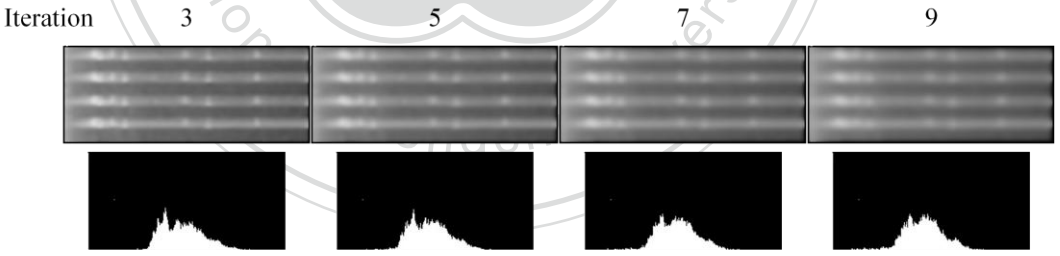


圖 3-14. 雙向濾波器(門鈴聲) -迭代次數之影響

然而，於原本雙向濾波器中，對於空間參數與空間權重其縱軸與橫軸同為空間意義，而在時間-頻率頻譜圖上的縱軸與橫軸卻分別代表不同的意義，於本實驗中縱軸為時間，而橫軸為頻率。為因應聽覺對於時間與頻率的反應對應至時間-頻率頻譜圖的空間座標，我們將空間參數與空間權重更細分為時間權重與頻率權重。其算式改寫如下。

$$BF[I]_p = \frac{1}{W_p}\sum_{q\in S} G_{\sigma_t}(\|p-q\|)G_{\sigma_f}(\|p-q\|)G_{\sigma_r}(|I_p - I_q|)I_q \tag{3.5}$$

其中 $W_p = \sum_{q\in S} G_{\sigma_t}(\|p-q\|)G_{\sigma_f}(\|p-q\|)G_{\sigma_r}(|I_p - I_q|)$

$G_{\sigma_t}(\|p-q\|)$ 為 Gaussian time weighting

$G_{\sigma_f}(\|p-q\|)$ 為 Gaussian frequency weighting

$G_{\sigma_r}(|I_p - I_q|)$ 為 Gaussian range weighting

透過改寫雙向濾波器在時間、頻率與強度三者間的關係，可更為完整的對應至計算式聽覺場景分析的主要原則。同時利用實作雙向濾波器之處理核心大小(kernel size)對應至反應時間、可感頻率與感知敏感頻率、音強變化等人類聽覺對於聲音變化的感知特性，於圖 3-15、圖 3-16 所示其結果，由實驗結果可得，將原本的σ_s 改寫成σ_t與σ_f ，可對結構選擇性的造成平滑化，對應於聽覺式場景分析中所提出人耳對時間與頻率之關係，對於人類較為不敏感的部分加以平滑化，針對敏感感受的區段更為仔細的處理。

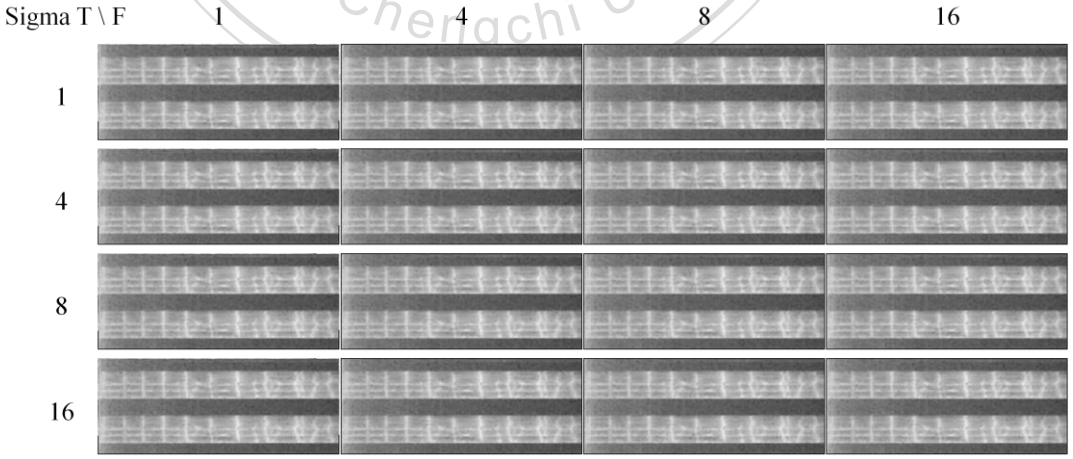


圖 3-15. 改寫雙向濾波器(火災警報聲)

設定σ_t與σ_f 所得結果 (σ_r 設定為 0.8)

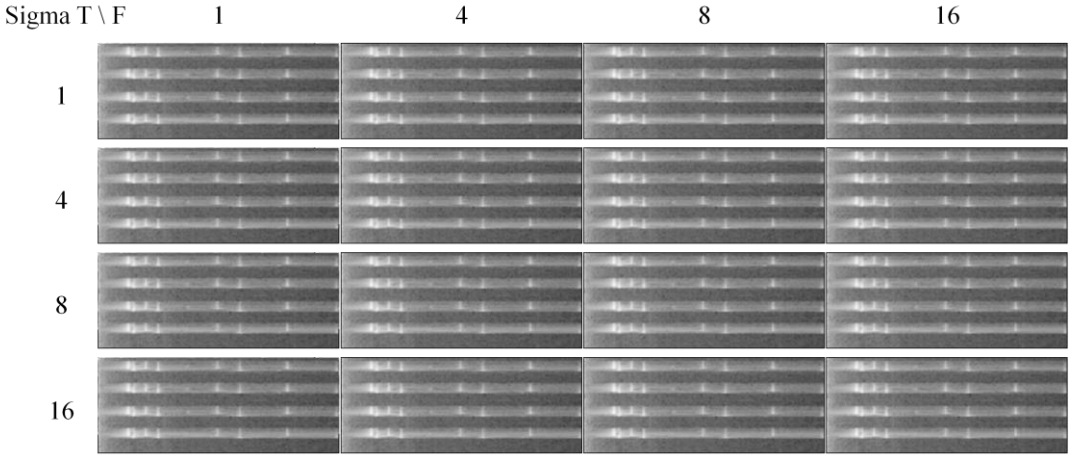


圖 3-16. 改寫雙向濾波器(門鈴聲)

設定σ_t與σ_f所得結果 (σ_r設定為 0.8)

雙向濾波器於本研究中做為起始點偵測之波動削減處理，主要目的在於降低雜訊與非音訊主結構之像素所造成的影響，將σ_r設定趨近於高斯濾波器之效果，以除去雜訊之用。根據聽覺心理學中對於人耳特性之闡述，我們企圖利用σ_t與σ_f的設定模擬人耳對於時間與頻率變化之感受，由於音訊於時間軸上的快速變動，故σ_t的設定傾向於將時間軸上之波動更為有效的削減，而對於頻率軸資訊則傾向設定σ_f以強調音訊事件之主體。經過雙向濾波器之影像，則接著由起始點偵測以判斷此段時間中，是否有音訊事件之產生。

3.2.2 音訊起始點偵測 (Audio Onset Detection)

音訊起始點偵測(audio onset detection)在音訊處理中是一個十分重要的角色[19]，透過偵測函數之計算，根據偵測所得之起始點(onset)與終止點(offset)，將此段音訊視為一個連續的音訊事件(audio event)，因為此段音訊事件中之音訊時間與頻率連續性，有利於針對這些特性加以分析。

一般而言，音訊起始點偵測之概念如圖 3-17 所示，可分為三個主要步驟：

1. 前處理(Pre-processing)

   將音訊訊號從現實中的類比訊號透過裝置加以取樣、量化，取得其數位訊號，並將此數位訊號轉換至頻率域上，有助於後面兩個步驟之分析。

2. 簡化(Reduction)

   將前處理後所得訊號，透過偵測函數用以削減波動，將此數位訊號轉換為有助於音訊起始點判斷之訊號。

3. 波峰選取(Peak-picking)

   將經過偵測函數計算的訊號，取出其中各區段的波峰，將這些波峰視為音訊事件之起始點候選。

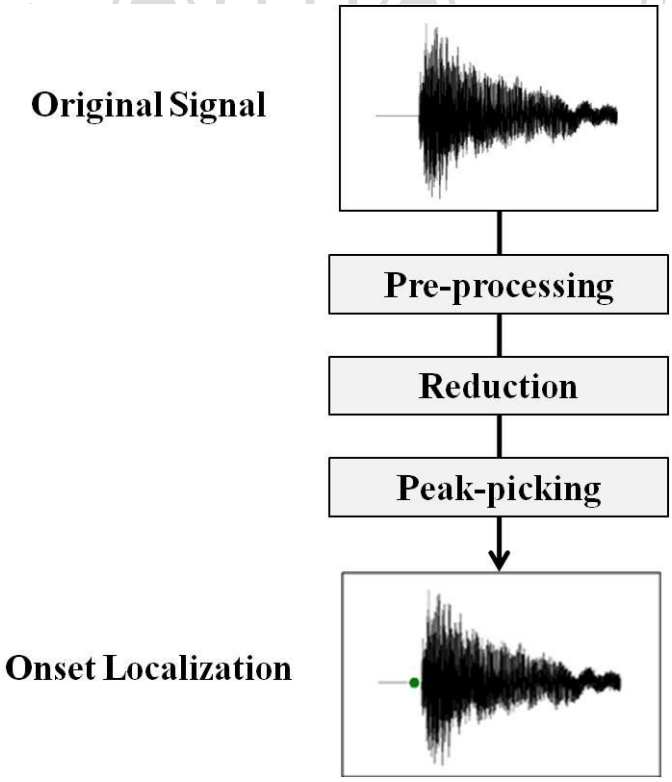


圖 3-17. 音訊起始點偵測

3.2.2.1 音訊起始點偵測實作

基於以上介紹的定義與方法，本研究針對音訊起始點偵測實驗設計如下：

1. 前處理：

   透過短時傅利葉轉換以取得此數位訊號之時間-頻率訊息。

2. 簡化：

   由於當音訊事件發生時，能量將產生明顯變化，因此我們將此運算對應至影像處理，針對能量變化對應於時間-頻率頻譜圖上之像素強度變化，透過雙向濾波器有效的對雜訊與獨立訊點平滑化，以達成波動削減之目的，並保留音訊事件之結構主體，以利於音訊起始點之偵測。

3. 波峰選取：

   Masri 等學者提出了頻帶權重的概念[24]，對於不同頻帶上所發生之音訊事件給予不同權重，而此總加權越高即越有機會是一個事件之起始點。Goto 等學者提出了將訊號分成七個頻帶，於此七個子頻帶上尋找音訊起始點的方法[25]。為此，我們對於音訊起始點之波峰選取採用多頻帶權重的做法。透過聽覺心理學於臨床實驗中所得人類聽覺感知之曲線，從圖 3-18 中，我們可以得知人類聽覺於不同頻帶上之不同敏感程度，取其中聽覺閾值曲線(threshold of hearing)做為本研究中多頻帶權重設定之依據，表 3-2 為設定之結果，圖 3-19 呈現多頻帶於時間-頻率頻譜圖上之區分。

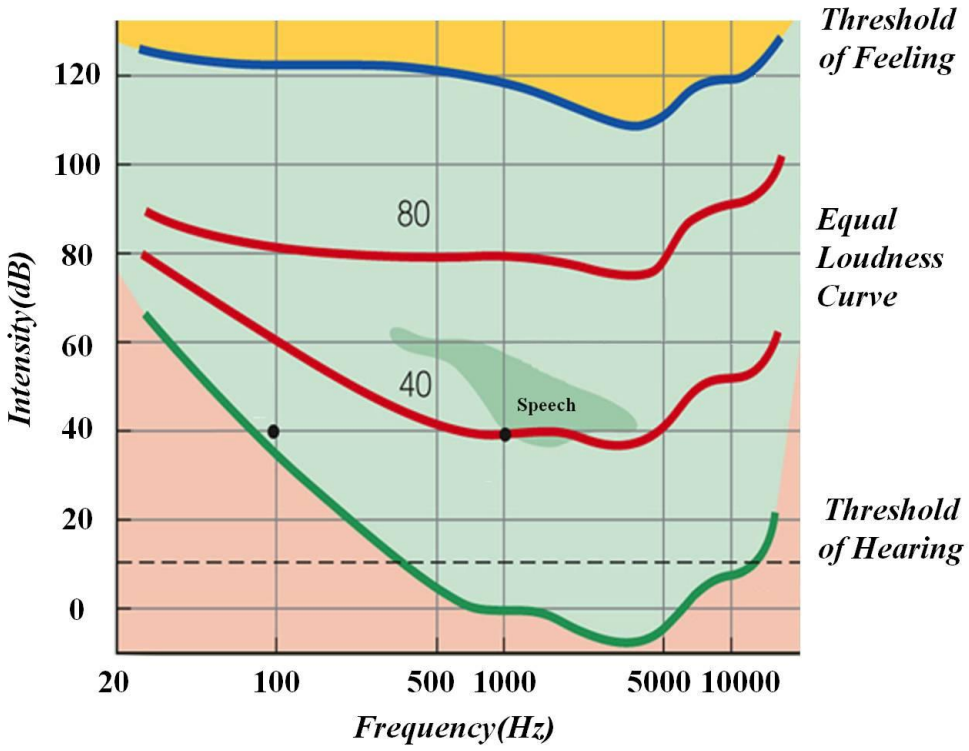


圖 3-18. 聽覺感知曲線[27]

表 3-2. 多頻帶權重

Frequency	Weight
0 ~ 200 Hz	1
200 ~ 400 Hz	6
400 ~ 800 Hz	11
800 ~ 1600 Hz	16
1600 ~ 3200 Hz	23
3200 ~ 6400 Hz	16
6400 ~ 11025 Hz	6

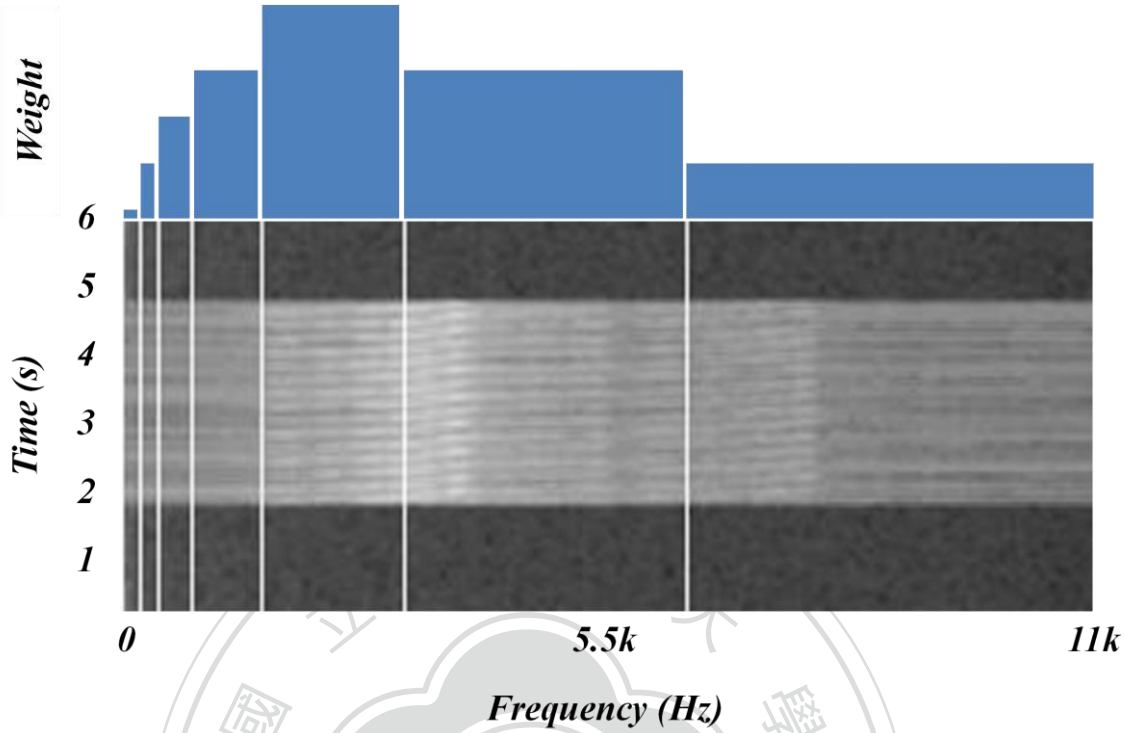


圖 3-19. 頻帶權重區分示意圖

透過多頻帶權重之設定，當超過測試環境總音量閾值時即視為一個可能的音訊起始點，並搜尋之後的能量曲線中是否產生遞減現象，當遞減超過閾值時即視為此事件之終止點，將此段時間內的訊號判定為一音訊事件，經實驗結果測試，設定環境總音量的前20%與後 20%為起始與終止之判定閾值。

3.2.2.2 音訊起始點偵測結果

如表 3-3、表 3-4 所列門鈴聲於音訊事件偵測之實驗結果，可以清楚看出基於多頻帶權重之偵測演算法對音訊事件有著不錯的效果，對於連續音訊事件，義可清楚的區分出單音的連續撥放與連續事件的差異，當測試環境具有其他音訊時之實驗結果，在本研究中之音訊起始點偵測仍能有效找出其起始點。於附錄 A 收錄所有分類與環境聲之偵測結果。

表 3-3. 音訊事件起始點偵測-單獨事件與連續事件

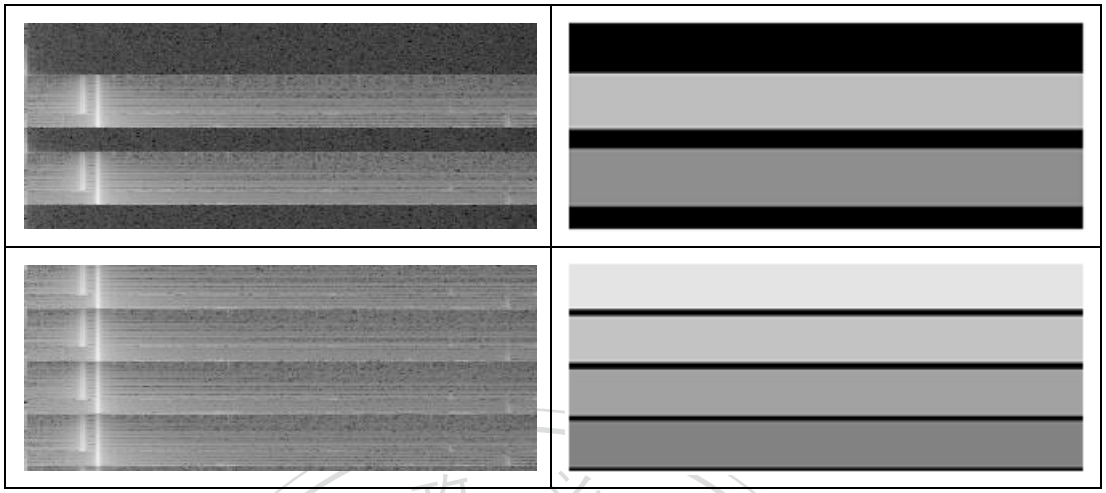
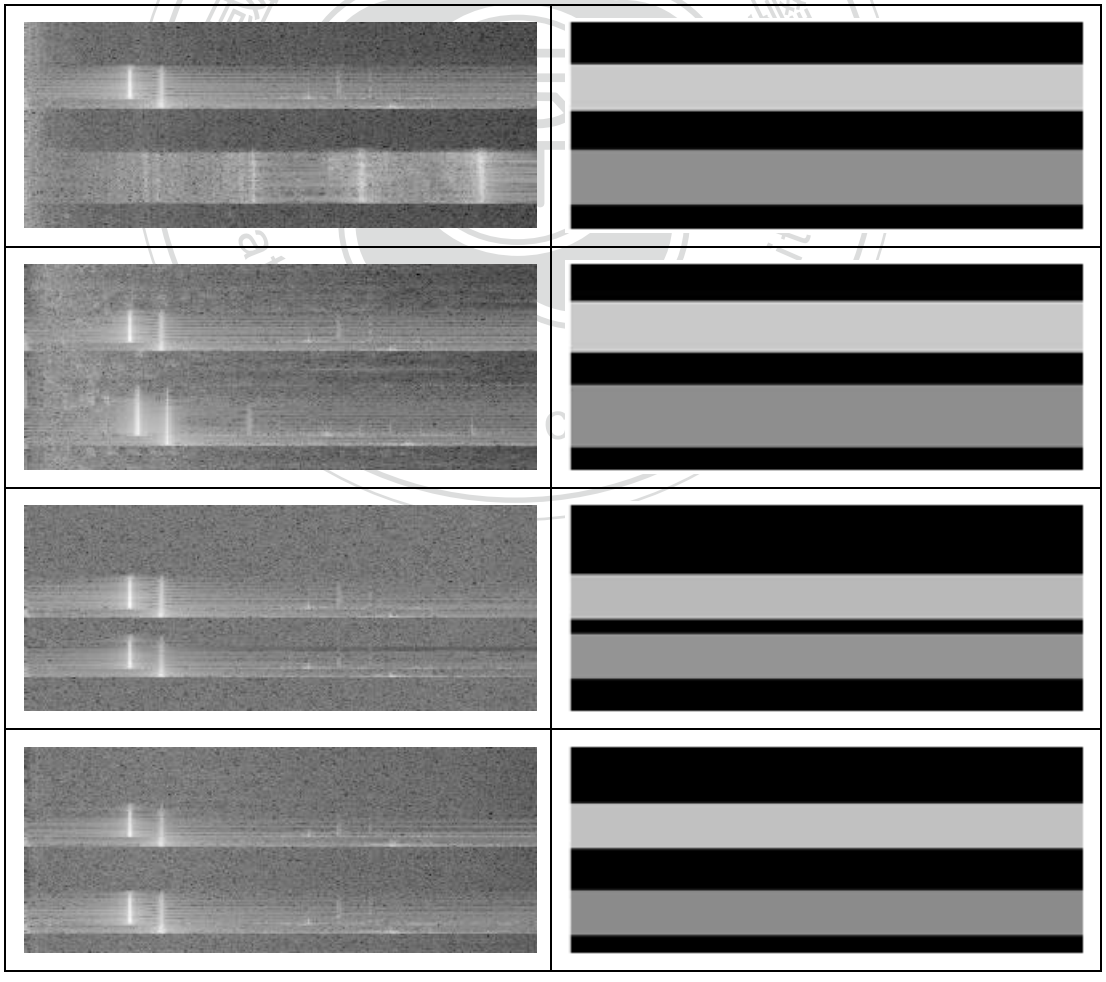
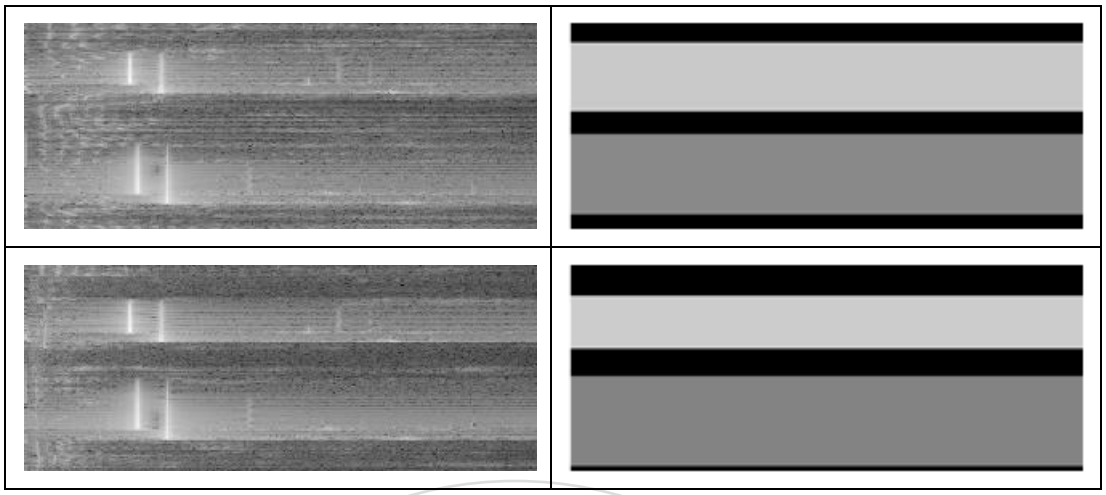


表 3-4. 音訊事件起始點偵測-於各種環境聲中





經過音訊事件之起始點與終止點偵測，找出這段時間中是否有音訊事件的發生，將此視為音訊事件影像，之後的處理將針對音訊事件影像加以處理，企圖找出此段時間中包含何種音訊。

3.2.3 閾值設定 (Thresholding)

由於閾值設定(thresholding)的實作簡單、計算快速及其直觀的特性，一直都是影像分割應用中廣泛使用且十分重要的技術之一。一般多以灰階做比較，針對影像中像素強度處理，將影像中物件(objects)與背景(backgrounds)依適當的閾值分隔出來，實現影像二值化。閾值設定基本概念如下，當影像強度 $f(x,y)$ > 閾值 T 時，則此點設定為物件點(object point)；反之，則設定為背景點(background point)。

$$g(x,y)=\begin{cases}1, f(x,y)>T\\ 0, f(x,y)\le T\end{cases} \tag{3.6}$$

影響強度閾值設定準確與否的五個主要因素：

1. 波峰(Peaks)間的分割

2. 影像中的雜訊

3. 物件與背景的相對大小

4. 光源的一致性

5. 影像中反射特性的一致性

3.2.3.1 基本全域閾值設定 (Basic Global Thresholding)

閾值設定中最為基本的方法之一，透過一個迭代的演算法估算此影像的全域閾值，此演算法流程如下：

1. 估算初始閾值T(必須在影像強度的最大值與最小值之間。一般而言，影像強度的平均是個好的初始選擇)。
2. 依閾值T將影像區分成兩個群組。G_1包含所有大於T的像素；而G_2則包含剩下小於等於T的像素。
3. 計算G_1、G_2所有像素的強度平均，分別為m_1、m_2。
4. 計算新的閾值：$T=\frac{1}{2}(m_1+m_2)$
5. 重複步驟2到步驟4，直到連續兩次迭代計算所得T之差值小於預先定義ΔT。

ΔT的定義將決定迭代計算的次數，換句話說，決定了閾值設定的準確與速度。基本全域閾值設定擁有十分快速的優點，對於物件與背景間有明顯分隔的影像，如圖3-20(左) 所示單閾值長條分布，可以得到良好的分割效果。然而，對於影像中具有多個閾值分布的情況，如圖3-20(右)所示雙閾值長條分布，便無法有效取得其適當的閾值。

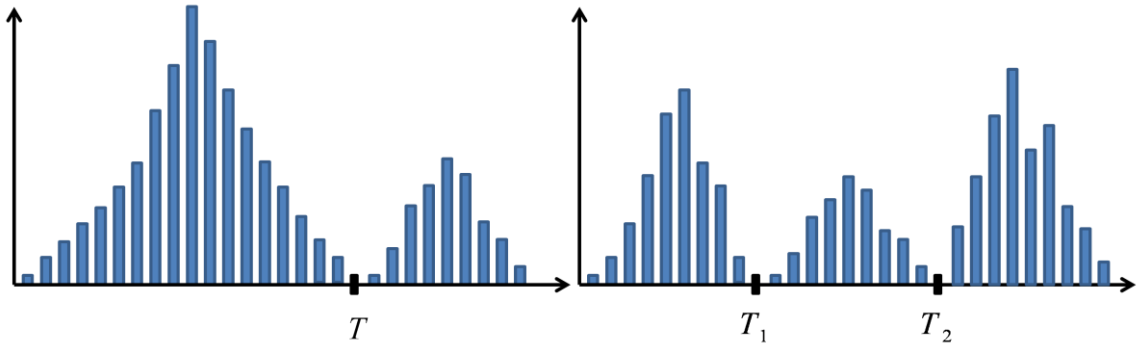
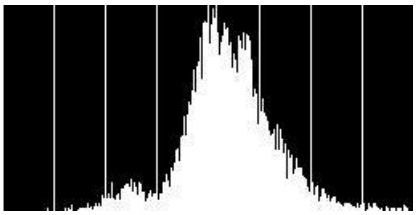


圖 3-20. 單閾值長條分布與雙閾值長條分布

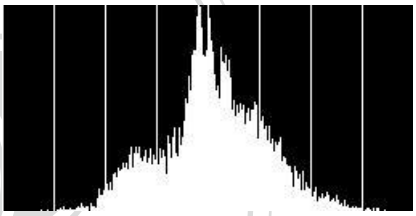
利用基本全域閾值設定於本次實驗之時間-頻率頻譜圖，以門鈴聲為例，如表 3-5、表 3-6 為基本全域閾值設定之實驗結果，從實驗結果中可得隨著 ΔT 設定伴隨著計算次數的增加，但可讓閾值運算更為精準，雖然此演算法在單閾值分布時，有著不錯的效果，但亦可發現當音訊事件為雙峰甚至多峰分布時，音訊事件之結構與紋理均未能有效的被分割出來，而音訊主體佔整個影像比例小的事件尤為明顯(如門鈴聲)，因過多非音訊事件之訊號造成閾值下降，過多非相關像素亦被視為事件的誤判比例過高。於附錄 B 收錄所有分類之基本全域設定實驗結果。

表 3-5. Class1 門鈴聲-1 之基本全域閾值設定實驗結果

音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	

ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

表 3-6. Class2 門鈴聲-2 之基本全域閾值設定實驗結果

音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

因為音訊事件主體大多會伴隨產生於不同頻帶上但強度較弱的諧波，故其強度多以雙峰甚至多峰分布，因此需要可以依音訊主體結構之時間與頻率連續性判斷是否為物件的演算法，以冀先保留音訊主體，去除確定不為事件之雜訊點，並依其連續性判斷其結構。

3.2.3.2 雙閾值設定 (Double Thresholding)

為上述之目的，設計一個雙閾值設定演算法以實現此概念，其概念如下圖與下式所示：

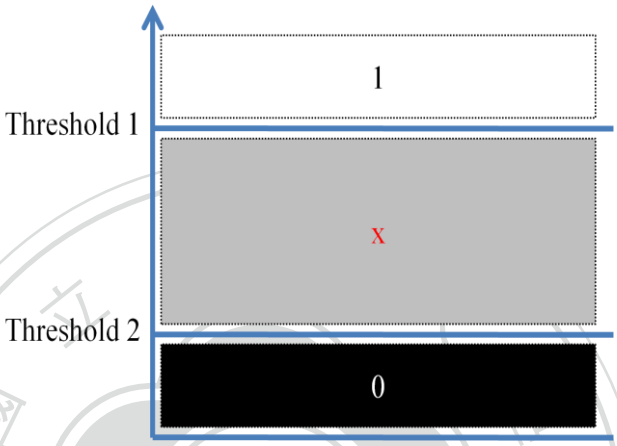


圖 3-21. 雙閾值設定示意圖

1. 設定 T_1 、 T_2 。

2. $$g(x,y)=\begin{cases}1, f(x,y)>T_1\\0, f(x,y)<T_2\\x, others\end{cases} \quad (3.7)$$

3. 考慮 x ，設目前考慮點為 p ，基於計算式聽覺場景分析的音訊事件定義中，時間與頻率的逐漸變化產生音訊事件於時間與頻率上的連續性，故此我們採 4-Neighbor 方式判斷點 p 上下左右四個鄰點。如果 p 點的四個鄰點有其一為 1，則此 p 亦為 1；反之， p 點的四個鄰均為 0，則此 p 點為一孤立點，視此 p 點為 0。

4. 重複進行步驟三，直到沒有點變化則停止。

關於參數 T_1 、 T_2 的設定，將 T_2 設定為全圖之平均與後八分之三，透過強度分布直方圖的觀察，對 T_1 進行調整實驗。實驗結果以門鈴聲為例，如表 3-7，表 3-8 所示，附錄 C 收錄所有分類之雙閾值設定實驗結果。

表 3-7. Class1 門鈴聲-1 之雙閾值設定實驗結果

音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

表 3-8. Class2 門鈴聲-2 之雙閾值設定實驗結果

音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

由上列實驗結果明顯可得，雙閾值設計之結果遠比基本全域閾值設定更為合適，當 T_1 為強度分布之前 25%、T_2 為強度分布之平均時，將可得較佳之閾值設定結果，不僅已可從此閾值設定結果觀察其音訊事件之主要結構，其諧波之結構亦大致成功保留，隨後即將此雙閾值設定之結果以區塊偵測擷取其中之音訊區塊。

3.2.4 區塊偵測 (Blob Detection)

在影像處理的領域上經常需要偵測物件並標示，這些物件通常由連續像素形成之部件(components)所組成。當物件從影像中成功的偵測出來，他們仍需特別的標示出來，因為上述理由，部件輪廓(component contour)是經常被使用在物件偵測上的方法。過去的研究中，鍊碼(chain codes)[23]與傅利葉描述子(Fourier descriptors)[27]被廣泛的使用，也有利用統計的模型[28]以配對的方法。上述方法在區塊偵測中均表現出不錯的效果及運算速度。於本研究中採用鍊碼方法，以期圈選出音訊事件影像中的音訊區塊。

3.2.4.1 鍊碼 (Chain Code)

鍊碼方法由 Freeman 於 1961 年提出[26]，用以表示物件輪廓的標示方法。將輪廓的追蹤方向分成四或八個方向，方向編碼如下：

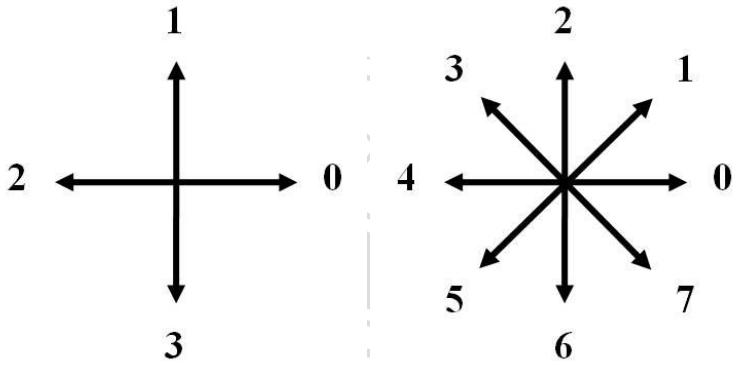


圖 3-22. 鍊碼四方向與八方向編碼

沿物件外部輪廓的單方向路徑追蹤一次，即可得此物件之輪廓鍊碼，例如圖 3-23 範例所得 Freeman 鍊碼為 1-2-0-7-0-6-4-3。

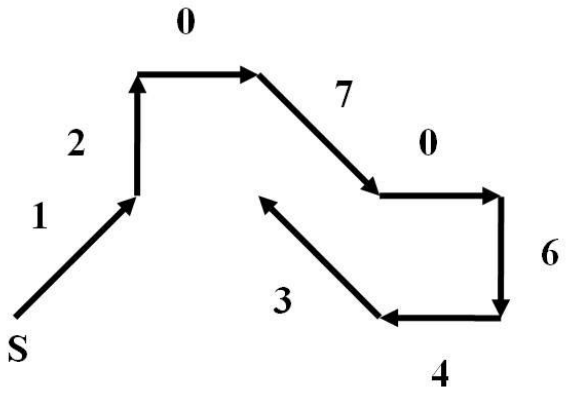
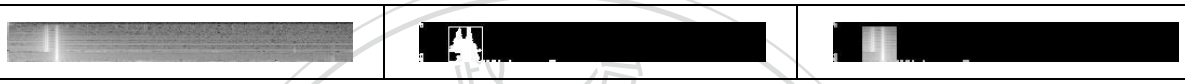
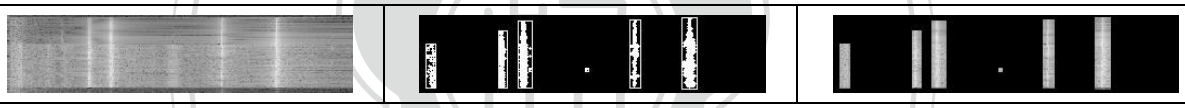


圖 3-23. 鍊碼編碼範例

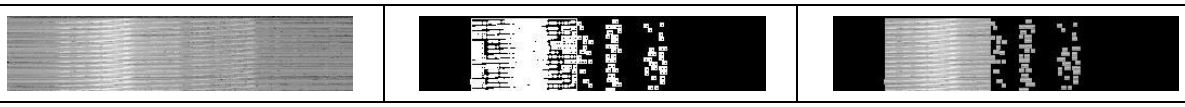
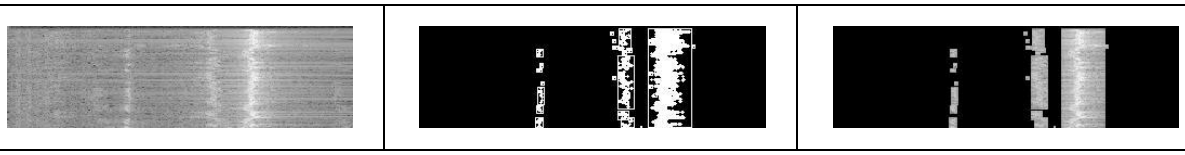
於本研究中採用八方向編碼的鍊碼，對於經過雙閾值設定所得音訊事件二值化影像，加以圈選出其中所含音訊事件。找出每個區塊之邊界框(bounding box)，用以代表此音訊事件的部件，並由不同頻帶上多個部件組成一個音訊區塊，如下列實驗結果所示，由左至右分別為音訊事件影像、音訊區塊偵測影像、音訊區塊影像。

表 3-9. 區塊偵測實驗結果

Class1. 門鈴聲-1 區塊偵測之結果

Class2. 門鈴聲-2 區塊偵測之結果

Class3. 電話鈴聲-1 區塊偵測之結果

Class4. 電話鈴聲-2 區塊偵測之結果

Class5. 嬰兒哭聲 區塊偵測之結果

Class6. 汽車警報聲 區塊偵測之結果

Class7. 水壺汽笛聲 區塊偵測之結果

Class8. 火災警報聲 區塊偵測之結果


由實驗結果可得，我們經過以上處理，完成了音訊主體之保留並加以標示。經過區塊偵測所得之音訊區塊影像，為此聲音於時間-頻率頻譜圖上所呈現之訊息，我們嘗試以影像描述子對此音訊區塊影像加以描述，用於之後的音訊分類。於此，針對此音訊區塊影像所呈現之紋理，並考量串流式音訊分類所需之即時性，我們選用區域二元化圖型以描述此音訊區塊影像。

3.2.5 區域二元化圖型 (Local Binary Pattern)

Ojala 等人於 2002 年所提出之區域二元化圖型特徵[17]已在過去許多相關研究中被證實在影像的描述上有十分優異的表現，是由材質區域性的定義所導出，具有不受影像灰階值比例改變而變化的算子；同時它也被發展成擁有不依影像旋轉而改變的特性。且為了適合不同的應用，區域二元化圖型延伸出各式各樣的形式。

區域二元化圖型是一種用以描述紋理變化的特徵計算方式[17][29]，運算簡單且快速，適合用於即時系統(real-time system)，而缺點在於平滑影像或轉換成灰階影像後紋理表現不明顯者的描述效果較差。

區域二元化圖型流程如下:

1. Ojala 於最初所提出的演算法為設定一 3×3 區塊用以標記影像。
2. 色彩轉換: 將彩色影像轉換成灰階影像。
3. 閾值(Threshold)設定:

   假設影像中標記區塊之灰階值如圖 3-24，則設定此 3×3 區塊的閾值為正中央的值，正中央灰色部分為 50。

98	26	23
85	50	36
12	13	50

圖 3-24. 3×3 區塊範例

4. 閾值運算:

周圍八個點分別與閾值比較大小，如大於或等於閾值，則設定為 1；如小於閾值，則設定為 0，運算結果如下圖。

1	0	0
1		0
0	0	1

圖 3-25. 區塊經 LBP 運算後結果

5. 權重(Weight)設定: 給予周圍八個點權重，如圖 3-26 所示。

2^0	2^1	2^2
2^7		2^3
2^6	2^5	2^4

圖 3-26. 區塊權重分布

6. 數值統計:

將運算後結果與權重分布相乘後再相加，如圖 3-24、圖 3-25 先相乘後相加，可得 $2^0 \times 1 + 2^4 \times 1 + 2^7 \times 1 = 145$。

7. 直方圖(Histogram)統計:

將影像以上述方式計算完畢後，即可累加計算統計出此影像的區域二元化圖型直方圖(local binary pattern histogram)，此直方圖中共有 256 個 Bin，故此區域二元化圖型直方圖即為一大小為 256 的陣列(array)。

8. 特徵描述: 最後將 256 個 Bin 的個數作為一 256 維的影像描述特徵。

3.2.5.1 Uniform Pattern

Ojala 也提出區域二元化圖型中，特定的圖型(如圖 3-27 所示)經過旋轉後將可得到生成此區域二元化圖型特徵之基底(Base)，稱之為 Uniform Pattern。於文獻實驗中，使用最初定義 8 個像素而半徑為 1 的 3×3 區塊，Uniform Pattern 的直方圖數量總和佔所有 8-bit (00000000~11111111)的 85%~90%，故將這 58 個 Uniform Pattern 獨立出來計算直方圖，作為特徵向量中的 58 個維度，並將剩下的直方圖 Bin 值加總後作為第 59 維。

圖 3-27. Uniform Pattern

於本研究中為實現即時音訊分類，希望可利用 Uniform Pattern 之特性，將描述用的 256 維空間減少至 59 維空間，以達到降低維度之效。故設計一實驗針對時間-頻率頻譜

圖是否亦具有 Uniform Pattern 分布之特性。將輸入之音訊透過預處理與起始點偵測，取得其音訊事件之影像，及其透過區塊偵測所得之音訊區塊影像進行 Uniform Pattern 分布性質測試，於此取訓練資料中 8 個分類各 5 個樣本(於附錄 D 收錄音訊事件與其音訊區塊之樣本)來進行實驗。其中音訊區塊影像僅處理非 0 部分(意即影像中黑色區塊不納入區域二元化圖型計算)。

實驗結果如下，我們可從實驗結果中表 3-10、表 3-11 得出，音訊事件與音訊區塊之時間-頻率頻譜圖於 Uniform Pattern 中的分布約略落於 75%~90%之間，意即 75%~90%的值分布此 58 維空間中，故我們可將原 256 維空間降低至此 58 維空間，以降低維度求取更為快速的計算，同時也因 Uniform Pattern 之特性，不因大幅度的降維而造成描述力的損失。其中因為 Uniform Pattern 編碼所呈現多為邊緣與紋理，故亦可從實驗結果觀察的得知，音訊區塊因保留較多的音訊事件主體及其紋理，故於此實驗中有較佳的表現。

表 3-10. 音訊事件之 Uniform Pattern 比例

	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**
Door Bell-1	80.67	80.71	80.01	81.51	78.08
Door Bell-2	77.37	76.51	77.96	75.69	78.99
Phone Bell-1	77.81	75.86	73.86	79.98	75.11
Phone Bell-2	78.64	76.75	71.2	74.41	75.88
Baby Cry	78.68	74.26	80.23	78.24	75.01
Car Alarm	82.78	81.19	82.04	79.36	79.45
Kettle Whistle	79.33	78.96	77.98	79.05	78.84
Fire Alarm	71.69	74.24	71.59	69.86	76.96

單位：百分比

表 3-11. 音訊區塊之 Uniform Pattern 比例

	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**
Door Bell-1	87.31	86.33	86.32	85.92	84.87
Door Bell-2	91.87	89.87	92.48	90.26	92.66
Phone Bell-1	84.83	84.86	85.97	85.48	93.72
Phone Bell-2	78.23	77.16	74.05	86.66	89.38
Baby Cry	78.74	74.32	83.61	81.98	78.46
Car Alarm	91.11	89.91	90.47	85.69	87.39
Kettle Whistle	91.63	88.26	87.36	90.15	88.17
Fire Alarm	87.22	89.26	85.15	84.36	90.57

單位：百分比

音訊事件影像經過區域二元化圖型運算並因其 Uniform Pattern 特性，可得一 59 維之編碼直方圖，於此我們對此直方圖正規化(Normalize)，將此正規化過後之直方圖視為此音訊事件的特徵向量。透過之後的直方圖距離定義與計算，對音訊事件加以分類。

3.3 相似度搜尋 (Similarity Search)

擷取出音訊事件其影像特徵後，以這些特徵加以分類，找出哪些資料其相似度較高，便有較高的機率可被分類於同一種類。最鄰近搜尋(nearest neighbor search)是相似度搜尋中最為廣泛使用的方法之一。

3.3.1 K 個最近鄰點分類器 (K- Nearest Neighbor, KNN Classifier)

解決最鄰近搜尋問題中最為著名的是 K 個最近臨點(K-Nearest Neighbor, KNN)，首先我們將資料透過其特徵投射至空間中，將每筆資料視為一點，如圖 3-28 所示，每當一筆新的資料進入系統，KNN 演算法便計算這筆資料與其他已知資料的距離，如 K 為 1，稱之為 1-NN，代表演算法將找出第一位距離最接近的點，將此新資料點與最近點視為同一種分類，如圖中所示，當新資料其最近似點為橘點，則此資料分類為橘點；K 為 3 時，代表演算法將找出前三距離最近的點，在這些點中計算哪些分類所佔比例較高，用以判斷此新資料點為何分類。以此類推可以有 5-NN、9-NN 等，端看問題需求而定義。

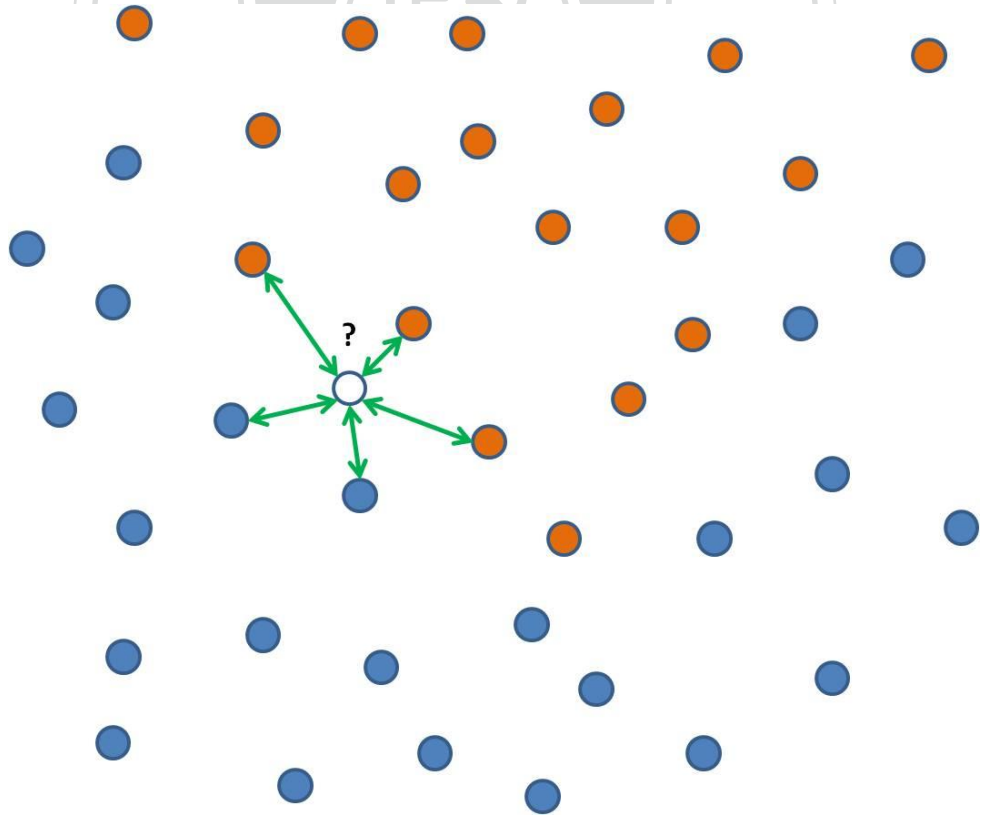


圖 3-28. KNN 概念圖

3.3.2 距離定義

針對每個音訊事件內包含的音訊區塊所產生之區域二元化圖型直方圖，透過其區域二元化圖型特徵進行相似度度量，於本研究中採用三種常見於直方圖比較之距離定義：

1. Chi-Square Distance

$$x^2(S,M)=\sum_{b\in B}\frac{(S_b-M_b)^2}{S_b+M_b} \tag{3.8}$$

S 為 Testing Data、M 為 Training Data。

用以計算 S 與 M 之間的距離，當所得越小，表二直方圖距離越近，其相似度越高。

2. Histogram Intersection

$$K_{\text{int}}(S,M)=\sum_{b\in B}\min(S_b,M_b) \tag{3.9}$$

當所得值越大，表二直方圖越相似，當二直方圖完全相同時可得最大值。

3. Log Likelihood Ratio

$$L(S,M)=\sum_{b\in B}(S_b\log S_b-S_b\log M_b) \tag{3.10}$$

此值以 0 為中心，向正負兩側以對數函數型態發散，故此值越接近 0，表二直方圖越相似，當二直方圖完全相同時，此值為 0。

透過上述三種常見於直方圖比較的距離定義，以 K 個最近似臨點方法計算音訊區塊間相似度，以達成音訊分類之目的，於下個章節呈現音訊分類結果。

第四章

實作與實驗結果

本章將說明實作之架構與辨識結果。辨識實驗將依三種距離定義透過 K 個最近鄰點演算法做為分類方式，計算本實驗之準確率。此外，為符合真實應用情境，我們將原本音訊加入家庭常見之環境音與雜訊，以驗證其分類能力。而由於本研究強調智慧家庭之應用，因此本章的最後也將討論本系統的計算複雜度與即時性。

4.1 系統實作

本系統以電腦視覺函式庫 EmguCV 進行實作，系統介面如圖 4-1 所示：

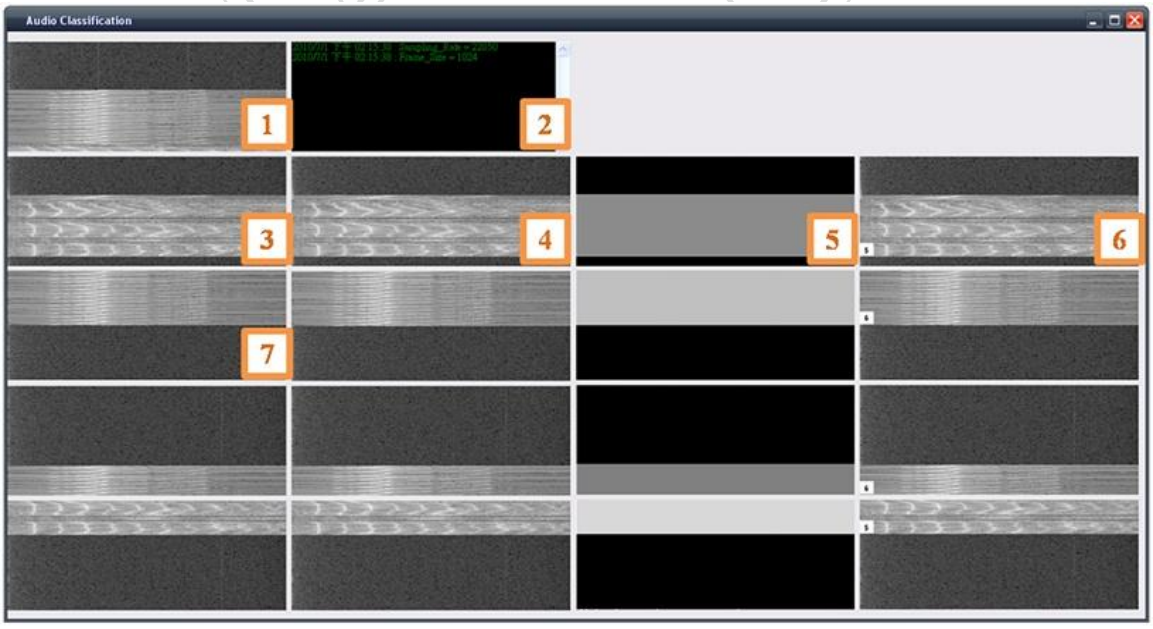


圖 4-1. 系統介面

1：目前測試環境之時間-頻率頻譜圖

2：系統資訊，包括系統時間、取樣率與短時傅利葉轉換之視窗大小

3：第一個音訊訊框，與第二個音訊訊框 7 有四分之一的重疊

4：音訊訊框經過雙向濾波器所得影像

5：音訊起始點偵測之結果

6：辨識結果，以數字表示其分類結果，分類對照表如 表 4-1 所示。

7：第二個音訊訊框，以此類推。

表 4-1. 分類代號

常見於家庭之音訊分類	類別編號
門鈴聲-1	1
門鈴聲-2	2
電話鈴聲-1	3
電話鈴聲-2	4
嬰兒哭聲	5
汽車警報聲	6
水壺汽笛聲	7
火災警報聲	8
不可辨識或無音訊事件	x

4.2 音訊分類

於本實驗採用 8 個分類各 15 個樣本做為訓練資料集，針對常見於家庭之不同情境

設計不同測試資料。

4.2.1 分類結果

測試集為 8 個分類各 7 個測試用資料，以三種不同距離定義加以比較。以下分別列出將 intra-class 視為不同 class 與相同 class 之辨識正確率(accuracy)。

表 4-2. 以 Chi-Square Distance 為距離之辨識結果(視 intra-class 為不同分類)

Class	1	2	3	4	5	6	7	8	平均
1-NN	57.14	100.00	71.43	85.71	100.00	100.00	100.00	42.86	82.14
3-NN	42.86	85.71	85.71	85.71	100.00	100.00	100.00	57.14	82.14
5-NN	57.14	85.71	71.43	85.71	100.00	100.00	100.00	42.86	80.36
平均	52.38	90.48	76.19	85.71	100.00	100.00	100.00	47.62	81.55

單位：百分比

表 4-3. 以 Chi-Square Distance 為距離之辨識結果(視 intra-class 為相同分類)

Class	1+2	3+4	5	6	7	8	平均
1-NN	100.00	100.00	100.00	100.00	100.00	42.86	90.48
3-NN	100.00	92.86	100.00	100.00	100.00	57.14	91.67
5-NN	100.00	100.00	100.00	100.00	100.00	42.86	90.48
平均	100.00	97.62	100.00	100.00	100.00	47.62	90.87

單位：百分比

表 4-4. 以 Histogram Intersection 為距離之辨識結果(視 intra-class 為不同分類)

Class	1	2	3	4	5	6	7	8	平均
1-NN	57.14	100.00	71.43	71.43	100.00	100.00	100.00	57.14	82.14
3-NN	42.86	71.43	71.43	57.14	100.00	100.00	100.00	42.86	73.21
5-NN	42.86	85.71	71.43	57.14	100.00	100.00	100.00	28.57	73.21
平均	47.62	85.71	71.43	61.90	100.00	100.00	100.00	42.86	76.19

單位：百分比

表 4-5. 以 Histogram Intersection 為距離之辨識結果(視 intra-class 為相同分類)

Class	1+2	3+4	5	6	7	8	平均
1-NN	100.00	85.71	100.00	100.00	100.00	57.14	90.48
3-NN	100.00	85.71	100.00	100.00	100.00	42.86	88.10
5-NN	100.00	78.57	100.00	100.00	100.00	28.57	84.52
平均	100.00	83.33	100.00	100.00	100.00	42.86	87.70

單位：百分比

表 4-6. 以 Log-Likelihood Ratio 為距離之辨識結果(視 intra-class 為不同分類)

Class	1	2	3	4	5	6	7	8	平均
1-NN	71.43	100.00	57.14	42.86	100.00	100.00	100.00	28.57	75.00
3-NN	85.71	100.00	71.43	57.14	100.00	100.00	100.00	42.86	82.14
5-NN	85.71	100.00	57.14	71.43	100.00	100.00	100.00	42.86	82.14
平均	80.95	100.00	61.90	57.14	100.00	100.00	100.00	38.10	79.76

單位：百分比

表 4-7. 以 Log-Likelihood Ratio 為距離之辨識結果(視 intra-class 為相同分類)

Class	1+2	3+4	5	6	7	8	平均
1-NN	100.00	71.43	100.00	100.00	100.00	28.57	83.33
3-NN	100.00	78.57	100.00	100.00	100.00	42.86	86.90
5-NN	100.00	85.71	100.00	100.00	100.00	42.86	88.10
平均	100.00	78.57	100.00	100.00	100.00	38.10	86.11

單位：百分比

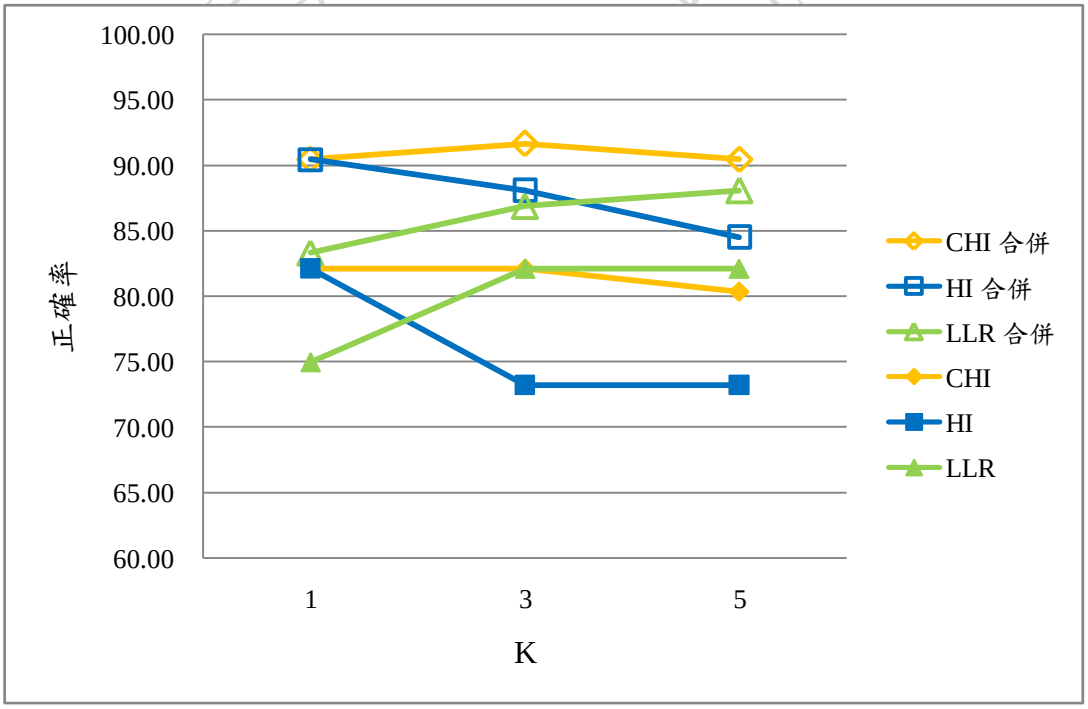


圖 4-2. 三種距離定義於不同 K 值時之正確率

圖中將 intra-class 視為相同 class 的實驗結果縮寫為合併。在本實驗中觀察得知當 K 為 3 時可得到最佳的結果，其中又以 Chi-Square Distance 的辨識正確率最好。

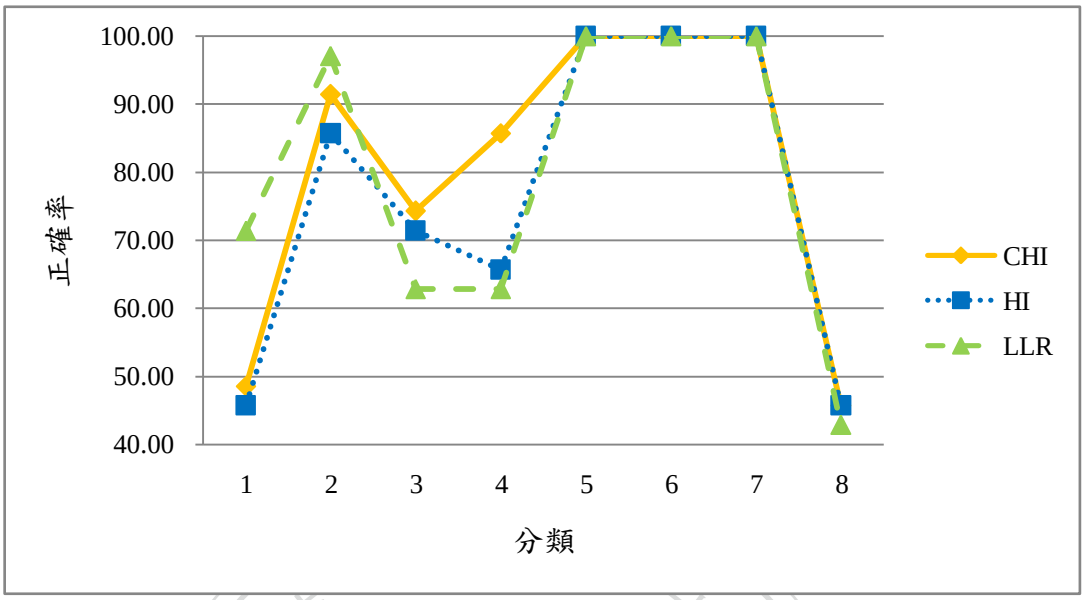


圖 4-3. 八種分類於不同距離定義之辨識正確率比較

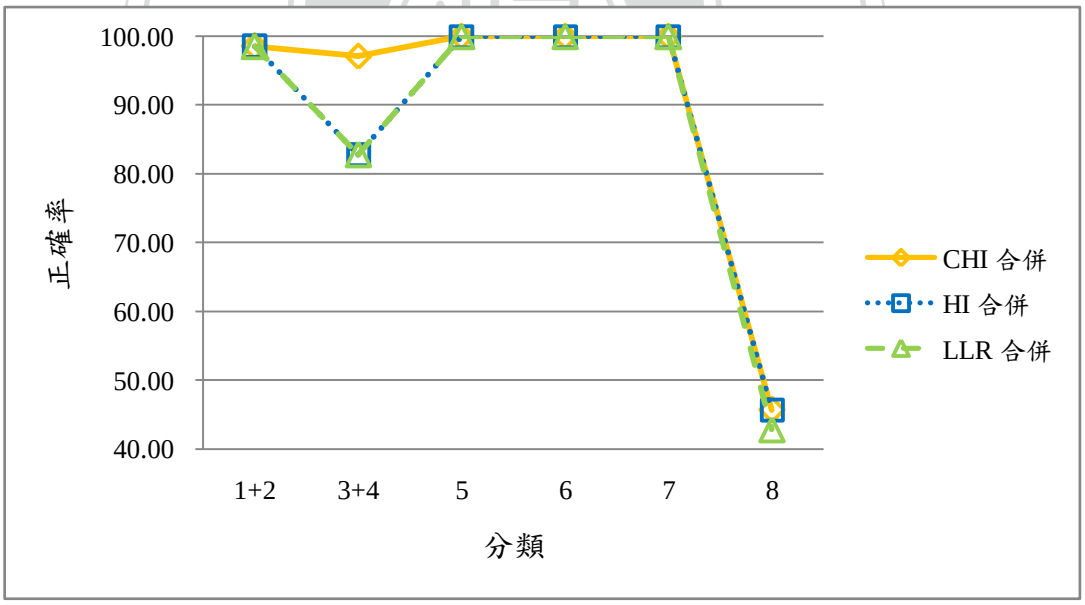


圖 4-4. 六種分類於不同距離定義之辨識正確率比較

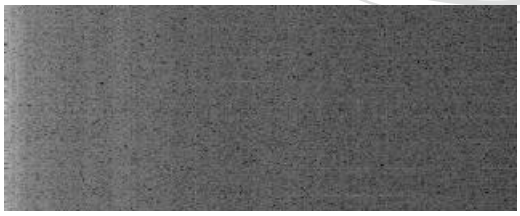
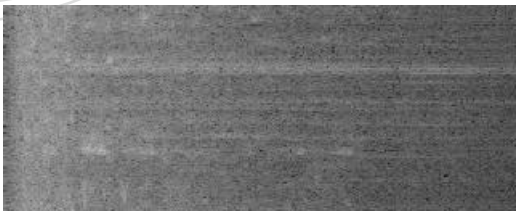
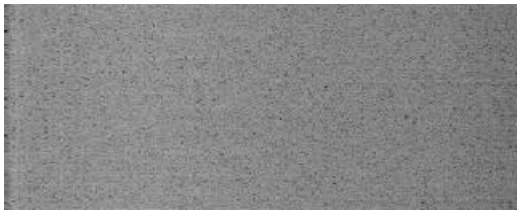
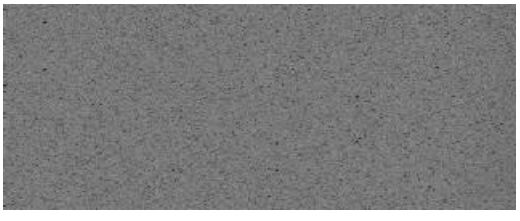
於圖 4-3、圖 4-4 可得，其中辨識正確率較低的分類 1 與分類 8，分類 1(門鈴聲-1)其時間-頻率頻譜圖與分類 2(門鈴聲-2)相當相似，因系統應用情境之故，對系統而言，分類 1(門鈴聲-1)與分類 2(門鈴聲)-2 所需做出之反應相同，故我們可以將這兩個分類視

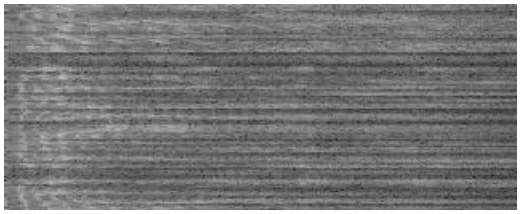
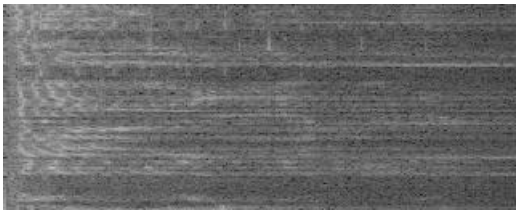
為同一個分類。而分類8(火災警報聲)則因為其機械發聲構造與分類3(電話鈴聲-1)相同，故產生此辨識結果。而分類 5(嬰兒哭聲)、分類 6(汽車警報聲)、分類 7(水壺汽笛聲)之高辨識正確率則是起因於其時間-頻率頻譜圖所呈現的樣式與其他分類有著非常大的差異，擁有相當高的辨識度。

4.2.2 情境環境聲

我們企圖將此研究應用於智慧家庭環境，將家庭常見之情境環境聲加入實驗，用以模擬我們的系統實際應用於測試環境時，是否對環境聲有其抗噪力並且仍能有效的對音訊事件加以分類。如表 4-8 所示，為本研究中所定義常見於家庭之情境環境聲，以下結果呈現當情境環境聲與音訊事件同時發生之分類結果。測試資料同為上個實驗中之 8 個分類各 7 個測試樣本，於 6 種情境環境聲中的抗噪實驗。

表 4-8. 常見於家庭之情境環境聲及其時間-頻率頻譜圖

情境環境聲分類	
冷氣空調聲	人群聲
	
高斯雜訊	雨聲
	

電視-新聞播報	電視-談話性節目
	

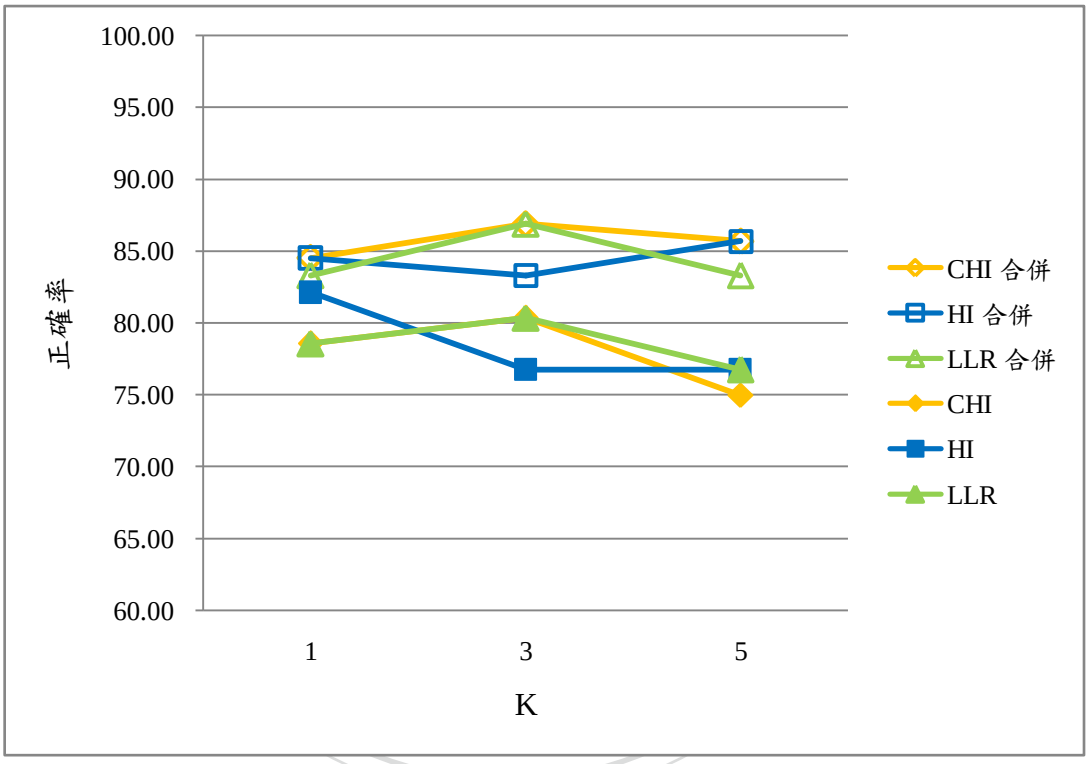


圖 4-5. 於冷氣空調聲中之辨識正確率比較

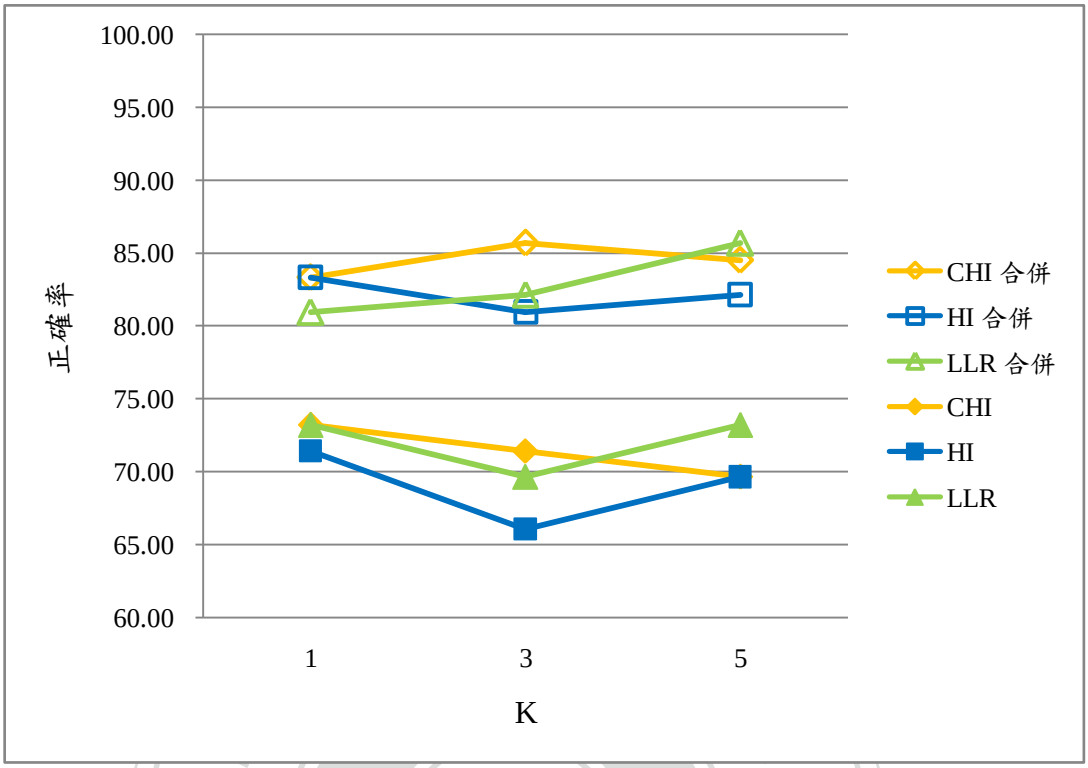


圖 4-6. 於人群聲中之辨識正確率比較

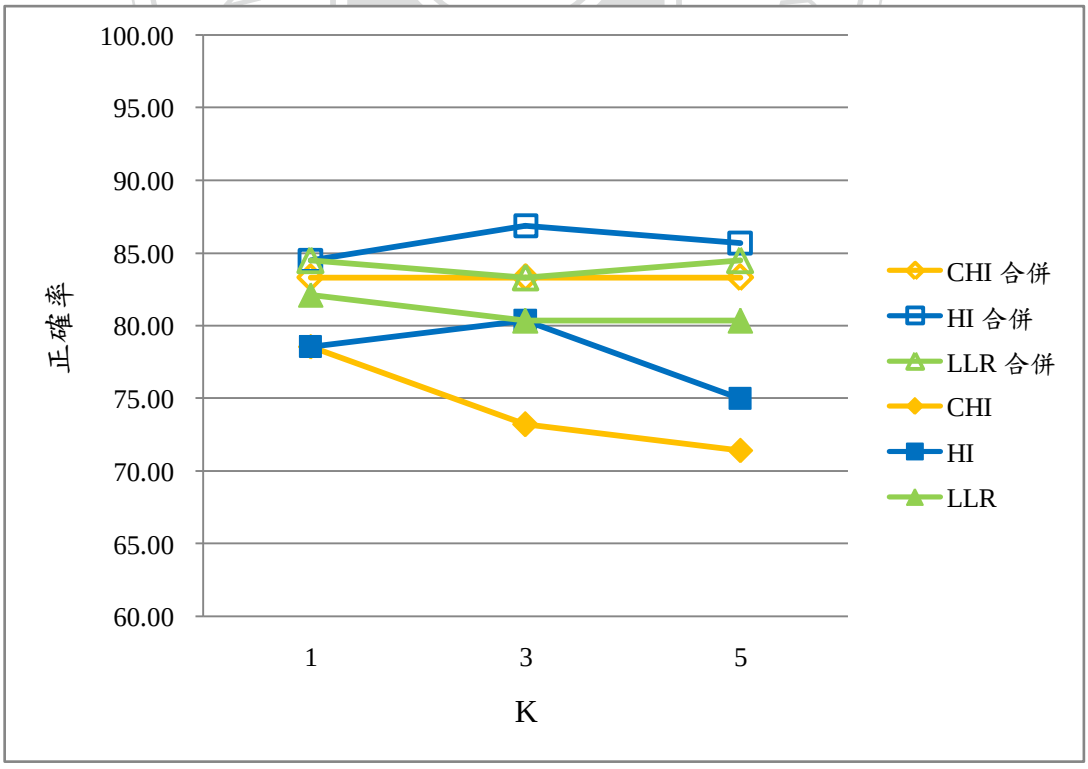


圖 4-7. 於高斯雜訊中之辨識正確率比較

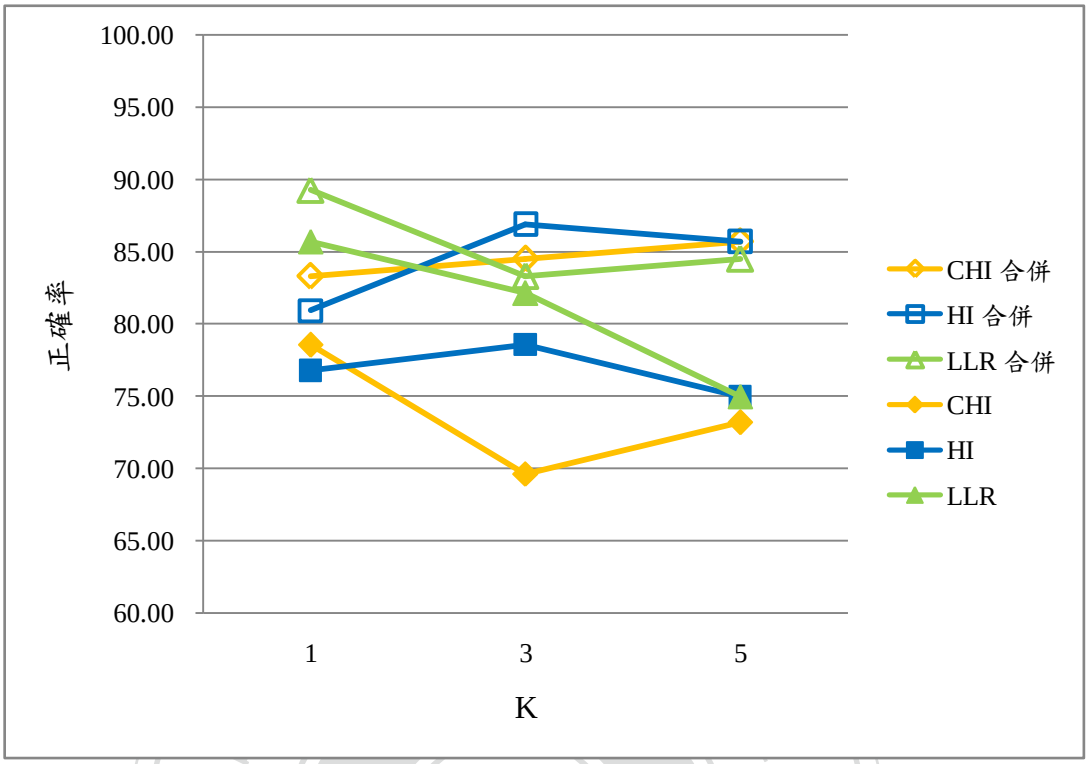


圖 4-8. 於雨聲中之辨識正確率比較

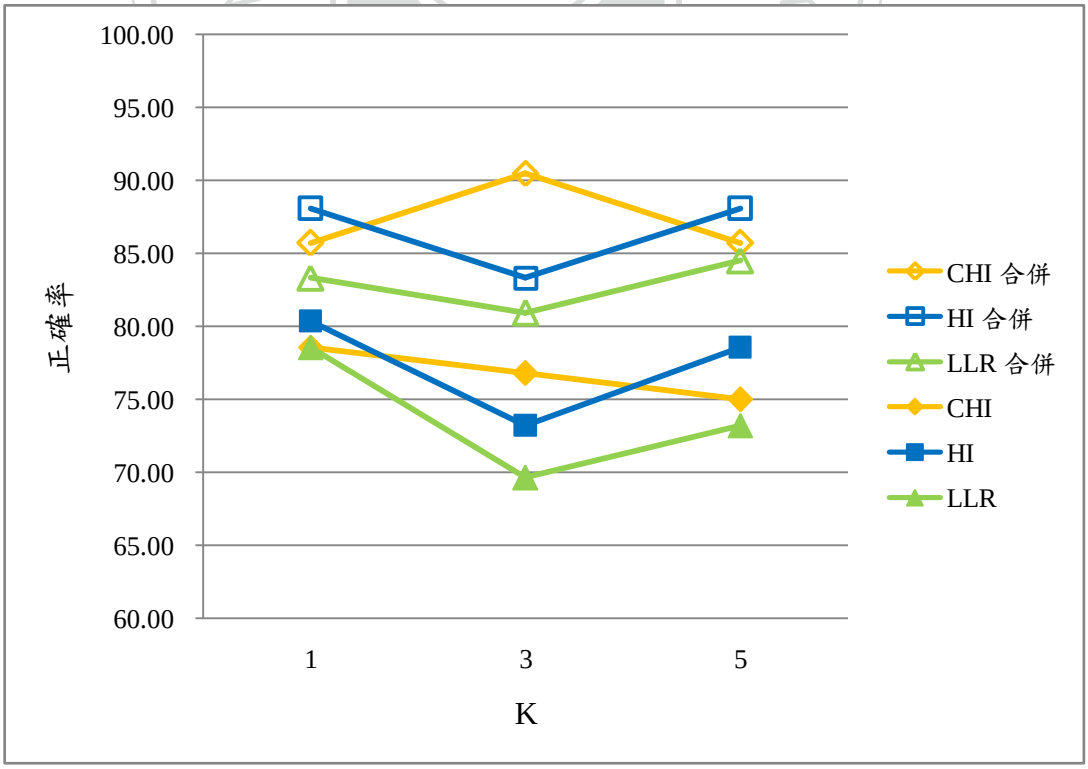


圖 4-9. 於電視情境-新聞聲中之辨識正確率比較

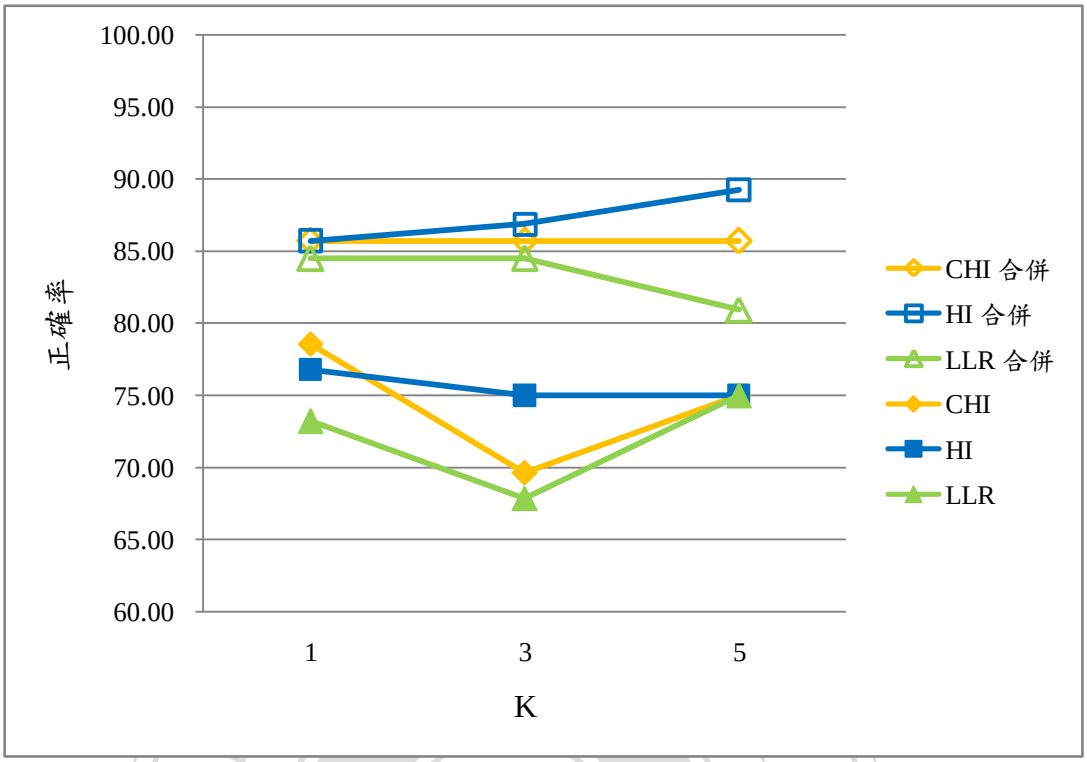


圖 4-10. 於電視情境-談話性節目聲中之辨識正確率比較

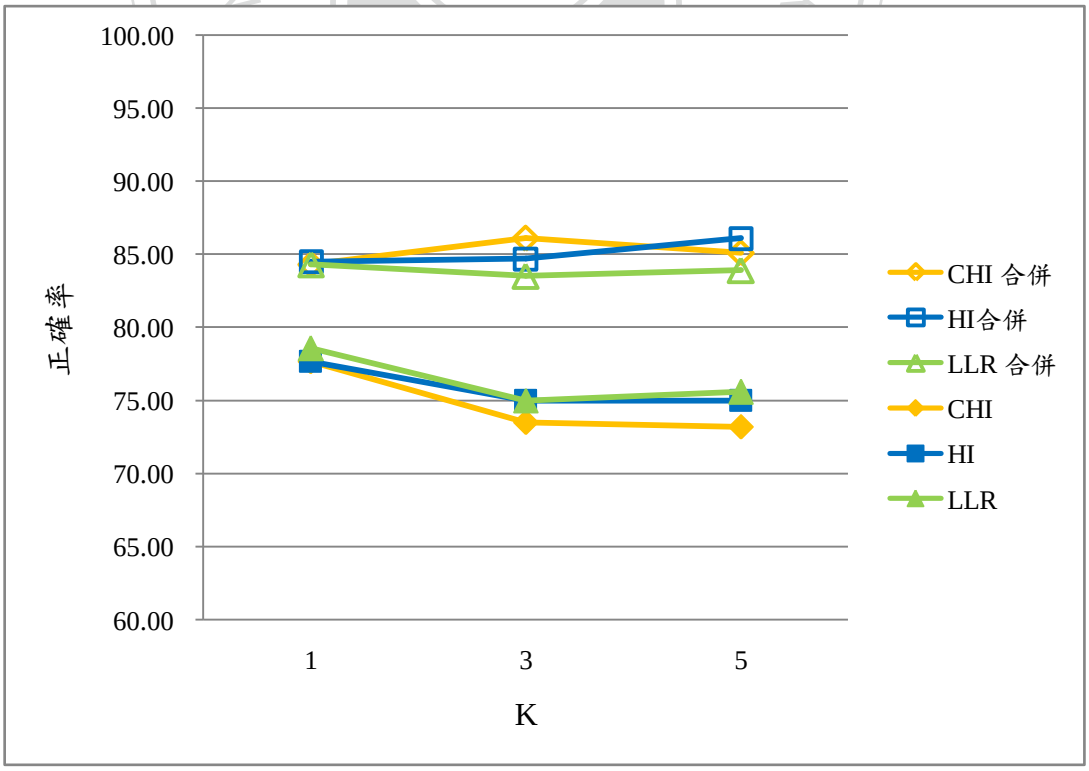


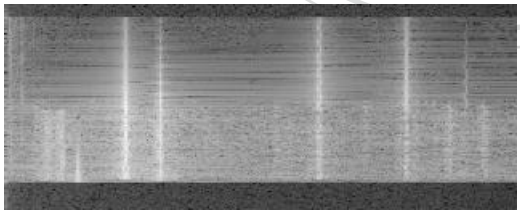
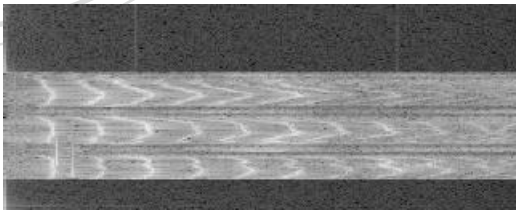
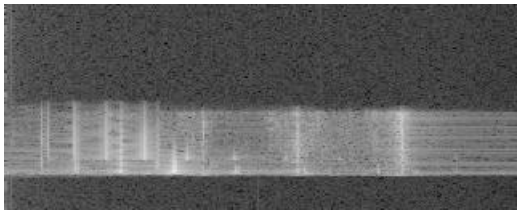
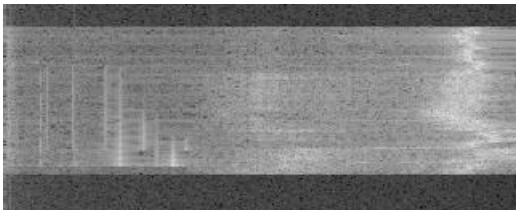
圖 4-11. 加入情境環境聲之辨識正確率平均比較

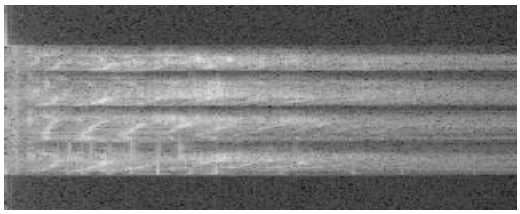
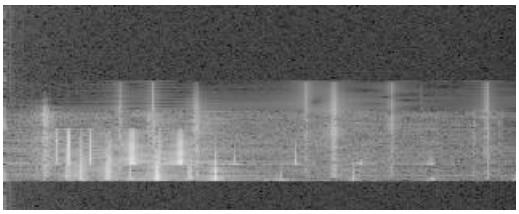
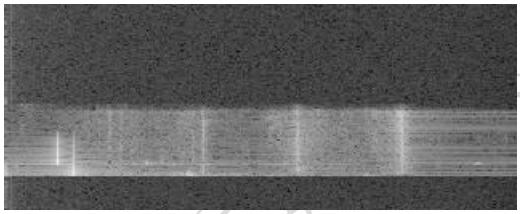
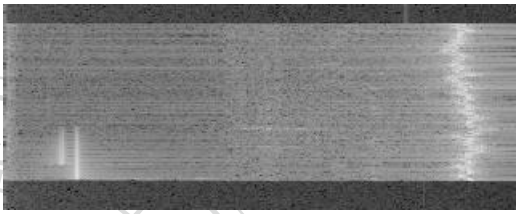
由以上實驗結果可得，當音訊事件發生於情境環境聲中，系統對於事件的辨識能力仍能維持在一定的程度。未加入環境聲與加入環境聲後約為 5~10%之差異，但整體仍維持在 70%以上。其中圖 4-11 可以觀察到一個有趣的現象，因為環境聲的影響，產生音訊事件結構上的變化，使得未合併之辨識結果會隨著 K 值的遞增，系統卻將結構變化過後的音訊事件辨識為另一個事件，造成 K 為 1 時，可以得到一個較佳的辨識結果，而其他情況仍維持 K 為 3 時有較穩定之辨識正確率。

4.2.3 音訊事件同時發生

由於音訊事件可能與其他音訊事件同時發生，我們改寫系統測試對於此情境的辨識能力，系統針對音訊事件其頻帶分布的不同，圈選其主要音訊事件結構用以辨識，並設計此實驗以測試系統對於同時發生之音訊事件的正確率。測試資料為 15 組兩個音訊事件同時發生之組合，每一組各 5 個樣本。如表 4-9 所列其中幾個樣本之時間-頻率頻譜圖。

表 4-9. 同時發生之音訊事件樣本

門鈴聲-1 + 電話鈴聲-1	門鈴聲-1 + 嬰兒哭聲
	
門鈴聲-2 + 火災警報聲	門鈴聲-2 + 水壺汽笛聲
	

門鈴聲-2 + 嬰兒哭聲	門鈴聲-2 + 電話鈴聲-1
	
門鈴聲-1 + 火災警報聲	門鈴-1 + 水壺汽笛聲
	

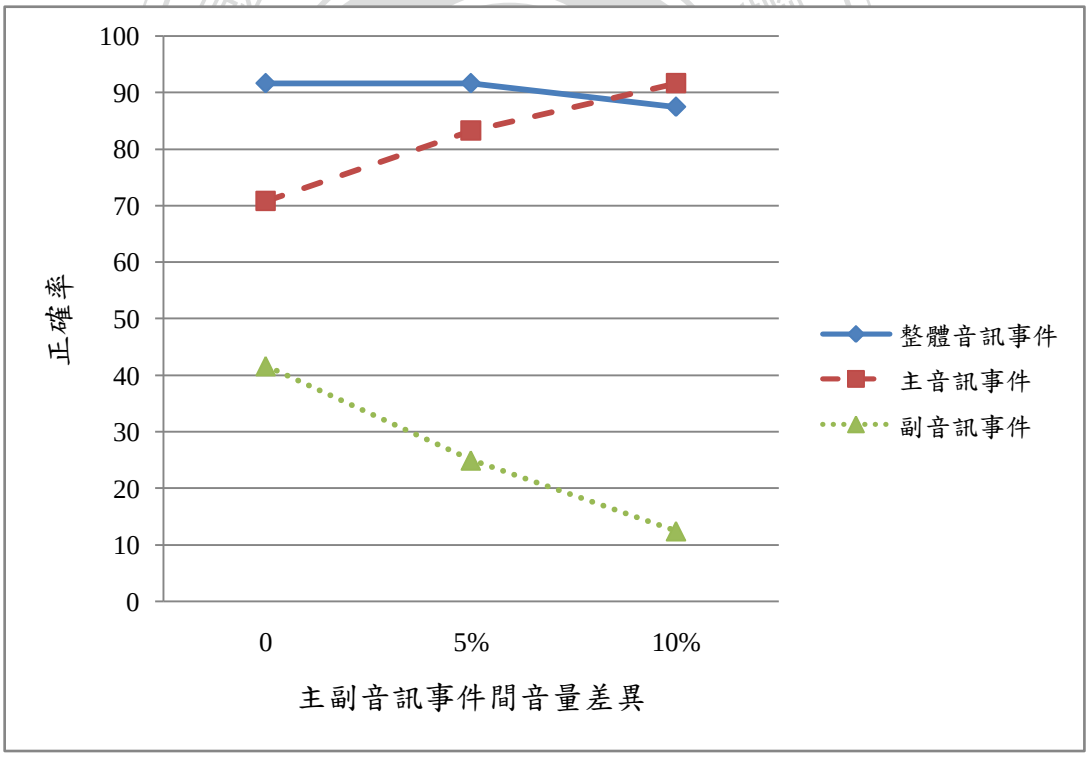


圖 4-12. 主副音訊事件音量差異之辨識正確率比較

我們可以依照音訊事件所產生時間-頻率頻譜圖上之紋理結構將音訊事件分為主音

訊事件與副音訊事件，於上表可得，兩個疊加所得之整體頻譜圖仍能有效的得到正確的分類，雖然副音訊事件則因音量的遞減而難以辨識，但主音訊事件亦隨著主副音訊事件之音量差異而大幅增加，更能有效的辨識出來。由於當音訊事件分布於相同頻帶上時，因為其結構難以區分，造成辨識結果不佳，故我們於此設計一實驗，當兩音訊事件頻帶分布有明顯差異時，系統是否能有效的將兩個音訊事件區分開來。測試資料集為 10 種組合，各 3 個測試樣本。表 4-10 為其中測試資料集之範例。

表 4-10. 頻帶分布差異明顯的音訊組合樣本

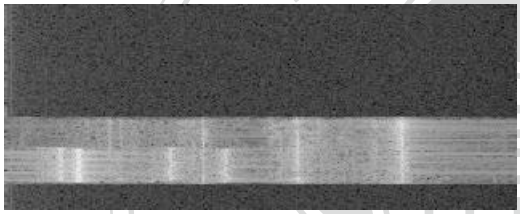
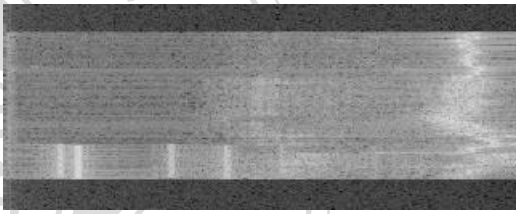
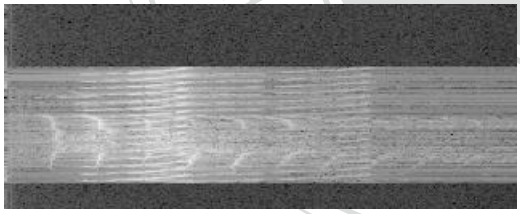
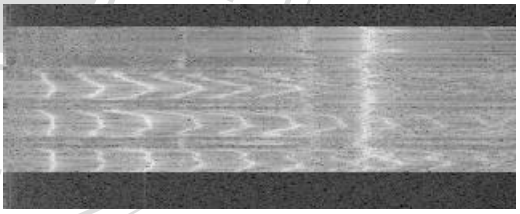
電話鈴聲-2 + 火災警報聲	電話鈴聲-2 + 水壺汽笛聲
	
嬰兒哭聲 + 汽車警報聲	嬰兒哭聲 + 水壺汽笛聲
	

表 4-11. 頻帶分布差異明顯音訊事件組合之辨識正確率

	整體音訊事件	音訊事件 1	音訊事件 2	平均
1NN	100.00	80.00	60.00	80.00
3NN	100.00	73.33	66.67	80.00
5NN	100.00	66.67	80.00	82.22
平均	100.00	73.33	68.89	80.74

單位：百分比

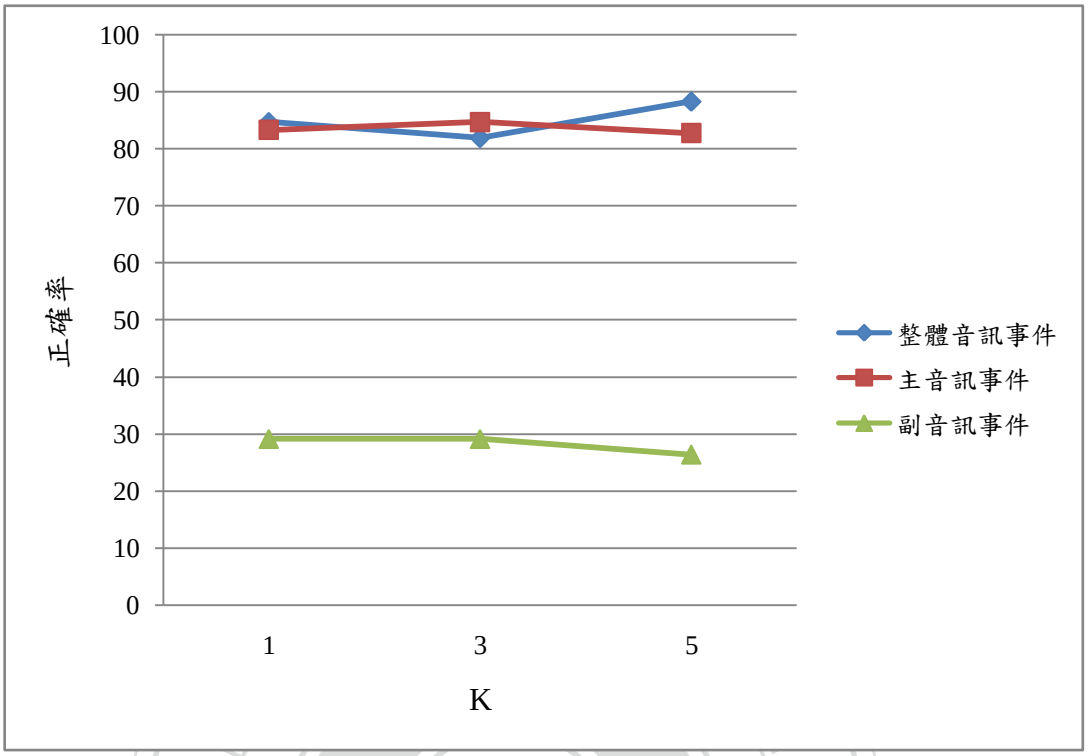


圖 4-13. 音訊事件於不同 K 值時之辨識正確率比較

於表 4-11 可得，當兩個音訊事件其頻帶分布有著較為明顯差異時，系統便能依其頻帶分布，將二者區分開來，整體辨識正確率約為 70%左右。圖 4-13 則針對 K 值討論，由表中可得 K 值於此實驗中並無明顯之差異。

4.2.4 音訊事件發生於不同音場位置

於智慧家庭的測試環境中，由於音訊事件與麥克風的相對位置均有可能因擺放與移動造成改變，於此，我們設計一個實驗來測試系統對於空間關係所造成辨識結果之影響。如下圖所示，我們針對當發聲音源位於麥克風的不同角度進行測試，距離麥克風均為 10 公分，測試資料集為 8 個分類各 7 個樣本。

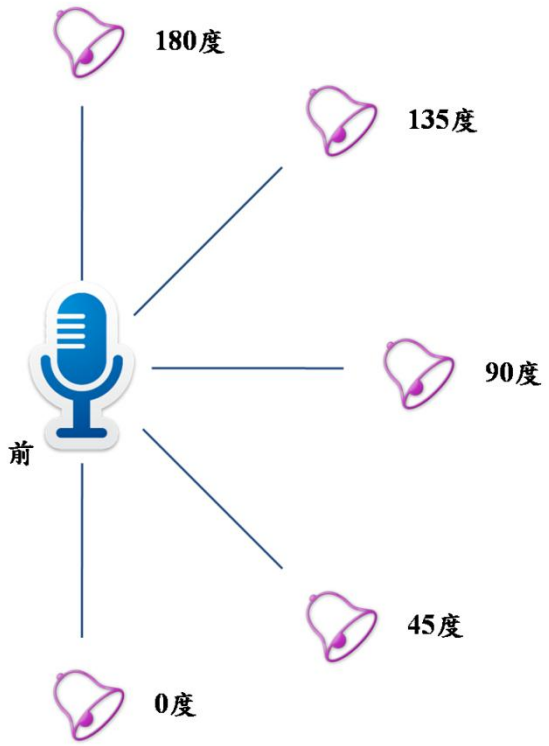


圖 4-14. 音訊事件發生於收音裝置不同方位

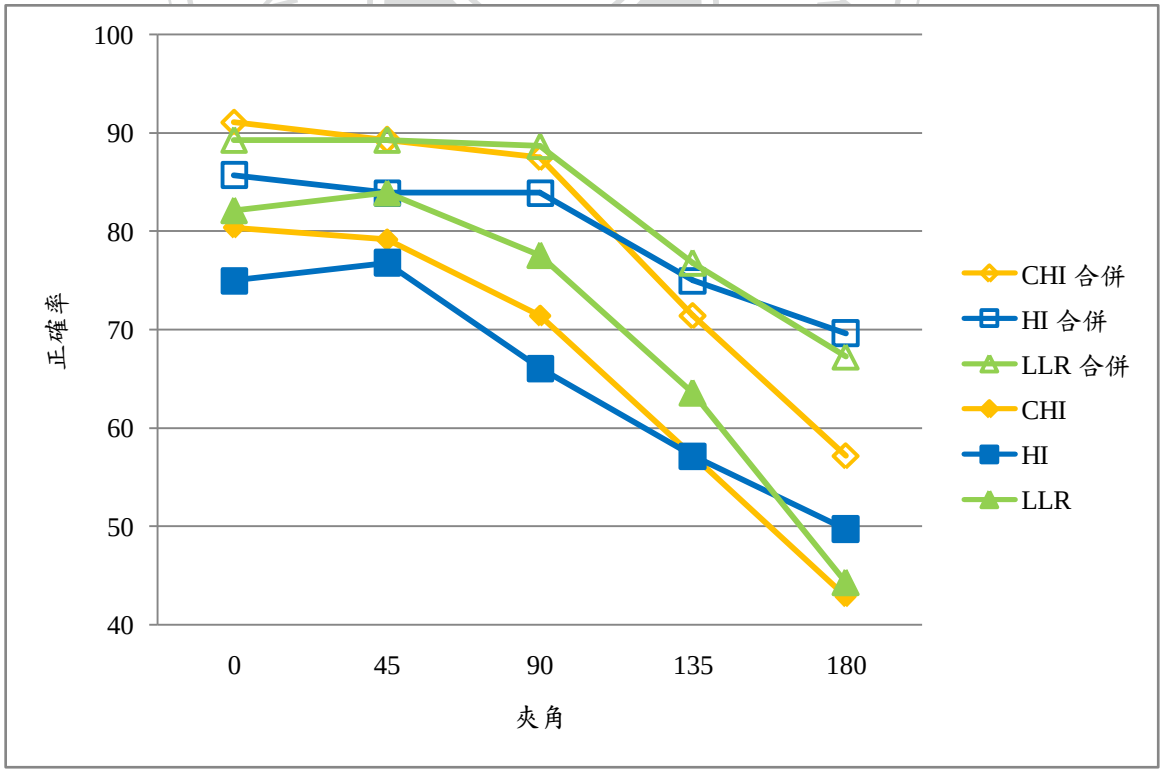


圖 4-15. 音訊事件發生於收音裝置不同方位的辨識結果比較

由上圖可得，由於收音裝置為半指向性麥克風，當音訊事件發生於麥克風的背面時，辨識結果有明顯下降的趨勢，音訊事件發生於麥克風的正面時，甚至側邊亦仍能維持辨識能力。

我們也針對音訊事件與收音裝置間的距離改變進行測試，設計一實驗以驗證系統對於距離變化時產生的辨識結果差異。如下圖所示，音訊事件發生方位均為收音裝置正前方，距離為 10 公分、1 公尺與 5 公尺處，測試資料集為 8 個分類各 7 個樣本。

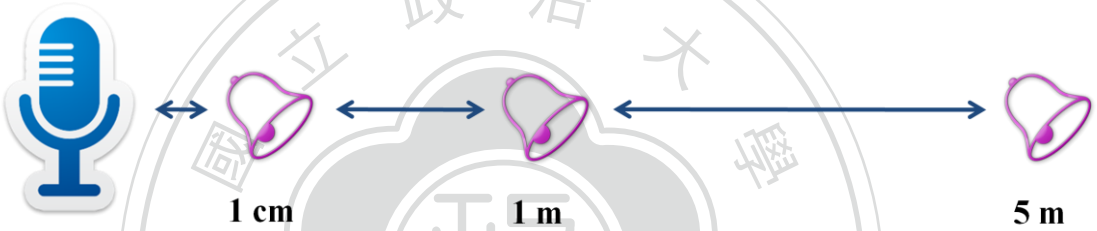


圖 4-16. 音訊事件發生於收音裝置不同距離處

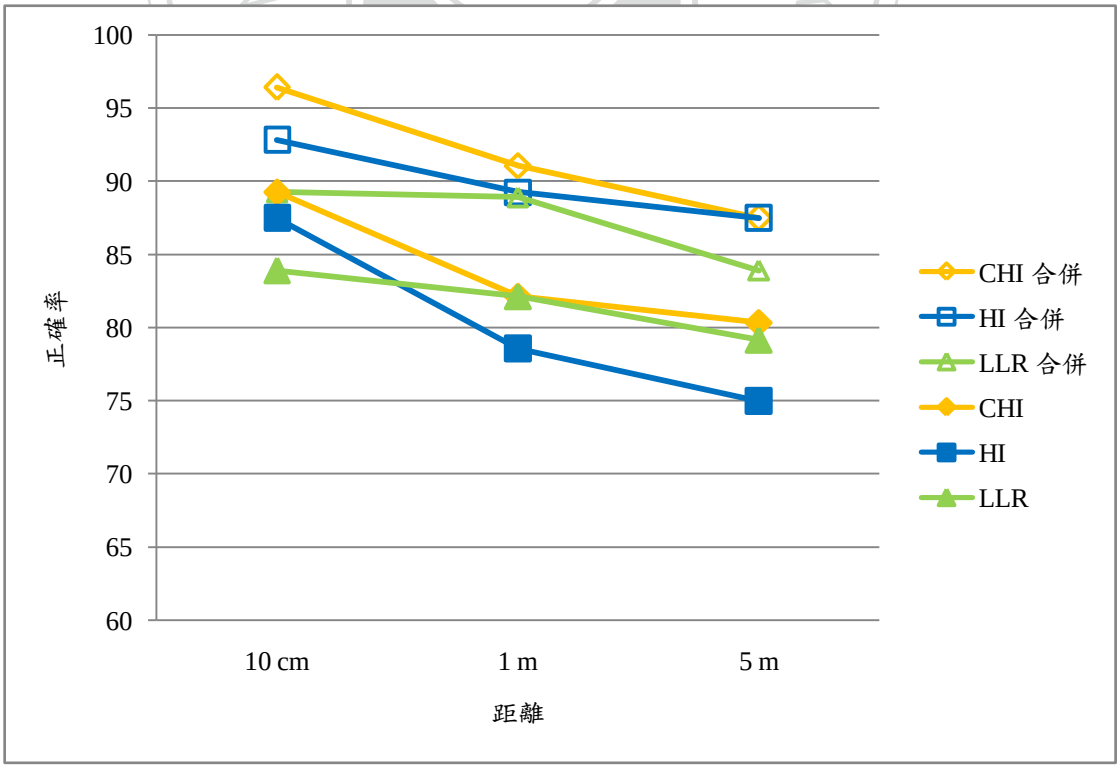


圖 4-17. 音訊事件發生於收音裝置不同距離處之比較結果

由上圖可得，由於系統設計是針對整體環境變化，故音訊事件並未隨著距離的增加造成辨識正確率有大幅的改變，雖辨識率有下降的趨勢，但在測試環境中合理的距離內，整體辨識正確率仍能維持於 95%到 80%之間。

4.3 即時性驗證

針對串流式音訊分類需要更為嚴謹的討論其即時性，經過實驗所得以下結果，反應時間為當音訊事件影像進入系統，經過各種運算顯示其辨識結果的總耗時，隨著音訊訊框中所偵測之音訊事件數量不同而有所增減，如下表所示，平均系統處理時間約為 0.7 秒至 1.2 秒之間。由於每個音訊訊框約 6 秒，彼此間有約 1.5 秒重疊，故最差情況僅發生於系統起始時，第一個音訊訊框所需之 6 秒再加上系統處理時間，約為 7 秒。初始訊框以外的最差情況則是 4.5 秒加上系統處理時間，約為 5.5 秒。而最佳情況則同為 200 毫秒加反應時間，200 毫秒為人耳後遮蔽效應之下限，故約為 1.2 秒。

表 4-12. 即時性驗證之實驗結果 (電話鈴聲-2)

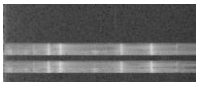
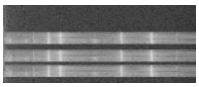
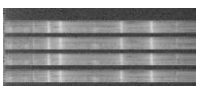
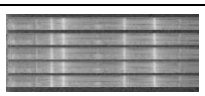
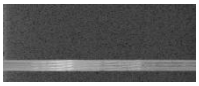
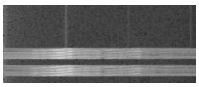
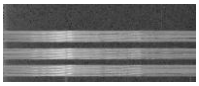
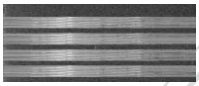
音訊事件	事件偵測數量	反應時間	雙向濾波器耗時
	0	703 ms	656 ms
	1	797 ms	641 ms
	2	875 ms	656 ms
	3	922 ms	656 ms
	4	984 ms	641 ms
	5	1047 ms	656 ms

表 4-13. 即時性驗證之實驗結果 (汽車警報聲)

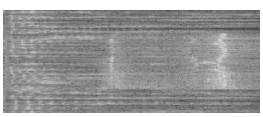
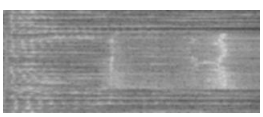
音訊事件	事件偵測數量	反應時間	雙向濾波器耗時
	0	719 ms	672 ms
	1	842 ms	656 ms
	2	922 ms	656 ms
	3	1031 ms	656 ms
	4	1140 ms	656 ms
	5	1250 ms	656 ms

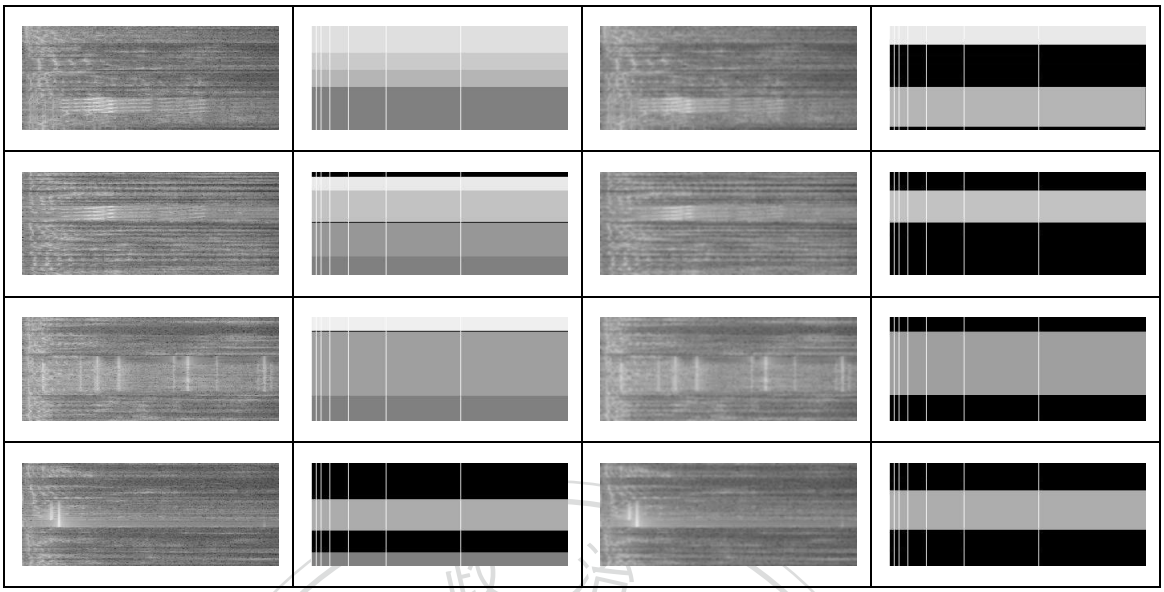
於實驗結果可以發現，雙向濾波器為反應時間之瓶頸，如果為了時間考量是否可以考慮放棄雙向濾波器，針對這個問題設計以下實驗，實驗結果如下表所示，是否使用雙向濾波器對於起始點偵測所造成之結果比較。

表 4-14. 即時性驗證

實驗樣本	反應時間	雙向濾波器耗時
638	1176 ms	656 ms

表 4-15. 有無雙向濾波器之起始點偵測結果比較

音訊事件	起始點偵測	雙向濾波器	起始點偵測
			



當測試環境中僅有音訊事件時，系統不需經過雙向濾波器亦能有效的進行起始點偵測，而當測試環境中有環境聲時，如上表所示，對音訊訊框直接進行起始點偵測將會造成過度切割(over segmentation)，尤其於電視情境聲中新聞播報與談話性節目特別明顯，對測試結果造成影響，由此可見雙向濾波器仍有其必要性。

第五章

結論與後續研究方向

本研究主要目的在實現串流式音訊分類於家庭中常見之音訊事件，有別於以往之研究，我們嘗試以影像處理與圖型辨識之技術為基礎，實現即時音訊分類。透過聽覺心理學中，對於人類聽覺感受特性的了解，設計影像處理技術以期達成音訊處理之結果。於實驗結果可得，我們成功的將音訊處理中十分重要的起始點偵測對應至影像處理之技術，將短時傅利葉轉換作為前處理，以雙向濾波器簡化時間-頻率頻譜圖並降噪，最後透過聽覺閾值曲線定義波峰選取之方式，以影像處理的技術達成音訊起始點偵測。

面對"串流式"處理的限制，必須對時間有著更為嚴謹之考量，也就是要發展快速且有效的處理方式，針對音訊事件影像中的音訊區塊，加以描述與分析。於本研究中，我們針對音訊事件發生與結束的特性，設計雙閾值設定，利用區塊偵測對此音訊區塊加以標記，為求速度上之考量，採用區域二元化圖型以描述此音訊區塊，最後透過 K 個最近鄰點計算，實現即時之音訊分類。

實驗結果顯示，對於各種常見於家庭環境中之聲音，系統擁有相當不錯的辨識能力，將 intra-class 視為不同 class 的分類正確率可達 80%，而將 intra-class 視為相同 class 的分類正確率更可提高至 90%以上。設計環境聲與多個音訊事件同時發生之實驗，以模擬真實家庭環境可能發生之情境，系統仍能相當程度的辨識出各種聲音並正確地加以分類，在不同情境環境聲下，辨識正確率仍能有 85%的表現。當兩個音訊事件同時發生時，對個別事件之辨識正確率亦可有 70%以上之辨識正確率。應用於擁有較高容忍程度之家庭

情境中，系統有效的對各種音訊事件分類。針對即時性的驗證，系統反應時間亦有相當不錯的表現，音訊訊框之處理時間約為 1 秒鐘，因最差情況僅發生於系統初始之時，故系統對於音訊事件發生之平均反應時間約為 1 秒至 5.5 秒之間，應用於智慧家庭的各種情境中，均能有效於合理之時間限制內有所回應。

然而為因應測試環境的變化，更為有效的音源區分是未來最為重要的努力目標之一，透過多個指向性麥克風或麥克風陣列，企圖將多個同時發生之音源加以定位，可有效的以空間位置為音源區分之依據，將發生於相同時間甚至發生於相同頻帶上之音訊事件加以區分。另一方面，透過收集更多的音訊訓練資料建立更為完整之資料庫，或以背景建模之方式針對環境聲加以排除，均可將本研究之系統推展至更大的測試環境，以實現更為完善之電腦聽覺技術。

參考文獻

[1] A. S. Bregman. “Auditory Scene Analysis”. The Perceptual Organization of Sound. Cambridge, MA: MIT Press, 1990.

[2] D. Rosenthal and H. Okuno, Eds.. “Computational Auditory Scene Analysis”. *Lawrence Erlbaum Associates*, 1998.

[3] D. Ellis. “Prediction-Driven Computational Auditory Scene Analysis”. *Ph.D. thesis, MIT*, 1996.

[4] 王小川，「語音訊號處理」，全華股份有限公司，2007年4月。

[5] 張智星，「音訊處理與辨識」，http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/ [retrieved July 2009]

[6] Wen-Hung Liao and Yi-Syuan Su. “Analysis and classification of human sounds”. *Master’s thesis, Department of Computer Science National Chengchi University*, July 2006.

[7] Yan Ke, Derek Hoiem and Rahul Sukthankar. “Computer Vision For Music Identification”. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[8] J. Haitsma and T. Kalker. “A Highly Robust Audio Fingerprinting System”. in *Proceedings of International Conference on Music Information Retrieval*, 2002.

[9] G. Hu and D.L. Wang. “Auditory Segmentation Based on Event Detection”. *In ISCA Tutorial and Research Workshop on Stat. and Percept. Audio Process.*, 2004.

[10] S.H. Srinivasan. “Auditory blobs”. *in IEEE ICASSP '04*, vol. 4, pp. iv–313 – iv–316, 2004.

[11] Valerie Pierson and Nadine Martin. “Comparison of Shape Descriptors For Feature Extraction of A Time- Frequency Image”. *CEPHAG-ENSJEG - BP 46 - 38402 ST-MARTIN-D’HERES C&Ex FRANCE.*

[12] Ruohua Zhou, Marco Mattavelli, and Giorgio Zoia. “Music Onset Detection Based On Resonator Time Frequency Image”. *IEEE Transactions On Audio, Speech, And Language Processing, Vol. 16, No. 8*, 2008.

[13] 王駿發，「多媒體影音檢索系統」，http://web1.nsc.gov.tw/ct.aspx?xItem=8460&ctNode=40&mp=1[retrieved July 2009]

[14] D. Li, I. Sethi, N. Dimitrova, and T. McGee. “Classification Of General Audio Data For Content-Based Retrieval”. *Pattern Recognition Letters,* vol. 22(5), pp. 533–544, 2001.

[15] Zhu Liu, Yao Wang and Tsuhan Chen. “Audio Feature Extraction And Analysis For Scene Segmentation And Classification”. Polytechnic University, Brooklyn, NY 11201, Carnegie Mellon University, Pittsburgh, PA 15213.

[16] Silvia Allegro, Michael Büchler and Stefan Launer. “Automatic Sound Classification Inspired By Auditory Scene Analysis”. Signal Processing Department, Phonak AG, Switzerland Department of Otorhinolaryngology, University Hospital Zurich, Switzerland.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution Gray-Scale And Rotation Invariant Texture Classification With Local Binary Patterns”. *IEEE Trans. On Pattern Analysis and Machine Intelligence,* vol. 24, pp. 971-987, 2002.

[18] L. Cohen. “Time-Frequency Analysis”. *Prentice Hall PTR, Englewood Cliffs* 1995.

[19] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. Sandler. “A Tutorial On Onset Detection In Music Signals”. *IEEE Transactions on Speech and Audio Processing,* 2005.

[20] S. Paris. “A Gentle Introduction To Bilateral Filtering And Its Applications”. In *ACM SIGGRAPH 2007 courses,* Course 13.

[21] V. Aurich and J.Weule. “Non-Linear Gaussian Filters Performing Edge Preserving Diffusion”. in *Proceedings of the DAGM Symposium,* pp. 538–545, 1995.

[22] C. Tomasi and R. Manduchi. “Bilateral Filtering For Gray And Color Images”. in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 839–846, 1998.

[23] F. Durand and J. Dorsey. “Fast Bilateral Filtering For The Display Of Highdynamic-Range Images”. in *Proceedings of the ACM SIGGRAPH conference*, 2002.

[24] Paul Masri and Andrew Bateman. “Improved Modeling Of Attack Transients In Music Analysis-Resynthesis”. in *Proceeding of International Computer Music Conference*, 1996.

[25] M. Goto and Y. Muraoka. “Beat Tracking Based On Multiple-Agent Architecture — A Real-Time Beat Tracking System For Audio Signals —” in *ICMAS-96*, pp. 103–110, 1996.

[26] H. Freeman, “Techniques For The Digital Computer Analysis Of Chain-Encoded Arbitrary Plane Curves”. in: *Proc. Nat. Electronics Conf.*, 1961, pp. 421-432.

[27] E. Bruce Goldstein. *Sensation and Perception.* Wadsworth Publishing Co., Belmont, California, 1980.

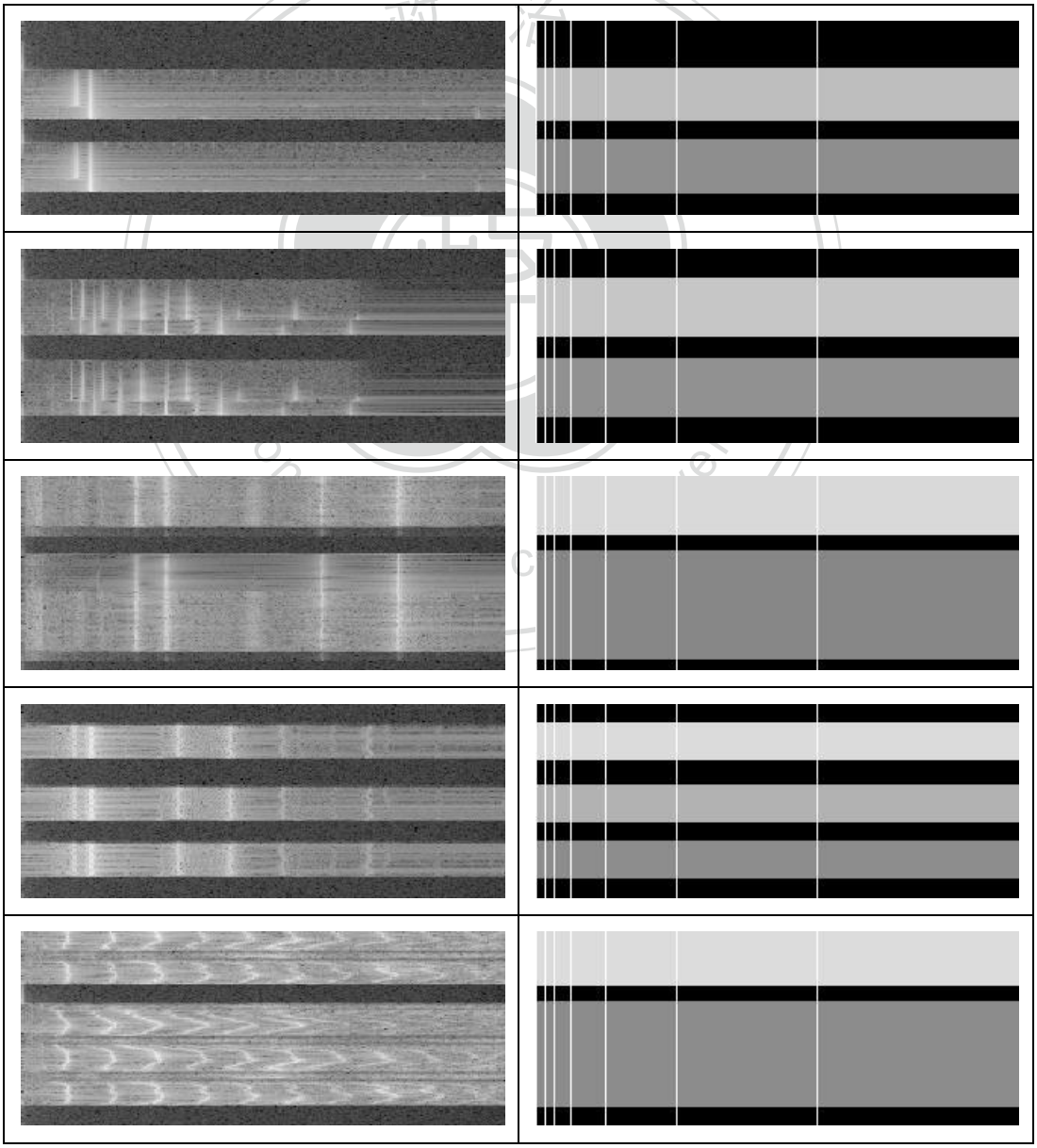
[28] Y. He and A. Kundu. “2-D Shape Classification Using Hidden Markov Model”. *IEEE Trans. Pat-tern Analysis and Machine Intelligence*, 13(1991) 1172-1184.

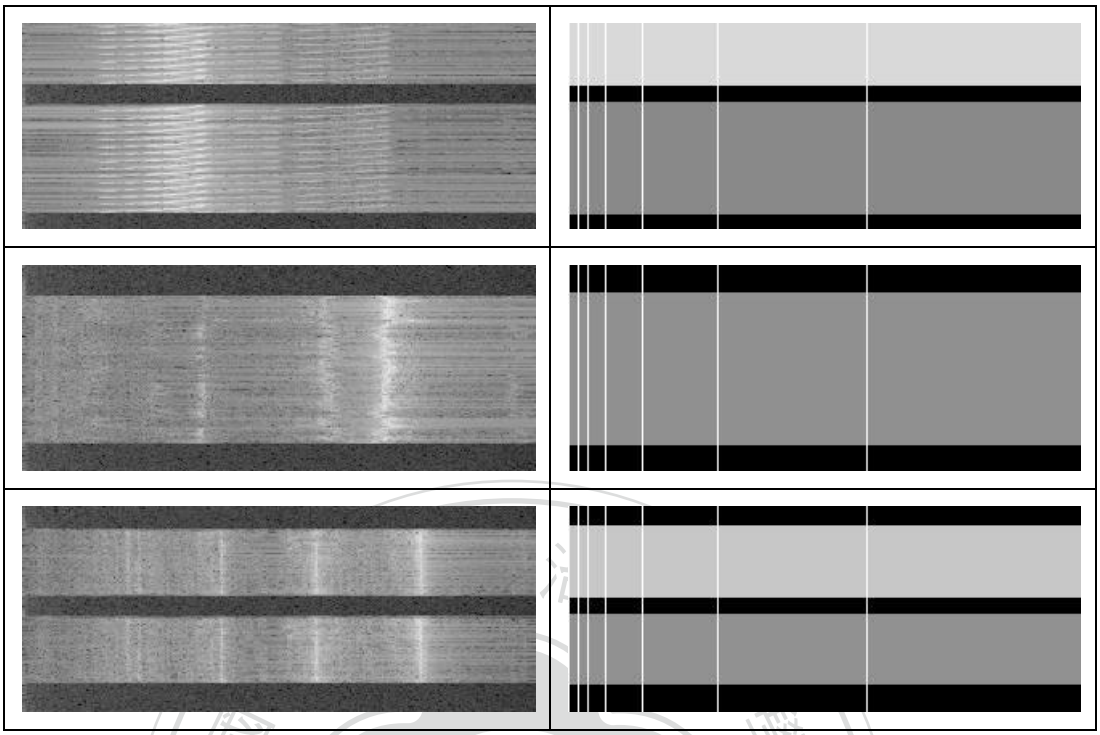
[29] Xu Qing, Yang Jie and Ding Siyi. “Texture Segmentation Using LBP Embedded Region Competition”. *Inst. of Image Processing & Pattern Recognition.*

附錄

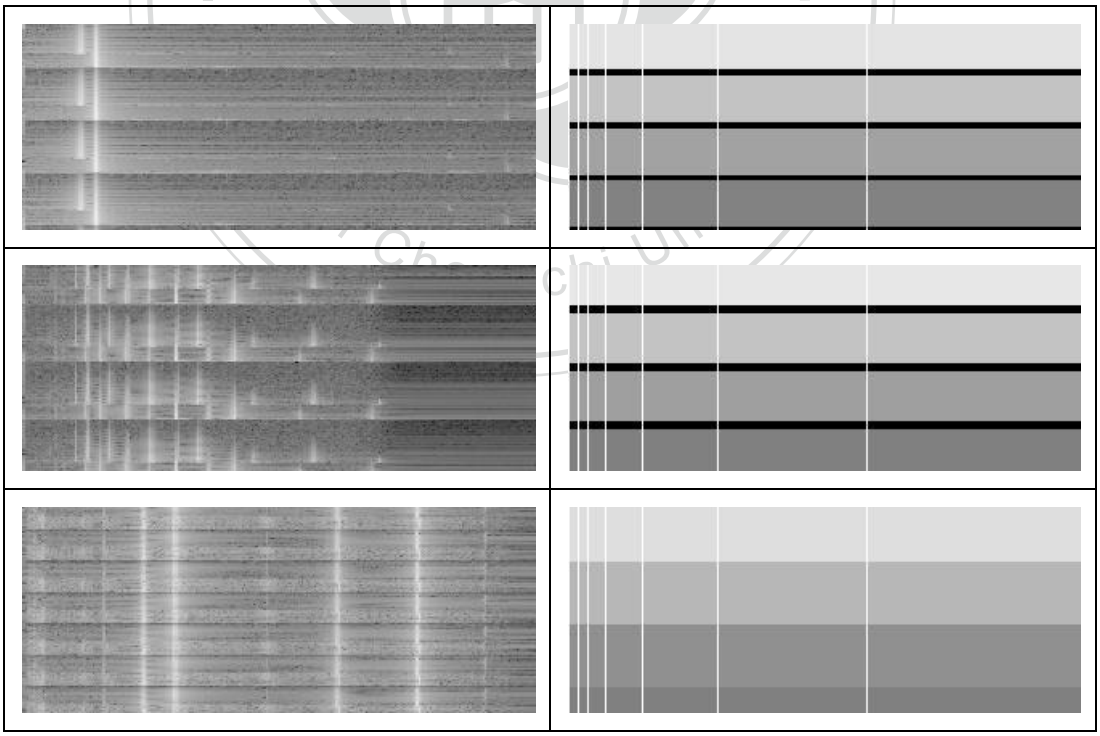
A. 音訊起始點偵測-音訊事件發生於環境聲中

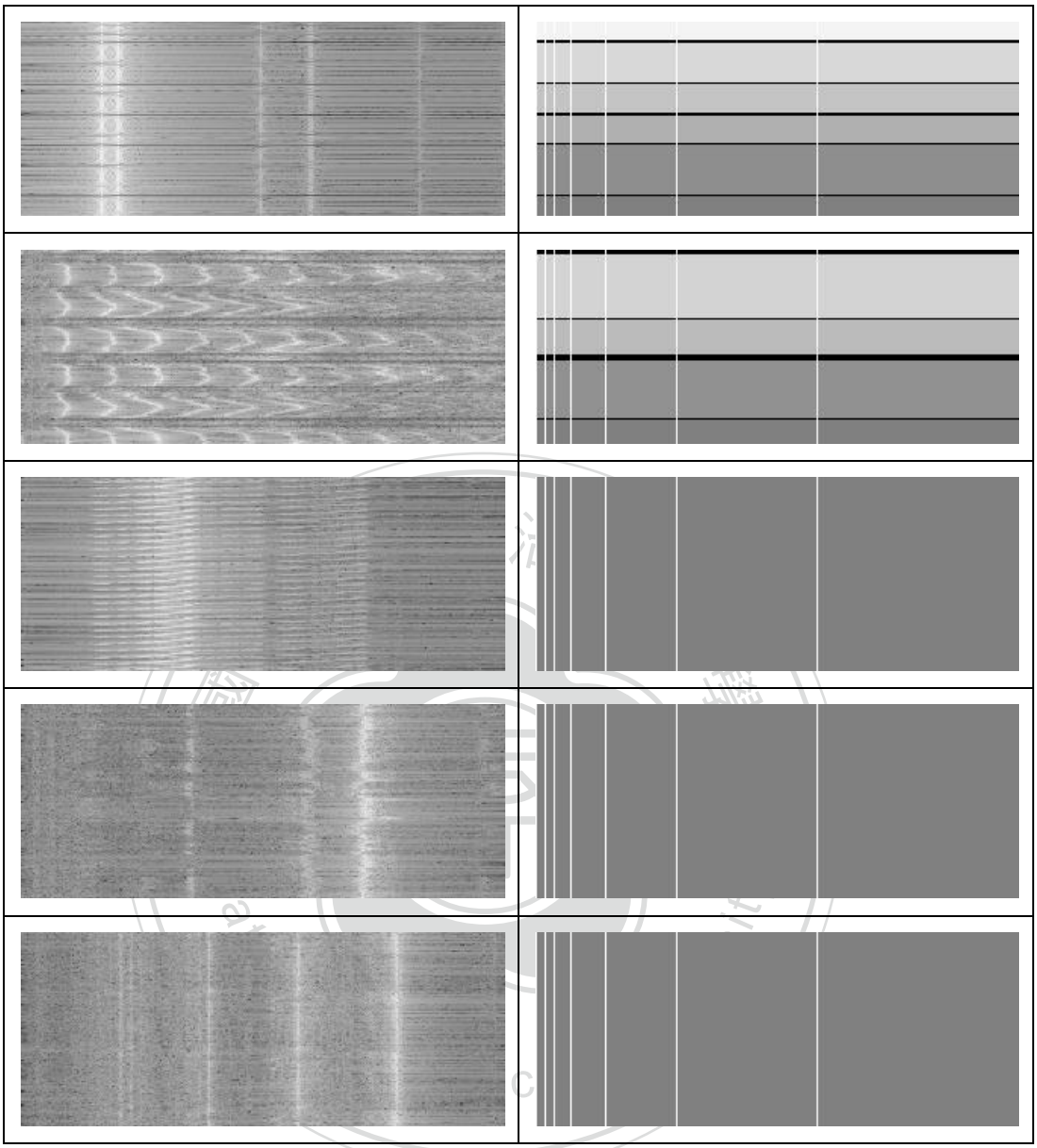
音訊起始點偵測-獨立音訊事件



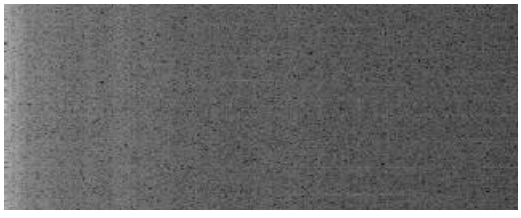
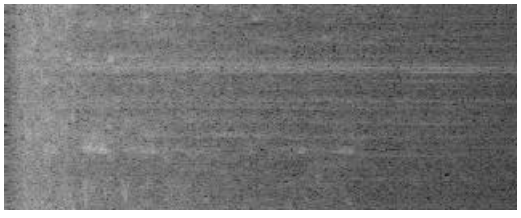


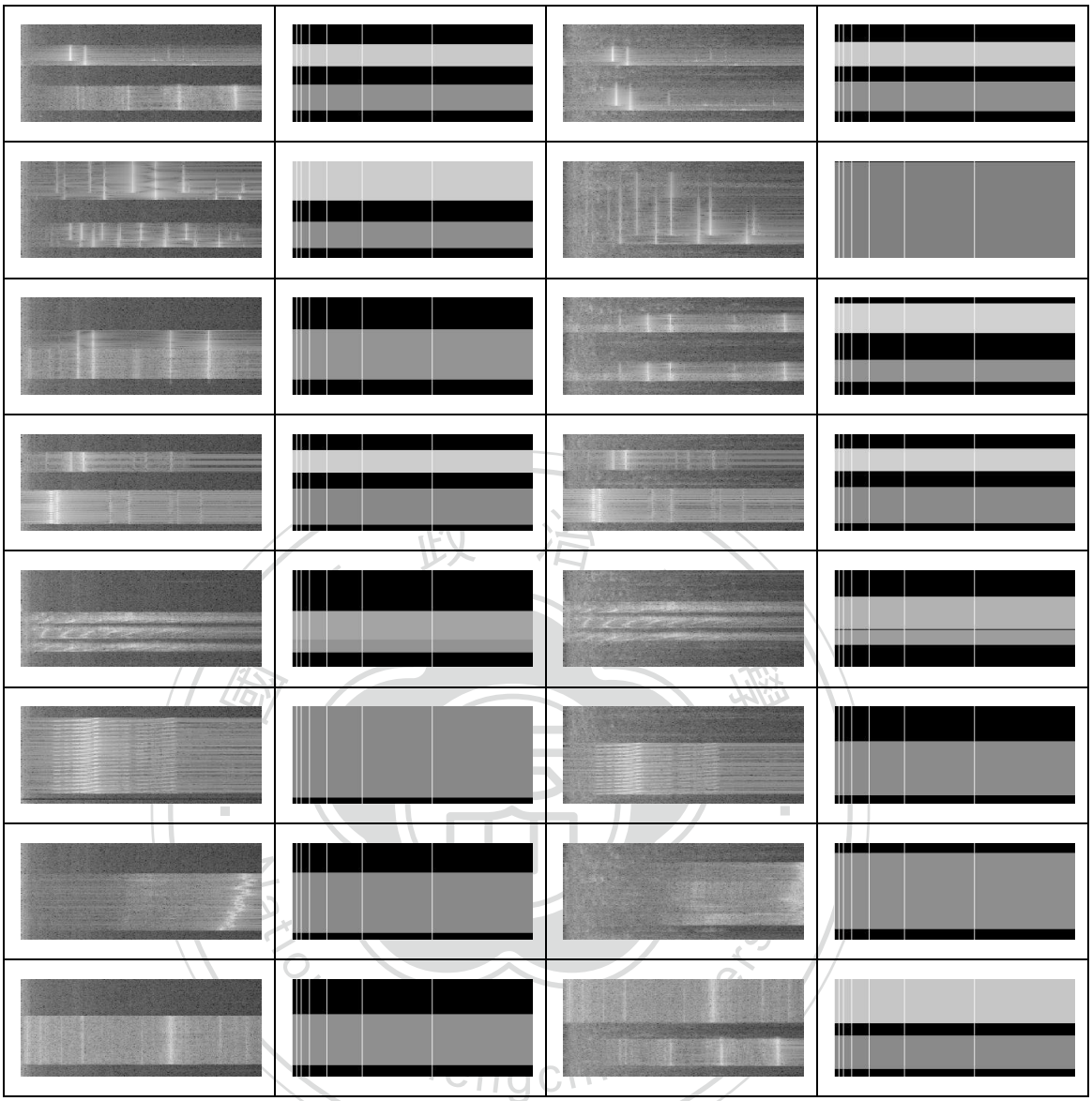
音訊起始點偵測-連續音訊事件



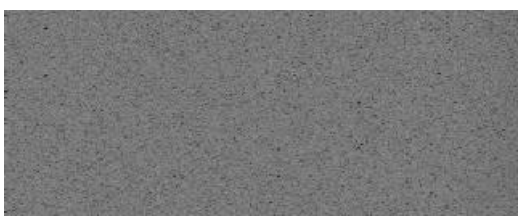
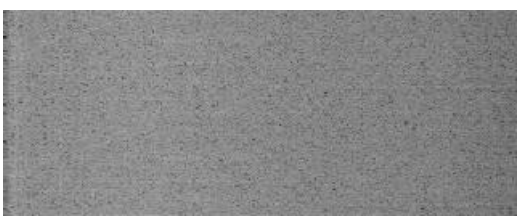
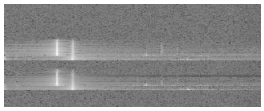
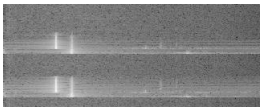


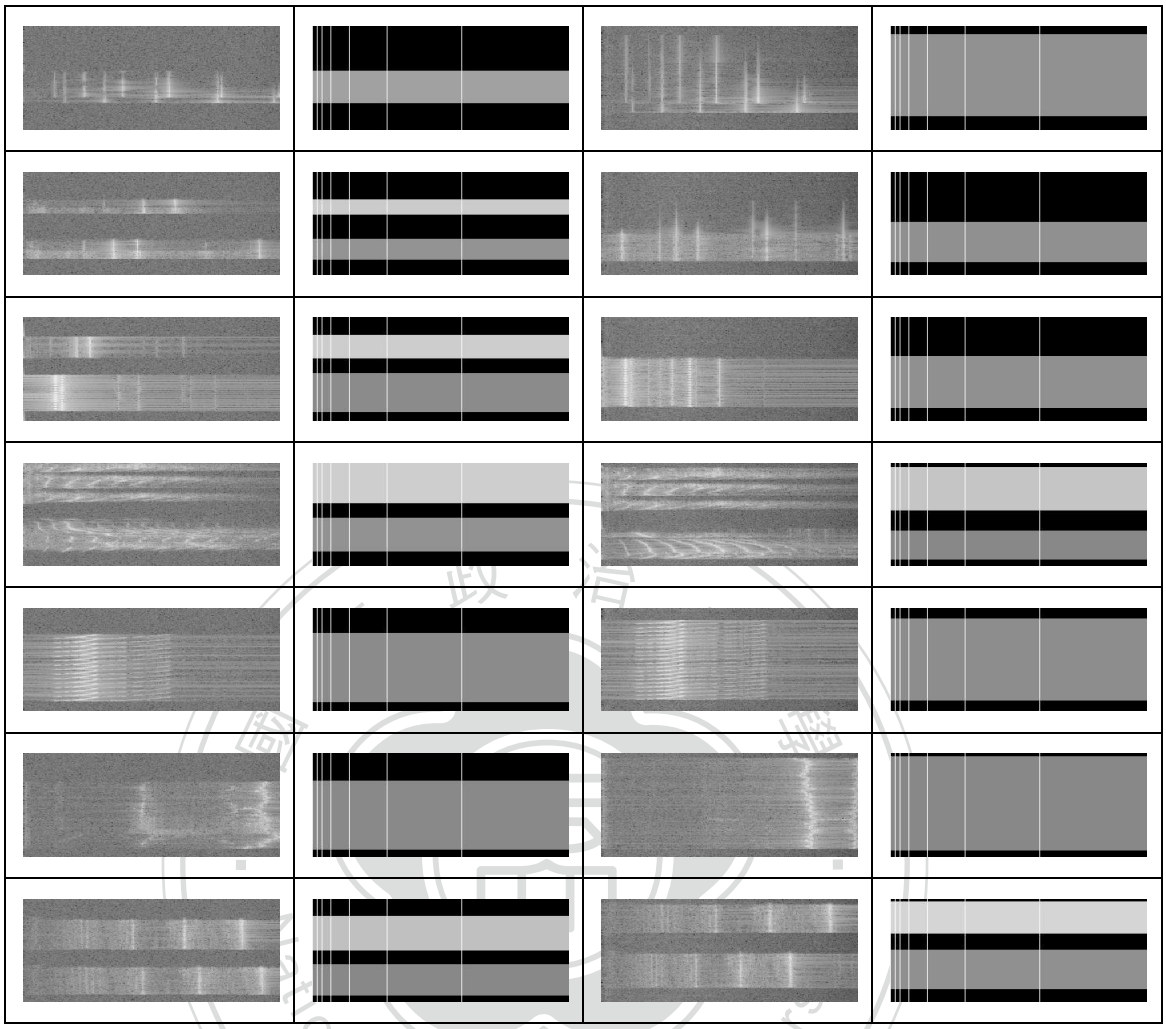
音訊起始點偵測-冷氣環境音與人群環境音

Air Condition	Crowd
	

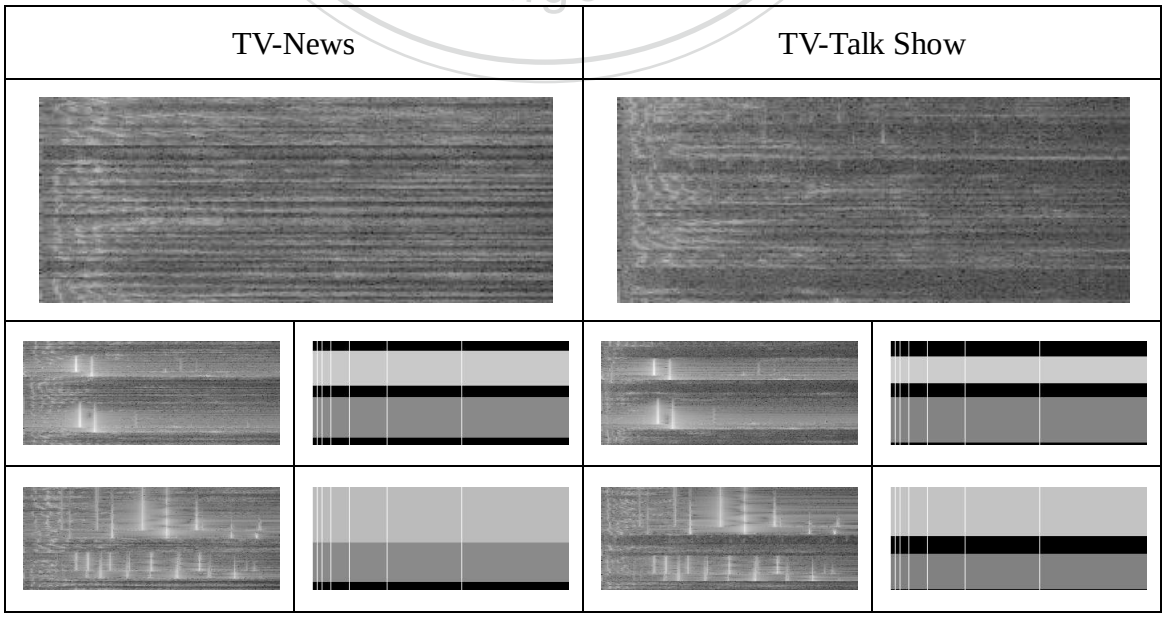


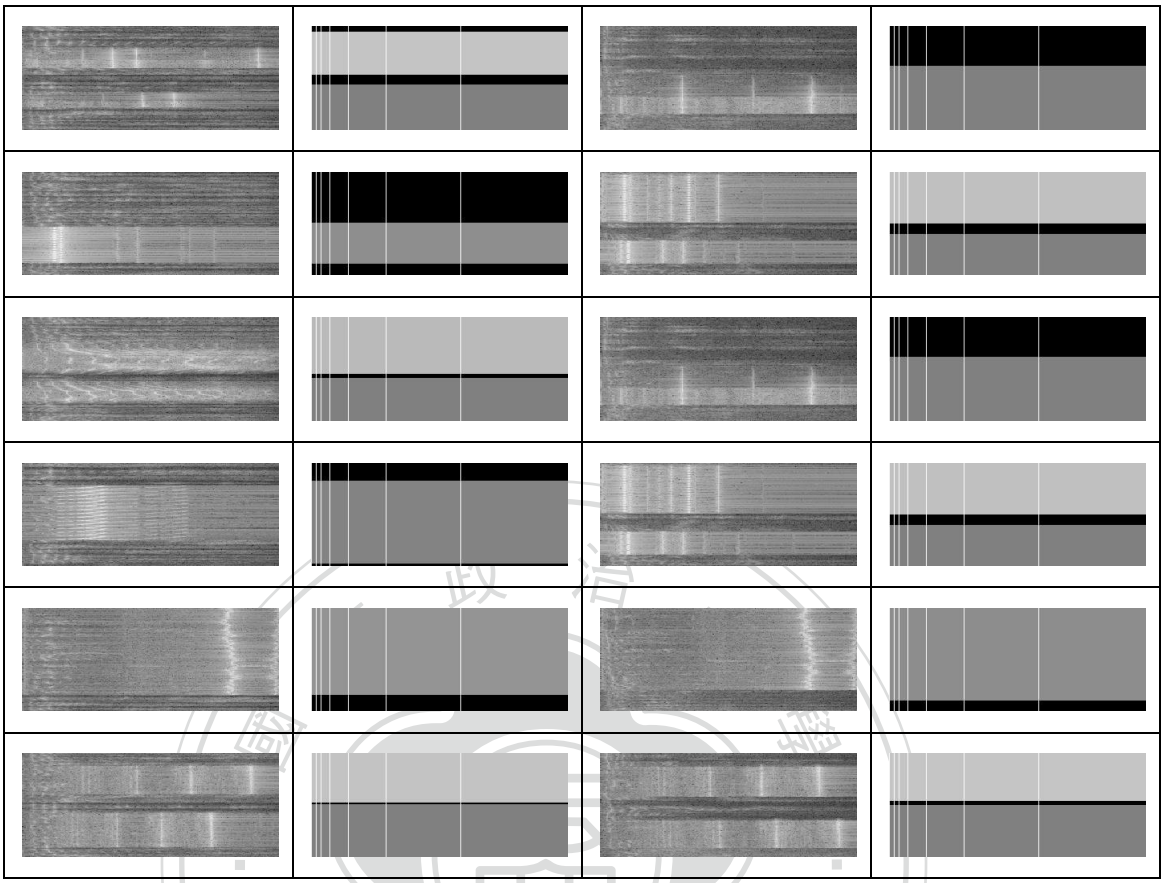
音訊起始點偵測-高斯雜訊與雨聲環境音

Gaussian Noise		Rain	
			
			



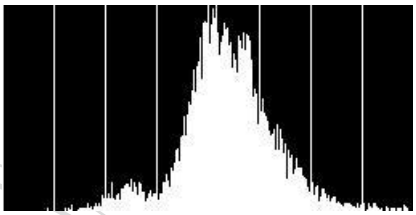
音訊起始點偵測-電視情境新聞環境音與談話性節目環境音



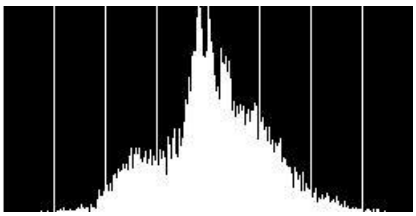


B. 基本全域閾值設定之實驗結果

Class1 門鈴聲-1 之基本全域閾值設定實驗結果

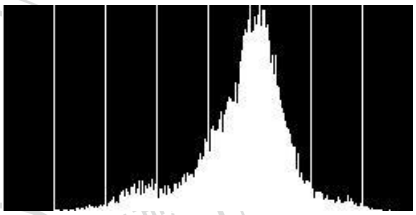
音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

Class2 門鈴聲-2 之基本全域閾值設定實驗結果

音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10

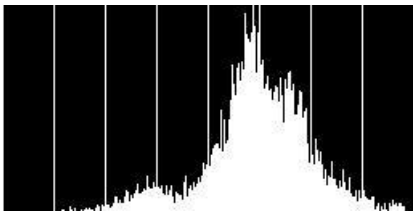
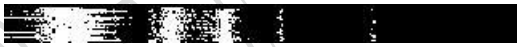
	
ΔT = 5	ΔT = 1
	

Class3 電話鈴聲-1 之基本全域閾值設定實驗結果

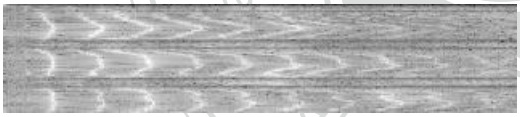
音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

Class4 電話鈴聲-2 之基本全域閾值設定實驗結果

音訊事件影像	音訊事件影像強度分布直方圖

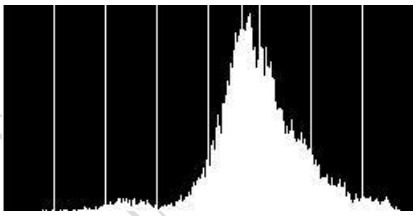
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

Class5 嬰兒哭聲之基本全域閾值設定實驗結果

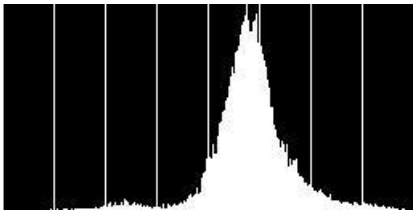
音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1

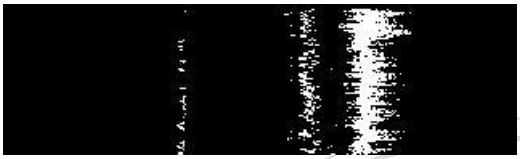
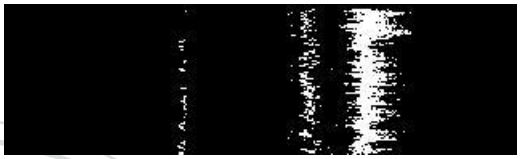
	

Class6 汽車警報聲之基本全域閾值設定實驗結果

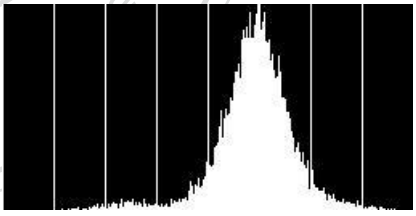
音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

Class7 水壺汽笛聲之基本全域閾值設定實驗結果

音訊事件影像	音訊事件影像強度分布直方圖
	

ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

Class8 火災警報聲之基本全域閾值設定實驗結果

音訊事件影像	音訊事件影像強度分布直方圖
	
ΔT = 50	ΔT = 30
	
ΔT = 20	ΔT = 10
	
ΔT = 5	ΔT = 1
	

C. 雙閾值設定之實驗結果

Class1 門鈴聲-1 之雙閾值設定實驗結果

	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

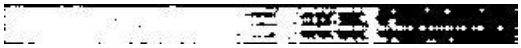
Class2 門鈴聲-2 之雙閾值設定實驗結果

音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

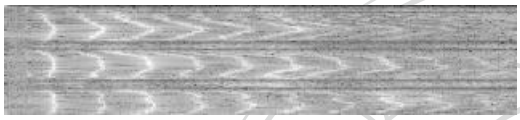
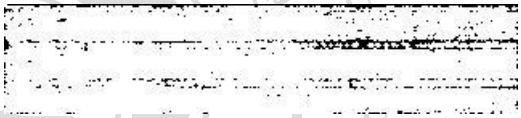
Class3 電話鈴聲-1 之雙閾值設定實驗結果

音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

Class4 電話鈴聲-2 之雙閾值設定實驗結果

音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

Class5 嬰兒哭聲之雙閾值設定實驗結果

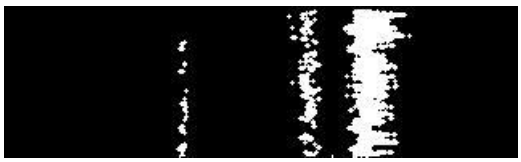
音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

Class6 汽車警報聲之雙閾值設定實驗結果

音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

Class7 水壺汽笛聲之雙閾值設定實驗結果

音訊事件影像	T_1 為前 1/4、T_2 為全圖平均

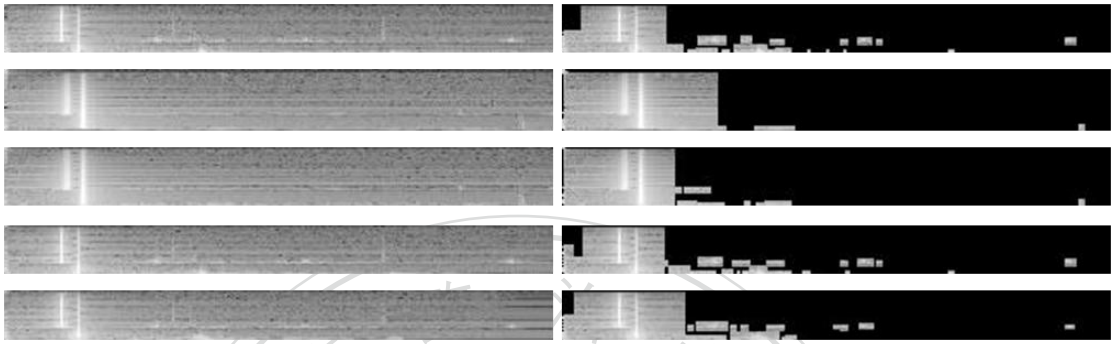
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

Class8 火災警報聲之雙閾值設定實驗結果

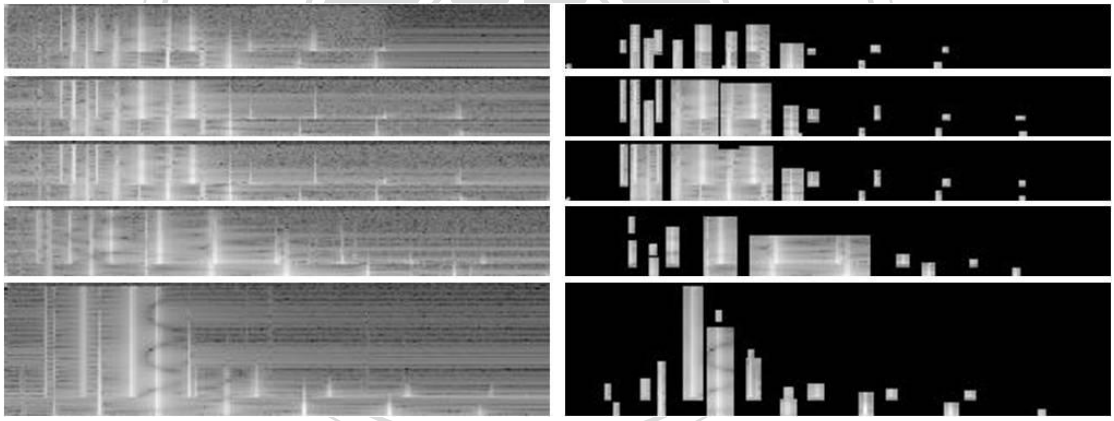
音訊事件影像	T_1 為前 1/4、T_2 為全圖平均
	
T_1 為前 1/8、T_2 為全圖平均	T_1 為前 3/8、T_2 為全圖平均
	

D. 音訊事件與音訊區塊用於Uniform Pattern實驗之樣本

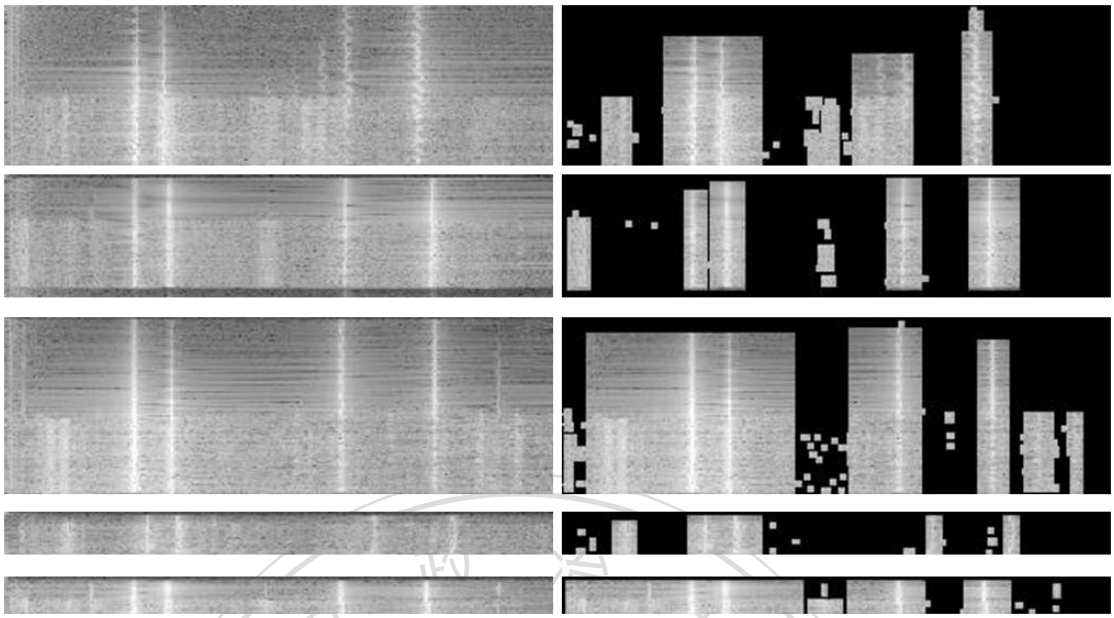
Class1. 門鈴聲-1 音訊事件與音訊區塊樣本



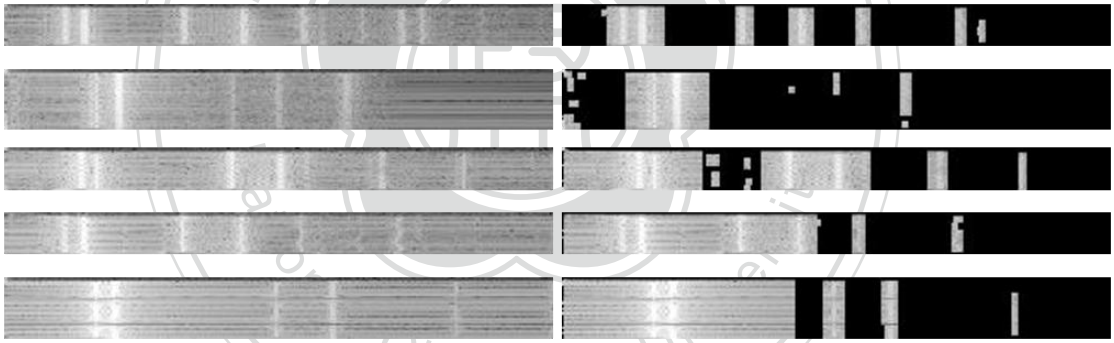
Class2. 門鈴聲-2 音訊事件與音訊區塊樣本



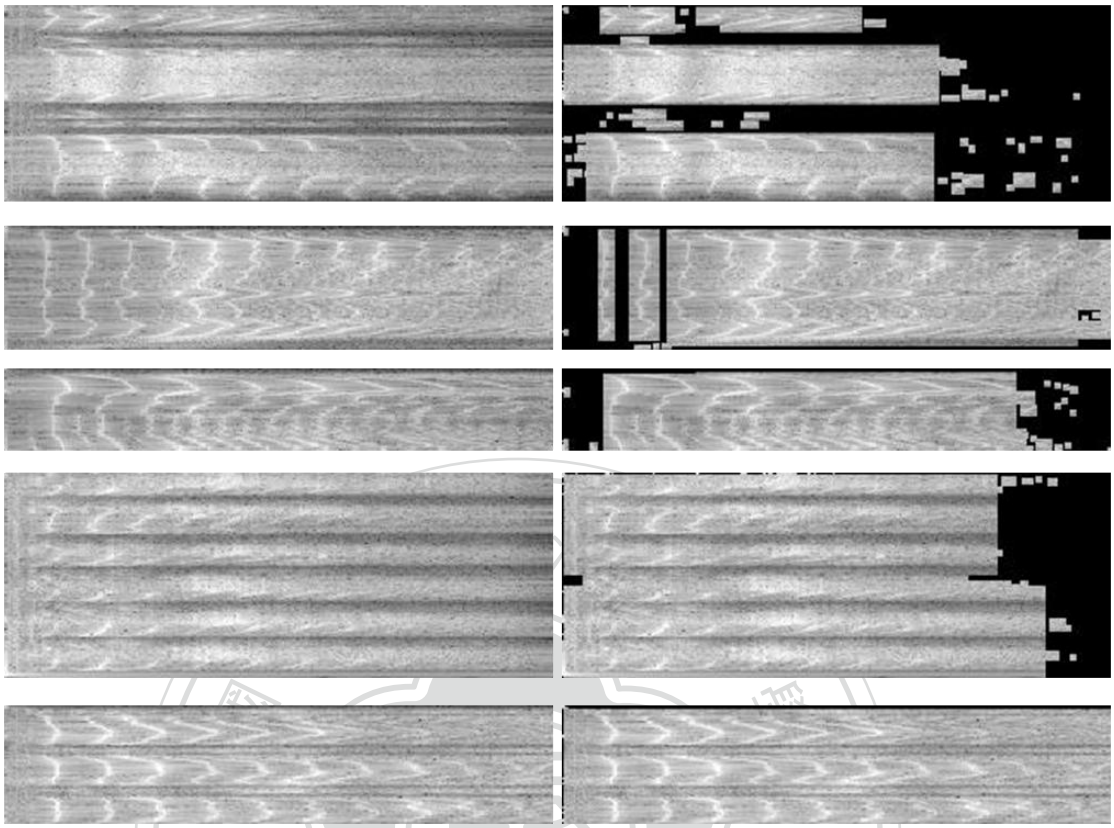
Class3. 電話鈴聲-1 音訊事件與音訊區塊樣本



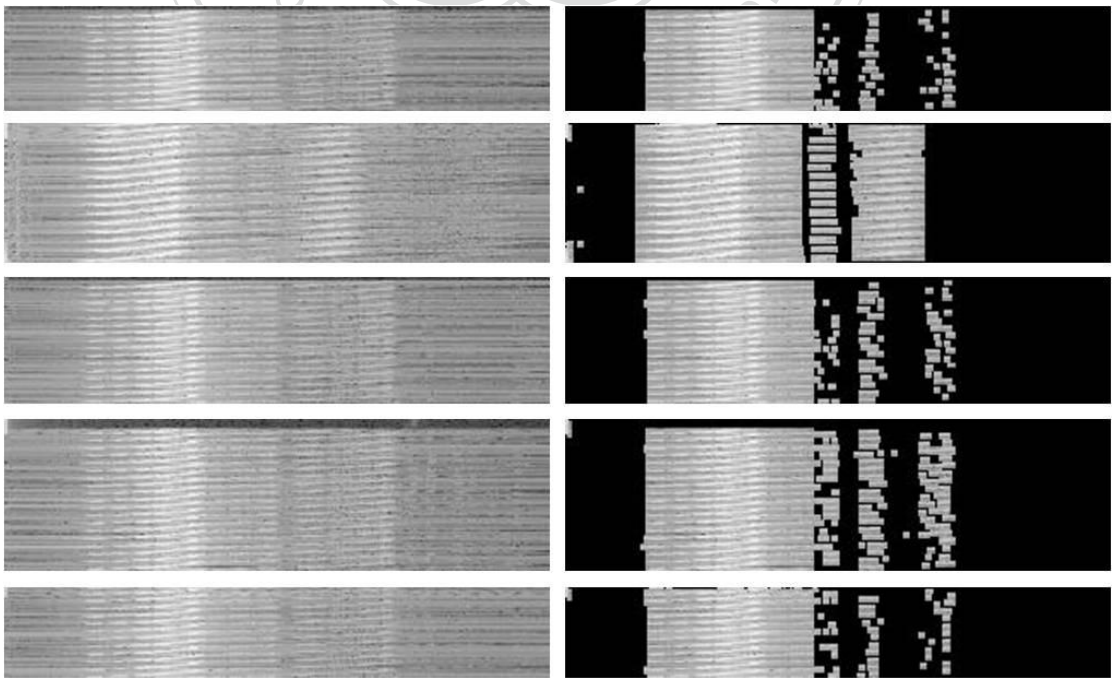
Class4. 門鈴聲-2 音訊事件與音訊區塊樣本



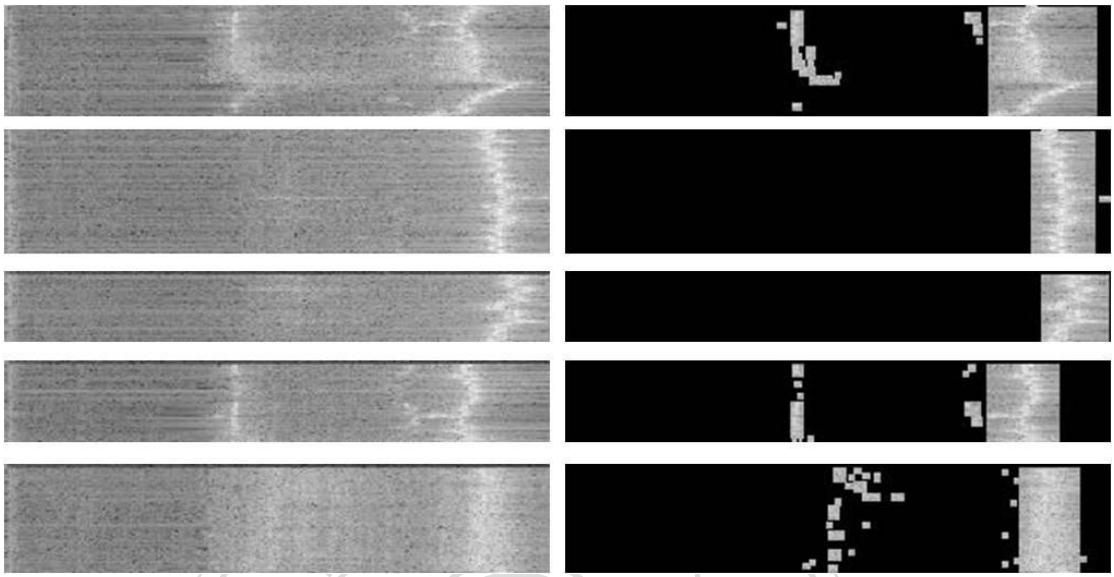
Class5. 嬰兒哭聲 音訊事件與音訊區塊樣本



Class6. 汽車警報聲 音訊事件與音訊區塊樣本



Class7. 水壺汽笛聲 音訊事件與音訊區塊樣本



Class8. 火災警報聲 音訊事件與音訊區塊樣本