國 立 交 通 大 學

電信工程學系碩士班

碩士論文

自組織映射圖應用於聽覺場景式語音分離

Self-Organizing Map on Auditory-Scene based Sound Segregation

研 究 生：吳柏宏　　Student: Po-Hung Wu

指導教授：冀泰石 博士　　Advisor: Dr. Tai-Shih Chi

中 華 民 國 九 十 七 年 九 月

自組織映射圖應用於聽覺場景式語音分離

Self-Organizing Map on Auditory-Scene based Sound Segregation

研 究 生：吳柏宏　　Student: Po-Hung Wu

指導教授：冀泰石 博士　　Advisor: Dr. Tai-Shih Chi

國立交通大學

電信工程學系碩士班

碩士論文

A Thesis
Submitted to Department of Communication Engineering
College of Electrical and Computer Engineering
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Communication Engineering
June 2008
Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 七 年 九 月

自組織映射圖應用於聽覺場景式語音分離

學生：吳柏宏　　指導教授：冀泰石 博士

國立交通大學電信工程學系碩士班

中文摘要

過去十年間，聽覺感知的一些細部的特性被大量的應用在語音處理的演算法中以提升效能。例如：在語音分離的領域中，使用多個麥克風的演算法如獨立成份分析(Independent Component Analysis, ICA)經常被使用而且有令人滿意的成果。然而，人類並只需要單耳便能將混合的聲音分開。本論文中，我們設計一個基於聽覺感知模型的單耳語音分離系統。我們從此模型中取出不同在時域-頻域上的一些使用於單耳語音分離系統的線索，之後，利用自組織映射圖來模擬神經元將混合的語音分組和歸類成分開的語音。最後，我們將比較分開語音和原來語音來顯示出本系統的效能。

Self-Organizing Map on Auditory-Scene based Sound Segregation

Student: Po-Hung Wu Advisor: Dr. Tai-Shih Chi

Department of Communication Engineering

National Chiao Tung University

Abstract

During the past decade, detailed characteristics of auditory perception have been largely incorporated into speech processing algorithms to enhance their performance. For example, in the field of sound segregation, algorithms good for the condition of multiple microphones, such as independent component analysis (ICA), are often used and show satisfactory performance. However, the truth is human has no problems in segregating mixed sounds with only one ear. In this thesis, we design such a monaural speech segregation system based on an auditory perceptual model. Various spectral-temporal cues extracted from the model are used for monaural speech segregation. Then, a self-organizing feature map neural network is utilized to mimic the neural function in segregating and clustering a mixed sound into separated sounds. At the end, we demonstrate our system's performance by comparing the separated sound with original sound.

誌謝

阿姆斯壯在登陸月球時說了一句經典名言：「我的一小步，是人類的一大步」。當初的我，為了自己的小小的宅男夢想—能做出有如漫畫「名偵探柯南」中，阿笠博士設計的變聲器，而開始跨出了自己的一小步—加入冀泰石老師的門下，鑽研更先進的語音處理技術。不過，對於要一年畢業的我來說，壓力著實不小，在這邊要感謝指導教授冀泰石老師。對於從大學專題就跟著老師的我來說，從老師的身上學習到了對研究要有熱情同時要有正確的態度—尋找物理意義而不是嘗試而已。而除了研究之外，老師更在我人生的抉擇上徬徨時，適時的提點了我，讓我能堅定信心，撐過壓力，繼續往我的夢想努力邁進，可以說，若沒有老師如此認真的指導，我是沒有資格站在這個所有碩士生的最終試驗場合來挑戰。真的很謝謝冀泰石老師這兩年來對於我的指導。

再來，要謝謝這間 711 實驗室裡的所有人，對於我來說，因為認識了你們，在我的人生當中，增添了許多的色彩及回憶，同時也很感謝你們長期能夠的忍受我的大嗓門噪音，以後想再聽到，可能要等我出名的時候吧。在這些人中，更要感謝的是我們自己實驗室的夥伴和學長們，經常的討論，使我在研究碰到瓶頸時能夠得到更多的靈感，謝謝你們。

最後，要謝謝我的父母親和我的外公外婆。我的求學的路程的大關卡上，總是很順利的過關，但是背後的過程，是相當的艱辛而且苦悶的，如果沒有父母親和外公外婆的信心勉勵，我不會有這樣的自信渡過種種的困難，感謝你們 23 年來的支持。

在交通大學的生涯即將畫上了最後的一筆，緊接而來的是人生新的空白的一頁，等著我去給他畫上幾筆呢!!不用多說，邁開大步向前走吧！！

目　　錄

表　目　錄

圖　目　錄

第一章

緒論

1.1 研究動機

在語音處理的研究當中，由於在一般自然環境下，目標語音的背景雜訊通常是其他人的語音(例如：雞尾酒派對問題(Cocktail Party problem))。因此如何將目標語音從多人的語音中取出來，就成了熱門的研究之一。一般所提出的方法，如：信號盲分離(Blind source separation)…等，皆需要兩個以上的輸入，才可以做處理；然而，在某些應用領域上面，如：電信通訊、語音的補償處理、語者辨別上面，僅能使用單一的輸入。因此，單耳語音分離(Monaural speech segregation)漸漸熱門起來。

近年來隨著科技的進步及研究越來越深入，數位信號處理的研究逐漸往生物的現象研究邁進。人類聽覺研究在這領域中逐漸重要起來，隨著人類在心理聲學和生理學上的研究，發現人類亦可用單耳即辨別出目標語音，因此，我們希望能應用此種現象，來達到單耳語音分離的目標。

1.2 聽覺場景分析概論

在一般常用的單耳分離技術當中，聽覺場景分析(Auditory Scene Analysis, ASA)是常用的技術之一。ASA 提出的觀念是：人類聽覺系統在分離語音時，而是利用兩個步驟來完成：

(1)分析階段(Analysis stage)：分割階段為將輸入語音藉由一些機制尋找出許多語音分離的線索，如；音高、頻率調變……等。

(2)分組階段(Grouping stage)：分組階段係藉由前面階段的分析出來的線索，將原語音依照線索的分佈做分組並將語音分離。

本論文所使用的語音分離線索為：音高、起始點和終止點、頻率調變、振幅調變。

1.3 研究方法

本論文主要的研究方向在結合一已知的聽覺感知模型，將兩個混合的語音，先經由在頻譜及在人類大腦上某一些時域-頻域區塊的能量反應，來找出語音的一些特徵，然後再利用類神經網路中的自組織映射圖網路(Self-Organizing feature MAP)來做分類，來達到語音分離的目的。

1.4 章節綱要

第一章 序論：本章說明研究之動機、研究方法以及各章節之綱要。

第二章 聽覺感知模型及系統之基本介紹：此章對本論文所使用之聽覺感知模型做一基本之介紹，同時介紹本論文所使用之語料庫。

第三章 語音特徵抽取：本章主要說明在聽覺模型中，介紹並說明如何抽取出語音分離

所使用的特徵，並特別針對音高(pitch)的抽取做說明及結果的驗證。

第四章 語音分離：本章主要介紹自組織映射圖網路，並說明在本論文中的應用方式及結果。

第五章 結論與展望：本章對總結本論文所提出之方法，並針對此方法做分析討論其未來可改進的方向。

第二章

聽覺感知模型及系統之基本介紹

本章將先介紹由 NSL(Neural Systems Laboratory)提出的人類聽覺感知模型；接著介紹本論文所使用的語料庫，最後簡單介紹本論文所使用的語音特徵及結果評斷的方式。

2.1 聽覺感知模型介紹

此聽覺模型是由 NSL 所提出的。由於哺乳類動物的聽覺系統皆應相似，因此，NSL 藉由研究哺乳動物的聽覺系統的生理實驗，求出人類聽覺系統處理聲音的路徑和模式，大致上可分為以下兩個部份：

(1)初期階段：此階段模擬人類由耳朵接受到聲音，將聲音訊號做轉換之後傳輸到中腦的神經元的轉換結果，此部份的模擬是頻譜估計。

(2)大腦皮質階段：此階段是人類將初期階段所輸出的東西做分析。經由觀察及生理實驗發現，此階段可以用一組時域—頻域的調變濾波器來完成。而初期階段的輸出在此可以用時域-頻域來分析。

以下將針對各階段做詳細的介紹。

2.1.1 耳朵基本構造簡介

耳朵的構造主要分為三個部份，外耳、中耳及內耳。就接受外界的聲音來論，外耳的功能在於接受外界聲波，利用特殊構造適當放大音量。大致可分為耳介、耳殼和外耳道，其中耳殼的共振頻率約在 2.5~3KHz。中耳主要是將外耳收到的聲波，經過耳膜、三小聽骨後抵達卵圓窗。此時聲波會轉成位移波的振動，而卵圓窗可以將振動傳至內耳的耳蝸，其生理學現象將於下一小節做介紹。外耳送入的空氣波動振動鼓膜連接著聽小骨，不但是傳送訊息，同時藉由聽小骨的升壓作用也保護了內耳。下圖 2-1 即是耳朵之基本構造圖：

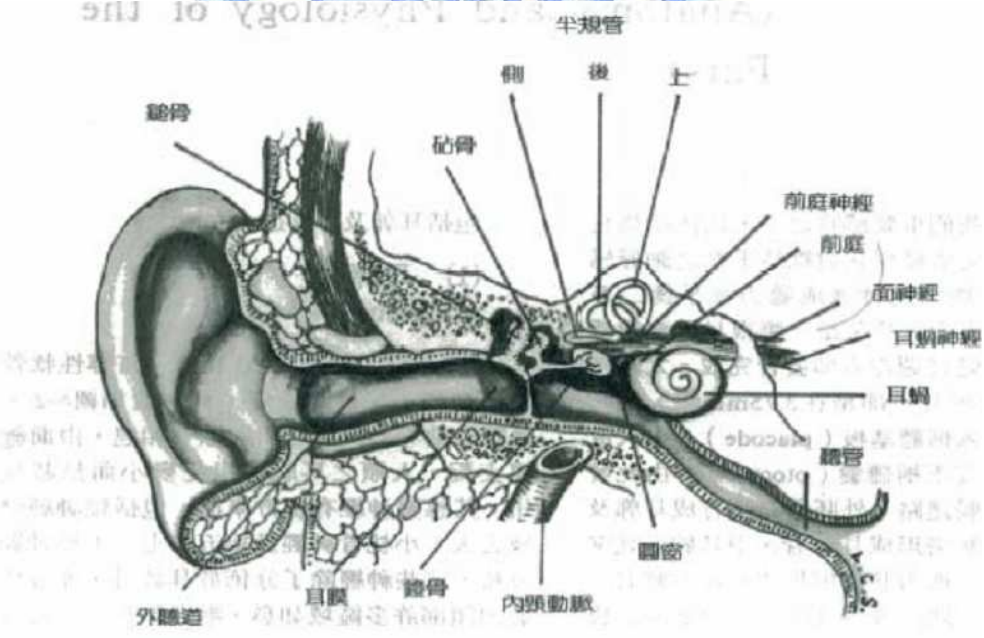


圖 2-1：耳朵基本構造圖

2.1.2 初期階段的生理學現象

初期階段中，主要主管著聽覺感知的受器，即是耳蝸。耳蝸是由三個空腔組織所組

成，各個空腔組織內充滿著淋巴液。而這些淋巴液被基底膜(Basilar membrane)分隔成兩部份。由卵圓窗傳來的振動會使內耳淋巴液振動，在基底膜上形成一行進波(Traveling Wave)，並在基底膜上各部份產生不同的振幅。如圖 2-2 所表示：

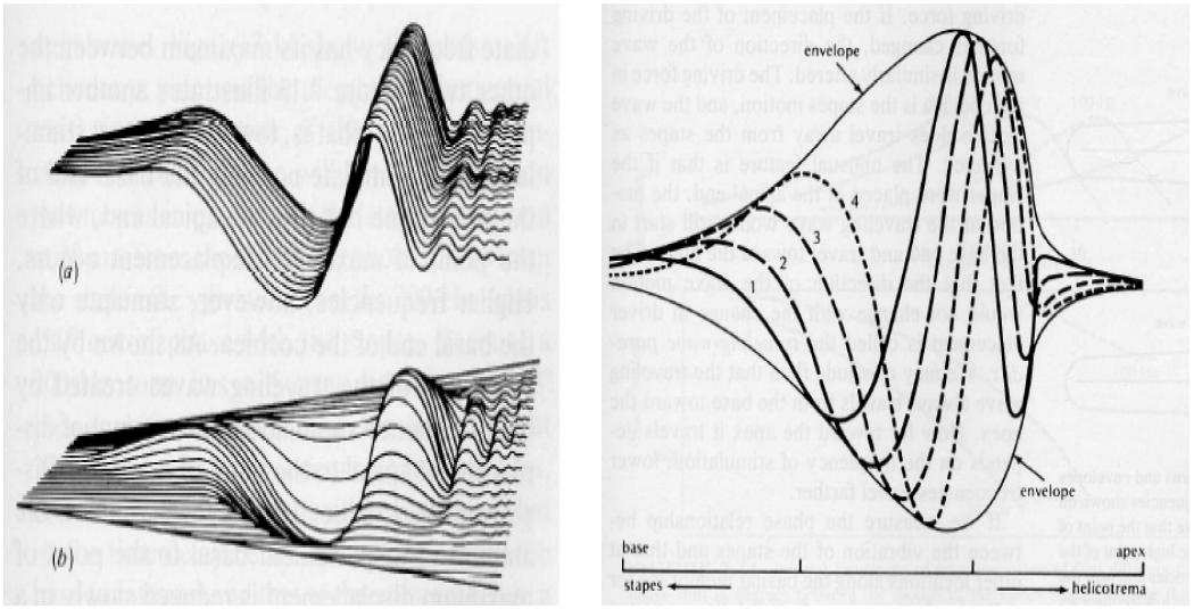


圖 2-2：基底膜上行進波示意圖

這些振幅會間接帶動上面毛細髮胞(Hair cell)的晃動，進而產生電流藉由聽神經傳至大腦做分析。毛髮細胞分成內毛髮細胞(inner Hair cell)和外毛髮細胞(Outer Hair Cell)。前述之轉換主要由內毛髮細胞所執行。內毛髮細胞會和若干聽神經形成突觸連結，將機械振動轉換為聽神經的動作電位；外毛髮細胞一般認為和增強聽神經之高度頻率選擇性、耳蝸的自我調節和保護有關。圖 2-4 即是內毛髮細胞的運作示意圖。

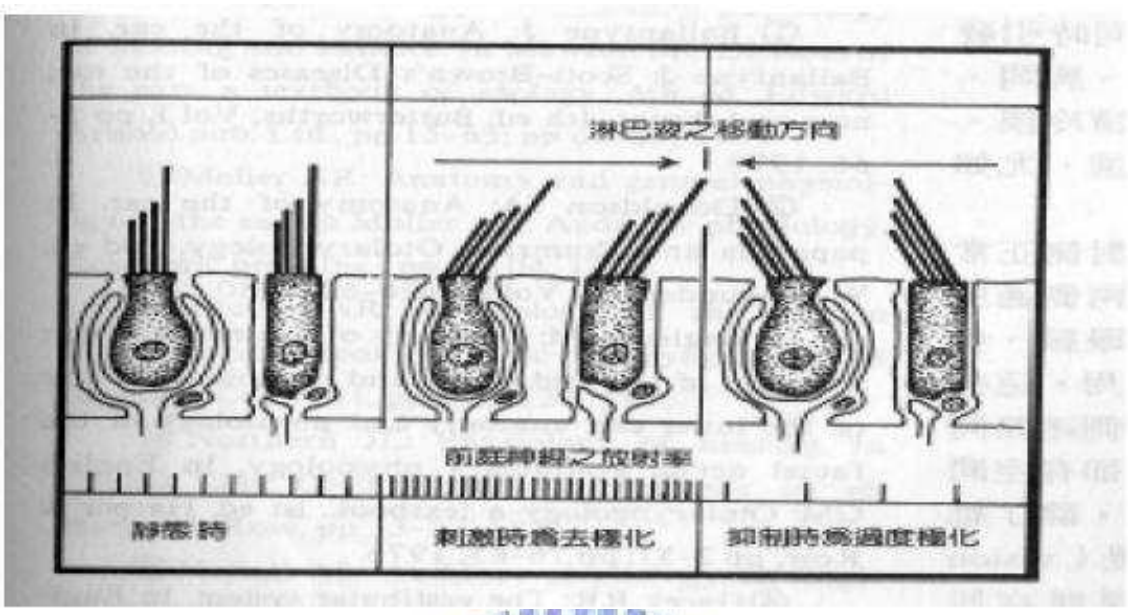


圖 2-3：內毛髮細胞的運作示意圖

當不同的頻率的聲音進入進入人耳時，會在基底膜上形成不同的行進波。基底膜從底部至頂部，寬度由窄變至寬、彈性則由軟變至硬，愈靠近窄端的可以感測愈高的共振頻率(或稱特性頻率(Characteristic Frequency)、最佳頻率(Best Frequency,BF))，愈遠離的可以感測愈低的共振頻率。一般人類聽覺可接受到的範圍約為 20~20000Hz，此即是基底膜的共振頻率範圍。圖 2-5 可清楚的表示出基底膜的分佈運作示意圖及對於不同頻率之共振反應。

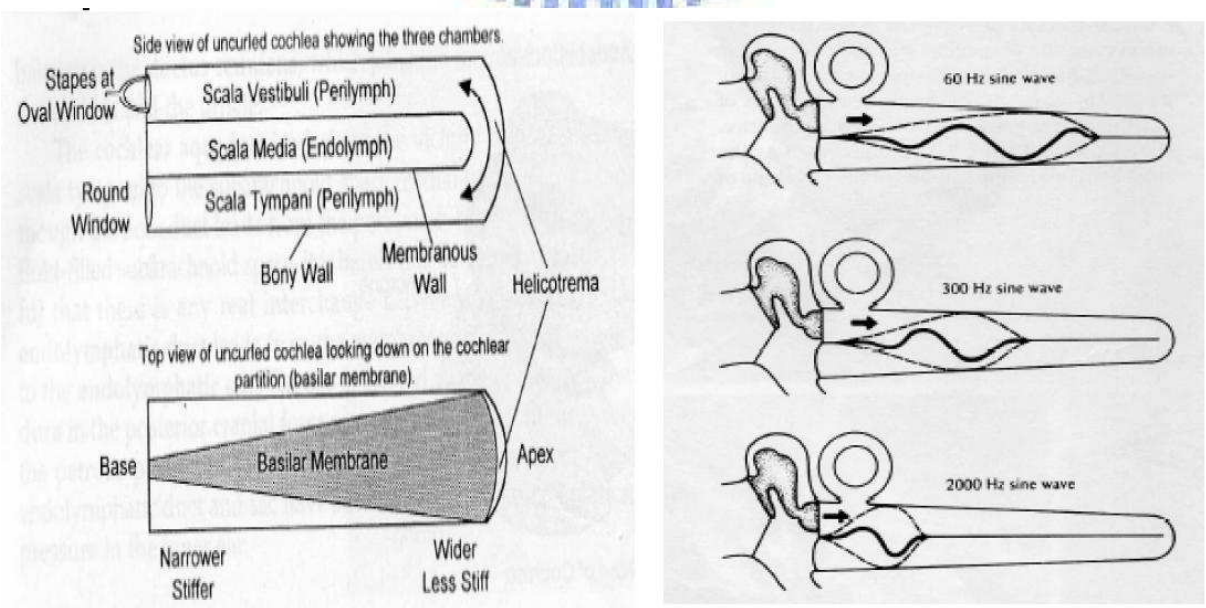


圖 2-4：基底膜的運作、分布及不同頻率之共振反應示意圖

自然界中的聲音，一般說來都是多頻所組成，因此當人類接收到一般自然界聲音時，會在基底膜上產生不同的行進波，造成對於鄰近位置之毛髮細胞的反應有壓抑的效果。在內毛髮細胞將機械振動轉換成電流時，訊息就會延著神經傳送上去，但是神經元在連續發射動作電位之後，必須進入靜止電位休息，此結果造成一旦輸入是一高頻信號時，神經的發射速率(Neural Firing Rate)會無法跟上，因此出現了最高的神經發射速率。內毛髮細胞的最高神經發射速率約莫 4~5KHz，而中腦聽神經，最高的發射速率只能到 1KHz。圖 2-6 即是聽覺神經發射動作電位的示意圖。

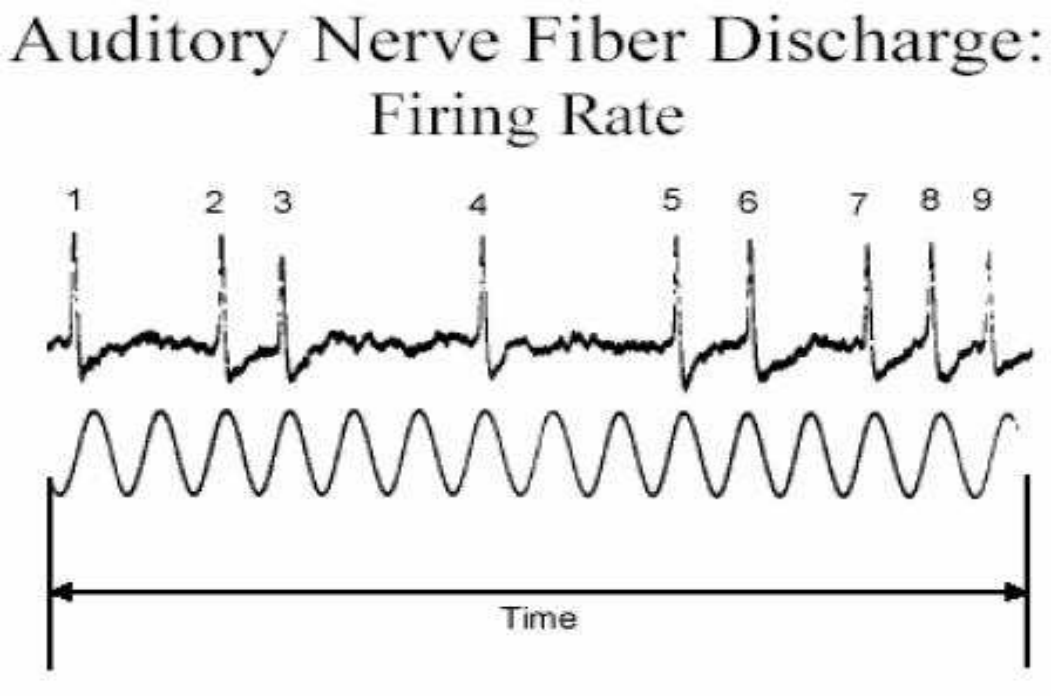


圖 2-5：聽覺神經發射動作電位之示意圖

2.1.3 聽覺感知模型—初期階段的模擬

在聽覺感知模型中，初期階段的模擬，即是模擬聲波在耳蝸中轉換成神經脈衝並傳輸到中腦。此部份是用將聲音換成聽覺頻譜(Auditory Spectrum)來模擬估計的，主要可分為三個部份：分析部份(Analysis Stage)、傳導部份(Tansduction Stage)和縮減部份(Reduction Stage)。圖 2-5 表示此階段的結構圖。

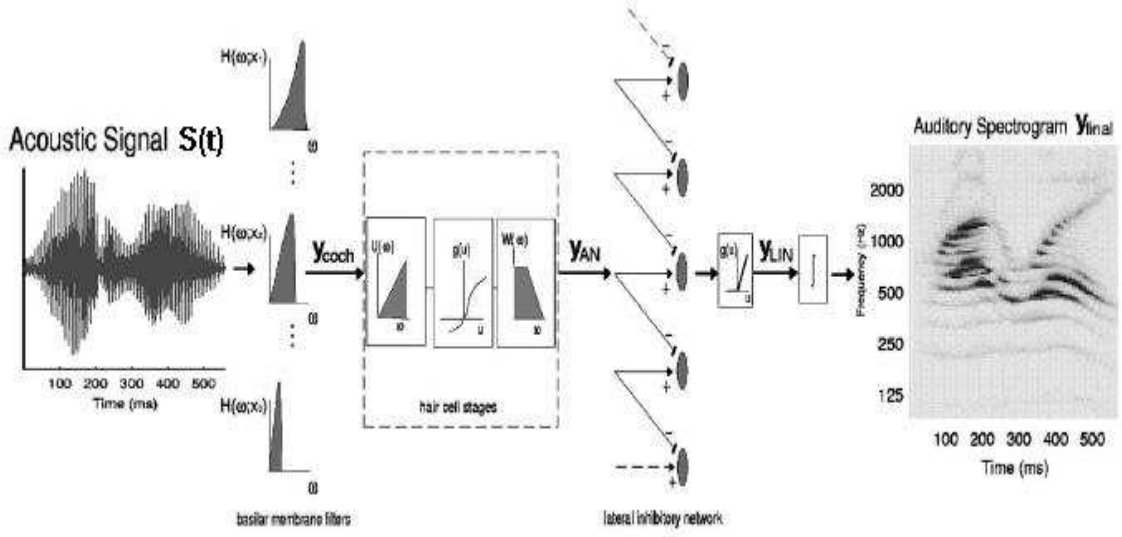


圖 2-6：模型中初期感知階段圖[3]

上圖之模型可以用以下四個數學式子來表示：

$$y_{coch}(t,x) = s(t) \otimes_t h(t;x) , \quad (2\text{-}1)$$

$$y_{AN}(t,x) = g(\partial_t y_{coch}(t,x)) \otimes_t w(t) , \quad (2\text{-}2)$$

$$y_{LIN}(t,x) = \max(\partial_t y_{AN}(t,x), 0) , \quad (2\text{-}3)$$

$$y_{final}(t,x) = y_{LIN}(t,x) \otimes_t \mu(t;\tau) , \quad (2\text{-}4)$$

式(2-1)表示的是分析部份。目的是在模擬時域信號 $s(t)$ 在基底膜上的共振反應。其中，$\otimes_t$ 表示在時間軸上的褶積(Convolution)；$h(t;x)$ 表示在某一離耳蝸底部距離 x 之脈衝響應，此 x 是在對數頻域軸(Logarithmic Frequency axis)均勻分佈，亦代表基底膜上不同據共振頻率的位置。在此的模型上，使用一濾波庫(Filterbank)去分別濾出聲音各頻率的成份，做基礎的聲音成份分析。此濾波庫係由 128 個不同中心頻率及不同頻率解析度的帶通濾波器(Bandpass filter)所組成，每個濾波器的頻寬和中心頻率有常數 Q(constant-Q)的的關係，且其中心頻率為均勻分佈在對數頻域軸上，其分佈範圍約為 5.3 倍頻(Octave)，即一個倍頻有 24 個濾波器來表示。圖 2-6 即是在取樣頻率 8KHz 下其濾波器的振幅響應；式(2-5)則是說明我們的頻寬和中心頻率的關係，‘由該式可知，頻寬會隨著中心頻率而增加。

$$f_{center} / Bandwidth = Q \quad (2\text{-}5)$$

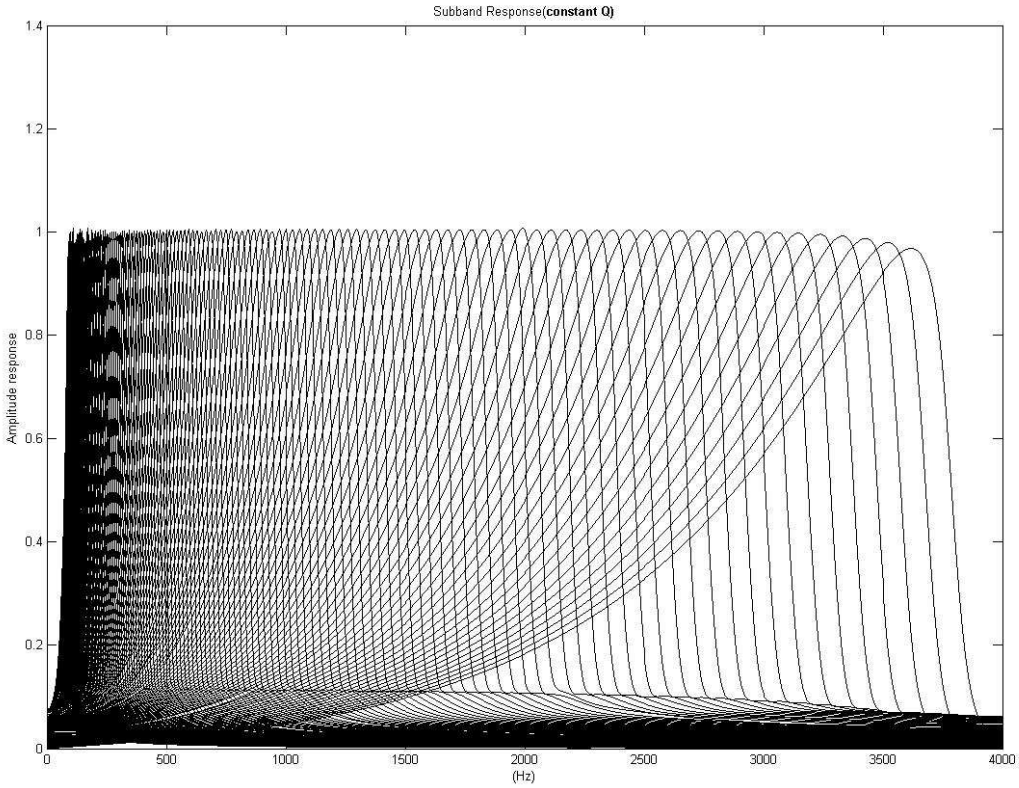


圖 2-7：濾波庫的振幅響應

在 Matlab 工具，此模型的 filterbank 屬於 IIR filter。如此可以減少處理的時間。

式(2-2)表示的是傳導部份。目的是在模擬內毛髮細胞的運作，先經過微分器(即'一高通濾波器)，將聲音大小造成的內毛髮細胞振動位移量變成速度。由於內毛髮細胞受到的刺激有飽和的狀態，因此經過 sigmoid 函數：$g(u) = 1/(1+e^{-u})$來達到。最後經過一 3dB 頻寬 4KHz 的低通濾波器，來模擬內毛髮細胞的最高發射速率，超過 4KHz 變化的會在此被壓抑。

式(2-3)及式(2-4)是屬於縮減部份。式(2-3)和內毛髮細胞運動特性有關，因為內毛髮細胞本甚會有左右抑制的現象，所以相鄰的部份都要比較相減，以達到彼此抑制的結果。此即前述所提在基底膜上的頻率壓抑效果。式(2-4)是模擬中腦聽覺神經元的神經發射速率，約 1KHz。此處用一時域上的積分視窗：$\mu(t;\tau) = e^{\frac{-t}{\tau}}u(t)$來模擬，此處之$\tau$是時間常數(Time Constant)，$u(t)$則是單位步階函數(Unit Step Function)。

經過此三部份的處理後，可以將結果化成時間-頻率的關係圖，這張圖表示經過此

模型分析過後的頻譜圖，稱為聽覺頻譜。圖 2-8 即是在取樣頻率 8KHz 下，輸入一英文語音(/Come home right away/)和其聽覺頻譜，其中顏色的深淺表示聲音成份的大小，且縱軸頻率軸屬於對數得形式去畫，屬於半對數的圖表。

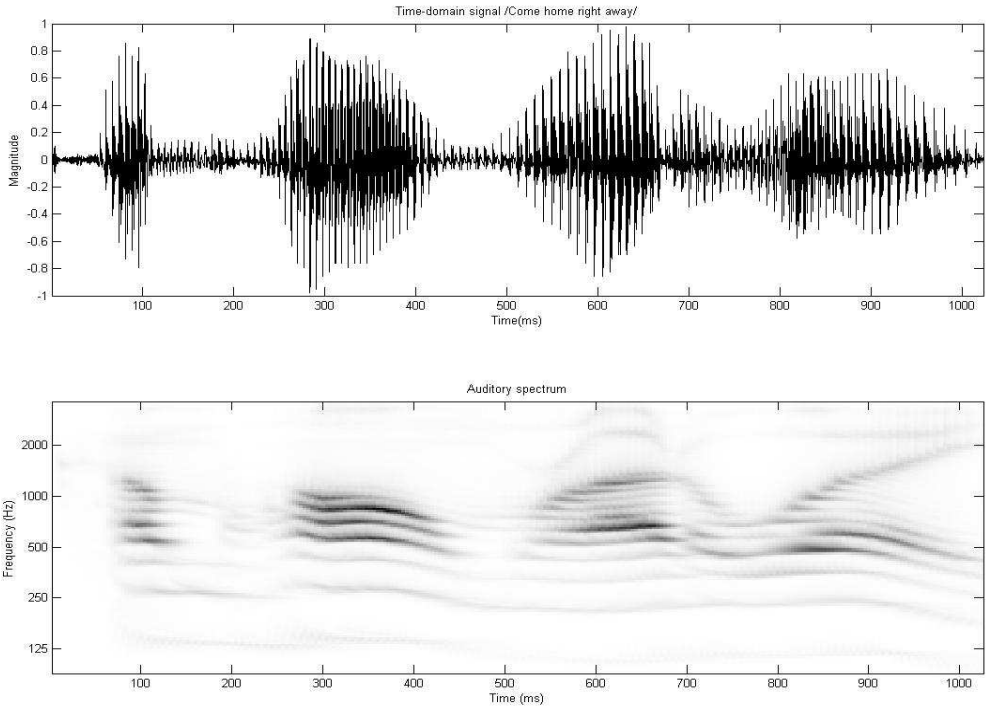


圖 2-8：英文語音/Come home right away/之時域波形及其聽覺頻譜

2.1.4 聽覺感知模型—大腦聽覺階段

由初期階段的所得到的聽覺頻譜圖，可繼續送到大腦聽覺階段做進一步的分析。此階段是在模擬大腦聽覺皮質(Auditory Cortex)的反應，它可以抓出聽覺頻譜圖中某時域(Temporal)和某頻域(Spectral)的調變。這個階段係由生物實驗而得到的。由於聽覺頻譜圖為一二維(時間-頻率)的成份，根據頻率響應測試的機制，當送入一在時間軸上和頻率軸上皆為固定週期弦波的組合信號(此信號稱”移動波紋刺激源〔moving ripple stimulus〕”）進去該系統，則得到的結果即為針對該固定於頻率軸與時間軸上週期之

脈衝響應。此脈衝響應即可代表該神經元的脈衝響應。圖 2-9 即表示一移動波紋刺激源之圖，圖中之單位 rate 之定義為：時間軸上的變化週期之倒數，單位為 Hz；而 scale 之定義為對數頻率軸上之變化率，單位為 cycle/octave，圖 2-9 之刺激源 rate 為 4Hz，sacle 為 0.5 cyc/oct：

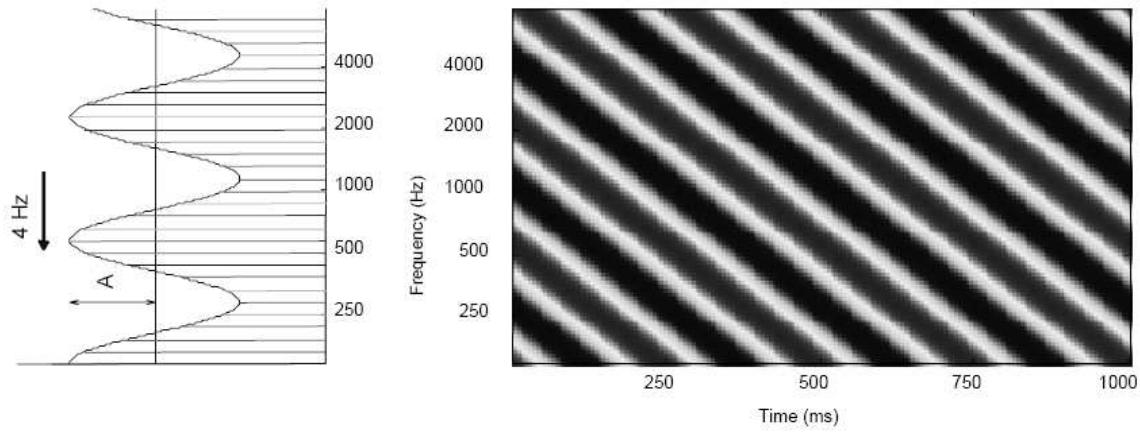


圖 2-9：移動波紋刺激源圖[3]

由生物實驗的證明，可以得知送入不同的移動波紋刺激源，會在不同位置的大腦皮質上有很強的反應，代表著聽覺頻譜圖可以由這些神經元的反應做組合而成，亦即表示每個神經元，其輸出之反應亦為一二維(時間-頻率)的成份，而每個送入測試之刺激源亦有在時間軸上之週期和頻率軸上之不同週期，因此在模型中，整個模擬大腦聽覺階段之輸出結果為一四維之成份，而其設計方式為將聽覺頻譜圖送入一組二維之濾波庫而得到不同時域-頻域解析度之分析結果。除了上述之四維之單位（時間-頻率-rate-scale）外，大腦亦對頻率調變(Freqeency Modulation)上升或下降有所反應，在本論文使用的模型中，對於頻率變化下降(downward)的，是用正的 rate 來代表；而對於頻率變化上升(upward)的，用負的 rate 來代表。圖 2-10 即為一英文語音/We‘ve done our part/的聽覺頻譜通過大腦聽覺階段的分析結果圖：

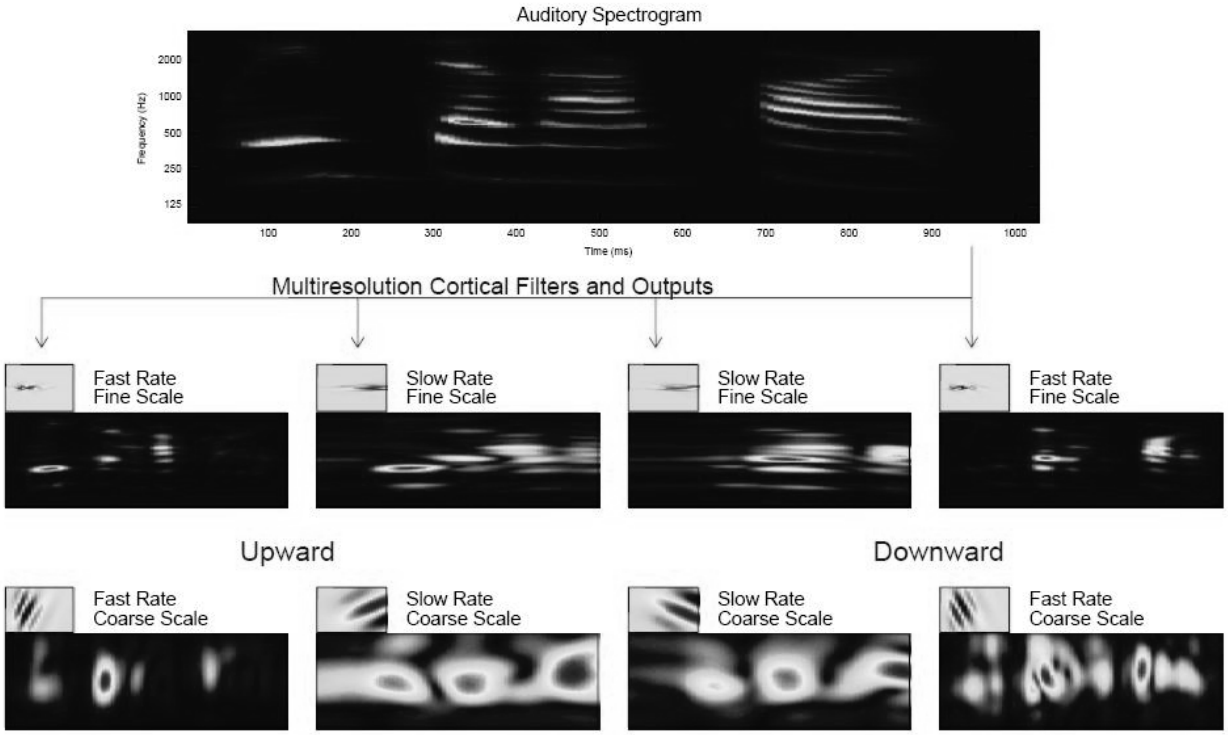


圖 2-10：大腦聽覺階段之分析

下半部的小圖即是模擬聽覺頻譜圖經過各神經元處理後的能量結果，左上角的小圖為時頻反應域(Spectral-Temporal Response Field, STRF)，即模擬各大腦神經元的脈衝響應(Impulse response)。由小圖種也可以發現，模擬的函式對於頻率調變的下降的反應比較強，表示在此句語音中，其頻率向下變化的趨勢較強。

這個部份的輸出結果，可以讓我們去藉由通過各個模擬不同神經元的濾波器處理後，去更容易取出語音分離所需要用的線索，例如：聲音的起始(Onset)/結束(Offset)、或是頻率調變……等，皆可經由此階段能更容易取出來。

2.2 系統之基本介紹

本節將介紹本論文在測試時所使用的語料庫，以及所使用的語音分離的線索、整個系統流程的簡介。

2.2.1 語料庫簡介

本論文使用之語料庫是使用 TIMIT 的語料庫。TIMIT 是由好幾個組織，如：國防高級研究計劃所—資訊科學與技術部門(the Defense Advanced Research Projects Agency - Information Science and Technology Office , DARPA-ITSO)、麻省理工學院(the Massachusetts Institute of Technology , MIT)、德州儀器公司(Texas Instruments, TI).. 等共同協力完成的語料庫。此語料庫是用來取得聽覺語音學的一些知識及用來測試改進自動語音辨識器，總共包含了 630 個語者，每個語者共 10 句，一共 6300 句的語料庫。這 6300 句依照美式英文的口音分成八個類別：新英格蘭口音(New England)、北方口音(Northern)、北中部口音(North Midland)、南中部口音(South Midland)、南方口音(Southern)、紐約市口音(New York City)、西部口音(Western)、Army brat 等共八種。每一句話 TIMIT 語料庫皆提供其每一句之句意、每一句之單詞在句中出現的時間，以及每一個音節在句中出現的時間。其所有的語料為取樣頻率 16KHz 的單一頻道的 PCM 檔案，我們使用時為了方便而將其取樣頻率降成 8kHz 來使用，圖 2-11 即為 TIMIT 之某句話之部份音節的聽覺頻譜圖：

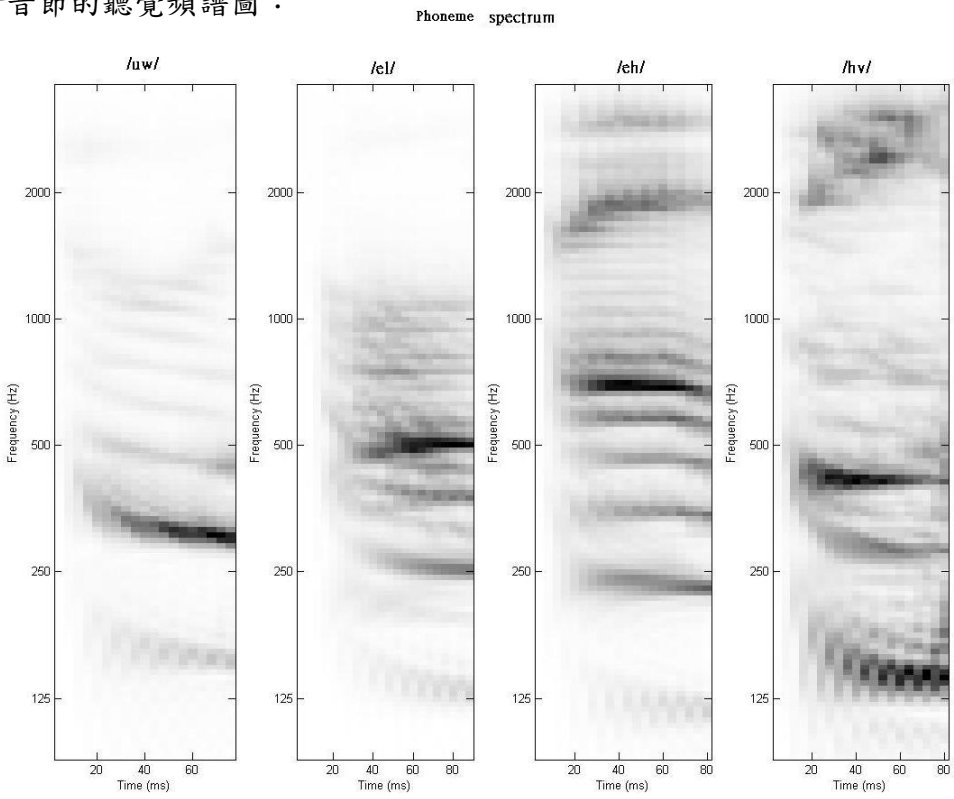


圖 2-11：TIMIT 之部份音節聽覺頻譜圖

2.2.2 系統流程簡介

本論文的系統，是使用聽覺場景分析(Auditory Scene analysis)的方式來做語音分離。聽覺場景分析是模擬人類聽覺系統處理和組織聲音的流程，它的觀念是當聲音進入人耳時，它會先被分析，之後再將分析後的聲音視結果做組合(integrated)或是分離(segregated)。本論文使用的方法就類似此種方式來做分離語音。下圖 2-12 即是本論文之系統流程圖：

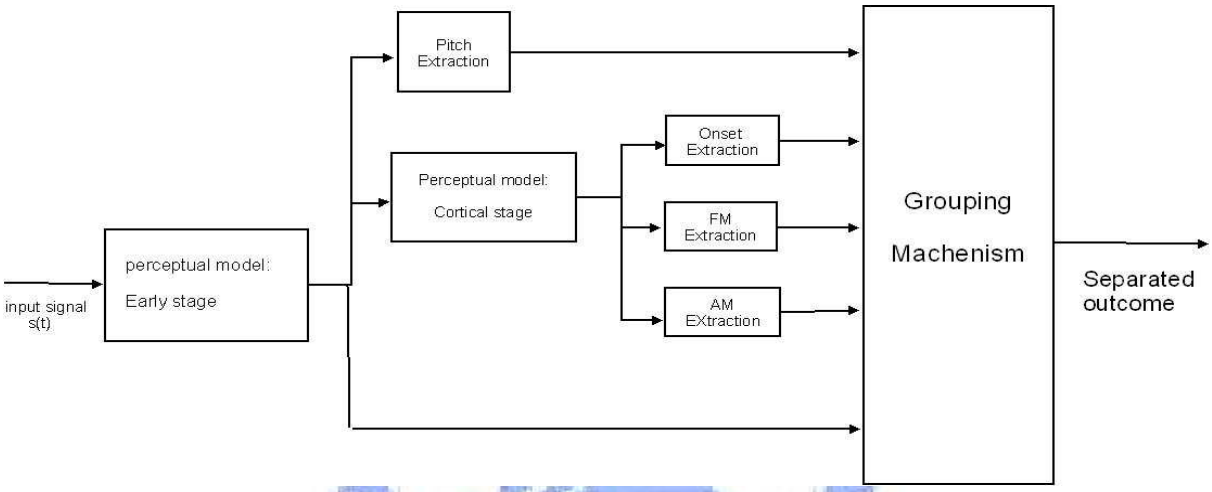


圖 2-12 系統流程圖

我們處理的步驟如下：

(1)先將混合語音送入聽覺模型的早期階段轉換成聽覺頻譜。

(2)這邊分成兩部份，一部份是將聽覺頻譜直接送入音高偵測機制求出混合語音的音高；另一部份是將聽覺頻譜送入大腦階段得到大腦聽覺的四維分析結果後，再各自送入起始點偵測、頻率調變偵測、振幅調變偵測求出此三個線索(Cue)。

(3)將全部線索送入分組的機制將原本的語音分開而得到分開語音的頻譜。

以下的各章中，會詳細介紹各線索之抽取機制及最後組合語音的機制及流程。

第三章

語音特徵之抽取

由前面一章的介紹，可以發現聽覺模型可以顯示出語音上時間-頻率的特徵。本章將介紹本論文所用於語音分離的一些語音的特徵(或稱為線索(cue))，並且說明從聽覺模型中，運用一些來找出語音分離所需要的語音特徵。

3.1 音高擷取

在本節中，首先將介紹語音之音高(Pitch)的定義，接著介紹心理聲學上人類對於音高感知的實驗及其結果。最後，介紹本論文所使用的音高抽取(Pitch Extraction)的機制。

3.1.1 音高之定義及相關心理聲學之實驗

音高之定義，根據 1960 年美國標準協會(American Standards Association)之定義，音高是一個在聽覺感知上面，可以在音樂級數上面做排序。換句話說，音高可以讓人感覺聲音在頻率上的高和低。音高同時也代表著語音在時間軸上的波形的重複性。而和音高最相關的就是泛音(Harmonic 或稱為音線(Partial))。一個複雜聲音(Complex Sound)其組成就是由多數的泛音組成，人類接受到聲音時，會感覺聲音的音高是這些泛音的基頻，故對於一個複雜聲音來說，音高即會是泛音的基頻，換言之，泛音會和音高有倍數的關係，因此在語音的泛音特性上，可由音高來找出該語音的泛音特性(Harmonicity)。人類在接收一個沒有基頻，但是有泛音特性的聲音時，仍就可以找出他的音高。根據 Goldstein 的實驗[6]，他將一段聲音只取後面連續三個高頻的泛音給聽者測試，發現當聽者假設的泛音位置不同時，Ex:取第 9、第 10、第 11 泛音及取第 10、第 11、第 12 泛音，其感覺上的音高會改變，因此 Goldstein 認為，人類在對音高的感知，應該是有一個在頻率上的泛音的模板，和語音的頻率軸的去做對應，有最好對應的模板，此時該模板的基頻即會是其音高，此即稱為”頻譜音高假說(the Spectral Pitch Hypothesis)”，下圖 3-1 即為頻譜音高假說之運作示意圖。

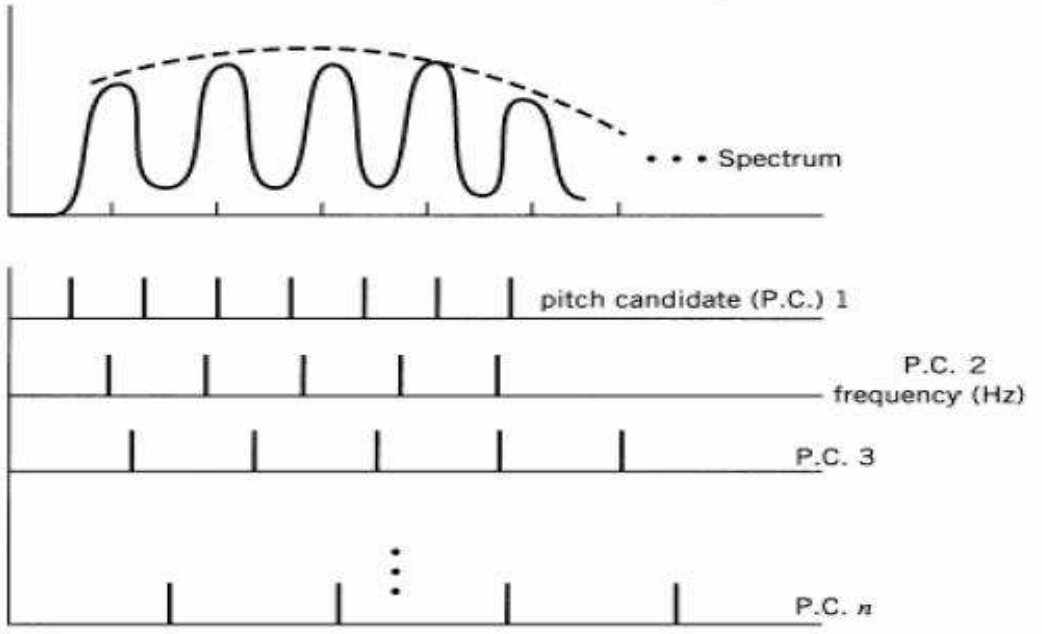


圖 3-1：頻譜音高假說之示意圖(Goldstein-Duifhuis 版本)

上圖上半部為頻譜，下半部為模板，最合適的模板其基頻即為該段語音之音高

3.1.2 泛音模板的建立

根據頻譜音高假說的運作及 Goldstein 的實驗可知，人類大腦系統會製造模板來做對應。按照泛音關係，若給定一個固定的基頻 f_0，則其第 n 泛音和基頻的關係如下式(3-1)：

$$f_i = nf_0 \ , n = 1.....N \qquad (3\text{-}1)$$

f_i 代表的是第 i 個泛音所在之頻率，N 則表示了泛音的總個數。由式(3-1)可知，若依據此特性建立模板，則其泛音的位置會隨著基頻的不同而改變，如此必須針對每一個測試基頻做一個模板。而此處使用的聽覺感知模型，在頻率軸上是以對數的方式分佈，因此泛音和基頻的關係轉換成下式(3-2)：

$$\log_2 f_i = \log_2 nf_0 = \log_2 n + \log_2 f_0 \qquad (3\text{-}2)$$

$$\log_2 f_i - \log_2 f_0 = \log_2 n \qquad (3\text{-}3)$$

上式中，以 2 為底的原因是，聲音中頻率的差異可以用倍頻(Octave)來表示，而且人類聽覺耳蝸上的頻率分佈也是這樣的分佈，例：如果 f_1 和 f_0 差兩倍，則 $\log_2(f_1/f_0)=1$。而由式(3-3)可知，泛音和基頻之間的關係變成了線性的關係，而且不論基頻之數值，第 n 個泛音和基頻的位置差距皆為固定的 $\log_2 n$，因此，針對不同的的基頻，該模板只須向前或向後平移，即可代表不同基頻的模版。因此我們可以利用聽覺模型來模擬出人類制造出來的模板，圖 3-2 即是模擬人類製造模板的流程圖：

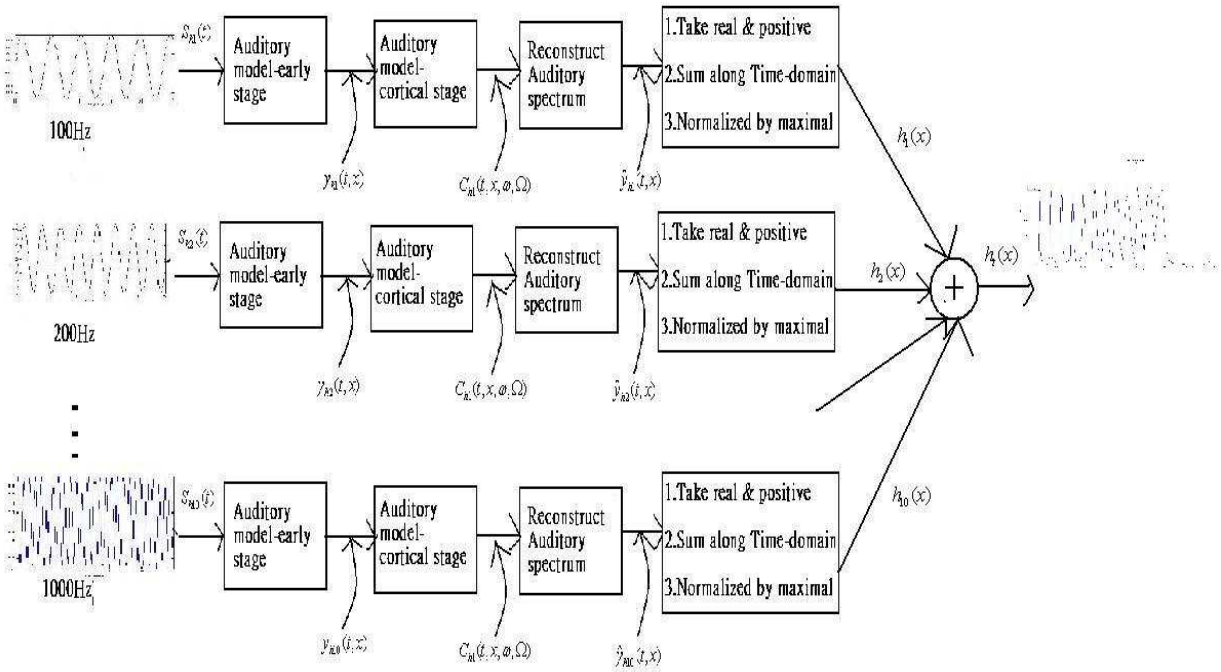


圖 3-2：模擬人類製造模板流程圖

上圖中 $S_{hi}(t)$，i=1…10，表示送入模型初期階段的第 i 個泛音在時間軸上的波形（以 100Hz 為基頻），$y_{hi}(t,x)$ 第 i 個泛音之聽覺頻譜，t 表示時間，x 表示在頻率軸上之位置；$C_{hi}(t,x,\omega,\Omega)$ 則是第 i 個泛音在大腦聽覺處理後的結果；$\hat{y}_{hi}(t,x)$ 則是由大腦聽覺階段重建頻譜之結果，$h_i(x)$ 則是將重建頻譜沿時間軸相加的結果，最後，$h_t(x)$ 即是我做出來的模擬的模版。其作法是先將 100Hz 的弦波送入聽覺初期階段後，其在聽覺頻譜上會在 100Hz 的地方出現波峰，其餘地方是平緩的，之後將此結果送入大腦階段裡，由於人類大腦可以對頻率軸上有不同的解析度，因此這邊在參數選擇上，rate 取 2、4、8、16、32、64Hz，scale 取 4、8 cyc/oct，rate 的範圍是因為目的在做頻率軸上的模版，在時間軸上不需要太精細的能量變化；scale 取 4、8 是因為頻率上的解析度比較高，所以頻率軸上的形狀比較清楚。之後再由大腦聽覺的反應重建回頻譜，此代表著該頻譜通過了大腦聽覺處理後所變回來的頻譜。之後將重建回的頻譜沿時間軸相加並標準化後，即可得到 100Hz 在我們的聽覺頻譜的頻率軸上應該會有的形狀，接著再以同樣方法去做 100Hz 兩倍頻、三倍頻…10 倍頻的模版形狀，最後把它們組合起來，即會是我的模版的形狀，用 100Hz 的原因是，100Hz 在聽覺頻譜的頻率軸上的形狀，會是一個完整的波形，而且可以到 10 倍頻的內容都是完整的波峰，如此一來，我們可以利用其平移來模擬出不同基頻的模版。下圖 3-3 即是不同基頻的模版圖及轉換到頻率軸上之模版圖

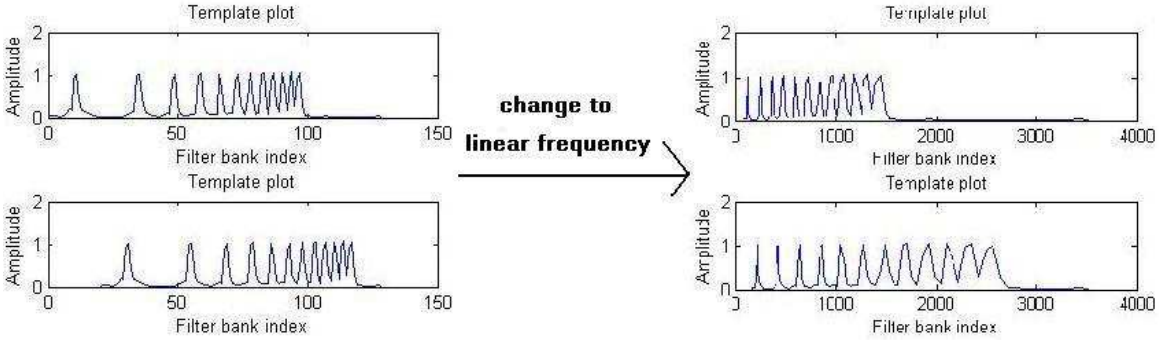


圖 3-3：不同基頻模版比較圖(濾波庫指標/頻率)

由上圖可以知道，由於聽覺模型在頻率軸上的設計是以對數，一個倍頻由 24 個點來代表，因此在濾波庫指標上，不論其基頻之位置在哪哩，倍頻之間的間隔接是固定的，換回頻率軸上，更可以看出平移即可做出不同基頻的模板。

3.1.3 音高抽取之機制

建立好模板之後，我們就利用下面式子來做計算：

$$R(t_c,\tau_x)=\frac{1}{N_{fc}}\sum_{x=1}^{128}y(x;t_c)h_t(x-\tau_x)\ ,\ \tau_x=1\cdots 84 \qquad (3\text{-}4)$$

式(3-4)是說明，在固定某個時間點下，模板平移距離為τ_x時，該時間頻譜和模板做相關性(Correlation)的結果。N_{fc}所有濾波庫的頻道個數。τ_x的範圍從 1 到 84，原因是因為音高通常不會超過 1KHz(在濾波庫的位置上為 84)。接著從$R(t_c,\tau_x)$取出波峰和波谷，算出每一個的峰谷比(Peak-to-Valley ratio)，峰谷比最大的位置，即是音高的所在位置，其原因為，當模板對應的泛音的地方時，其在相關性的數值上相較於週圍得地方會比較大，而模板對到的泛音越多，則其相關性之值和週圍的值相差會越多，因此峰谷比最大表示該點和週圍其他點的差距最大(亦即波峰的變化最急促)，即代表著模版對應聽覺頻譜的最佳對應位置，此時之模板之位移，即代表該時間點的音高。下圖 3-4 即是從英文語句\Come home right away\中的第 100 個音框的交互相關性之圖：

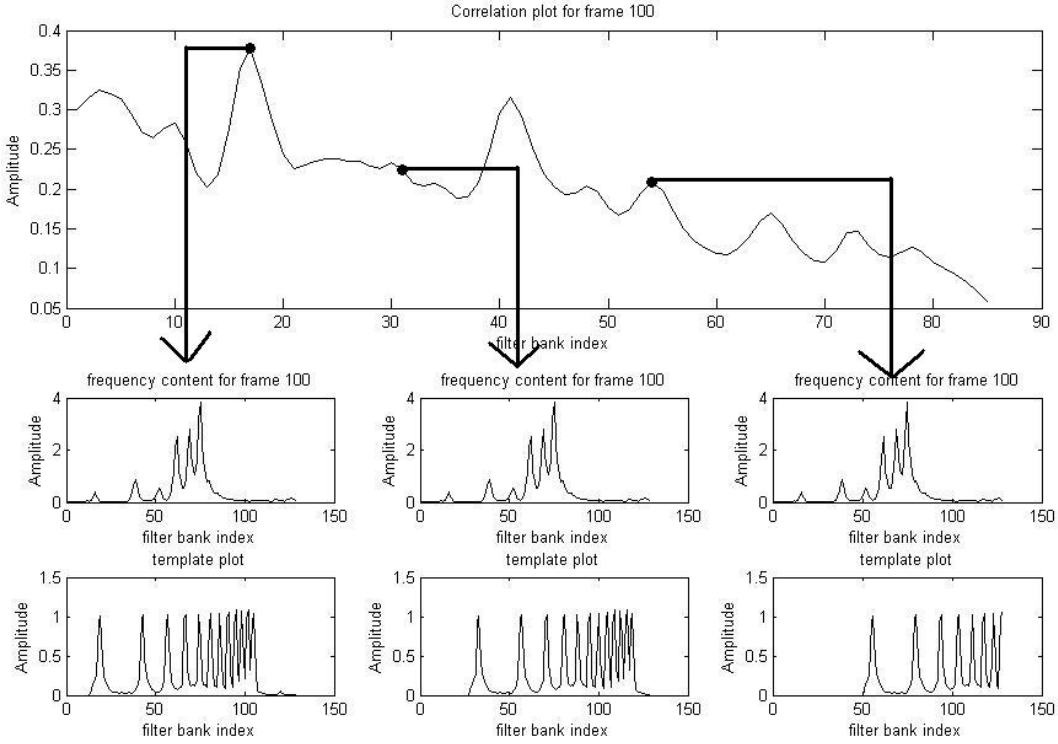


圖 3-4：英文語句\Come home right away\第 100 個 frame 的交互相關性圖

由圖 3-4 左邊的圖，模板正好對到基頻的位置，反應在相關性的結果圖上，形成波峰而且峰谷比是最大的；中間的圖，模板並沒有對到任泛音的位置，所以峰谷比就很小；右邊的圖，模板對到第三個泛音，因此它在相關性結果圖上有一個波峰，但是其峰谷的比就沒有像左邊的圖這麼高。所以整個流程圖如下圖 3-5：

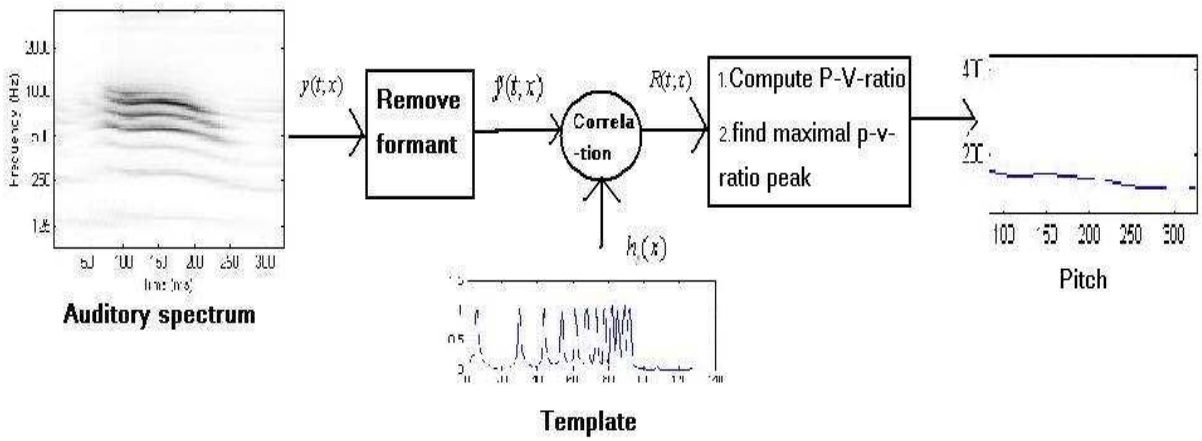


圖 3-5：音高抽取機制流程圖

首先，根據[12]，我們可以利用大腦聽覺階段找出語音共振峰(formant)，將 scale 設

定為 1 cyc/oct，之後將該音框之頻率軸上的成份送入，即可得到該音框的共振峰的大致情形。去除掉共振峰的原因是因為語音的第一共振峰通常會使 300Hz~900Hz 的頻率成份放大，因此在沒有把共振峰盡量去除的情況下，用模板來擷取音高很容易出現“倍頻錯誤(Octave error)”。圖 3-6 即是用大腦聽覺階段求出共振峰：

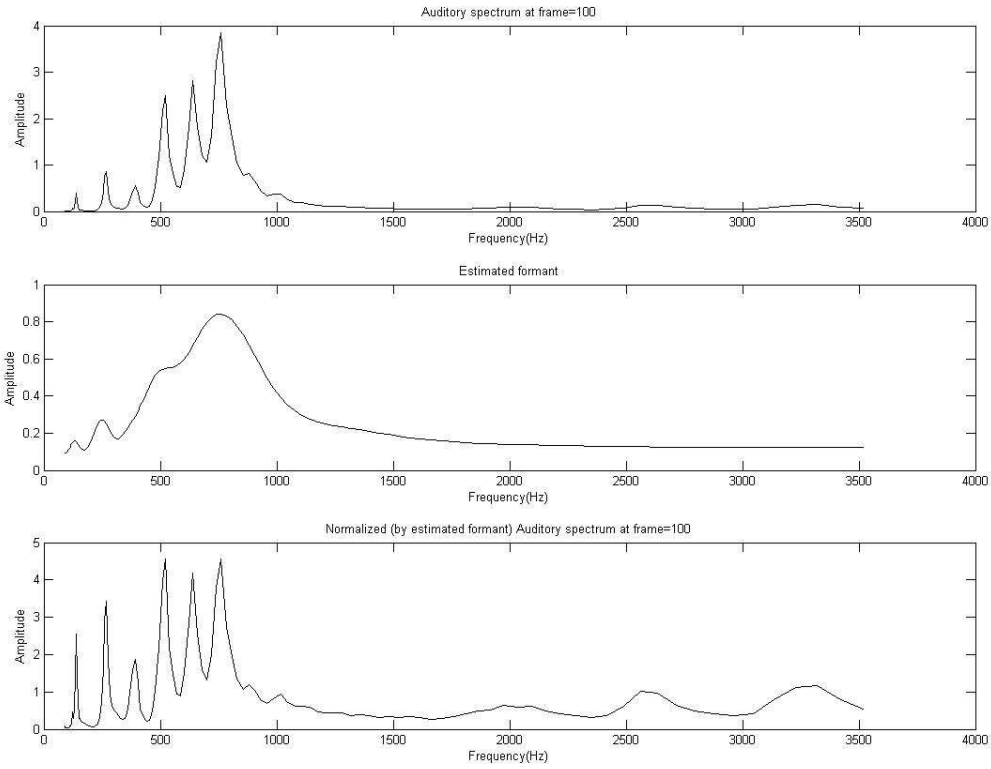


圖 3-6：大腦聽覺階段所求出之共振峰

將共振峰去除後，接著和模板做交互相關性計算，再從結果中取出峰谷比最大的即是我音高在頻率軸上的位置。下圖 3-7：即是以英文語音\We‘ ve done our part\來做測試之結果：

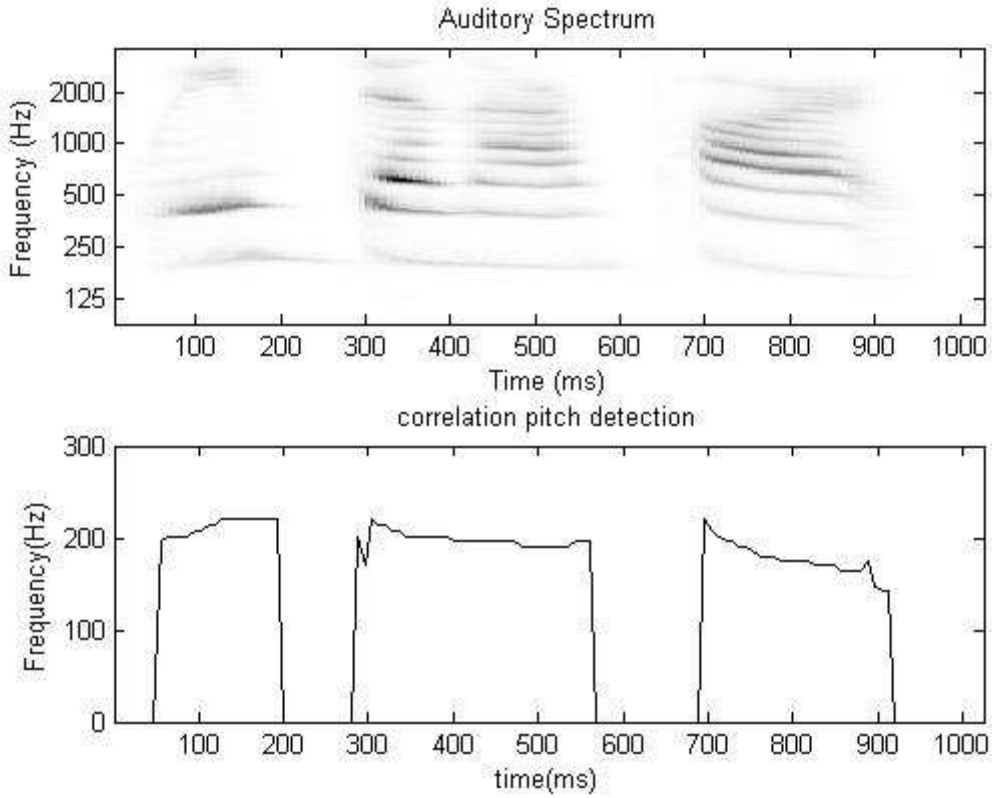


圖 3-7：英文語音\We have done apart\的測試結果

3.1.4 音高抽取機制之實驗結果

本節之實驗使用 TIMIT 中使用新英格蘭地區的口音去取做模擬。我們從所有新英格蘭地區的語料中取出十大美國母音(American vowel)：aa、ae、ah、ao、eh、er、ih、iy、uh、uw 去做我們的音高抽取，並和一常用的音高抽取的方法—平均振幅差異函數(Average Magnitude Difference Function)來比較。由於平均振幅差異函數是一已知在母音上面的音高抽取常用而且穩定之方法，因此我們要用我們的音高抽取藉由計算相關係數來比較相關性，結果如表 3-1 和圖 3-8、圖 3-9 所示：

表 3-1：和 AMDF 之相關係數分佈

correlation coefficient	母音個數	百分比(%)
0.95~1	1534	75.45%
0.9~0.95	126	6.20%
0.85~0.9	83	4.08%
0.8~0.85	89	4.38%
0.75~0.8	77	3.79%
0.7~0.75	66	3.25%
0.65~0.7	13	0.64%
0.6~0.65	11	0.54%
0.55~0.6	16	0.79%
0.5~0.55	7	0.34%
0.45~0.5	6	0.30%
0.4~0.45	2	0.10%
0.35~0.4	0	0.00%
0.3~0.35	3	0.15%
總計個數	2033	100.00%

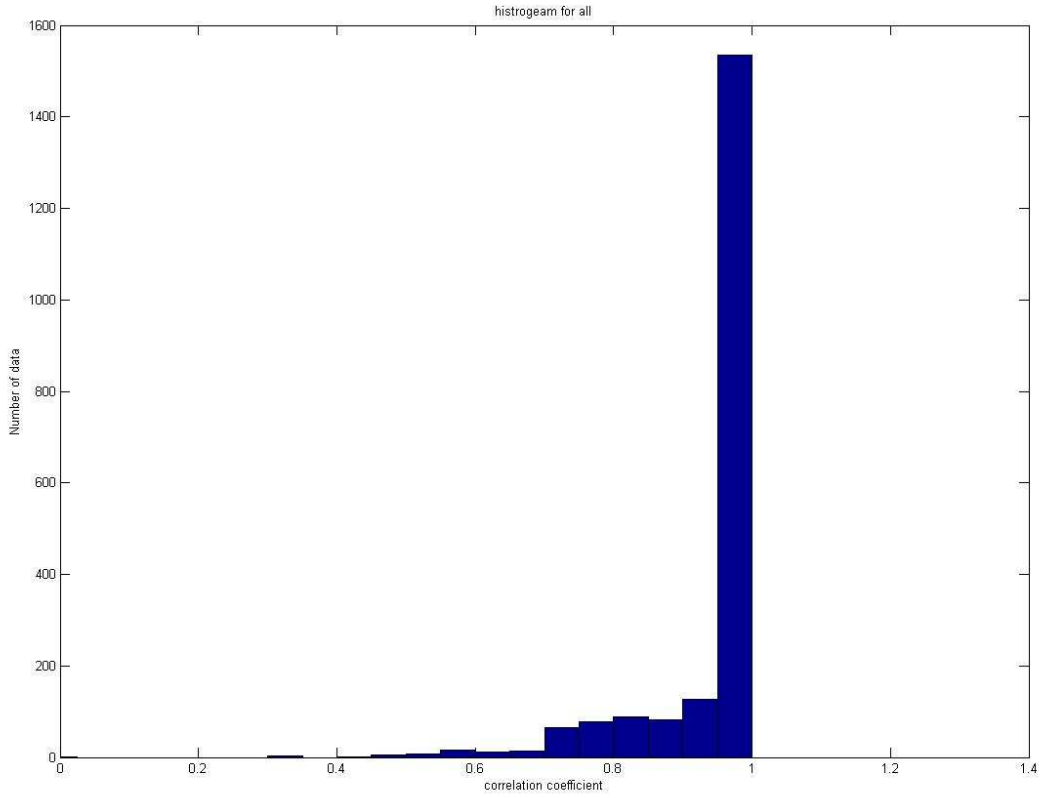


圖 3-8：和 AMDF 之相關係數之長條統計圖

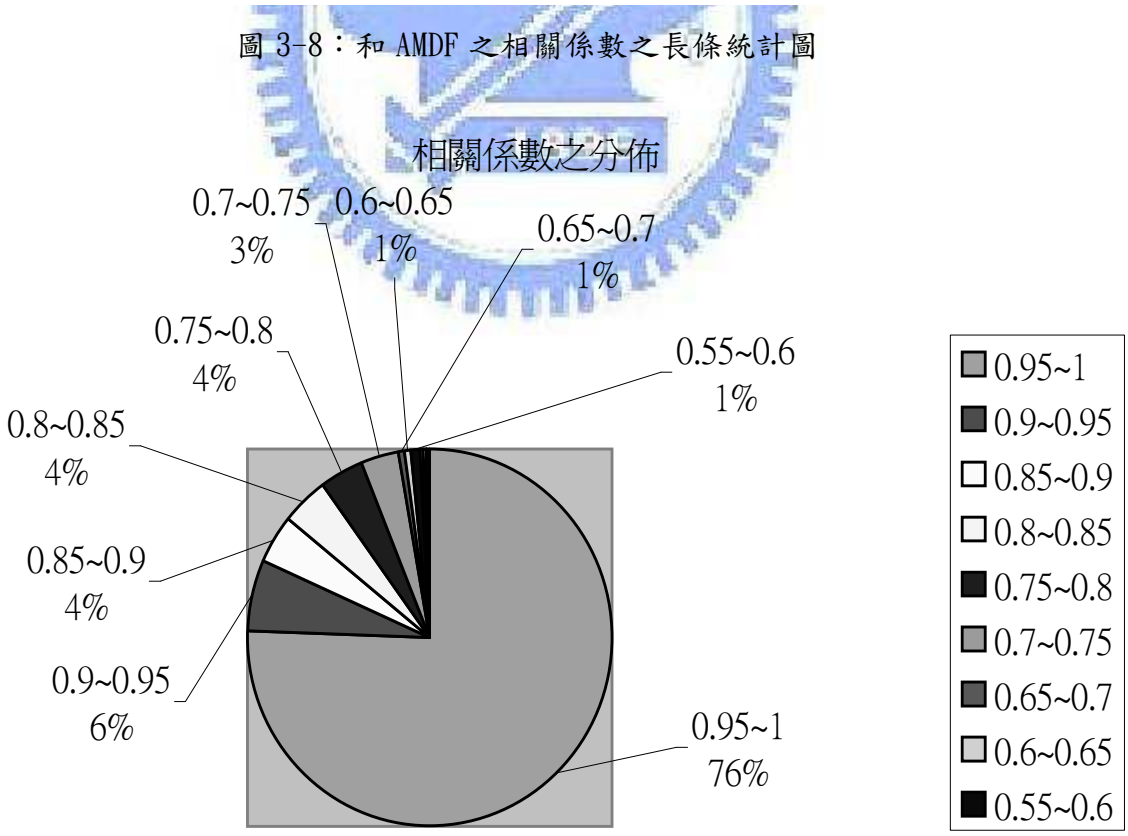


圖 3-9：和 AMDF 之相關係數百分比分部圖

由上面的圖表，我們可以發現，在這個測試語料中，本論文的音高抽取機制，和 AMDF 之相關係數達 0.85 以上的約佔全部比例的 85%以上，因此我們的音高抽取機制基本上是具有一定正確性。如此在面對多人語音混合的時候，我們也可以利用同樣的方法將多人得音高抽取出來。

3.2 頻率調變擷取

在本節中，將介紹頻率調變在語音上的定義及本論文中所使用的擷取方式。

3.2.1 頻率調變之定義

頻率調變(Frequency Modulation)，或稱頻率轉移(Frequency Transition)在語音處理上指的是一個語音的頻率隨時間的變化量。根據一些研究發現[9][10][11][21]，人類聽覺系統對於在同一時間的頻率變化會有感知，而且在多個聲音混合的狀況下，人類聽覺系統會去將同一時間內頻率變化相同的視作是同一個聲音來源，在一些的語音分離或語音分組的系統當中皆有使用到頻率調變來做分離或分組的線索[17]。

3.2.2 頻率調變的擷取-運用聽覺模型

在聽覺模形的大腦聽覺階段裡，定義了 rate 和 scale 兩個參數。Rate 的定義代表時間上的能量變化，亦可視為移動波紋刺激源每秒鐘在頻譜的低頻邊界上通過的波紋週期，又稱為波紋速度(Ripple velocity)[23]。而 scale 另一個含意是移動波紋刺激源在頻率軸上每個倍頻內有幾個週期，又稱為波紋密度(RIpple density)。圖 3-10(a)、圖 3-10(b)即可以看出移動波紋刺激源和 rate 及 scale 各自的關係：

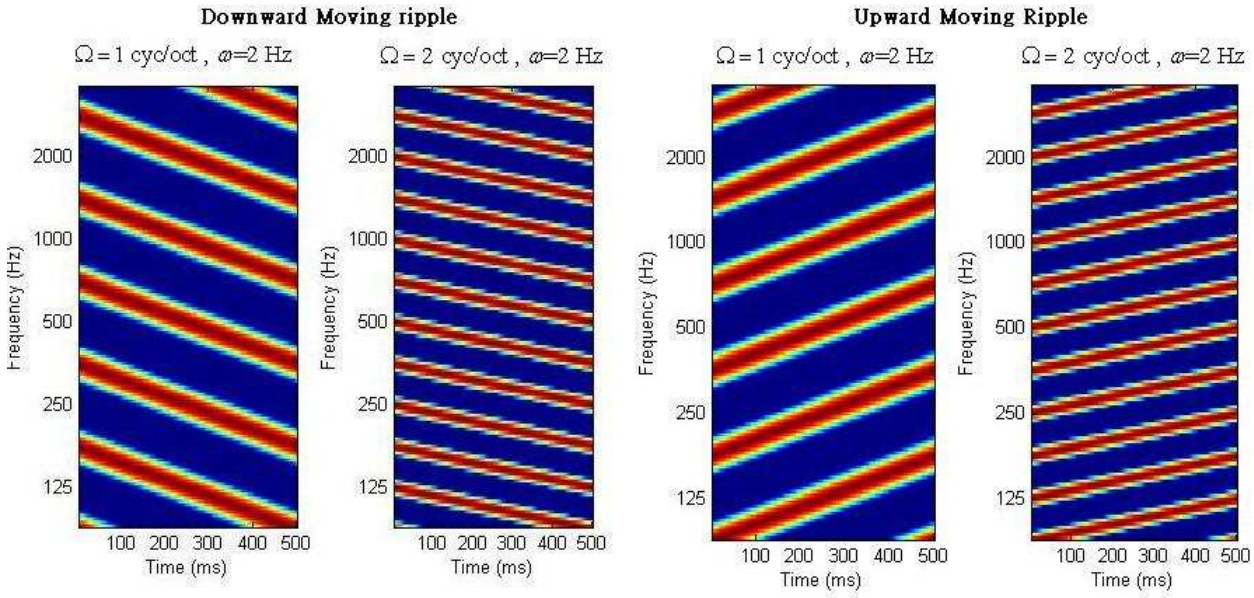


圖 3-10(a)：rate 固定下改變 scale 的移動波紋刺激源比較圖

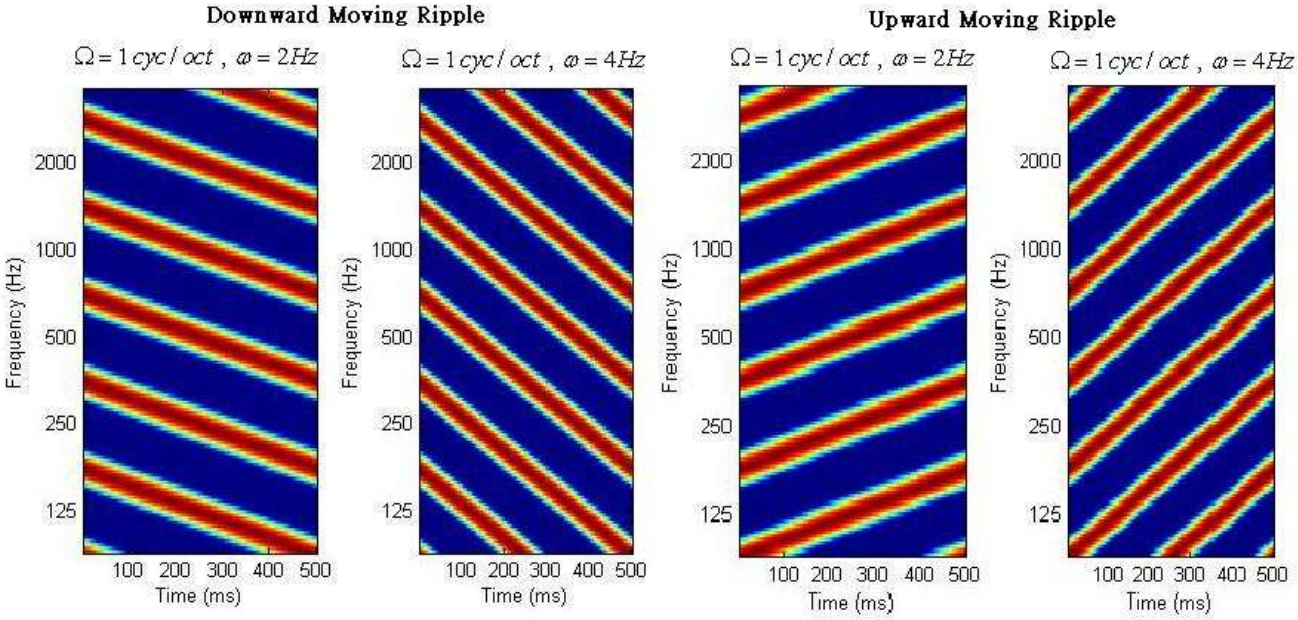


圖 3-10(b)：scale 固定下改變 rate 的移動波紋刺激源比較圖

由圖 3-10(a)上可以看出，改變 scale，會改變頻率軸上的密度，scale 越高，頻率軸上的密度越密;由圖 3-10(b)可以看出，在頻率軸的密度固定下，通過低頻邊界的週期改變，會改變移動波紋刺激源的波峰在每單位時間(ms)內的頻率變化，若將每個時間音框上的頻率成份拿出來比較，則可以很明顯的看出動波紋刺激源的波峰在頻率軸上的移動，且其移動速度會是 rate，下圖 3-11 即表示波紋刺激源的波峰在頻率軸上的移動情形：

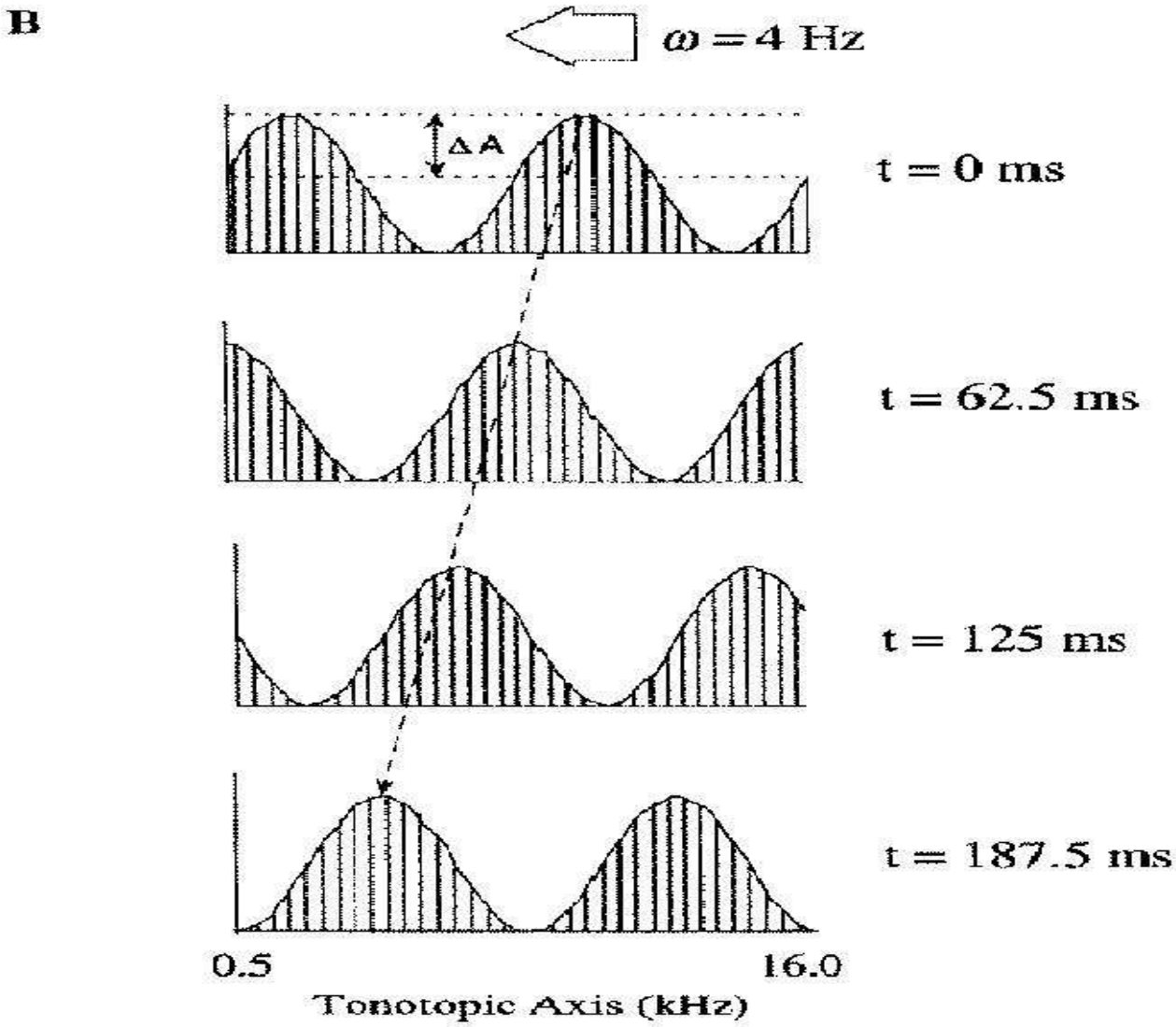


圖 3-11：移動波紋刺激源在 rate=4Hz 時，波峰移動之情形。[19]

由上圖之結果可以看出，頻率軸上的波峰隨時間改變頻率軸上的位置，此即為聽覺頻譜中隨時間的頻率變化，因此我們可以利用此來求出頻率調變之線索，下圖 3-12 即是擷取頻率調變的流程圖：

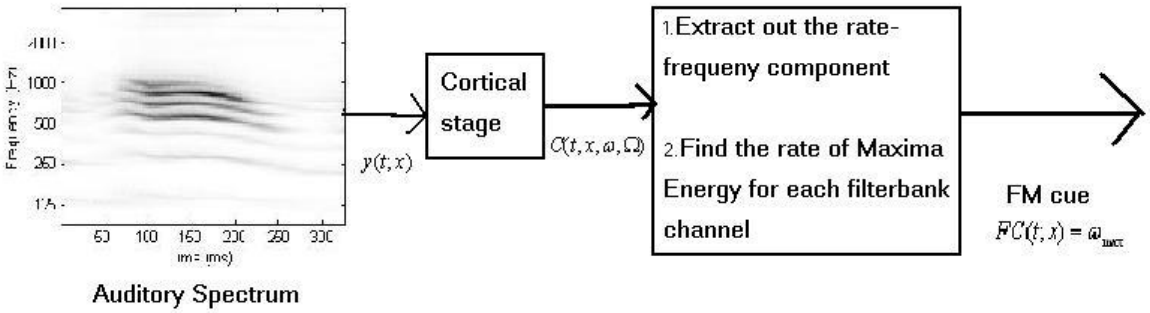


圖 3-12：某一時間 t 的頻率調變擷取的流程圖

處理的步驟如下：

(1)將某一個時間 t 的聽覺頻譜 y(x;t)送入大腦聽覺階段分析，得到四維(時間-頻率-rate-s-cale)的結果。這邊 rate 取 0.125~16Hz，中間共有 80 點，scale 取 4 cyc/oct，原因是這個位置的反應會將聽覺頻譜上 500~1000Hz 的成分解析清楚；rate 的範圍選擇是因為，在大腦聽覺分析中，大部份的語音經過此階段分析後，在這段區域內有比較強的反應，因此取這個範圍內，來看我的聽覺頻譜圖上的變化[3]。

(2)從大腦聽覺階段的四維結果取出 rate 和頻率的結果出來。

(3)從(2)之結果中取出能量最大的 rate，此用意即類似拿不同 rate 的移動波紋刺激源去和該時間附近的聽覺頻譜去做摺積。最後即可得某一個時間和頻率位置上的頻率調變線索 FC(t;x)。

因為摺積代表著有從進來的訊號中抓出相似的成份，因此將聽覺頻譜和移動波紋刺激源做摺積的最大值，即代表著我的聽覺頻譜裡，具有和對應的移動波紋刺激源最多的成份，因此該移動波紋刺激源的 rate 值，即可代表該時間附近的聽覺頻譜的頻率變化。圖 3-13(a)(b)即是將原聽覺頻譜和反應最大之 rate 的移動波紋刺激源來比較：

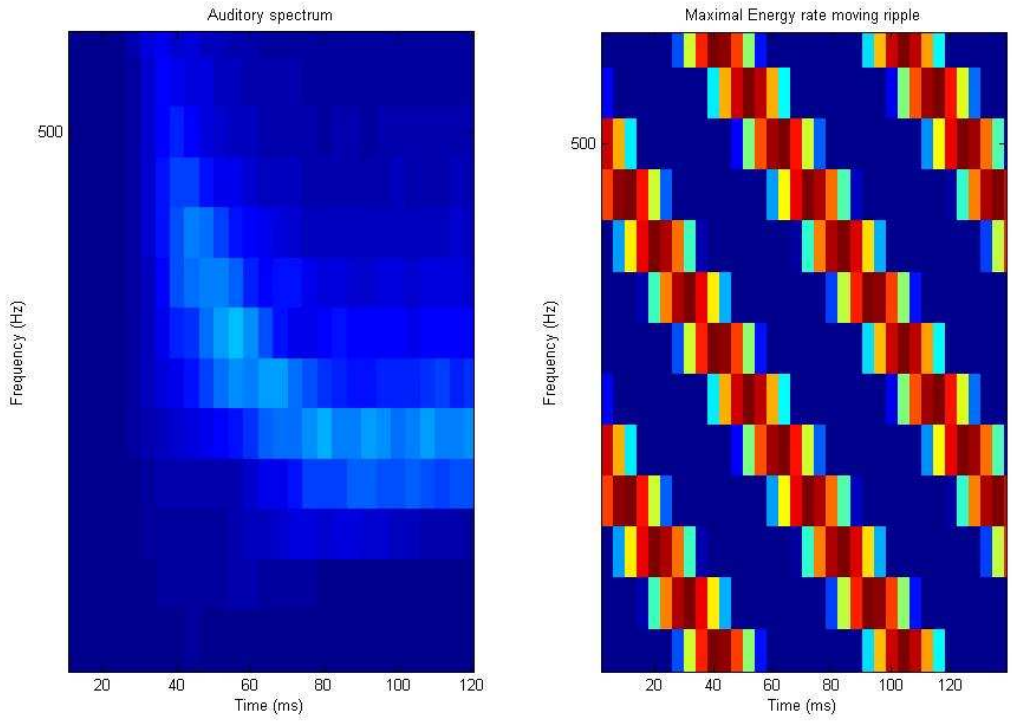


圖 3-13(a) 聽覺頻譜和反應最大之 rate 的移動波紋刺激源來比較圖(頻率下降)

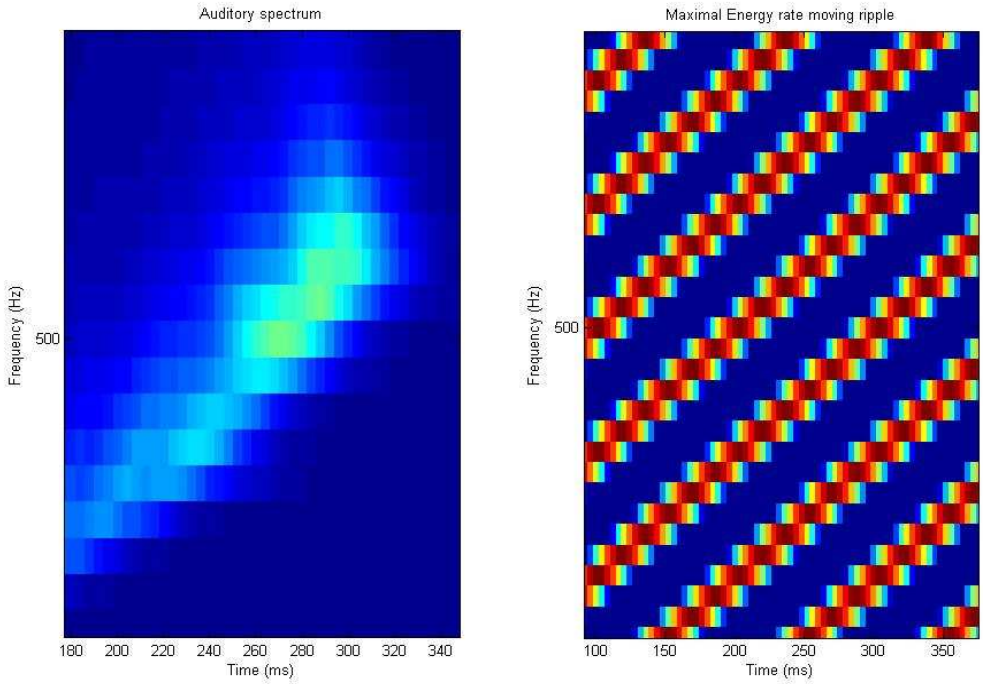


圖 3-13(b) 聽覺頻譜和反應最大之 rate 的移動波紋刺激源來比較圖(頻率上升)

由上面的結果，可以看出，rate 反應最大的移動波紋刺激源，其隨時間的頻率變化，會和聽覺頻譜上的頻率變化很相近。因此我們可以拿來利用為語音分離之一個線索。下圖 3-14(a)(b)即是頻率調變的線索圖。

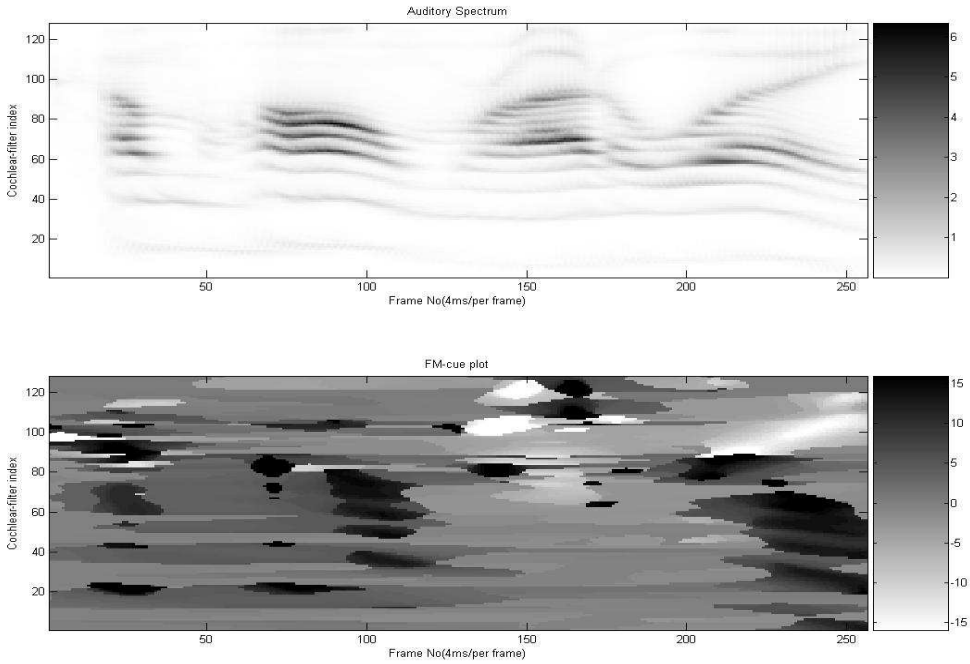


圖 3-14(a)：頻率調變的線索圖(單一語音)。

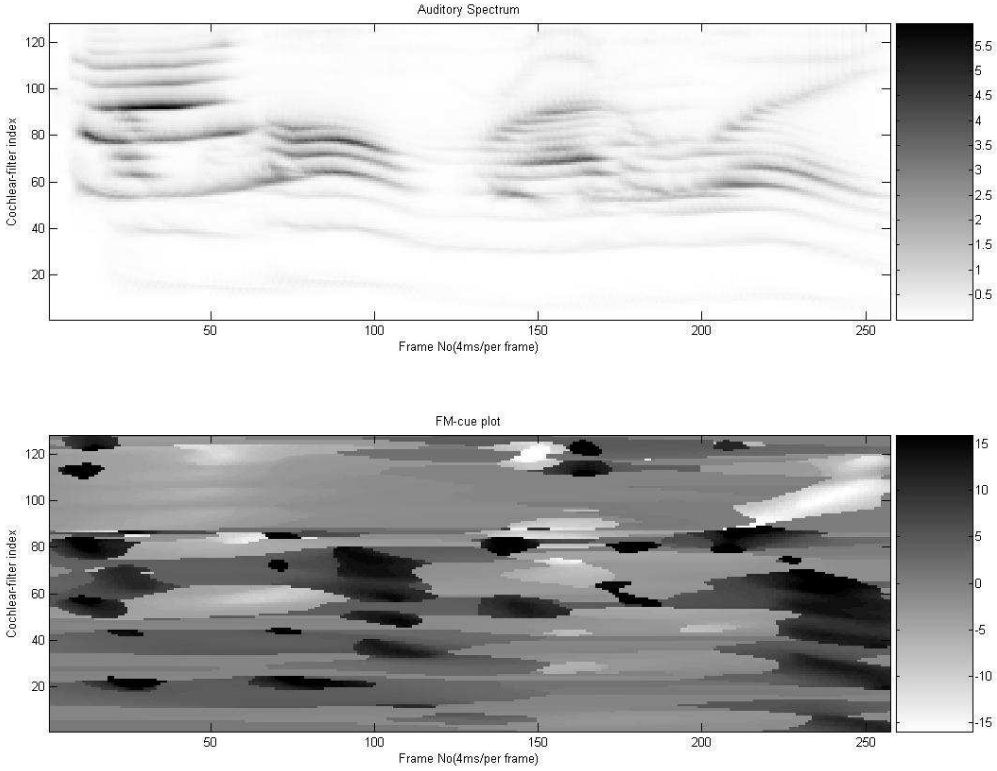


圖 3-14(b)：頻率調變的線索圖(混合語音)。

由上圖可以看出，在同一個語音的泛音位置下，其頻率調變的數值是相似的，因此證明此方法有其正確性及物理意義。

3.3 聲音起始點和終止點擷取

在本節中，將介紹語因起始點和終止點的定義及本論文中所使用的擷取方式。

3.3.1 起始點和終止點之定義

聲音起始點(Onset)和終止點(Offset)的定義是，短時間內的能量上升和下降，一

般是指在 30ms 之內的能量上升或下降的變化[22]。在本論文中，聲音的起始點和終止點的作用比較像是語音偵測(Voice Activity detector)，偵測混合語音中，哪裡是非語音的地方，哪邊是語音的地方，辨別出語音的地方再用本論文的語音分離的機制去處理。以便減少不必要的運算。

3.3.2 起始點和終止點的擷取-運用聽覺模型

由於本論文之大腦聽覺階段，可以解析出在時間不同的變化量，因此在這邊運用這個特性來找出聲音的起始點和終止點。圖 3-13 即是起始點和終止點的擷取的流程圖：

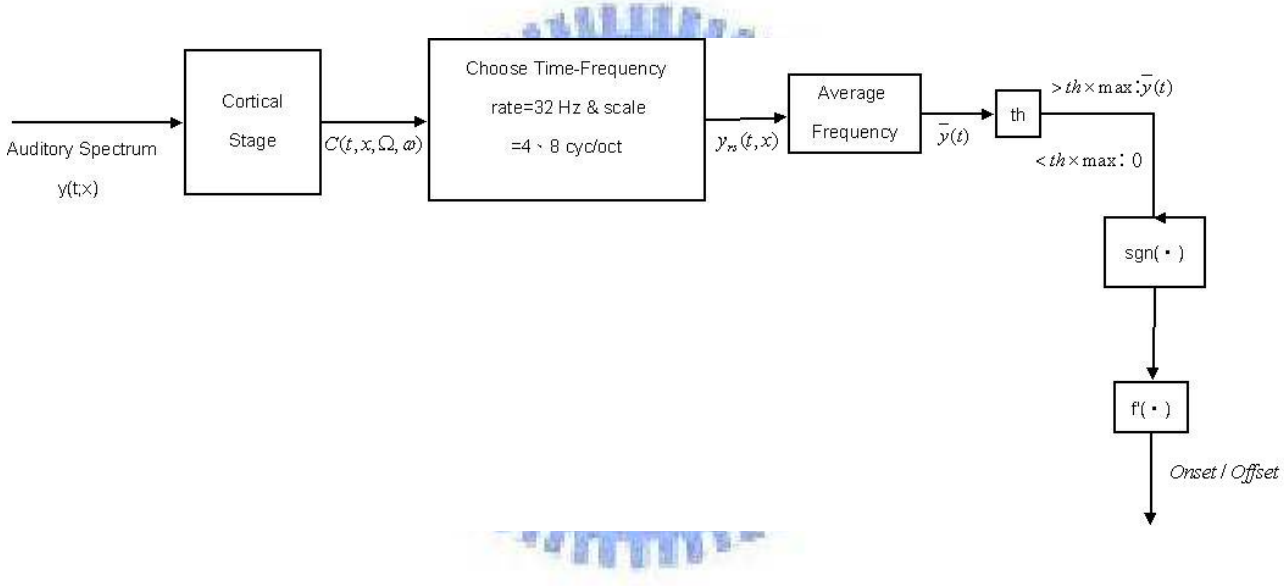


圖 3-15 起始點和終止點的擷取的流程圖

處理的步驟如下：

(1)將全部混合的語音的聽覺頻譜送入大腦聽覺階段，這邊的 rate 取 32Hz，這樣取的原因是因為 rate 32Hz 可以抓出 30ms 左右的變化，scale 取 4、8 cyc/oct，其原因是因為 4、8 cyc/oct 在頻率軸上的解析度很高，因此可以清楚的抓出泛音所在的位置和能量。

(2)從四維的結果中取出時間-頻率的成份出來，然後對頻率軸做平均，因此在有泛音的地方，能量會比較大。

(3)先將聽覺頻譜送入一個中位數濾波器(Median filter)將聽覺頻譜平緩化之後，再將

其切成以 250ms 為一區塊，原因是通常人口腔的變化約 4Hz，剛好為 250ms 的時間長度間的變化。

(4)我們設定一個門檻值 th=0.16，當在 250ms 時間長度內的點之能量值小於該長度的最大值的 0.16 倍時，視為沒有語音的部份，設為 0。

(5)通過 sgn(t)的函數之後，得到的值 1 是有聲區塊，0 是無聲區塊。Sgn 函數如式(3-5)：

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \qquad (3\text{-}5)$$

之後再微分，正值者即為起始點，負值者即為終止點。

上面的步驟做完後，就可以得到時間的起始點和終止點，得到此之後就可以將有聲區塊的部份送入我們的機制去處理。圖 3-16(a)、3-16(b)即是單一和混合的聲音所求出的起始點和終止點：

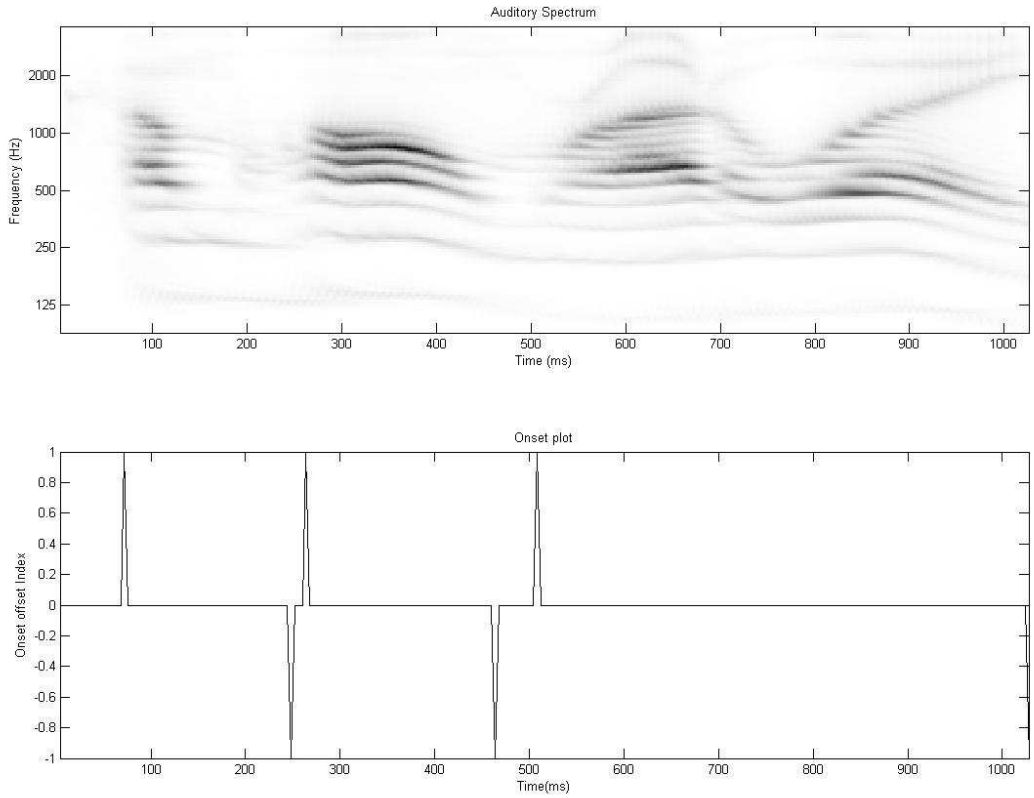


圖 3-16(a) 單一語音之起始點和終止點

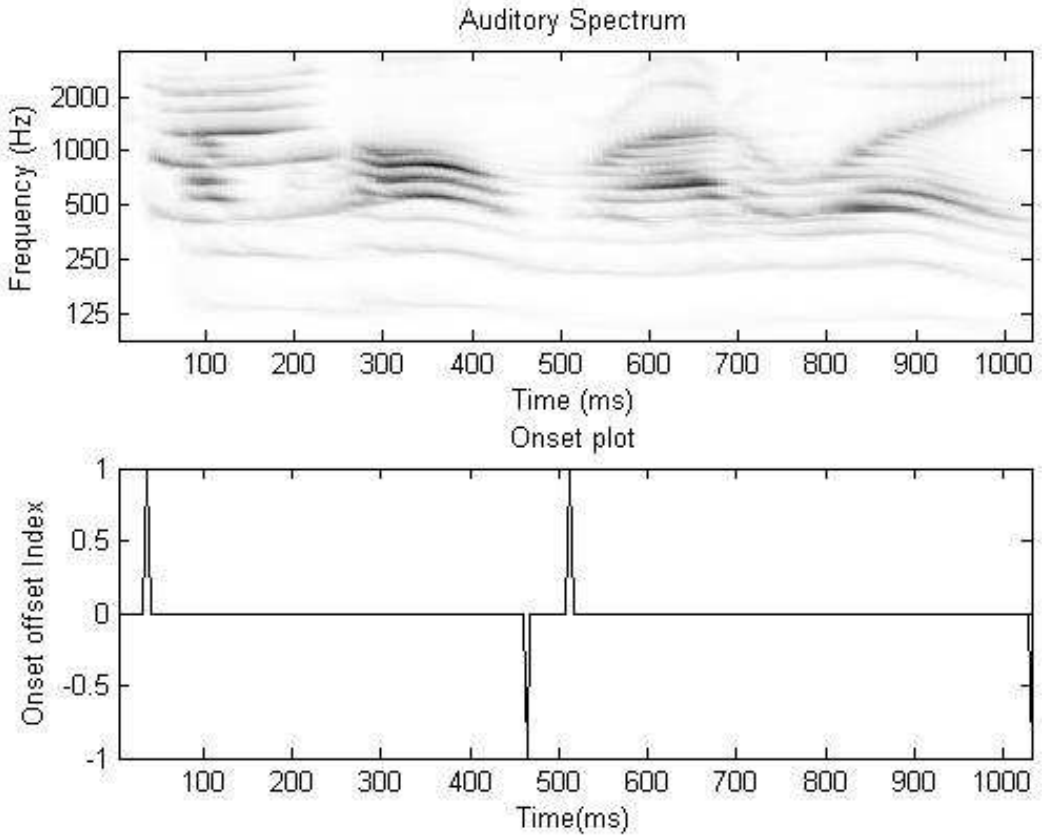


圖 3-16(b) 混合語音之起始點和終止點

3.4 振幅調變擷取

在本節中，將介紹振幅調變在語音上的定義、一些心理聲學之實驗結果及本論文中所使用的擷取方式。

3.4.1 振幅調變之定義

振幅調變(Amplitude Modulation)指的是聲音隨時間上的振幅變化。根據[8]和[14]，人類聽覺系統對於具有同樣振幅變化或是同樣振幅變化速率的不同的頻率成份，會將其視為同一個聲音來源。此和 3.2 節所提之頻率調變類似，此種會以相同變化來當做語音分離依據者稱為共同結果分組(Common Fate)。

3.4.2 振幅調變之擷取-運用聽覺模型

在這邊的振幅調變，主要是針對語音在各頻率成份的波封(Envelope)變化。此波封變化會和所講的音節不同而有不同的變化，所以不同語音加成後波封的變化應該會有所改變，而我們的大腦聽覺階段的模型，根據前一章所敘述的，大腦聽覺階段的反應區域，是由送入移動波紋刺激源所量測的結果，換言之，大腦聽覺模型對於送入聽覺頻譜的處理，是從聽覺頻譜中找出和移動波紋刺激源相似的部份把他取出來，又根據[4]，移動波紋刺激源當 scale=0 代表的是最單純的時域調變，因此這邊利用這樣的特性。圖 3-17 即是振幅調變擷取的流程圖：

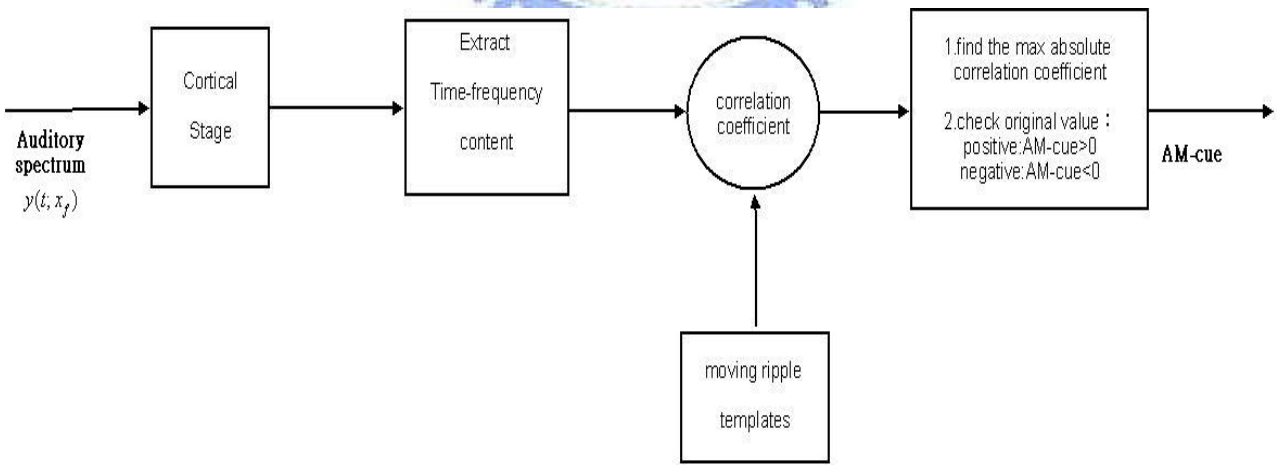


圖 3-17 振幅調變擷取的流程圖

我們處理的步驟如下：

(1)將某一頻率頻道 x_f 的聽覺頻譜送入大腦聽覺階段分析，這邊的 rate 取的是 16Hz，因為語音混合後，用以抓出混合後比較細微的能量變化，以利於辨別。Scale 取 2

cyc/oct，原因為我們不需要太過細微的頻率軸上的解析度。在四維的中我們取出時間-頻率軸的成份出來。

(2)做出不同 rate 的移動波紋刺激源模板(moving ripple templates)，長度 12ms(約 3 個音框)，scale 設為 0，這邊移動波紋刺激源的 rate 的範圍是 2~64Hz，做出來會是代表著這個移動波紋刺激源的能量變化，圖 3-18 即是代表著不同移動波紋刺激源的能量變化：

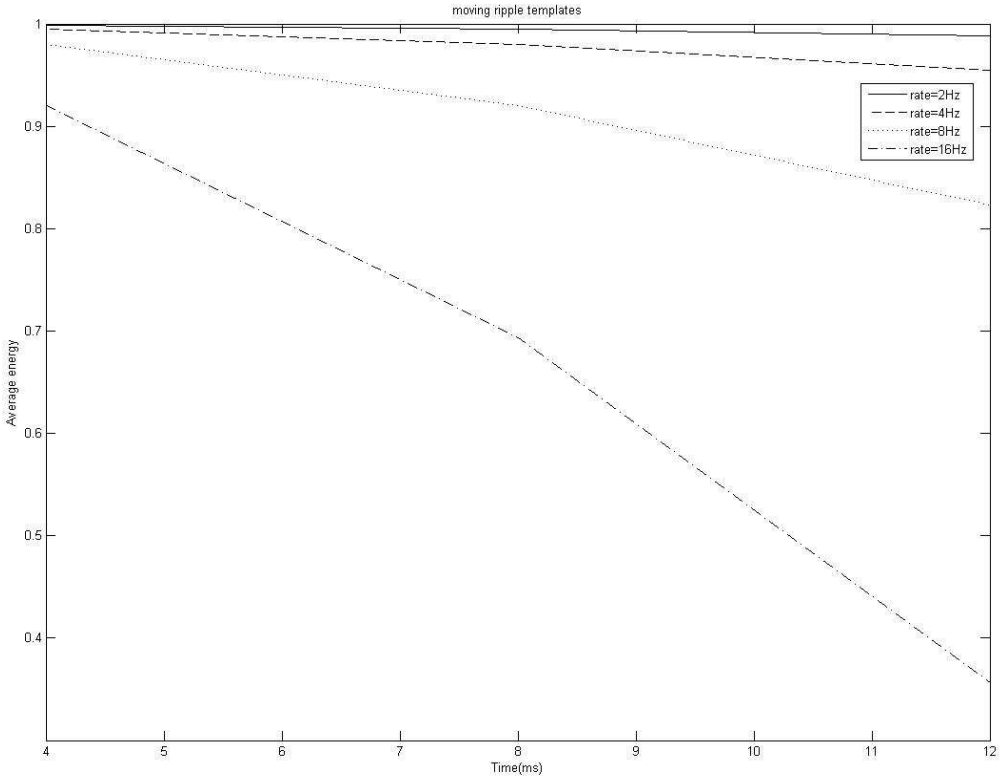


圖 3-18 代表著不同移動波紋刺激源的能量變化

(3)將聽覺頻譜取和移動波紋刺激源同樣點數，然後去做相關性係數。找出相關性係數絕對值最大者之 rate 值，即是代表和該 rate 之移動波紋刺激源的能量變化最相近。相關係數絕對值最大可正可負，若絕對值最大為正，則表示其能量是下降的，絕對值能量為負，則表示其能量是上升。圖 3-19(a)(b)即是振幅調變的線索的圖。

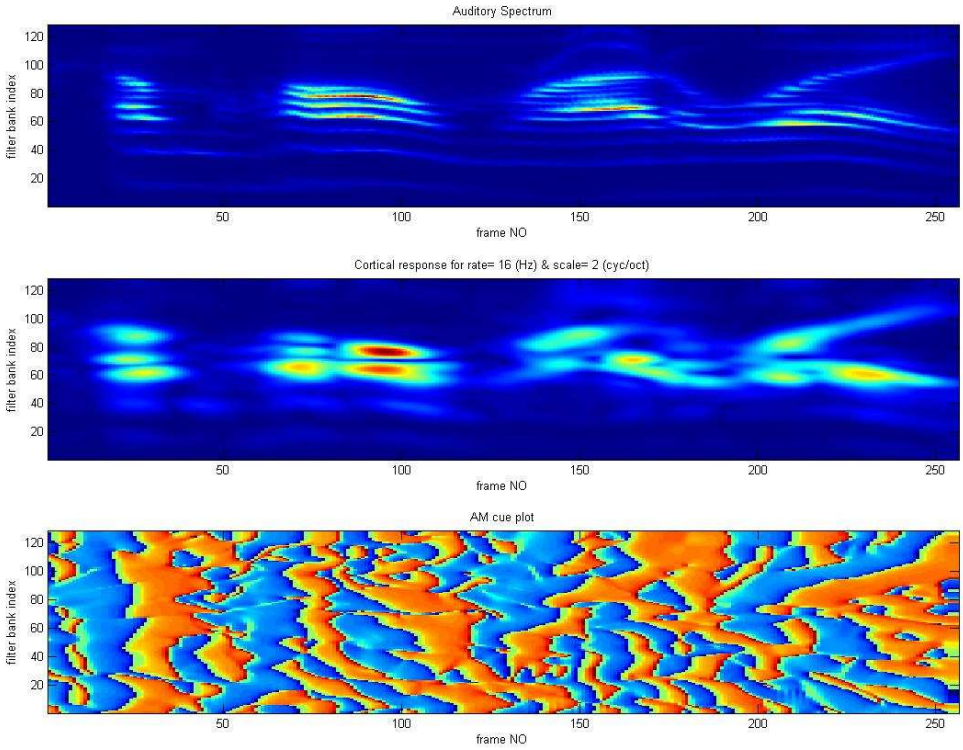


圖 3-19(a)：振幅調變的線索圖。(單一語音)

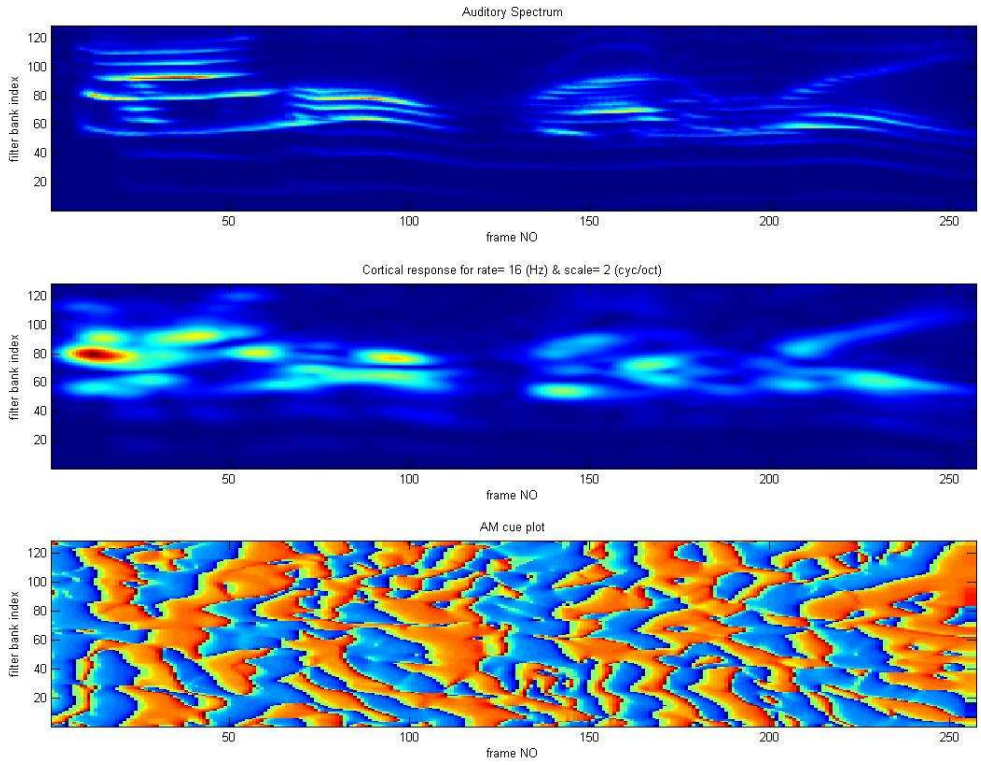


圖 3-19(b)：振幅調變的線索圖。(混合語音)

上面之振幅線索圖中的顏色，即代表了該位置的振幅變化情形。由上面兩圖可以看的出來，振幅調變的線索上可以看出每個頻率頻道上的能量變化情形，由圖上可以看出原語音的泛音的分佈，而且在語音每個泛音的位置附近，皆有相似的振幅變化，因此可以拿來利用為語音分離的線索之一。

第四章

語音分離

本論文已於上一章經說明了，如何藉由聽覺感知模型擷取出語音分離的線索。在本章中，將介紹系統所使用的語音分離機制—類神經網路(Artificial Neural Network, ANN)中的自組織映射圖(Self-organizing Map)。首先將會簡單介紹類神經網路、自組織映射圖以及本論文所使用之聽覺感知模型的語音重建機制。其後，介紹本論文如何使用自組織映射圖來達到分離語音的目的，最後比較語音分離的效果。

4.1 類神經網路簡介

類神經網路是以電腦來模擬人類神經細胞網路行為。人類神經細胞網路具有累積經驗、儲存知識、傳遞訊息等功用。類神經網路也有同樣的功能。類神經網路是由許多非線性的人工神經元(或稱運算單元)和人工神經元之間的連結所組成，這些人工神經元是以平行且分散，所以類神經網路可有效的分析大量的資料，而且其又具有學習的特性。本節將簡單介紹類神經網路的關鍵核心—人工神經元的架構、類神經網路的架構以

及類神經網路的學習方式。

4.1.1 人工神經元

人類的神經元(或稱神經細胞)，是神經系統的基本功能單位。下圖 4-1 即是人類神經元的示意圖：

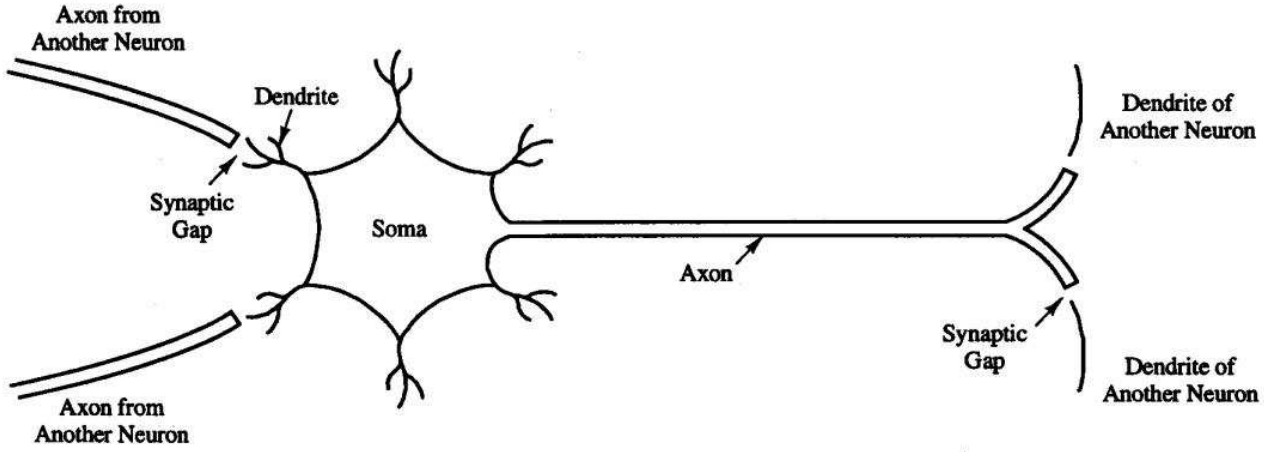


圖 4-1 人類神經元的示意圖[27]

由上圖可知，一般的生物神經元包含下列的部份：

(1)細胞體(Soma)：負責處理輸入及輸出訊息的核狀細胞。

(2)樹突(Dendrites)：負責將來自其他神經元的訊息接收後送入細胞體內。

(3)軸突(Axon)：負責將細胞體的訊息傳送至其他神經元的樹突。

(4)突觸(Synapse)：軸突的末端和目標細胞的接觸處。

基於人類神經元的架構，人工神經元採用了類似的架構。圖 4-2 即是人工神經元的架構。

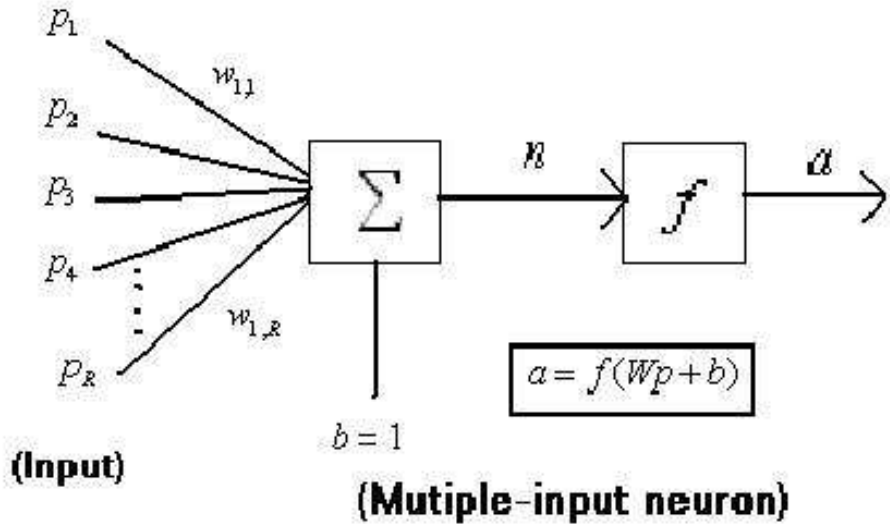


圖 4-2 人工神經元的架構。

由上圖可知，一個人工神經元組成，是由一組輸入向量(P)、權重向量(W)(Weight vector)、活化函數(f)(Activated function)，以及輸出向量(a)。一個人工神經元的架構，可分為以下三部份：

(1)權重向量：此在模擬突觸之行為。代表不同神經元間有不同的連結強弱。

(2)加法器：此在模擬細胞體之行為。代表生物神經元受到來自各方的刺激時膜電位的總變化量。所以這邊會將乘上不同權重的刺激源加總來代表膜電位。

(3)活化函數：又稱門檻值。用來轉化刺激源疊加後的的輸出值範圍。

所以上面之步驟可用下面之數學式來表達：

$$n_j = \sum_{i=1}^{R} w_{ji} p_i + b_j \qquad (4\text{-}1)$$

$$a_j = f(n_j) \qquad (4\text{-}2)$$

式(4-1)中， p_i 代表是人工神經元的輸入訊號。 w_{ji} 它的連結權重； n_j 是模擬膜電位的改變量； b_j 則是偏權值，大於 0 表示對輸入是增益，小於 0 是對輸入壓抑。式(4-2)中， $f(n_j)$ 是活化函數，是用來轉換 n_j 的數學函數；通常活化函數的輸出值範圍會在[-1 1]之間，使得神經元的輸出值維持在合理範圍內。下圖 4-3 即是一般常用的活化函數。

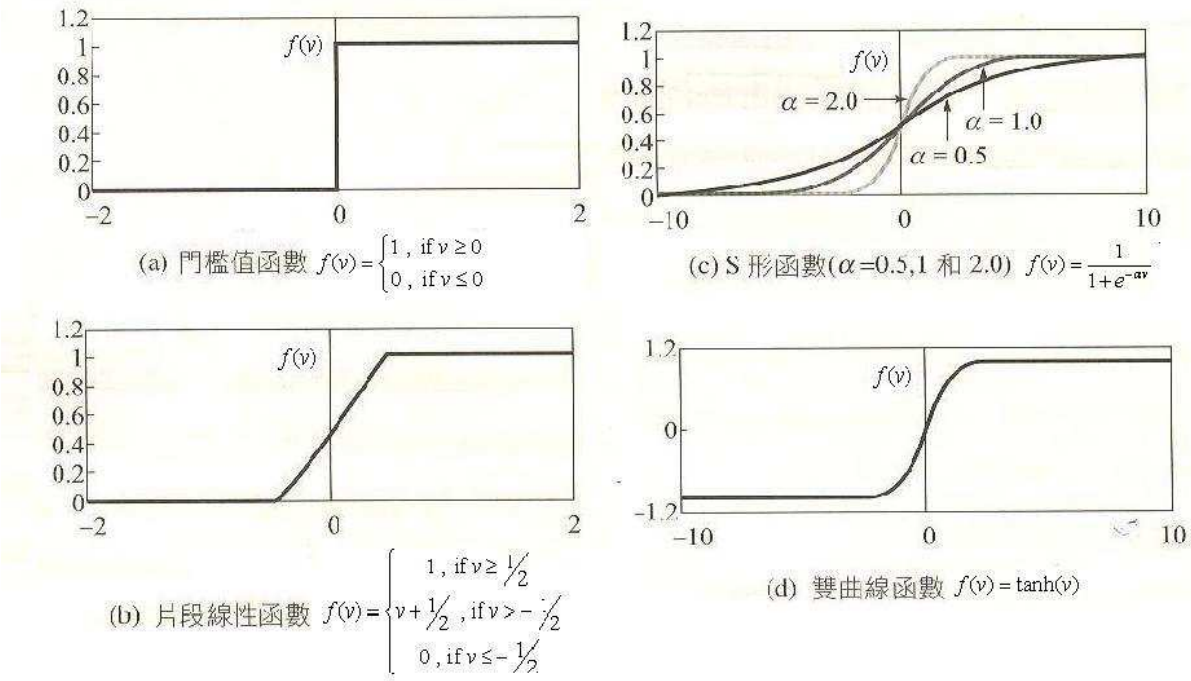


圖 4-3 一般常用的活化函數[29]

4.1.2 類神經網路系統架構

類神經網路的系統架構類型可分為以下兩種：

1. 前饋式類神經網路：

   前饋式(Feedforward)類神經網路，其神經元連結方式為單一方向向前傳遞，其中的網路神經元，皆無後向或側向的傳遞。下圖 4-4(a)(b)即是兩種常用的前饋式類神經網路系統：

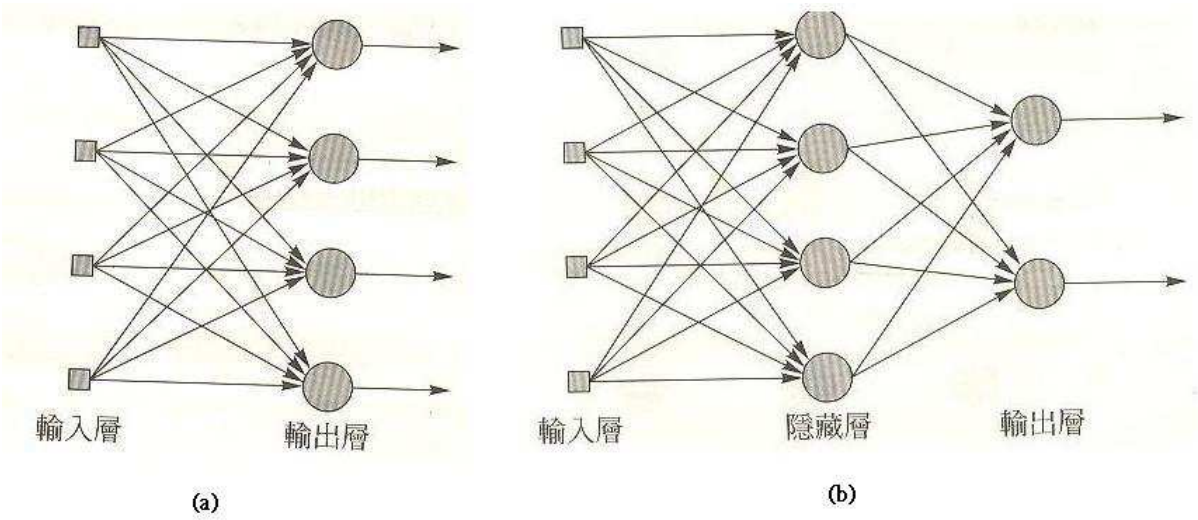


圖 4-4：兩種常用的前饋式類神經網路系統[29]

圖 4-4 中，(a)為單層前饋式，(b)圖為多層前饋式。兩者的差異性在於多層前饋式的網路在輸入層和輸出層之間還多了一層或多層的隱藏層，因此可以處理更複雜的問題，例如：複雜的高維度非線性問題等。本論文使用的自組織映射圖屬於前者。

2. 回饋式類神經網路：

回饋式(Feedback)的類神經網路，其特徵為至少含有一個回饋圈，在某一層的神經元會各自將其輸出訊號回傳給同一層或前一層的其他神經元作為該回饋之神經元的輸入資料。回饋式的網路可藉由遞迴加強網路的學習表現，所以常用於動態的系統中。圖 4-5 即是回饋式類神經網路系統。

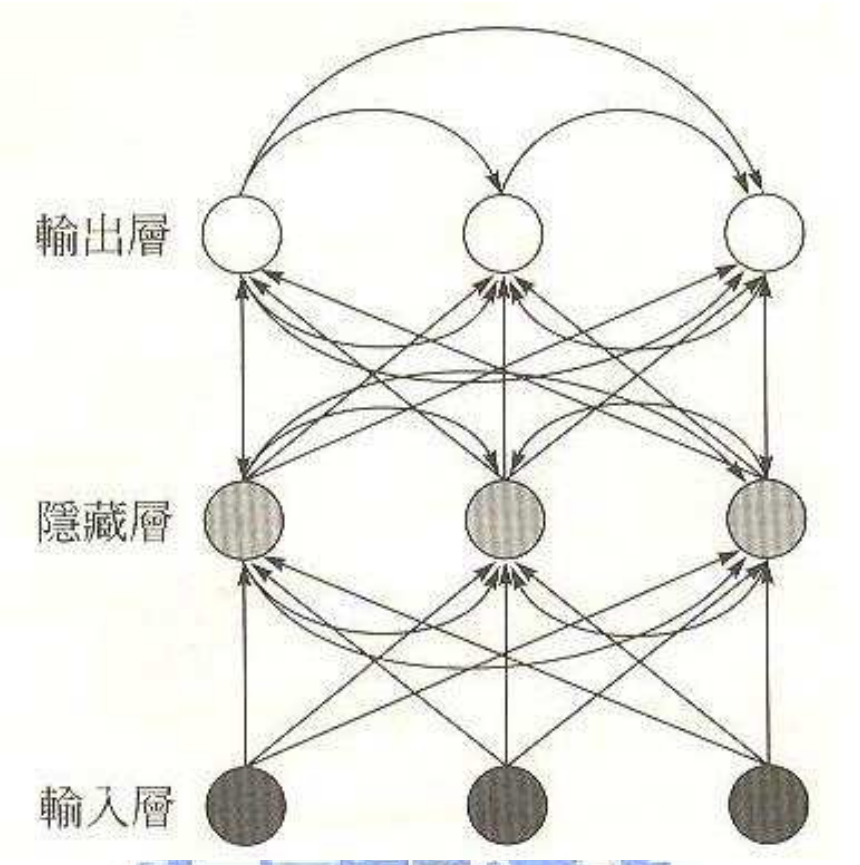


圖 4-5 回饋式類神經網路系統[29]

4.1.3 類神經網路學習演算法

類神經網路的學習演算法是類神經網路的重要核心。其是藉由訓練的過程來調整神經元之間的連結權重，此意在模擬將知識放入神經元的過程。學習演算法可分為下列兩種：

1. 監督式學習：

   監督式學習(Supervised learning)的方法是我們給予訓練範例，會包含輸入項和解答值，藉由輸出項和解答值的差距，來調整網路神經元的連結權重值，使輸出項和解答值越來越近。

2. 非監督式學習：

   非監督式學習(Unsupervised learning)的方法是在我們提供的訓練範例中，只提供

輸入資料，演算法會找出這些輸入資料的規律性或相關性，來改變自己的連結權重。常用於聚類的演算法。下圖 4-6 即是兩不同學習演算法之示意圖。

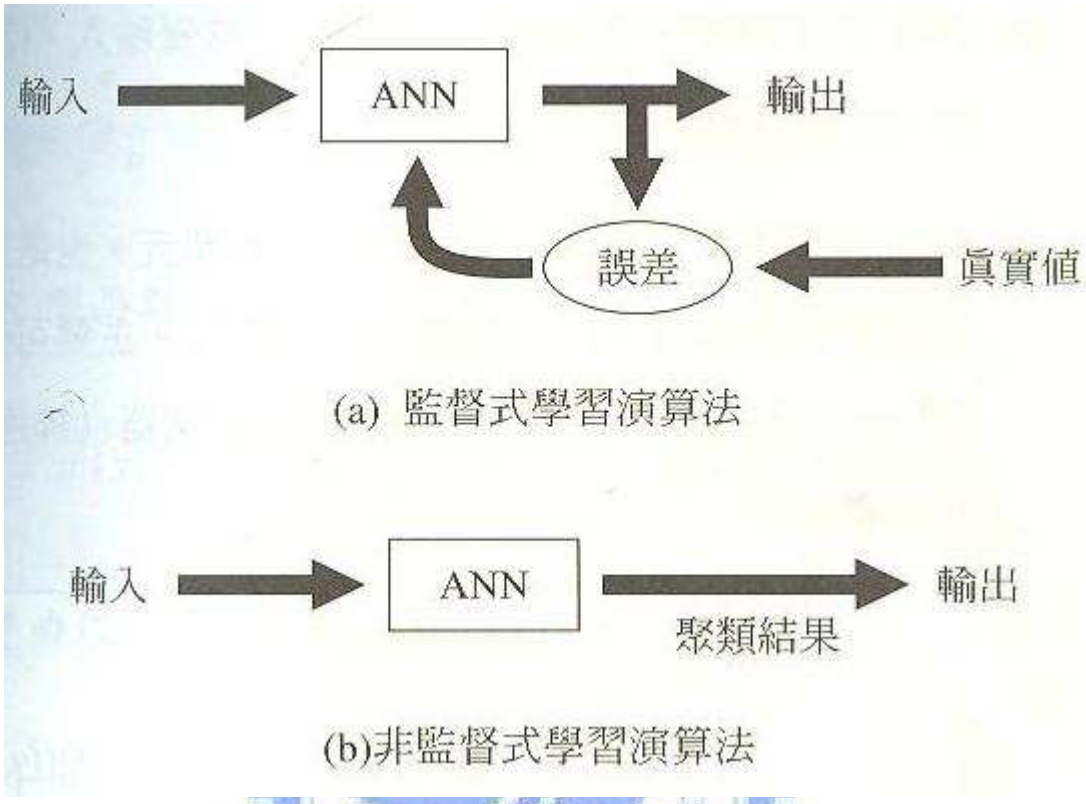


圖 4-6：學習演算法的示意圖[29]

由於本論文設計是要能處理任何的語音，而非某些特定之語料庫，因此我們採用自組織映射圖的這種非監督式學習的方法來處理。

4.2 自組織映射圖簡介

在本節中，將簡單介紹自組織映射圖的基本觀念、基本的架構以及演算法的介紹。

4.2.1 自組織映射圖之基本觀念

自組織映射圖(Self-organizing Map , SOM)是由 Kohonen 於 1982 年提出的[4]。由於人類大腦細胞具有某區塊是負責專門管理某一種類感知訊號，換言之，大腦細胞具有功能相似之細胞放在一起的情形。SOM 就是基於此種觀念下誕生的。他是屬於前饋式的非監督式的網路。它是以特徵映射的方式，將任意維度的向量，降低至比較低的維度，形成具有拓樸架構(Topological Structure)的特徵映射圖，此亦即它可以將多維度的輸入向量群以一低維度的點來代表。這個圖可以反應出所有不同輸入之值的分布關係。換句話說，自組織映射圖可以將一群零散的輸入資料，找出其相似性或規則性，再依此規則性，將零散的輸入資料中，具有該相似特性的資料聚集成一類，此種演算法稱為聚類演算法(Clustering algorithm)。本論文即是利用 SOM 的這種特性來達成語音分離的目的。

4.2.2 自組織映射圖之基本架構及參數

自組織映射圖的基本架構，可由下面圖 4-7 來說明：

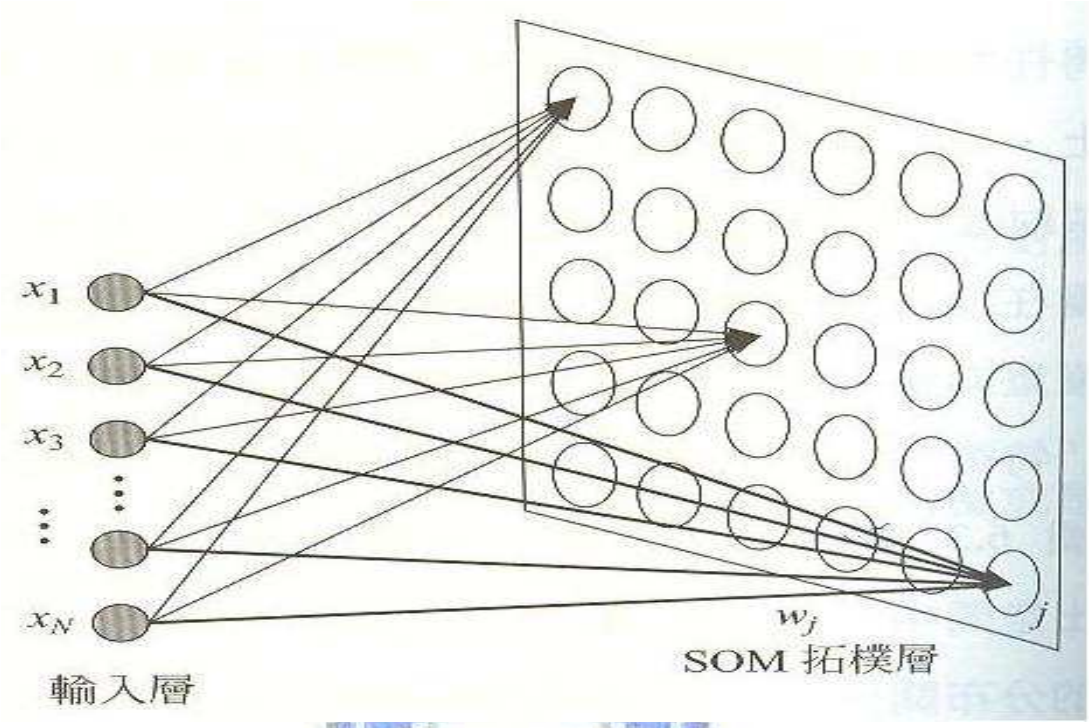


圖 4-7：二維 SOM 架構圖[29]

由圖 4-7 中輸入層 X_i 為輸入向量，做為訓練之語料；w_{ji} 為 X_i 和第 j 個神經元的連結權重；由圖上可以看出，SOM 只有輸入層和輸出層(即 SOM 的拓樸圖)，拓樸層上的每一連結到的點即代表一個神經元，這些神經元的所在位置點稱為拓樸座標，拓樸座標之用意只在標明是第幾個神經元接收訊息，每次訓練中更改的是連結權重。而控制這些神經元之間關係的參數，如下圖 4-8 顯示：

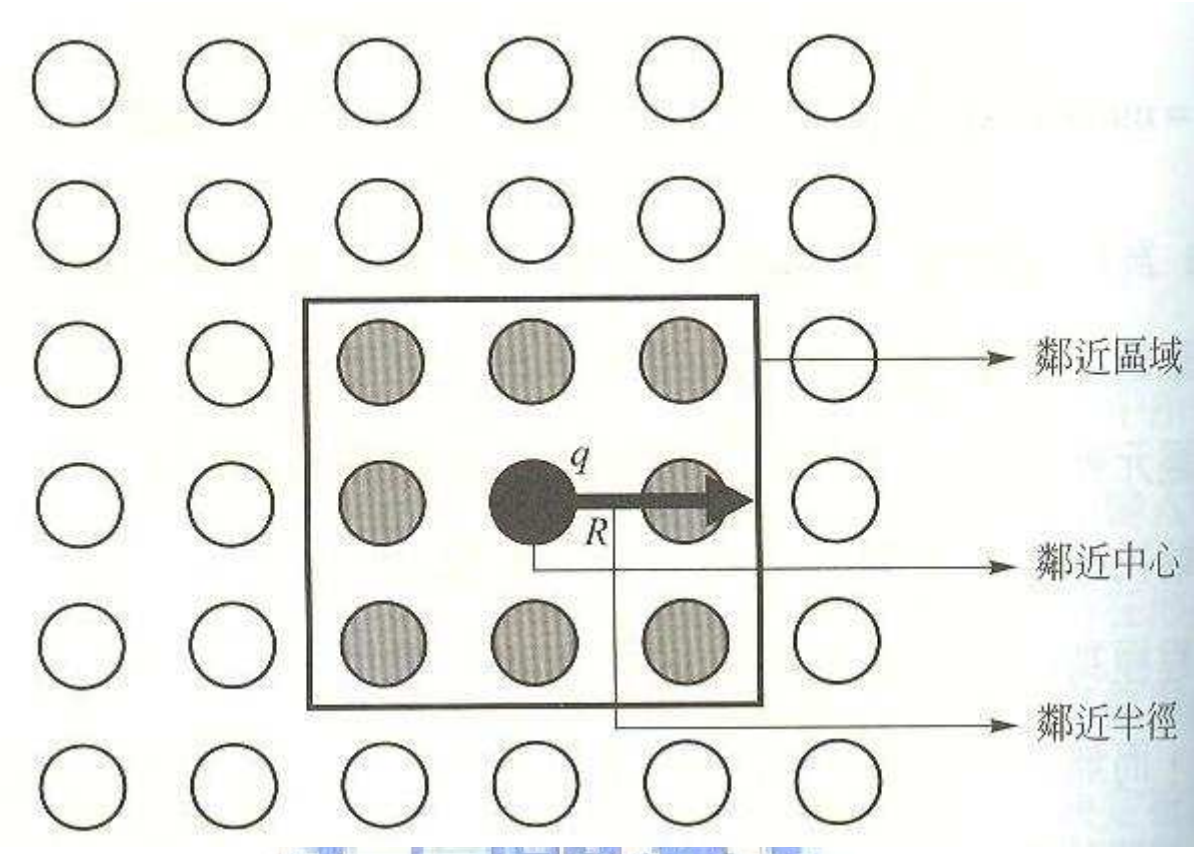


圖 4-8：優勝神經元和鄰近神經元的關係圖[29]

控制影響神經元之間的參數有：

1. 鄰近中心：圖 4-8 中深黑色的部份。鄰近中心即為鄰近區域的中心，一般和某一當時之輸入向量最接近的神經元就會做為鄰近中心，此中心亦稱為優勝神經元。
2. 鄰近區域：以鄰近中心為主，半徑為鄰近半徑 R 的區域，稱為鄰近區域。鄰近區域不一定是要方形，可以為其他的正多邊形，下圖 4-9 即為不同鄰近區域形狀圖。

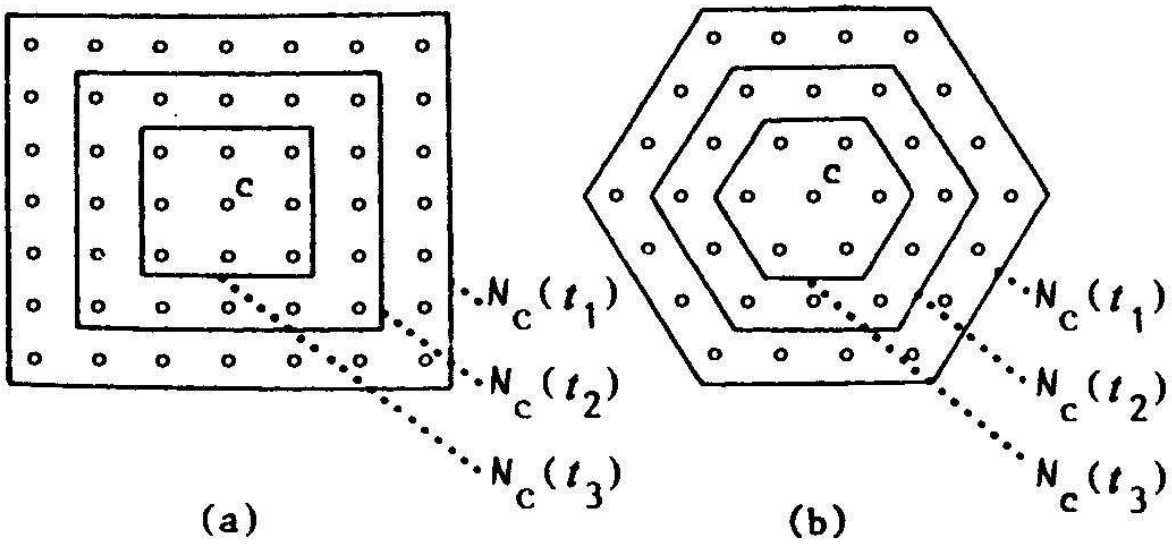


圖 4-9：不同鄰近區域形狀圖[25]

圖(a)為方形區域，圖(b)為正六角形區域。區域或隨著時間或訓練的次數而不減少。

3. 鄰近函數：鄰近函數是用來表示鄰近區域內各神經元之間的關係。算式如式(4-3)表示：

$$\eta_{qj} = \exp(-\left\| r_j - r_q \right\|^2 / R^2) \qquad (4\text{-}3)$$

η_{qj} 表示鄰近區域第 j 神經元和第 q 神經元之鄰近關係值，r_j、r_q 代表是第 j 及第 q 個神經元在拓樸圖上的位置座標；所以 $\left\| r_j - r_q \right\|$ 代表的即是鄰近區域內第 j 和第 Q 的神經元之間的距離。當 $\left\| r_j - r_q \right\|$ 大，則 η_{qj} 就會比較小，表示彼此之間的關係不強；當 $\left\| r_j - r_q \right\|$ 小，則 η_{qj} 就會比較大，表示彼此之間的關係較強；，R 即為鄰近半徑。由於鄰近半徑 R 會隨時間而縮小，因此鄰近函數之值也隨時間而改變。

4. 學習速率：學習速率是用來調整每一次訓練時的連結權重的變化；在演算法開始之初，尚未抓出輸入資料的規則性，因此學習速率要大；當演算法執行到一個段落之後，輸入資料的規則性已經成立，因此學習速率要小來微調。

4.2.3 自組織映射圖之演算法

自組織映射圖為一無監督式的類神經網路，其演算步驟如下：

1. 設定拓樸形狀及拓樸上神經元的的座標。以亂數的方式產生連至各神經元的連結權重 $\overline{w}_j = [w_{j1}, w_{j2}, ..., w_{jN}], j = 1.....M$ 。
2. 設定好鄰近半徑、學習速率、終止條件……等參數。
3. 輸入訓練資料 $X = [x_1, x_2, ..., x_N]^T$ ，一次輸入一筆訓練資料。利用 L2 距離來求出優勝神經元，如式(4-4)所示：

$$\left\| X - w_c \right\| = \min_j \left\| X - w_j \right\| \qquad (4\text{-}4)$$

上式中，c 表示是輸入向量和所有連結權重中距離最近之神經元，而 w_c 則是該優勝神經元之連結權重，此神經元 c 就稱為優勝神經元(Winner)或稱最佳對應神經元(Best Matching Neuron)。

4. 以優勝神經元為中心，修正鄰近區域內的所有神經元的連結權重值。權重值修改如式(4-5)、(4-6)所示。

$$\Delta w_j = \mu(k)\eta_{qj}(k)[x(k) - w_j(k)] \qquad (4\text{-}5)$$

$$w_j(k+1) = \begin{cases} w_j(k) + \Delta w_j, j \in N_c \\ w_j(k), j \notin N_c \end{cases} \qquad (4\text{-}6)$$

式(4-5)中，$\mu(k)$ 代表第 k 次訓練的學習速率，$\eta_{qj}(k)$ 即為式(4-3)。式(4-6) Nc 表示鄰近區域的範圍，在範圍內的才會更動連結權重。

5. 調整學習速率及鄰近區域的範圍，若達終止條件則停止。

圖 4-10 即是 SOM 的執行前和執行後的權重比較圖。

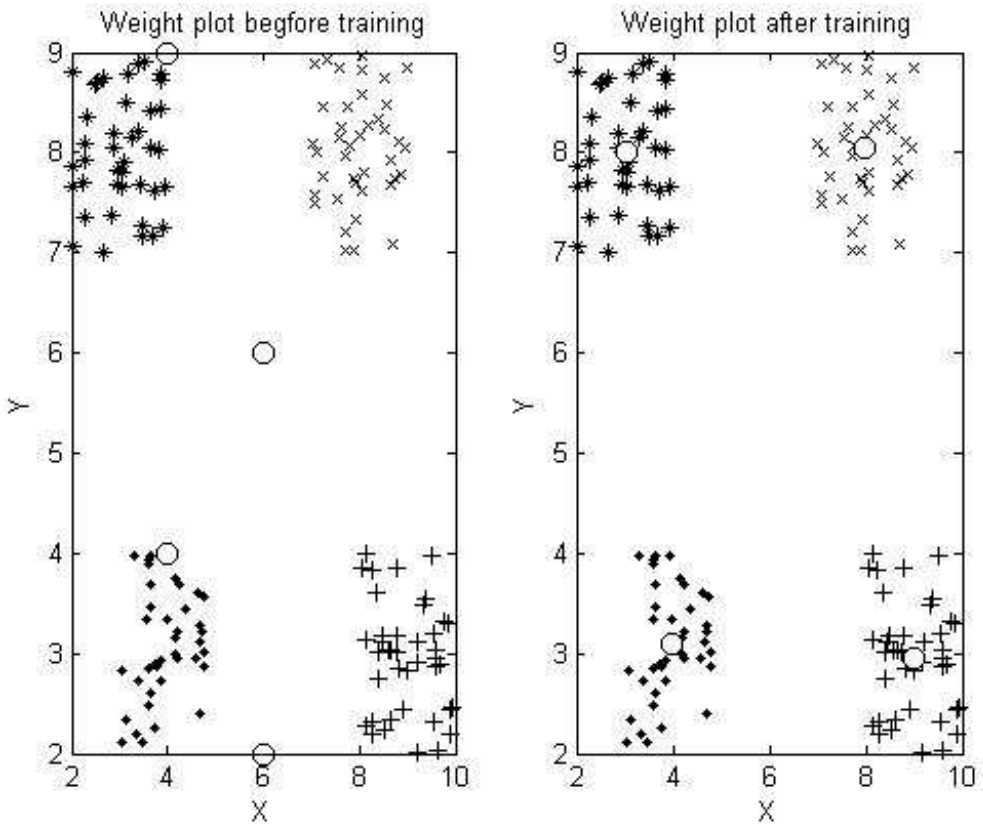


圖 4-10：SOM 的執行前和執行後的權重比較圖

上圖是在 X-Y 座標上有四群的資料送入 SOM 訓練。由上面的圖可以看出，連結權重經過 SOM 的演算法訓練後，神經元的連結權重值(在此為座標 X-Y)，分別變成各群資料的中心點，由此結果，我們可以得知：

1. 而後如果送入新的資料進來，SOM 系統可以依照現有的結果對新進入的資料做歸類的動作。
2. 原先這些訓練的資料可由 SOM 的神經元權重來代表，換言之，這些原先用來訓練的資料群，也被分群，因此我們可以在訓練完後將訓練的資料直接分群。

我們的語音分離系統，運用的就是後面這項特性。

4.3 語音分離機制

在前面介紹完 SOM 的基架構後，在本節將介紹本論所使用的語音分離機制。首先將簡單介紹我們用於最後的語音重建機制，接著介紹如何將 SOM 應用在語音分離機制，最後則和原頻譜來比較結果。

4.3.1 語音重建的基本觀念

將語音從聽覺頻譜重建的觀念，建立在式(2-1)上。式(2-1)是做摺積，屬於線性運算，因此可以做反運算回來得到原來在時間軸上的信號。然而，由於我們通過濾波庫後還有經過其他的非線性運算以及左右壓抑的機制才得到結果的頻譜，要由結果頻譜變回去通過濾波庫後的結果有困難，因此這邊我們利用凸面投影(Convex Projection)的方式將剛通過濾波器的結果估計出來後，再做式(2-1)的反運算而得到重建回時間軸上的語音信號[3]。圖 4-11(a)(b)即是原來的訊號和重建訊號的時間和聽覺頻譜圖上的比較。

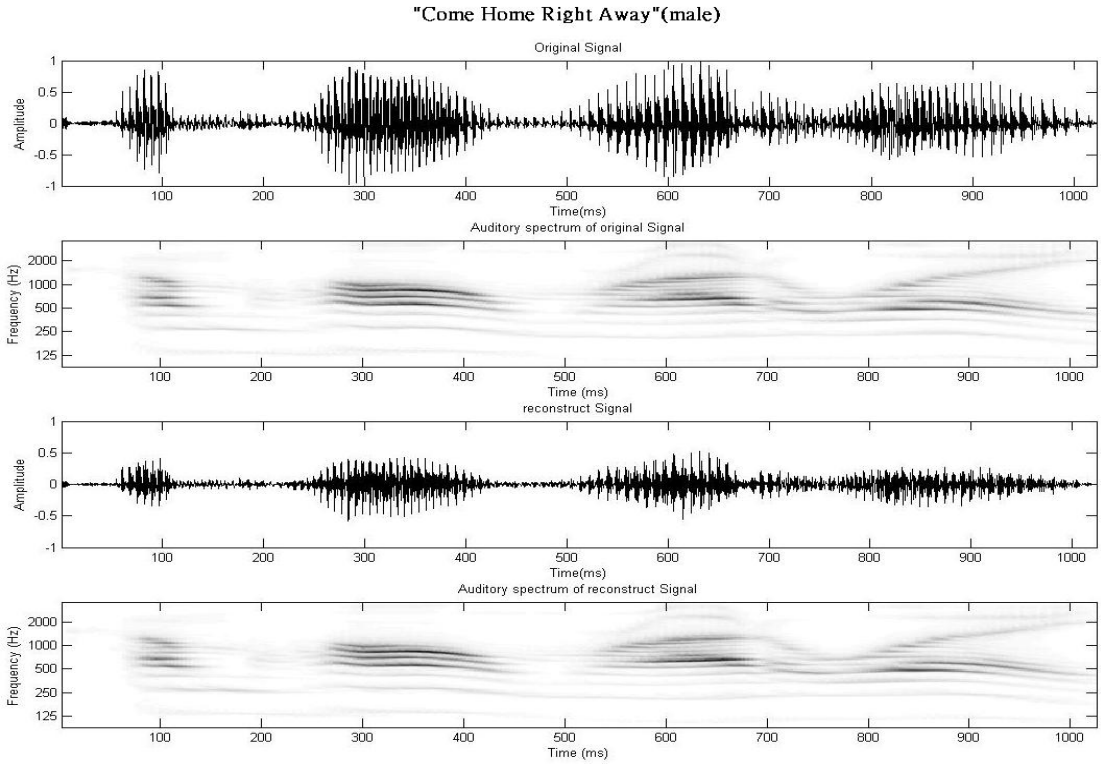


圖 4-11(a)英文語音”Come home right away”的原來訊號和重建訊號的頻譜圖

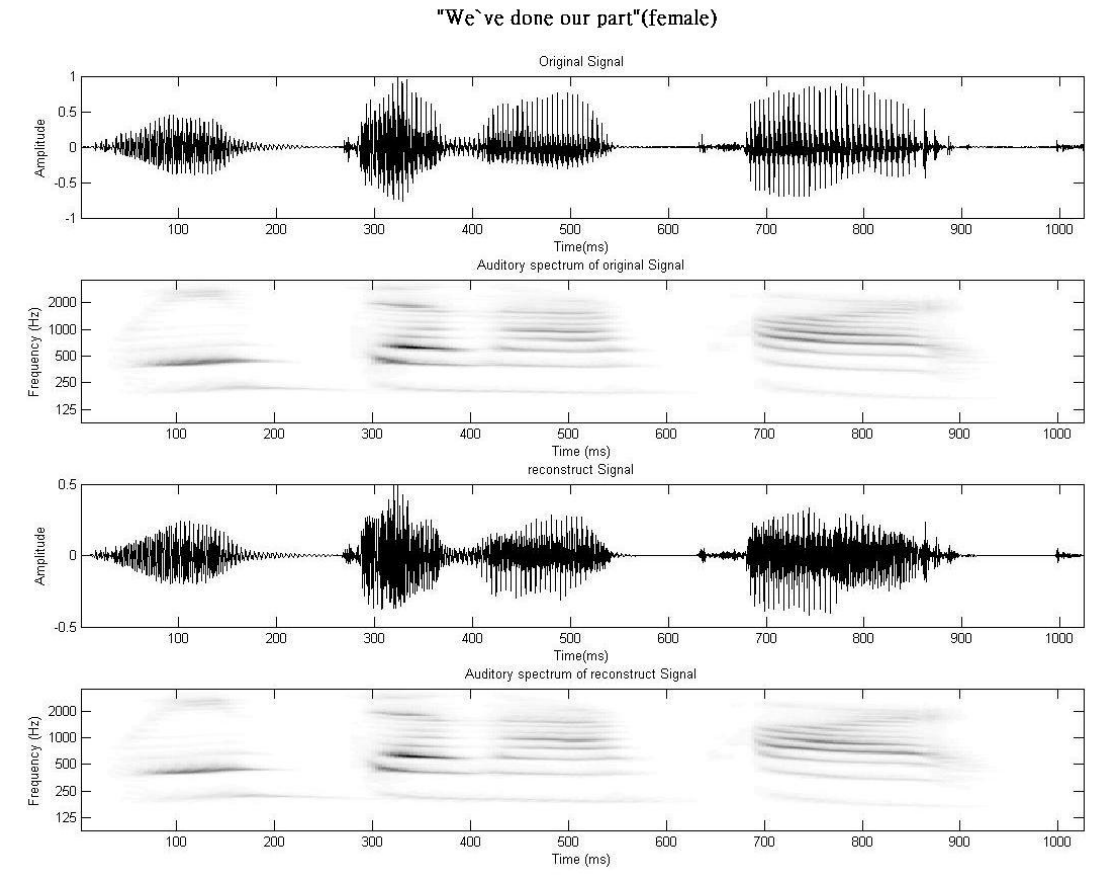


圖 4-11(b)英文語音”We‘ve done our part”的原來訊號和重建訊號的頻譜圖

圖 4-11 的重建語音是遞迴做 200 次後的結果。由上面之圖可以看出，經過重建後的聲音訊號，會比原來的訊號在振幅變化比較和緩，因此我們可以拿來對分開的頻譜做處理。

4.3.2 語音分離—利用 SOM

由 4.2 節介紹的 SOM 的測試結果可知，SOM 具有把相似的輸入資料做分群和聚類的特性。因此 SOM 的有一部份應用層面是在移動軌跡的偵測上[26]通常，在這類型的資料聚類的方法，皆是將能夠代表該資料的特徵值(Feature)送入，SOM 將具有相似特徵值的指標(index)歸類成一群，本論文所使用的也是類似的方法。圖 4-12 即是本論文運用 SOM 做語音分離的流程圖。

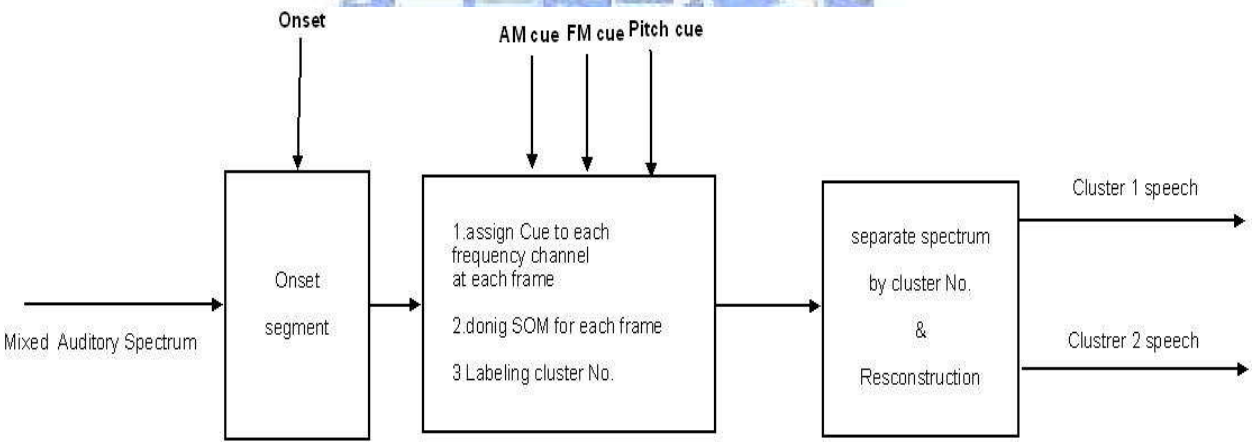


圖 4-12：運用 SOM 語音分離的流程圖

我們的機制步驟如下：

1. 根據之前所做的起始點/終點線索，將送入的混合語音的聽覺頻譜沿著時間軸切割出有語音和沒有語音的區域。我們將屬於有語音的區域再送入下一段來做處理。
2. 在屬於語音的區域中，我們隨時間一個音框一個音框送入，一次處理一個音框的內容，做以下的步驟：

   (1)根據音高線索，決定該音框內頻率軸上所有出現的泛音的寬度。我們的做法類似[13]所使用的交互頻道相關性(Cross-Channel correlation)，如式(4-7)：

$$Cr(x_h, x, t) = \frac{\sum_{\tau=0}^{4} y(x_h; t-\tau) y(x; t-\tau)}{\sqrt{\sum_{\tau=0}^{4} (y(x_h; t-\tau))^2} \sqrt{\sum_{\tau=0}^{4} (y(x; t-\tau))^2}}, x = 1, ..., 128 \qquad (4\text{-}7)$$

上式中 x_h 為某一泛音所在之位置，我們將泛音所在位置上沿時間軸取 5 點和其他頻率軸上之值來做交互相關性。根據[13]和[17]我們可以得知，對於屬於同一泛音的頻率軸位置的成份，比此之間的的相關性應該會很高，所以我們對每個頻率軸上的位置去做式(4-7)的運算，然後在最大值(相關性最大值必在 $x = x_h$ 的地方)往前和往後尋找區域最小值(Local Minimum)，則此段區域即會是該泛音的寬度。此用意在於當相關係數低的地方，應該與泛音的相關性比較低，但是泛音和其他泛音之間的相關性也很高，因此找尋最近轉折點的範圍，即是該泛音的範圍。圖 4-13 即為用此法找出泛音之寬度和原來的聽覺頻譜的比較圖。

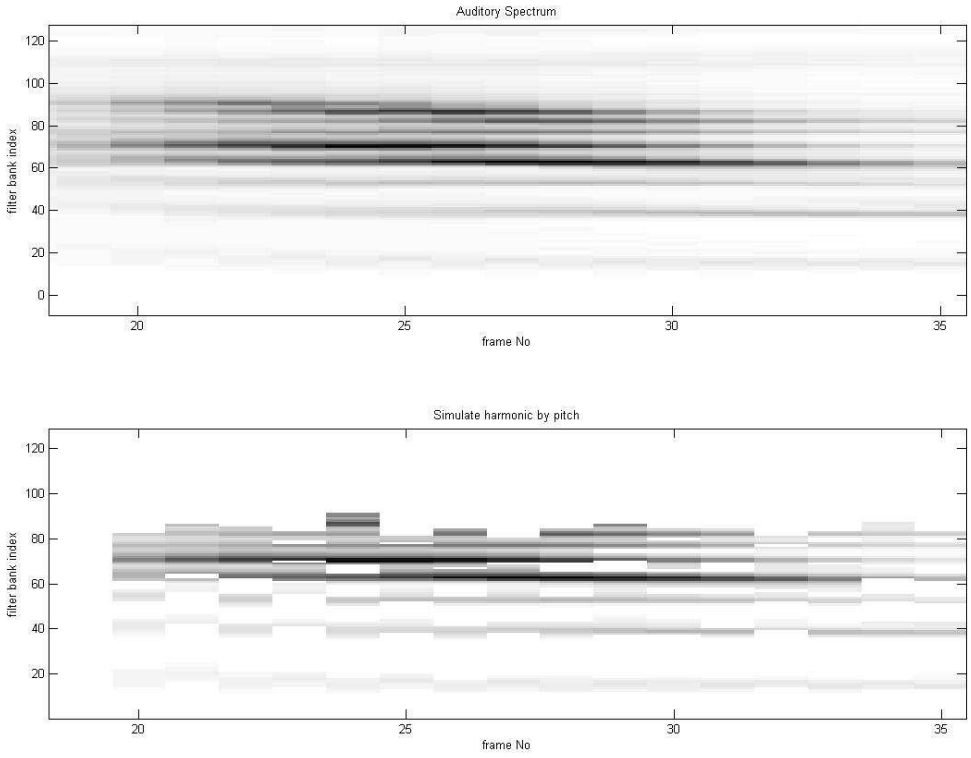


圖 4-13：估計之泛音寬度和原來的聽覺頻譜的比較圖

由上圖可以看出，大致上和原來聽覺頻譜的泛音的寬度類似，唯一些地方會有變動，原因是因為這個泛音寬度的重建，必須先找出音高，藉由音高來決定其全部泛音的位置，因此會和原來的比在位置上有一些誤差，才會導致看起來會有點不同。

(2)　在每個泛音其寬度內的位置，放入該泛音代表之音高、該泛音位置頻率調變線索值，以及該泛音位置振幅調變線索值。這邊在填入線索值時，只針對解析的(resolved)泛音頻率位置做填值，通常解析的泛音約莫在 1000Hz 以下，其原因在於 1000Hz 以上的濾波庫的頻寬會包含一個以上泛音成份，所以我們不易將其指定線索值，因為沒辦法正確找出其泛音的位置。

(3)　將填完值的該音框的內容送入 SOM 去訓練。SOM 的設定如下表

表 4-1 SOM 參數設定表

神經元個數	2 個
鄰近區域半徑/半徑遞減速率	1/0.7
學習速率	(1)音高：1 (2)FM：0.55 (3)AM：0.45
學習速率變化率	(1)音高：0.95 (2)FM：0.6 (3)AM：0.55
鄰近函數	$\eta(i)=\begin{cases}1, i<R\\0, else\end{cases}$
鄰近區域形狀	方形

我們使用 SOM 的目的，是希望能利用 SOM 的特性，將進來的資料做分群，因此我們希望能用神經元的權重連結代表一個人的語音群組。因為本文測試是使用兩個人混合的語料，因此神經元個數取 2，而又確定兩個神經元是幾乎要沒關係，因此我們這邊採取勝者全拿的方式，除了一開始神經元權重是亂數時需要一次修改兩個神經元之權重，以便分出兩個不同的群組，剩下的皆只要修該優勝神經元的權重即可。至於學習速率的設定，主要是考慮人類在做語音分離或是語音分組時，音高是最重要的因素[8][21]，其次的是頻率軸上的線索(這邊使用頻率調變)，再來就是時間軸上的線索[15](這邊使用的是振幅調變)。同樣考量亦反應在決定優勝神經元的條件上。原來的 SOM 使用的是如式(4-4)的距離公式，在這

邊為了強調各線索的重要性，我們改用下列式(4-8)：

$$\left\| X - w_c \right\| = \min_j \left\| A^T (X - w_j) \right\| \qquad (4\text{-}8)$$

上式中之 X 為輸入向量，在這邊即是我語音分離的三個線索；而 A 即是我這三個線索有不同的強調度。其中音高的權重值最大，其次為頻率調變，最後為振幅調變。式(4-8)可以看出音高線索和當時神經元權重的距離會影響到優勝神經元的決定最多。經過 SOM 訊練後，我們即可將音框上有線索之頻率位置(即那些泛音所在之位置)標記群組。

(4)當所有屬於語音的音框都經過 SOM 訓練之後，由於是各個音框獨自訓練，因此會出現順序不一的情形。因此這邊我們將所有訓練的音框內，每個神經元的權重去做重新排序。排序的標準為和前一個音框的神經元權重間的 L2 距離最近的視為同一群的成份。如此即完成了頻譜分離，之後就將分離好的聽覺頻譜去利用本論文所提之語音重建機制，重建回時間軸上的訊號，如此即完成語音分離。下圖 4-14 即為測試之結果，圖由上到下依序是：混合前之聽覺頻譜、混合時的聽覺頻譜、經過 SOM 做歸類後所得之結果、分離語音經重建後的聽覺頻譜。

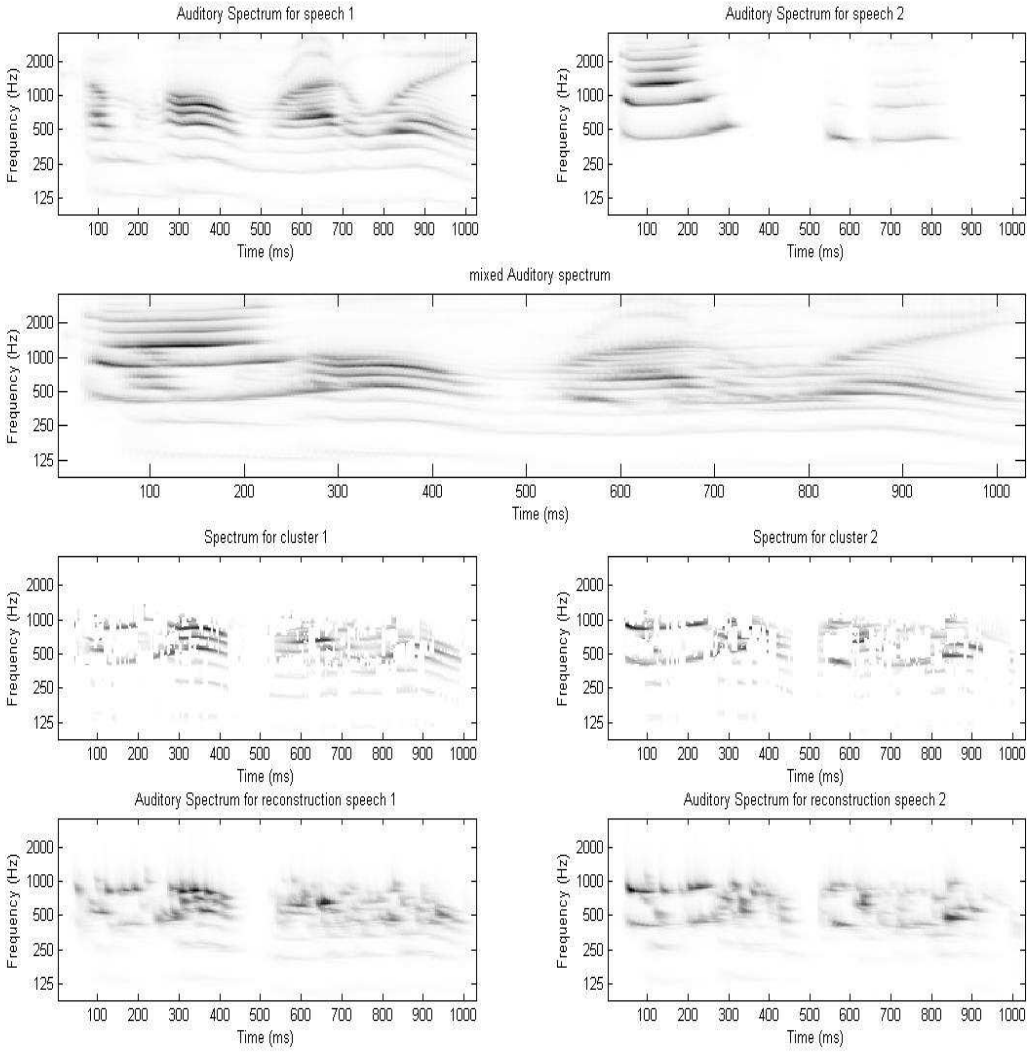


圖 4-14：語音分離機制的測試結果

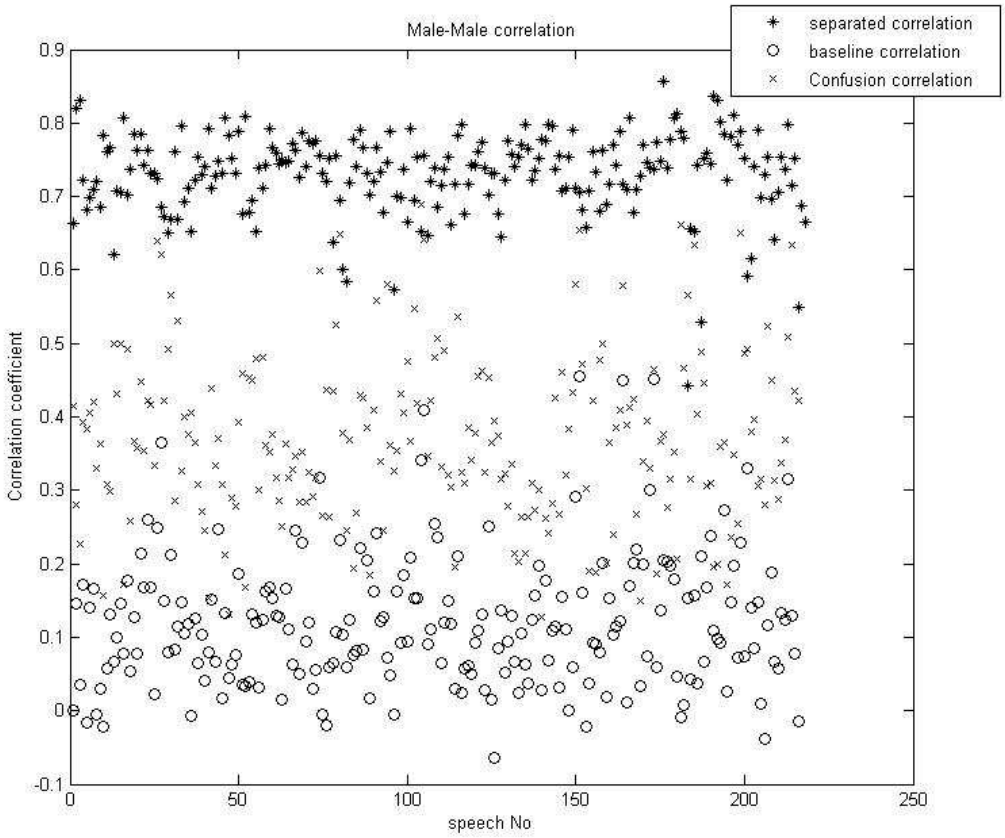
4.3.3 實驗設定

本論文將以TIMIT為語料庫，從新英格蘭地區口音的語料庫中取出男生24人，女生14人，去製作成混合的語音，從每個語者中挑出三句，每一句會和任意挑的其他語者的三句進行混合，在混合的時候，我們將欲混合進來的語音能量降至和目標語音差距0.5倍來進行混合。之後通過本論文之語音分離系統後，將目標語音取出來後和混合前之頻

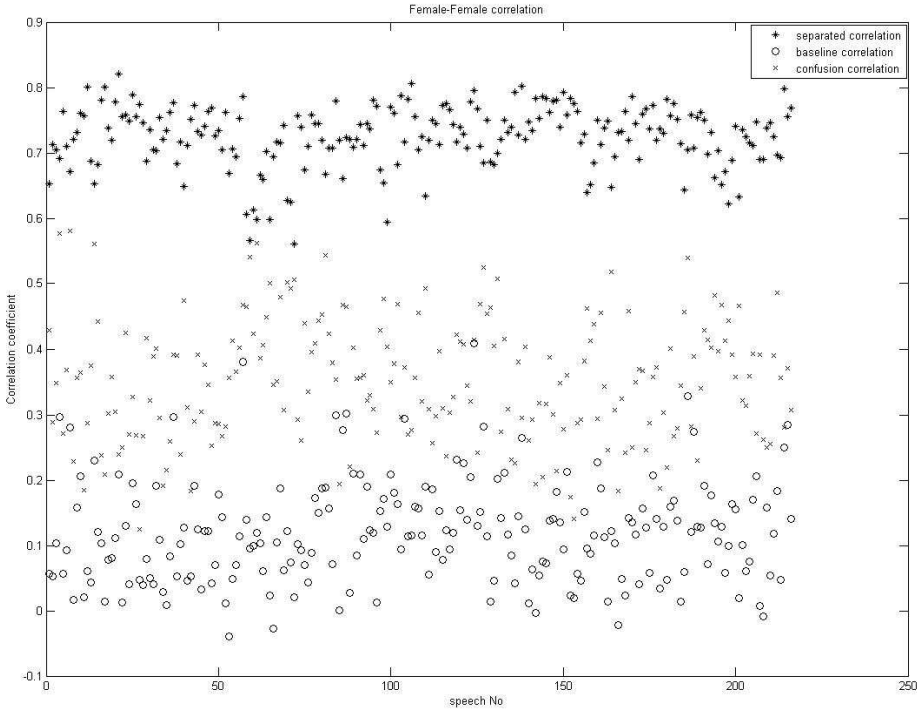
譜取解析泛音的地方來計算出其相關係數，來看其相似性有多高。我們將分成男生和男生混、女生和女生混、男生和女生混(男生為目標語音)、女生和男生混(女生為目標語音)四種情況來做測試，每個情況總共會有216句，女生因為語料庫中只有14人，因此我們目標語音會有重複的語者。我們另外又設定了相關係數的基本值，此基本值是將原頻譜和另一個隨機混合的頻譜去做相關係數，如果我們分出來的頻譜和原來頻譜相關係數約接近此基本值，表示我分出來的成份中，和原頻譜越不相似而且可能摻雜了比較多的競爭語音信號。另外一個數值是將分出來的頻譜和混合進來的語音的頻譜去做相關係數，稱為疑惑相關係數(Confusion correlation)，和結果的相關係數來做比較，此值越高代表我分出來的結果其實和兩邊都很相似，表示其實並沒有分的很好。

4.3.4 實驗結果

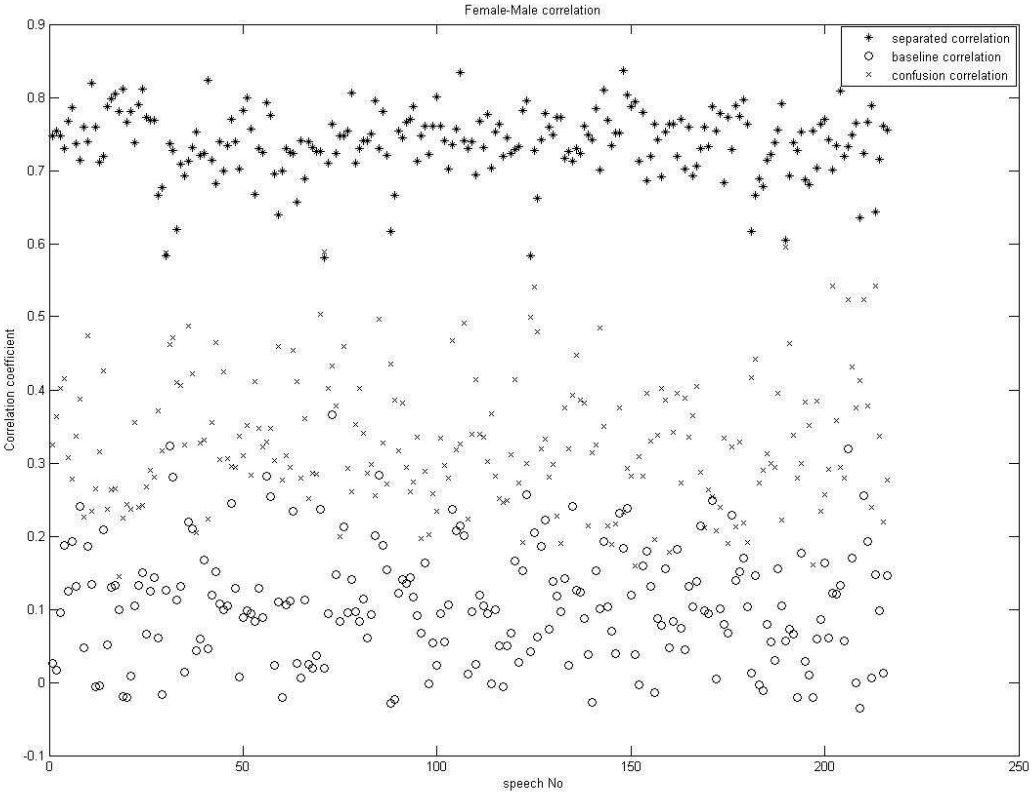
下圖 4-15(a)(b)(c)(d)及表 4-2、表 4-3 即是我們實驗的結果：



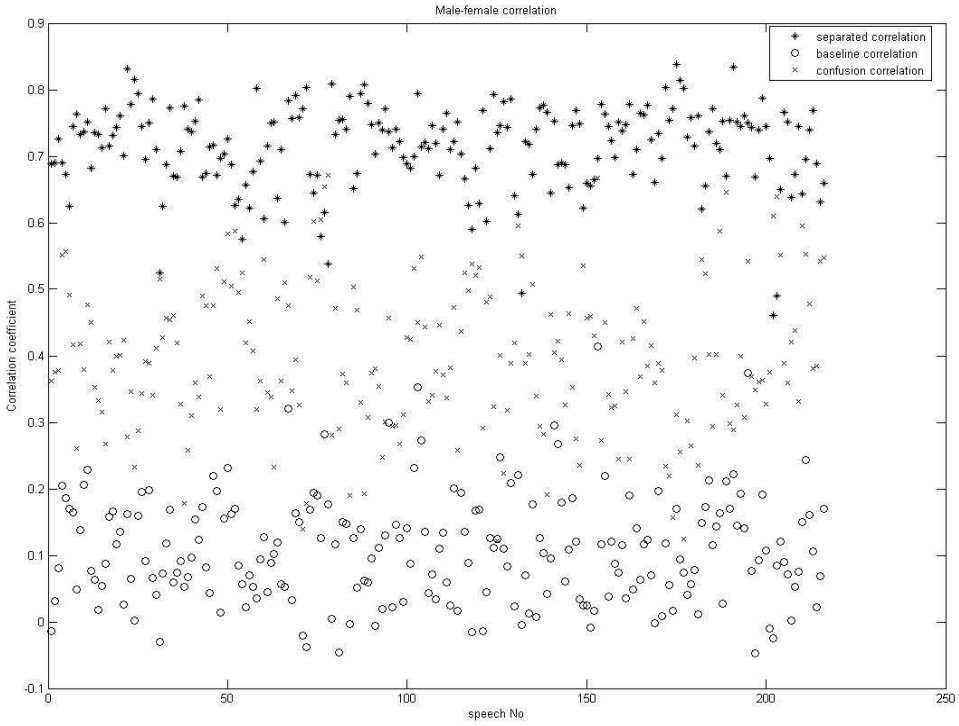
4-15(a)：分開語音和原語音的頻譜相關係數圖(男生-男生)



4-15(b)：分開語音和原語音的頻譜相關係數圖(女生-女生)



4-15(c)：分開語音和原語音的頻譜相關係數圖(女生-男生)



4-15(d)(左)：分開語音和原語音的頻譜相關係數圖(男生-女生)

表 4-2：男生 v.s 女生得平均相關係數

性別	男生	女生
相關係數(%)	72.30±6%	73.14±4.65%
困惑相關係數(%)	38.27±11.35%	34.02±9.14%

表 4-3：各狀況之平均相關係數

混合狀況	男生-男生	女生-女生	男生-女生	女生-男生
相關係數(%)	73.14±5.57%	72.5±4.73%	71.45±6.28%	73.78±4.48%
困惑相關係數(%)	36.98±11.56%	35.17±9.19%	39.58±11.01%	32.88±8.97%

由圖 4-15、表 4-2、表 4-3 的結果可以得知：

(1)由表 4-2 平均分離出來的目標語音聽覺頻譜和原語音聽覺頻譜的相關係數，不論目標語音的性別為何，都約在 70%以上，而女生的平均值約比男生平均值高 1%左右。而男生的相關係數的離散程度(標準差)，比女生相關係數的離散程度大約 0.1%。

(2)由困惑相關係數，女生的平均值比男生的平均值低約 4%。

(3)同性別之間的語音分離結果，男生-男生和女生-女生的結果差不多。但是女生-男生的相關係數比男生-女生高約 2%左右，困惑相關係數女生-男生比男生-女生低約 7%，會有此差別的原因在於我們的音高偵測系統對於男生比較容易出現倍頻錯誤，使得在重建泛音和填入線索上容易出現錯誤。因此僅管女生-男生混合和男生-女生混合，在泛音的部份都會有大量重疊的情況發生，但是相關係數女生-男生的情況就比較好。

(4)由表 4-2、表 4-3 的標準差，可以得知對於男生來說，本系統的結果不是非常穩定，其相關係數值變動比較大；而對於女生，其相關係數的結果就大致穩定一些。

第五章

結論與未來展望

本章將會將本論文所做的一些研究成果及貢獻做再一次的整理及說明，並且針對本論文中實驗姐結果不足之地方，提出可修正的方向及建議，以其未來能將此系統修正更好。

5.1 結論

本論文主要的研究內容是利用一已知的感知聽覺模型，擷取語音分離所需要的一些線索，如音高、頻率調變……等線索，並建立出語音分離的方法。首先我們先介紹了感知聽覺模型的初期階段和大腦階段。初期階段為頻譜估計，大腦階段為時域-頻域估計。其後我們利用這兩階段的分析加上一些簡單的運算後而得到語音分離的線索。

其後，我們利用 SOM 的系統的特性—相似的資料將其聚集在一起的觀念，利用在語音分離上面，將頻率軸上具有相似線索的位置聚集在一群，而後，根據分群的標記將兩個語音分開。其後，再將語音重建回時間訊號。跟據實驗結果，我們可以發現分離出

來的語音，雖然其聽覺頻譜在時間軸上有不連續的情況發生，但是其聽覺頻譜和原來語音的聽覺頻譜相關性還保有一定程度。我們的語音分離的線索除了音高之外，皆是經由大腦階段模型處理後而得。由許多心理聲學的實驗已知，人類在處理混合語音會有類似分組(Grouping)的現象，因此我們可以試著由大腦聽覺模型來取出各種的線索。

5.2 未來展望

本論文中，起始點/終止點只是拿來利用判斷是否為語音的線索，而實際上，人類對於起始點/終止點也是一個振幅調變的線索之一，因此未來可以嘗試加入這個振幅調變線索來輔助系統能達更好的效果。再來，由於線索會隨時間而有不同的變化，同一群的聲音，其線索在經過一段時間後，可能會有一段不小的差異。在此情況下，利用 L2 距離將每個音框 SOM 所訓練出來的權重做排序，會使得不連續的情況容易出現。而人類本身對於接收到的語音會去預測，因此可以有效的將語音從其他語音中分離，因此未來可以將系統的 SOM 做一些修正，使其有預測下一音框線索的能力，如此應可以將本論文系統效能大幅提高；本論文系統目前還未具備處理非解析(Unresolved)泛音的能力，因此未來能夠加入此能力的話，系統將更加完備。

以上目標皆可當作努力的方向，我們最後希望能從多了侵入語音中將目標語音近乎完整的抽取出來

參考文獻

[1]. Neural System Laboratory, http://www.isr.umd.edu/Labs/NSL/.

[2]. TIMIT Acoustic-Phonetic Continuous Speech Corpus, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[3]. T. Chi, P. Ru and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887-906, August 2005.

[4]. T. Chi, Y. Gao, M. C. Guyton, P. Ru and S. A. Shamma, "Spectro-temporal modulation transfer function and speech intelligibility," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719-2732, November 1999.

[5]. H. Duifhuis, L.F. Willems and R. J. Sluyter, "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," *Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1568-1580, June 1982.

[6]. J. L. Goldstein, "An optimum processor for the central formation of pitch of complex tone," *Journal of the Acoustical Society of America*, vol. 54, no. 6, pp. 1496-1516, 1973.

[7]. T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911-918, October 1976.

[8]. N. Grimault, S. P. Bacon and C. Micheyl, "Auditory stream segregation on the basis of amplitude-modulation rate," *Journal of the Acoustical Society of America*, vol. 111, no. 3, pp. 1340-1348, March 2002.

[9]. S. MacAdams, "Segregation of concurrent sounds I.:Effect of frequency modulation coherence," *Journal of the Acoustical Society of America*, vol. 86, no. 6, pp. 2149-2159, December 1989.

[10].J. F. Culling and Q. Summerfield, “The role of frequency modulation in the perceptual segregation of concurrent vowels,“ *Journal of the Acoustical Society of America*, vol. 98, no. 2, pp. 837-846, August 1995.

[11].C. J. Darwin, V. Ciocca and G. J. Sandell, “Effect of frequency and amplitude modulation on the pitch of a complex tone with mistuned harmonic,“ *Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2631-2636, May 1994.

[12].K. Wang and S. A. Shamma, “Spectral shape analysis in the central auditory system,” *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 382–395, September 1995.

[13].G. Hu and D. Wang, “Monaural speech segregation based on tracking and amplitude modulation,” *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, September 2004.

[14].Q. Summerfield, J. F. Culling and A. J. Fourcin, “Auditory Segregation of Competing voices: absence of effects of FM or AM coherence,” *Philosophical Trans. Royal Society Lond.B* 336, pp. 357–366, 1992.

[15].S. Rosen, “Temporal information in speech:acoustic, auditory and linguistic aspects,” *Philosophical Trans. Royal Society Lond.B* 336, pp. 367–373, 1992.

[16].M. Elhilali and S. A. Shamma, “A Biologically-inspired approach to the cocktail party problem,” *In Proc. ICASSP*,vol. 5, pp. 637–640, 2006.

[17].G. J. Brown and M. Cooke, “Computational auditory scene analysis,” *Computer Speech and Language*, vol. 8, pp. 297-336, 1994.

[18].M, Cooke and D. P. W. Ellis, “The auditory organization of speech and other sources in listeners and computational models,” *Speech Communication*, vol. 35, pp. 141-177, 2001.

[19].P. Ru and S. A. Shamma, “Representation of musical timbre in auditory cortex,” *Journal of New Music Research*, vol. 26, pp. 154-169, 1997.

[20].A. Palmer and S. A. Shamma, “Physiological Representations of speech,” in S*peech Processing in the Auditory System*, S. Greenberg, W. A. Ainsworth, A. N. Popper and R. R. Fay, Eds.: Springer 2004.

[21].C. J. Darwin, “Pitch and Auditory Grouping,” in *Pitch Neural coding and perception*, C.

J. Plack, A. J. Oxenham, R. R. Fay and A. N. Popper, Eds.: Springer 2005.

[22]., D. K. Mellinger and B. M. Mont-Reynaud, “Scene Analysis,” in *Auditory Computation*, H.L. Hawkins, T.A. McMullen, A.N. Popper, and R.R. Fay, eds. Springer-Verlag, New York, 1996.

[23].S. A. Shamma, “Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method,” in *Network:Computation in Neural System,* vol. 3, no.7, pp. 439-476, 1996.

[24].A. S. Bregman, *Auditory Scene Analysis,* MIT Press,1990.

[25].T. Kohonen, *Self-organizing Maps*, Springer Verlag, 1995

[26].W. Hu, D. Xie and T. Tan “A Hierarchical Self-organizing approach for learning the patterns of motion trajectories,” *IEEE Trans. Neural Networks*, vol. 15, no. 1, pp. 135–144, January 2004.

[27].M. T. Hagan, H. B. Demuth and M. H. Beale, *Neural Network Design*, PWS Pub. Co. 1995.

[28].L. V. Fausett, *Fundamentals of Neural Networks; Architectures, Algorithms And Applications (Pie)*, Prentice Hall 1993.

[29].張斐章，張麗秋，類神經網路，東華書局，台北，民國九十四年。

[30].陳桂霞，黃重光，自組織映射圖網路簡介，國立台中師範學院教育測驗統計研究所。