國 立 交 通 大 學
電控工程研究所

碩 士 論 文

利用空間域特徵空間一致性及共鳴曲線相似度之喚醒關鍵字偵測方法

A Wake-Up-Word Detection Method Using Spatial Eigenspace Consistency and Resonant Curve Similarity

研 究 生： 王 庭 昭
指導教授： 胡 竹 生 博士

中 華 民 國 九 十 九 年 七 月

利用空間域特徵空間一致性及共鳴曲線相似度之喚醒關鍵字偵測方法

A Wake-Up-Word Detection Method Using Spatial Eigenspace Consistency and Resonant Curve Similarity

研 究 生：王 庭 昭　　Student：Ting-Chao Wang

指導教授：胡 竹 生 博士　Advisor：Jwu-Sheng Hu

國 立 交 通 大 學
電控工程研究所
碩 士 論 文

A Thesis
Submitted to Institute of Electrical and Control Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Master
in

Electrical and Control Engineering

July 2010

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 九 年 七 月

利用空間域特徵空間一致性及共鳴曲線相似度之喚醒關鍵字偵測方法

研究生：王　庭　昭　　指導教授：胡　竹　生　博士

國立交通大學電控工程研究所碩士班

摘　要

本論文提出了一套使用麥克風陣列偵測喚醒關鍵字的方法，本方法運用聲源於麥克風陣列在不同頻率下的特徵空間一致性(Spatial Eigenspace Consistency)，以及喚醒關鍵字語音共鳴曲線相似度(Resonant Curve Similarity)作為判別關鍵字的特徵，並藉由貝氏風險評估與串聯式偵測器的結合做為偵測的機制。此方法在極低訊噪比下仍保有相當強健的辨識率，因而可以適用在遠距關鍵字語音偵測或者在吵雜的環境下作為關鍵字語音喚醒機制。除了偵測關鍵字外，本方法還能同時估測出關鍵字的聲源方向，並保有串接其他偵測器的能力。在有額外的語音特徵或空間特徵可以加入時，能夠簡易的設計新的偵測器，串接到原本的架構上以持續增進辨識率。

A Wake-Up-Word Detection Method Using Spatial Eigenspace Consistency and Resonant Curve Similarity

Student： Ting-Chao Wang Advisor： Prof. Jwu-Sheng Hu

Institute of Electrical and Control Engineering

ABSTRACT

This thesis proposes a method to detect keywords using microphone array. The consistency of the spatial eigenspaces formed by the speech source at different frequencies and the resonant curve similarity of the keyword are used as the features for detection. These features are processed and detected separately and the result is determined by cascading individual outcome using Bayes risk estimate. This proposed method can keep a high recognition rate under very low SNR conditions. Therefore, it is suitable for the applications such as distant wake-up and keyword detection in noisy environments. In addition, this method can estimate the direction of arrivals of the sound source, and the proposed architecture is easy to expand by adding other feature detection methods in the cascaded manner to further improve the recognition rate.

誌謝

兩年的碩士生涯，不知不覺就結束了，這應該是學生生涯的尾端了吧。之後也即將踏入另一個開始，另一個旅途。這兩年發生很多事情，大起大落，很感謝很多人的陪伴跟幫忙，最後完成了碩士學業。由衷的感謝我的指導老師，在研究上給予很多想法與建議，老師積極的態度、廣大的視野，也讓我看到不一樣的格局跟企圖，很值得去學習的榜樣。

感謝實驗室的大家，首先是唐哥，總是不厭其煩的回答我的問題，在研究上幫了許多忙，期待你的吉他速彈再現；還有興哥引領我進入聲音領域，在研究上的指點，祝你博班順利畢業；計畫一直跟著的博班學長永融，在你身上看見認真的男人最帥氣，很值得學習，希望小永融出生可以健康的成長茁壯，並像爸爸學習；什麼都會什麼都不奇怪的阿吉，希望還能吃到你帶來的豬腳；最近成為人妻的鏗元學姊，引領我進入三國殺的領域，有機會再喝酒聊天，並祝妳婚姻幸福；還有業界突然轉入學術界的昌言學長，我會永遠記住這句話”是大將，就不缺戰場”，真是太會鼓勵人了，開公司記得找我去打雜；會做小點心的阿法，新髮型真的不錯喔，總覺得你以後的女朋友一定會很幸福；聯誼大師兄丸子學長晉源，認真覺得你的性感小鬍子應該繼續留著，祝你博班和感情路上順遂；散打之神 JUDO，還沒有現場看過你的比賽真是太可惜了；偶爾閒聊的得洋學長，祝你在外事業順利。

還有已經畢業同樣是 Rocker 的 Lundy，在剛進實驗室時有你的陪伴，祝你研替順遂；戰神 DODO，交接了讓我學到很多的居家照護計畫給我，祝你申請國外博士順利；邊打籃球邊聊天的肉鬆，讓我了解很多業界的狀況，可謂先驅啊；接下來是同梯的僑生戰友阿 him，看你不管做什麼都沒有問題的啦，有空一起去衝浪吧；蛋糕社的小菜，做的蛋糕有職業級的水準了，以後都吃不到了怎麼辦；還有一同衝刺的沛錡，我們終於要畢業了！有空再來打打撞球吧；還有巴西同胞 Roldofol，因為你我才知道原來破碎的英文也是可以溝通的，你踏進實驗室的一小步，是我英文成長的一大步，希望你在台灣發展順利。

最後是碩一、碩零的學弟妹們，祝你們研究順利，從中找到興趣與成就。直屬學弟學文，以後家聚記得要找我；幽默的 Macaca，對 3C 跟 IT 產業的視野真是令人印象深刻；新文你該結婚了吧，看你感情很穩定；計畫一起努力的昀軒，接下來 Wake-Up-Word 就靠你了；喜歡聊天的學妹湘筑，穿夾腳拖不錯看啊；強大的新秀建安，希望你找到一個題目大展身手；散打冠軍昭男，帥氣的耕維，聽說好色的建廷，還沒鬥牛的宗翰，還有不得不說的吉他教學哲鳴，在最後都幫了許多忙，感謝你們。

更感謝我的家人和女朋友阿雅，持續給我鼓勵與支持，並陪伴我度過各個難關。最後感謝孕育我成長的交大。希望大家一切順遂。

目　錄

表　列

圖　列

第一章　緒論

1.1 研究動機

在一般語音訊號處理上，關鍵字偵測或是擷取(Keyword Detection or Spotting)是語音辨識 (Speech Recognition)相當重要的一環，辨識步驟主要為先擷取語音特徵參數(LPC、MFCC、PLP 等)[1-1] [1-2] [1-3]、為語音特徵建出模型(HMM 等) [1-4] [1-5]及設定特徵參數比對方法(計算距離或相似度)。儘管語音辨識技術已經發展多年，在訊噪比高的情形下對大型詞彙庫的辨識率已經相當不錯，然而面對環境的雜訊干擾或是多人同時發聲的情況，即使是單一關鍵字的辨識率，也大多很難維持一定的水準[1-6] [1-7] [1-8]。在現實環境中，各種不同的聲音干擾是無法避免的，因此如何能在吵雜的環境中對關鍵字仍保有極高的辨識率，仍是目前相當重要的研究課題。

在自動語音辨識系統中(Automatic Speech Recognition，ASR)，何時可以開始進行辨識是其中一項重要的功能，該功能通常稱作 push button 或是 wake-up。Wake-up 功能運用得宜可以大量降低辨識錯誤率。一般在如電腦或手機的介面中往往以觸控或按鈕來實現，但是這個前提是所面對的裝置或機器需要在使用者的手邊。如果與使用者有一段距離，使用者往往必須配戴一無線裝置以提供可靠的 wake-up 訊號，在許多實際應用上這仍有其障礙。例如要命令智慧型居家服務機器人提供服務，若使用者必須一直配戴一無線裝置，在居家的情境中幾乎是不可行。因此，如何能夠在無需配戴任何裝置的情形下有效的實現 wake-up 功能，就成為一個實用且富挑戰性的研究題目。因為使用者不能配戴任何裝置，且提供語音辨識介面的機器很可能不在視野範圍內，因此無可避免的必須回歸到以語音來執行 wake-up 的功能。簡單來說，這即是單一關鍵字的辨識問題，但是其所面臨的問題是語者可能距離相當遠，因此訊噪比通常很差。其次是如同按鈕或觸控，以語音關鍵字實現 wake-up 也必須有幾乎 100%的 detection rate 以及接近於 0 的 false positive

rate，否則將產生誤動作或反應遲鈍，誠如前段所述，以目前通用的語音辨識方法，仍無法達成這個效能。

本論文針對上述的 wake-up 問題，嘗試藉由加入相對抗雜訊能力強的空間資訊和語音辨識器結合以保持極高的辨識率。並且在所加入的空間資訊中考慮到語音特徵，使得判別空間資訊不僅僅考慮聲源來向的一致性，亦能區隔非 wake-up 關鍵字的字詞。其基本想法是由預先設定的 wake-up 關鍵字(如機器人的名字)，判斷其特徵頻率的空間特性的一致性，藉以在低訊噪比的情形下大量提升其辨識穩定度。此想法的空間特徵資訊獲取則必須使用麥克風陣列。

1.2 文獻回顧

語音在做為人機介面(Human-Machine Interface)上擁有相當多的優點與特性，以下先介紹人機介面的發展。在傳統上人機介面為使人去適應機器，學習並習慣人機介面的設計與機構，以傳達控制命令。舉如圖形化使用者介面(GUI)中的滑鼠、鍵盤等，需要適應鍵盤按鍵的位置，並於近距離與機器互動；而在遠距中則以無線控制裝置，如遙控器、無線搖桿等裝置溝通。

近年來的人機介面趨勢已逐漸朝向機器去適應人的觀念去發展，如觸控螢幕。但以智慧型人機介面而言，還需要更方便、更快速、更簡易使用的方法，同時還要讓使用者能空出自己的雙手和視覺的專注力，以能夠處理其他的事情或避免危險，例如在駕駛時與導航機溝通目的地等。多年來在許多科幻電影或小說中早已勾勒出未來以語音方式與機器溝通的情境，其終級目標為使人和機器間的溝通就像人和人之間的溝通同樣的簡易、方便。

在眾多人機介面技術中，語音辨識擁有以上的優點與特性[1-9][1-10]。語音為人類的本能，不需要額外再去訓練學習，快速且有效率，可以很方便的傳達訊息給另一方，也不需要面對著機器或為了與機器溝通必須使用雙手。雖然在電腦視覺中，也可以透過手勢或姿態辨識來和電腦溝通互動，但

語音具有繞射的特性，在和機器間無法直視的情況下，視覺會喪失其功能，若要在環境中安裝多支攝影機，則又會有隱私及成本的問題。

然而以目前語音辨識的技術，仍存在著許多技術瓶頸導致其無法大量應用於人類的生活中。其中最困難的問題是其在各種應用環境下的穩定度。語音辨識技術已發展多年，雖然在訊噪比高的情形下對大型詞彙庫的辨識率已經相當不錯，然而在面對環境雜訊的干擾或是多人同時發聲的情況，則很難維持一定的水準，以致於誤動作頻繁[1-6] [1-7] [1-8]。對於穩定度的提升，在傳統上常用的解決方法就是仰賴一個穩健的開關(Push Button or Wake-Up Button)，如滑鼠、觸控螢幕等。在需要辨識時按下開關後開始辨識，減少持續辨識產生的錯誤及誤動作，以增進語音辨識器的辨識率，如果辨識錯誤則按下開關開起重新辨識。若使用一個穩健並獨一無二的語音關鍵字作為開關，就為語音關鍵字喚醒機制(Wake-Up-Word)。

例如在居家機器人的應用上，預先設定一個 wake-up 關鍵字(如機器人的名字)，並只需要持續的偵測關鍵字是否有發聲。當使用者呼喚機器人時，機器人走近使用者並可開啟語音純化與語音辨識，亦或者再開啟其他應用功能，如人臉偵測、姿態辨識等，詢問並等待使用者發出進一步的指令。從此應用上可以看出四個優點，第一，若語音辨識器持續運作，則在長時間運作中可能時常辨識錯誤並產生誤動作，尤其在機器人與使用者距離過遠或者環境吵雜下更容易發生錯誤。第二，若不使用語音關鍵字作為喚醒機制，則使用者必須每次要走近機器人以啟動服務，不然就需要隨身攜帶遙控器，且需擔心電池電力用罄的問題，在方便性上就大打折扣。第三，在使用者呼叫後，機器人可以藉此得出使用者的資訊(如方位)，藉由此資訊為語音辨器創造一個更為良好的聲場環境，例如走近使用者或者開啟對使用者方向的空間純化器等。第四，機器人平時不用開啟多項功能，節省能源並增加運作時間，減少充電次數。

誠如以上所說，語音關鍵字喚醒有其存在的必要。在“*Nonlinear Analysis: Theory, Methods & Applications*”期刊中的一篇文獻有一個完整的定義[1-11]：語音關鍵字喚醒(Wake-Up-Word)為一個語音關鍵字或詞句的偵測機制，藉由挑選出獨一無二的關鍵字，並只對此字做偵測，達成穩健、準確、有效率的開關以喚醒機器平台(如機器人、電腦)，並當需要喚醒機器人時擁有將近 100%的偵測率(Detection Rate)，且阻絕其他任何的字、詞句、雜訊、音樂等任何聲音，達成將近 0 的誤報率(False Positive Rate)。此機制在語音辨識中是一個新的範疇，並且尚未被廣泛的探討，可視為一種語音關鍵字偵測方法(Keyword Detection or Keyword Spotting)，但又和其功能與設計目的有所不同。目前主要方法大致和關鍵字偵測相同[1-12]，而後也有對語音特徵進行多種評分並分類的方法[1-13] [1-14]，然而目前都還是在純淨語音訊號的語料庫下發展，未有見過在實際環境中可能遇到低 SNR 下的文獻探討。

1.3 研究目標

基於上述的說明，本論文以麥克風陣列訊號處理為基礎，將研究目標分為：

1. 測試特徵頻率其空間性特徵空間一致性是否具有區別不同字的能力。
2. 抽取可運用於空間資訊鑑別之語音特徵。
3. 測試同一發聲者跟不同發聲者在所使用的特徵之差異。
4. 發展一套可於低 SNR 下運作的 wake-up 關鍵字偵測演算法。

1.4 本研究創新說明

本研究的創新為相較於一般通用的語音辨識器，增加了空間判別的資訊，用關鍵字中特徵頻率其空間性特徵空間一致性來做進一步篩選，使得可以在低訊噪比(SNR)下有相當的辨識率，因而可以適用在遠距關鍵字語音偵測或者在嘈雜的環境下作為關鍵字語音喚醒機制，並且還能同時估測出關鍵字的聲源方向。本方法的優點是只需要累積 5 個音框即可進行辨識，因此可

以達到即時的反應。基於貝氏風險(Bayes Risk)理論的門檻值(Threshold)判別以及利用串接多個偵測器的組合，使本方法得以在極低的訊噪比之下仍保有非常高的辨識率。經大量的語料測試，本方法可以在-3.82dB 的訊噪比之下達成 100%的 detection rate 以及 10.32%的 false positive rate，且所測試其穩定度的語料為與目標關鍵字相當接近的語句，在目前文獻中幾乎未見此高效能的展現。同時，本研究串聯式偵測器保有串接其他偵測器的能力，在有額外的語音特徵或空間特徵可以加入時，能夠簡易的設計新的偵測器，串接到原本的架構上以持續增進辨識率。

1.5 論文架構

本篇論文架構包含了三個主要的部份，分別是麥克風陣列訊號處理與語音辨識的背景理論、本論文提出的演算法及實驗成果與分析。以下描述各章節的主要內容：

第二章：將介紹麥克風陣列訊號處理、Eigenspace Method 的聲源方位估算演算法 MUSIC、語音的特性與特徵、線性預測編碼(LPC)、貝氏風險及串聯式偵測器。

第三章：介紹本論文的演算法，如何利用空間域特徵空間一致性與共鳴曲線一致性偵測喚醒關鍵字。

第四章：實驗的結果與分析。

第五章：研究成果及未來展望。

第二章　背景技術介紹

2.1 麥克風陣列訊號處理

在傳統數位訊號處理研究中，大多著重於時域及頻域的處理技巧，通常先將連續訊號進行取樣，接著轉換成頻域以分析訊號或者通過濾波器來區分訊號中不同的成分。

多個麥克風排成一個固定的形狀，接收來自空間中傳遞的訊號，由於麥克風位在空間中不同處，因而對於同一發聲源的訊號會擷取到不同的能量變化及時間差，而對多個麥克風間擷取到的訊號差異進行處理分析的技術，稱為麥克風陣列訊號處理。在此領域中，依照其目的不同，大致可以將其研究領域分為兩大類：

第一類：著重於估測訊號的數量或在空間中的方位，一般稱為聲源到達方位估測(Direction of Arrivals Estimation，DOA)。

第二類：利用訊號的空間關係，能夠對不同方向的訊號作出不同的增益，以達到空間濾波的效果，藉以分離空間中不同方向聲源的訊號，一般稱之為波束形成理論(Beamformer)，也就是一種空間濾波器(Spatial Filter)。

當給定一個麥克風陣列及一個參考點(通常為某顆麥克風)，array manifold vector 定義出每個麥克風擷取到的聲源訊號相對於此參考點的時間關係。在波束形成理論中，array manifold vector 被用來補償入射訊號到不同麥克風之間的相位差；而在聲源到達角度估測理論中，很多方法藉由比對 array manifold vector 跟聲源訊號 eigenspace 的相似度來求出聲源方向，此類方法統稱為 Eigenstructure Methods DOA。

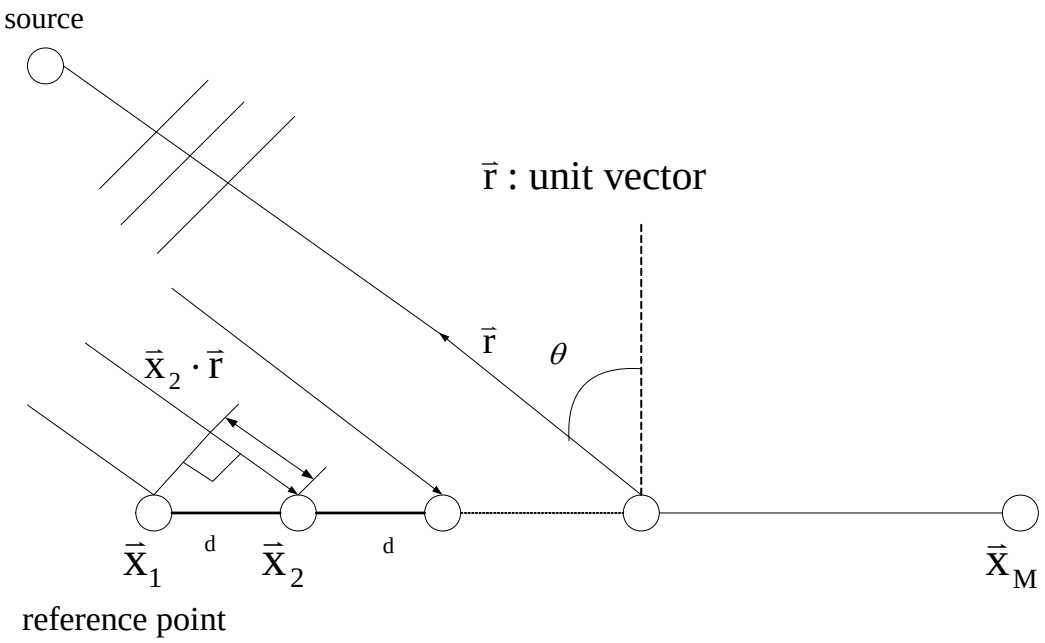


圖 2.1 均勻線性陣列架構

通常在不同的情況下會選用不同的陣列排列形狀，以下推導最常見的均勻線性陣列(Uniform Linear Array，ULA)，其架構圖如圖 2.1 所示。假設聲源訊號為遠場平面波(Far field plane wave)，s(t)為原始訊號，n(t)為雜訊，則M個麥克風輸出可以寫成下列向量的形式：

$$\boldsymbol{x}(t)=\begin{bmatrix} x_1(t) \\ \vdots \\ x_M(t) \end{bmatrix}=\begin{bmatrix} s(t)e^{jw_c\frac{\vec{x}_1\cdot\vec{r}}{c}} \\ \vdots \\ s(t)e^{jw_c\frac{\vec{x}_M\cdot\vec{r}}{c}} \end{bmatrix}+\begin{bmatrix} n_1(t) \\ \vdots \\ n_M(t) \end{bmatrix}$$

$$=\begin{bmatrix} e^{jk_c\vec{x}_1\cdot\vec{r}} \\ \vdots \\ e^{jk_c\vec{x}_M\cdot\vec{r}} \end{bmatrix}s(t)+\begin{bmatrix} n_1(t) \\ \vdots \\ n_M(t) \end{bmatrix}=\boldsymbol{a}(\vec{r})s(t)+\boldsymbol{n}(t) \qquad (2.1.1)$$

$k_c=\frac{w_c}{c}=\frac{2\pi}{\lambda_c}$ ，k_c 稱為 wavenumber，而 λ_c 為波長，c 為波速，其中 $\boldsymbol{a}(\vec{r})$ 稱為 array manifold vector，包含了訊號傳遞到麥克風之間的時間關係，可以再簡化為：

$$\boldsymbol{a}^{\mathrm{T}}(\theta)=\begin{bmatrix} 1 & e^{jk_cd\sin\theta} & \cdots & e^{jk_c(M-1)d\sin\theta} \end{bmatrix} \qquad (2.1.2)$$

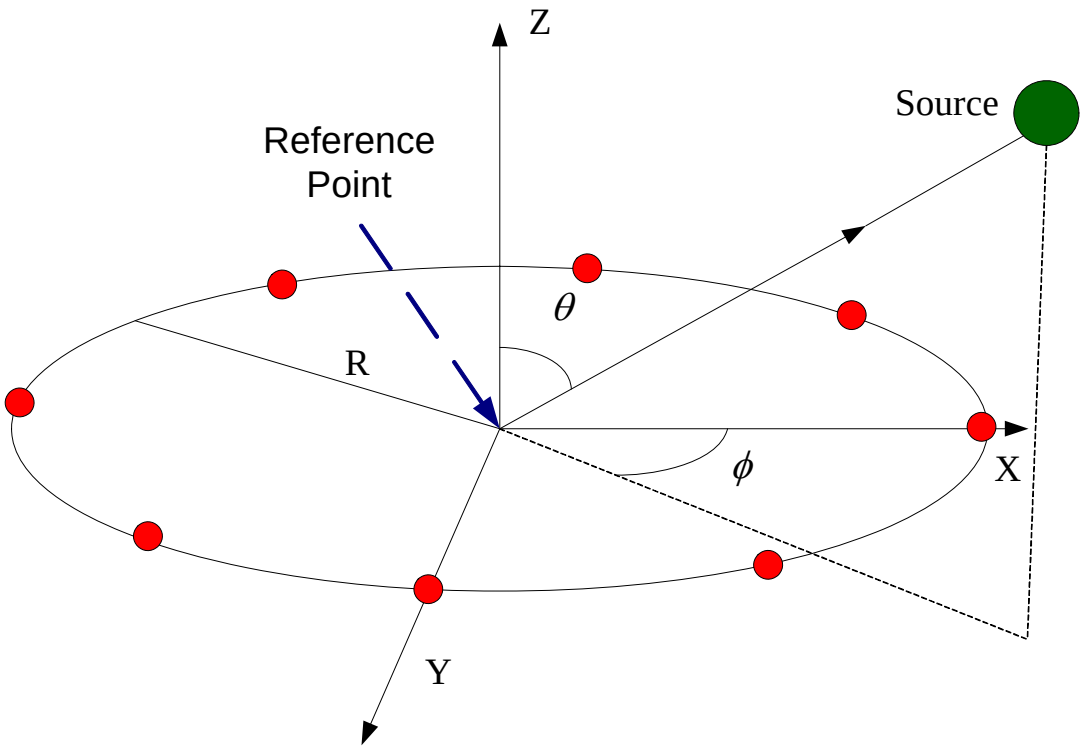


圖 2.2 環型陣列架構

圖 2.2 是等距的環形陣列排列方式，具備有 2-D 維度的角度搜索能力，本論文也是以此環型陣列架構為基礎。

設定圖 2.2 的圓心為參考點，可推得 array manifold vector 為：

$$\boldsymbol{a}^T(\theta)=\left[1 \quad e^{jk_c*R\sin\iota\cos\theta} \quad e^{jk_c*R\sin\phi\cos(\theta-2\pi/M)} \quad e^{jk_c*R\sin\phi\cos(\theta-2(M-1)\pi/M)}\right] \qquad (2.1.3)$$

R：圓形半徑　　M：天線個數

2.2 Multiple Signals Classification Method (MUSIC)

在聲源到達方位(DOA)估測方法中，依照技術層面不同巨觀上可以分成兩大類，第一類是利用聲音從空間中傳遞的特性所造成的時間差來作方位估測，稱為 TDE(Time Delay Estimation)[2-1]；第二類是利用不同訊號源間特徵向量的分布關係，以互相投影或判斷相似度的方式來估測出訊號的方位，稱為 Eigenstructure Method[2-2]。

MUSIC[2-3]是一個簡單有效率的聲源到達方位估測方法，在兩大類中屬於 Eigenstructure Method。此方法利用不同訊號源其特徵向量的分布關係，以訊號子空間與雜訊子空間互相正交的特性來估算聲源的方位，也因此可以同時偵測多個聲源的方位。

使用 MUSIC 演算法必須要滿足兩個基本假設：

I. 訊號相關矩陣(Source Correlation Matrix)是滿序(Full Rank)，且序等於訊號來源的數目 D。

II. Array manifold vector $a(\theta_i)$，$i=1,\cdots,$D 彼此之間是線性獨立，滿足 array manifold matrix 是滿序，且序也等於訊號來源的數目 D。

滿足以上兩個基本假設之後，把(2.1.1)式改寫成多個訊號源的形式：

$$\boldsymbol{x}(t)=\sum_{i=1}^{D}\boldsymbol{a}(\theta_i)s_i(t)+\boldsymbol{n}(t)$$

$$=\mathrm{A}\boldsymbol{s}(\mathrm{t})+\boldsymbol{n}(\mathrm{t})\ ,\ \mathrm{A}=\left[a(\theta_1)\ \cdots\ a(\theta_D)\right]\ ,\ \boldsymbol{s}(t)=\begin{bmatrix}s_1(t)\\ \vdots\\ s_M(t)\end{bmatrix} \qquad (2.2.1)$$

經過 short time fourier transform(STFT)轉換成頻域：

$$\boldsymbol{X}(\omega_f,k)=\mathbf{A}(\omega_f)\boldsymbol{S}(\omega_f,k)+\boldsymbol{N}(\omega_f,k)\quad ,f=1\cdots F \qquad (2.2.2)$$

其中 k 代表不同的時間間隔，F 代表 FFT size。

假設聲源和雜訊彼此不相關，則資料相關矩陣(Data Correlation Matrix) $\mathbf{R}_{XX}$ 為：

$$\begin{aligned}\mathbf{R}_{XX}(\omega_f,k) &= E\left(\boldsymbol{X}(\omega_f,k),\boldsymbol{X}(\omega_f,k)^H\right)\\ &= \mathbf{A}(\omega_f)\mathbf{R}_{SS}(\omega_f,k)\mathbf{A}(\omega_f)^{\mathbf{H}} + \sigma_N^2(\omega_f,k)\mathbf{I}\end{aligned} \tag{2.2.3}$$

將資料相關矩陣特徵值分解(Eigenvalue Decomposition，SVD)：

$$\mathbf{R}_{XX}(\omega_f) = \sum_{i=1}^{M}\lambda_i(\omega_f)\boldsymbol{V}_i(\omega_f)\boldsymbol{V}_i^H(\omega_f) \tag{2.2.4}$$

雜訊相關矩陣(Noise Correlation Matrix)也可以寫成為：

$$\sigma_N^2(\omega_f)\mathbf{I} = \sum_{i=1}^{M}\sigma_N^2(\omega_f)\boldsymbol{V}_i(\omega_f)\boldsymbol{V}_i^H(\omega_f) \tag{2.2.5}$$

則純訊號相關矩陣(Signal-Only Correlation Matrix)就可以由方程式(2.2.4)-(2.2.5)得出：

$$\begin{aligned}\mathbf{C}_{XX}(\omega_f) &= \mathbf{A}(\omega_f)\mathbf{R}_{SS}(\omega_f)\mathbf{A}(\omega_f)^H\\ &= \sum_{i=1}^{M}\left(\lambda_i(\omega_f)-\sigma_N^2(\omega_f)\right)\boldsymbol{V}_i(\omega_f)\boldsymbol{V}_i^H(\omega_f)\end{aligned} \tag{2.2.6}$$

因為$\mathbf{C}_{XX}(\omega_f)$的序為 D，可以發現其 Range space 為：

$Rs(\mathbf{C}_{XX}(\omega_f))$=$span\{\boldsymbol{V}_1(\omega_f),\cdots,\boldsymbol{V}_D(\omega_f)\}$, Rs : Range Space

$$Rs(\mathbf{A}(\omega_f)) = span\left\{\boldsymbol{a}(\theta_1,\omega_f),\cdots,\boldsymbol{a}(\theta_D,\omega_f)\right\} = span\left\{\boldsymbol{V}_1(\omega_f),\cdots,\boldsymbol{V}_D(\omega_f)\right\}$$

$$Rs(\mathbf{A}(\omega_f))^{\perp} = span\{\boldsymbol{V}_{D+1}(\omega_f),\cdots,\boldsymbol{V}_M(\omega_f)\} \tag{2.2.7}$$

因此定義出訊號跟雜訊子空間：

1. $\boldsymbol{V}_1(\omega_f),\cdots,\boldsymbol{V}_D(\omega_f)$ 稱為訊號的特徵向量

   並且以 $span\{\boldsymbol{V}_1(\omega_f),\cdots,\boldsymbol{V}_D(\omega_f)\}$ 為訊號子空間(Signal Subspace)。

2. $\boldsymbol{V}_{D+1}(\omega_f),\cdots,\boldsymbol{V}_M(\omega_f)$ 稱為雜訊的特徵向量

   並且以 $span\{\boldsymbol{V}_{D+1}(\omega_f),\cdots,\boldsymbol{V}_M(\omega_f)\}$ 為訊號子空間(Noise Subspace)。

再來利用訊號子空間與雜訊子空間正交(Orthogonal)的特性：

$$\boldsymbol{V}(\omega_f)_j^H \boldsymbol{a}(\theta_i,\omega_f)=0 \quad ,i=1,2,\cdots,D \ , \ j=D+1,\cdots,M \tag{2.2.8}$$

建立一個投影到雜訊子空間的投影矩陣 $\mathbf{P}_N$ ：

$$\mathbf{P}_N(\omega_f)=\sum_{i=D+1}^{M}\boldsymbol{V}(\omega_f)_i\boldsymbol{V}(\omega_f)_i^H=\mathbf{P}_A^{\perp}(\omega_f) \tag{2.2.9}$$

而聲源到達角度 $\theta_i,i=1,\cdots,D$ ，就可利用方程式(2.2.9)得到的投影矩陣 $\mathbf{P}_N$ 去解方程式(2.2.10)就可求得：

$$\mathbf{P}_N(\omega_f)\boldsymbol{a}(\theta,\omega_f)=0 \tag{2.2.10}$$

為了更方便地可以找到 θ ，取方程式(2.2.10)的大小：

$$\left\|\mathbf{P}_N(\omega_f)\boldsymbol{a}(\theta,\omega_f)\right\|_2^2=\boldsymbol{a}^H(\theta,\omega_f)\mathbf{P}_N(\omega_f)\boldsymbol{a}(\theta,\omega_f)=0 \tag{2.2.11}$$

$$\boldsymbol{S}_{MUSIC}(\theta,\omega_f)=\frac{1}{\boldsymbol{a}^H(\theta,\omega_f)\mathbf{P}_N(\omega_f)\boldsymbol{a}(\theta,\omega_f)} \tag{2.2.12}$$

所以最後的角度估測就為：

$$\hat{\theta}(\omega_f)=\arg\max_{\theta}\boldsymbol{S}_{MUSIC}(\theta,\omega_f) \tag{2.2.13}$$

在方程式(2.2.12)式子中，$\boldsymbol{S}_{MUSIC}(\theta,\omega_f)$ 稱為 MUSIC spectrum，或者稱為 Space spectrum。另外在實際上方程式(2.2.11)因為雜訊、誤差的緣故並不會真的等於零，所以不會看到無限高的 MUSIC spectrum。

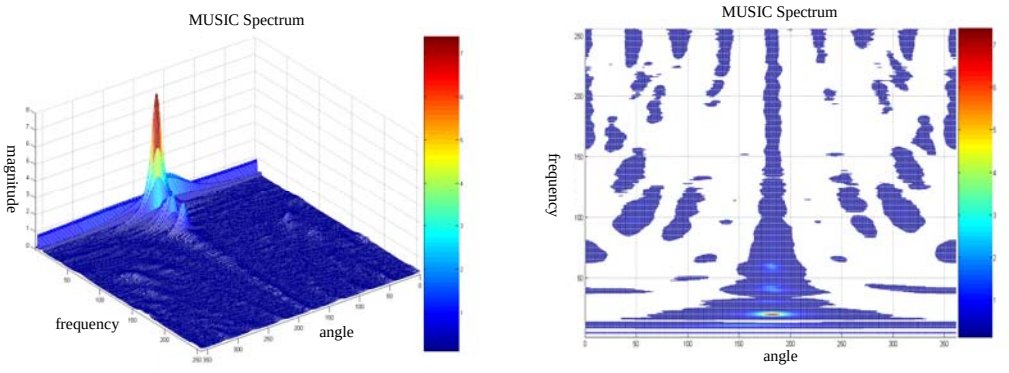


圖 2.3 各個頻率下的 MUSIC Spectrum

從統計的觀點來看，MUSIC 就相當於在各個頻帶下做主要成分分析(Principle Component Analysis，PCA)。在不考慮過低頻 space coherence 及過高頻 space spectrum aliasing 無法使用外，比較好的訊號到達方位估測結果，通常在某些頻域上 SNR 較高的頻帶，不過反過來卻不一定成立，因為空間的聲場環境過於複雜。

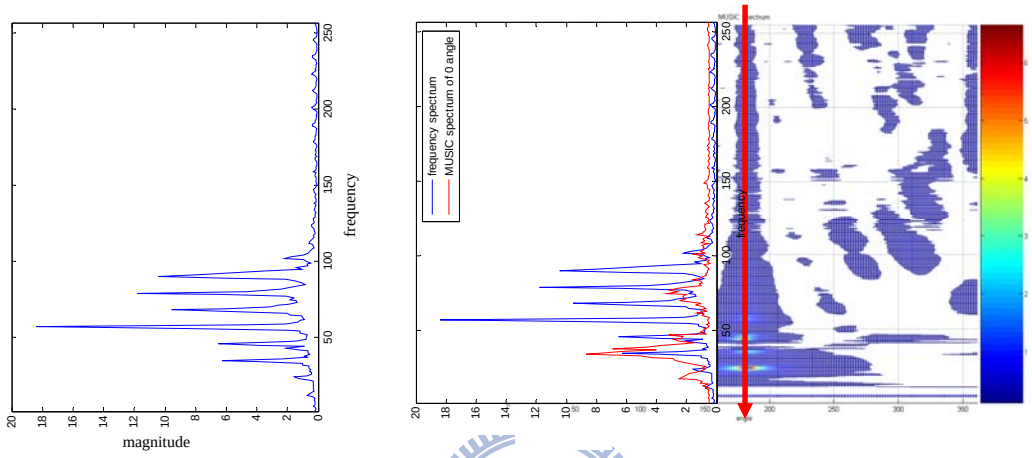


圖 2.4　頻率頻譜與空間頻譜關係圖(SNR=16.14)

在圖 2.4 的左圖，藍線為某個字的能量頻譜(此圖已經被逆時鐘翻轉了 90 度，所以縱軸是頻率，橫軸是能量大小)。右圖紅線為把空間頻譜中估測出正確角度的量值畫在能量頻譜上。

SNR=16.14　　SNR=-2.84　　SNR=-16.24

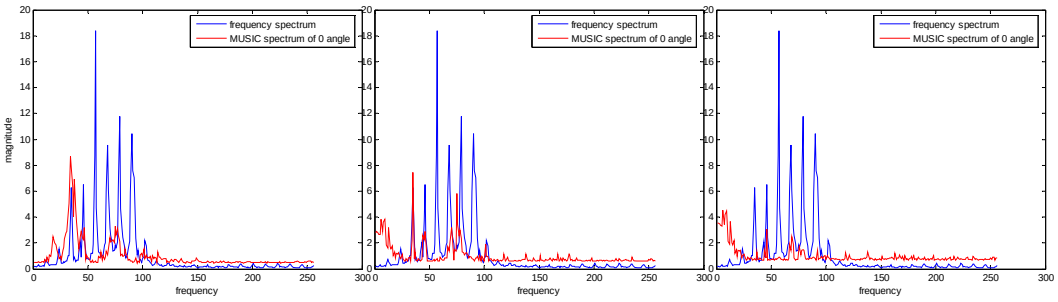


圖 2.5　頻率頻譜與空間頻譜關係圖

從圖 2.5 的三個小圖可以清楚看到，當訊號源未受到干擾時，大部分的頻帶 SNR 都夠高，足以保存空間資訊。但當環境雜訊增加時，空間資訊只會保持在能量夠高的頻帶上，而在語音中，也就是聲紋的頻帶上。

另外值得一題，在對資料相關矩陣做完特徵值分解後，最大的特徵值已經具有空間濾波器純化的效果。以下為推導：

假設雜訊的能量遠低於訊號的能量且聲源只有 1 個的情況下，根據理論則將只會有 1 個不為 0 的特徵值，依照(2.2.7)式知道其所對應的特徵向量即是聲源方向 array manifold vector 的估測：

$$span\left\{\hat{\boldsymbol{a}}(\theta_1,\omega_f)\right\} = span\left\{\boldsymbol{V}_1(\omega_f)\right\}$$
$$\boldsymbol{a}'(\theta_1,\omega_f) \equiv \boldsymbol{V}_1(\omega_f) = \frac{\hat{\boldsymbol{a}}(\theta_1,\omega_f)}{C} \qquad C\text{: normailzed factor} \tag{2.2.14}$$

在實作上以(2.2.15)式近似(2.2.3)式：

$$\begin{aligned}\mathbf{R}_{XX}(\omega_f) &= \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{X}(\omega_f,k)\boldsymbol{X}^{\mathrm{H}}(\omega_f,k) \\ &= \sum_{i=1}^{M}\lambda_i(\omega_f)\boldsymbol{V}_i(\omega_f)\boldsymbol{V}_i^{H}(\omega_f)\end{aligned} \tag{2.2.15}$$

N 代表計算 correlation matrix 的音框數(frame number)，則最大的特徵值 $\lambda_1(\omega_f)$ 可以求出：

$$\begin{aligned}\lambda_1(\omega_f) &= \boldsymbol{V}_1^{H}(\omega_f)\mathbf{R}_{XX}(\omega_f)\boldsymbol{V}_1(\omega_f) \\ &= \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{V}_1^{H}(\omega_f)\boldsymbol{X}(\omega_f,k)\boldsymbol{X}^{\mathrm{H}}(\omega_f,k)\boldsymbol{V}_1(\omega_f) \\ &= \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{a}'^{\mathrm{H}}(\theta_1,\omega_f)\boldsymbol{X}(\omega_f,k)\boldsymbol{X}^{\mathrm{H}}(\omega_f,k)\boldsymbol{a}'(\theta_1,\omega_f) \\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{y}(\omega_f,k)\mathbf{y}^{\mathrm{H}}(\omega_f,k) \ , \mathbf{y}(\omega_f,k) \equiv \boldsymbol{a}'^{\mathrm{H}}(\theta_1,\omega_f)\boldsymbol{X}(\omega_f,k) \\ &= \frac{1}{N}\sum_{k=1}^{N}\left\|\mathbf{y}(\omega_f,k)\right\|^2\end{aligned} \tag{2.2.16}$$

從(2.2.16)式看到，最大的特徵值就是依照 PCA 估測出的 array manifold vector 依照所內含的相位差去補償各個麥克風，並把各個麥克風的訊號加總，此效果同等於 Delay and Sum Beamformer[2-4]，具有空間濾波器純化語音訊號的效果，最後對 N 個音框取平均，。

2.3 Speech Feature

語音特徵(Speech Feature)為語音辨識鑑別出不同字的依據，每一個字、每一個發音都有其不同的特徵，並根據這些特徵區別出不同的字。圖 2.6 Source-Filter Model 為最常描述語音發聲的模型[2-5]，此模型假設激勵源和濾波器為互相獨立的。

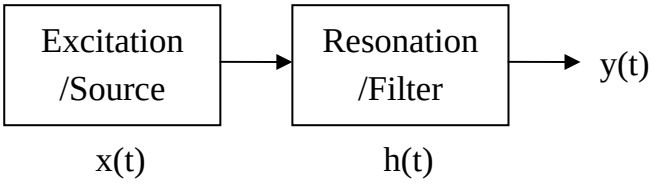


圖2.6 Source-Filter Model

$y(t)$ 為環境中聽到的訊號，$h(t)$ 為共鳴濾波器，$x(t)$ 為激勵源訊號，$\boldsymbol{Y}(w)$、$\boldsymbol{X}(w)$ 各為其能量頻譜，$\boldsymbol{H}(w)$ 為共鳴濾波器的頻率響應(Frequency Response)，又稱為共鳴曲線(Resonant Curve)[2-6]。圖 2.6 的關係可表示為：

$$y(t) = x(t) * h(t)$$
$$\boldsymbol{Y}(w) = \boldsymbol{X}(w)\boldsymbol{H}(w) \qquad (2.3.1)$$

在語音中激勵源(excitation)為人的聲帶，在弦樂器中激勵源就為琴弦，在管樂器中激勵源就為簧片。從激勵源出來的訊號經過了聲道、口腔、鼻腔、胸腔、嘴型等身體器官的共鳴放大就成為了環境中被人耳所聽到的人聲；在樂器中就是俗稱的共鳴箱。

從聲音的三要素來看，響度定義了聲音的強弱，由聲波的振幅來決定，振幅越大，表示聲波的能量越高，聲音也就越大聲。音調定義了聲音的高低，由振動的頻率決定，頻率越高，聲音也就越尖銳。音色定義了聲音的波形，不同的波形聽起來的感覺就不同。

因此，可以說激勵源控制了音高，共鳴濾波器控制了音色，而不同的語音文字則被歸類為不同音色的呈現。

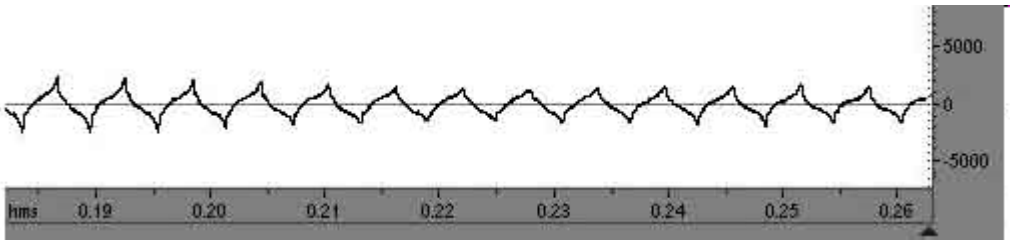


圖 2.7 聲帶震動產生的激勵訊號 $x(t)$

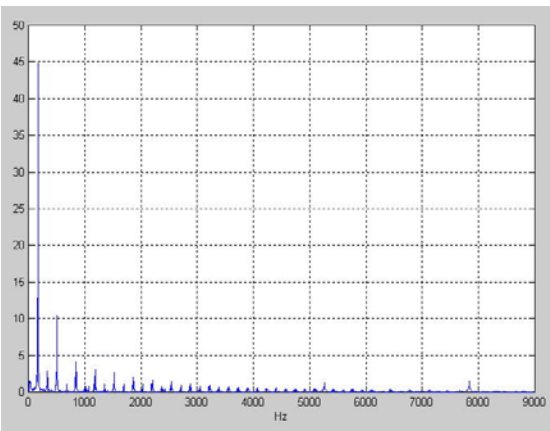


圖 2.8 激勵訊號的能量頻譜 $\boldsymbol{X}(w)$

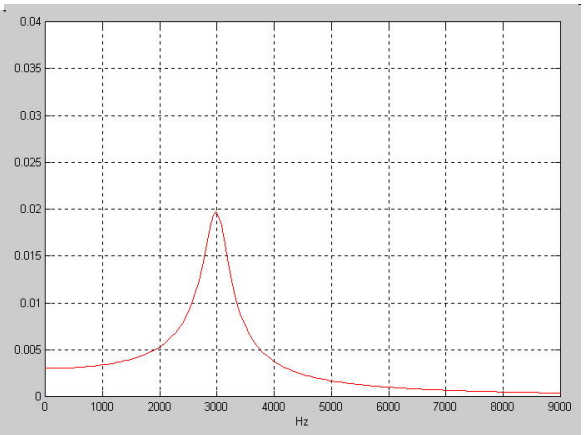


圖 2.9 共鳴濾波器的頻率響應(共鳴曲線) $\boldsymbol{H}(w)$

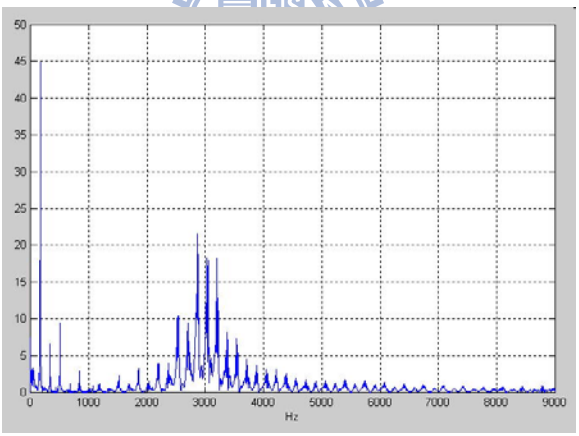


圖 2.10 合成訊號的能量頻譜 $\boldsymbol{Y}(w)$ [2-7]

圖 2.7 為人的聲帶震動後所產生的訊號，通常為一個三角波。圖 2.8 則為此波型的能量頻譜，可以看出激勵源的訊號主要為基頻與其諧波成分。此訊號經過了共鳴濾波器的共鳴放大，最後變為環境中所傳遞的訊號，如圖 2.10。

為了進一步了解語音的特性，以下測試同一個字、不同的發音音高來作比較，如圖 2.11(共鳴曲線為使用 LPC 所估測。藍色線為能量頻譜;紅色線為共鳴曲線)。

可以發現不同音高使基頻跟諧波的位置不同，而共鳴曲線只有些許變動，並注意共鳴曲線峰值處的頻率位置變異也不大。事實上共鳴曲線峰值處的頻率位置也為語音的重要特徵，稱為共振峰(Formant)[2-8]。

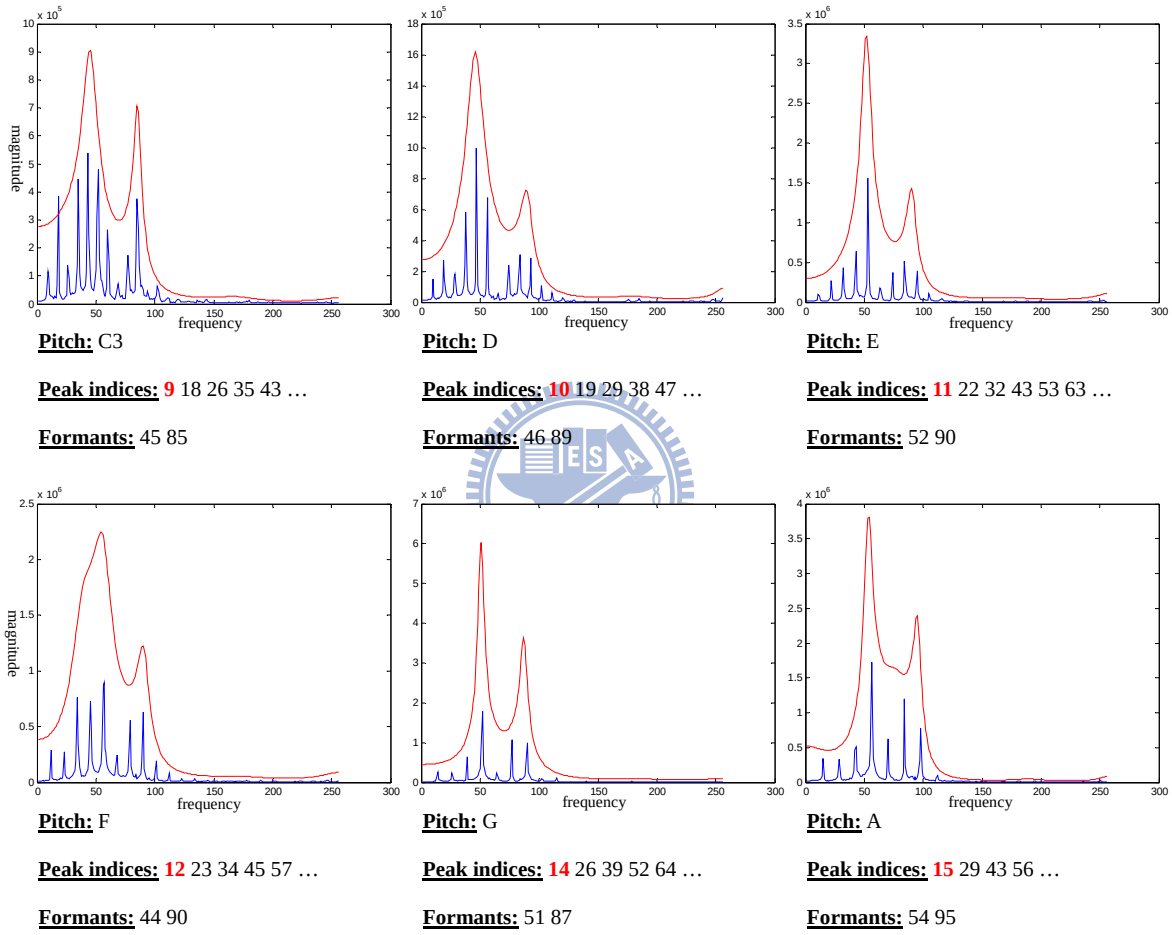


圖 2.11 不同音高下同字的能量頻譜與共鳴曲線

	C3	D	E	F	G	A
f (Hz)	130.81	146.83	164.814	174.614	195.998	220
Freq. index	8.37	9.4	10.55	11.18	12.54	14.08

表 2.1 實際頻率跟 FFT 後頻率索引對照表 (取樣頻率 8K)

再拿不同的字、同樣的發音音高來作比較，如圖 2.12。這時就可以發現共鳴曲線的變異就很大了，共振峰的位置也不同，而基頻跟諧波的位置沒有變動。在早期的語音辨識就是依照前兩個共振峰的位置來當作語音特徵，當時可以初步區別出不同的母音[2-9]。

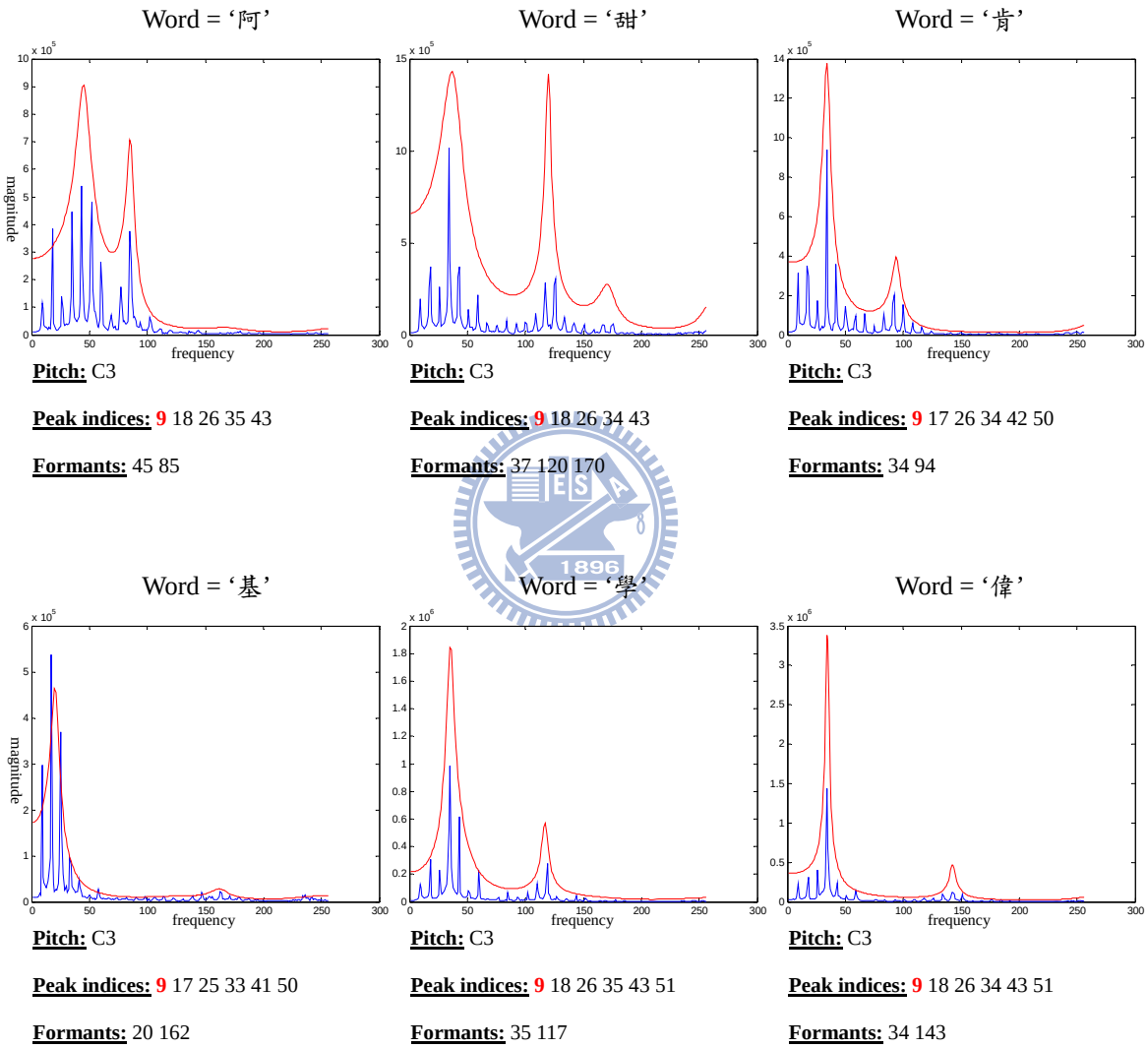


圖 2.12 同音高下不同字的能量頻譜與共鳴曲線

最後再拿同一個字、同樣的發音音高，但不同的人發音來做比較，如圖2.13。可以發現基頻跟諧波的位置沒有變動，而共鳴曲線的變異相對於不同字來說較小，共振峰的位置也只有稍微變動。共鳴曲線變動的差異是來自於不同人的口音、發音方式及講話習慣的不同所造成。

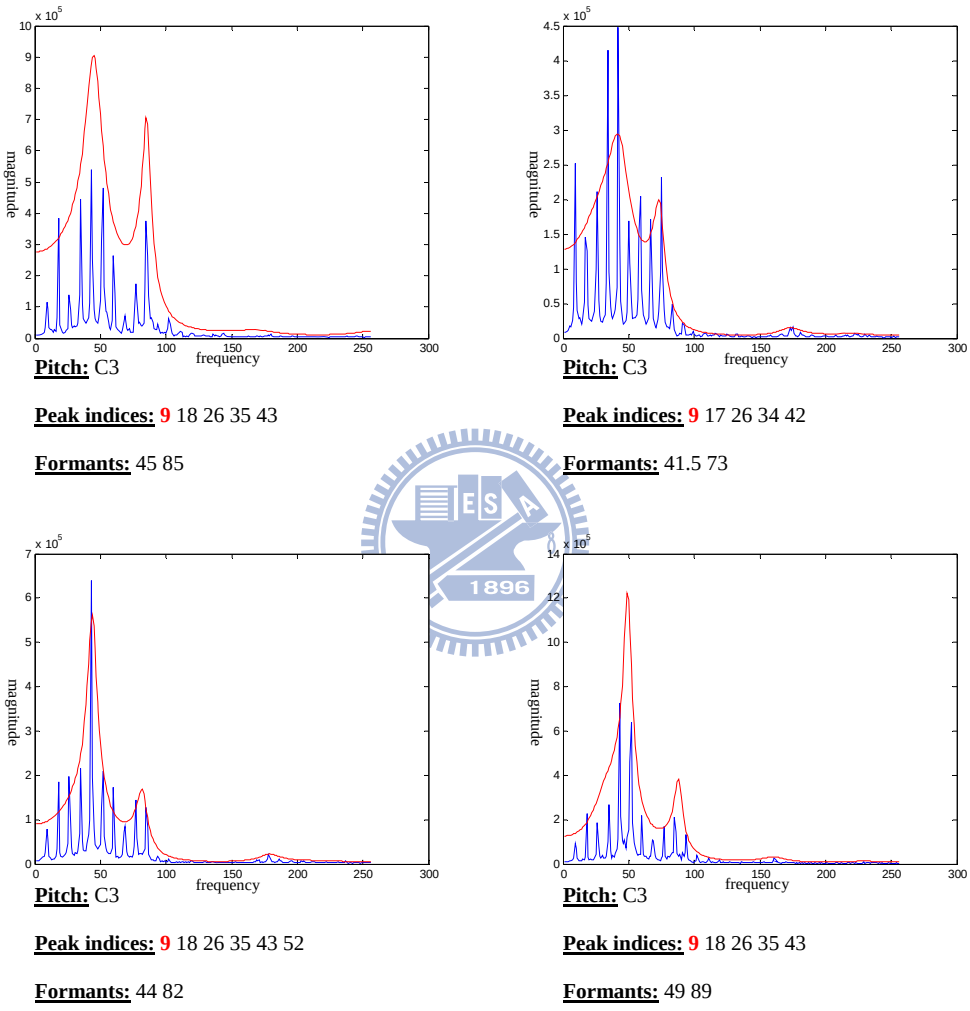


圖 2.13 同音高、同字但不同人的能量頻譜與共鳴曲線

經由以上的觀察，可以再次確定語音特徵就是整個共鳴濾波器(在人聲中共鳴濾波器也稱為聲道濾波器)，也就是此共鳴濾波器的頻率響應－共鳴曲線，而共振峰也為其中的特徵之一。至於激勵訊號則控制了基頻跟諧波的位置，不為語音的特徵。所以可以藉由判別共鳴曲線的相似性，或者共振峰的位置來區別出不同的語音文字，本論文也是依據這兩個語音特徵，來作後續的分析處理。

從語音辨識的觀點(Speech Recognition)，估測出的共鳴曲線代表著不同的發音方式，也就是不同的語音文字；從語音變換(Speech Transform)和語音合成(Speech Synthesize)的觀點，則對共鳴濾波器係數和激勵訊號做調整，合成出想要的語音。

而擷取語音特徵的問題，就會轉變成為估測共鳴濾波器係數的問題，或者直接蒐集大量資料建立共鳴曲線的模型。其中常見的方法為線性預測編碼(Linear Predictive Coding，LPC)、倒頻譜(Cepstrum)、梅爾倒頻譜係數（Mel-scale Frequency Cepstral Coefficients，MFCC）、線性感知預測(Perceptual Linear Predictive，PLP)幾種。

2.4 Linear Predictive Coding (LPC)

線性預測編碼(LPC)是一種最普遍被使用來描述共鳴濾波器的方法。它的概念是，目前的取樣點可以藉由之前幾個取樣點的線性組合來加以預測，也就是對過去的資料建立自迴歸模型(Autoregressive Mode，AR Model)來預測目前的資料，並假設共鳴濾波器為一個全極點濾波器(All-pole Filter)。由於語音的訊號在短時距內的變化緩慢，取樣點之間的自相關性(Autocorrelation)很高，所以LPC特別適合語音的處理。

把(2.3.1)式改成離散訊號並加入雜訊項改寫Source-Filter Model：

$$y(n) = x(n) * h(n) + e(n) \tag{2.4.1}$$

假設共鳴濾波器為全極點濾波器：

$$\boldsymbol{H}(z) = \frac{1}{1 - \sum_{k=1}^{p} z^{-k} a_k} \tag{2.4.2}$$

其中 p 為LPC 的階數(order)，a_k 為線性預測係數，$k = 1 \cdots p$ 。再把(2.4.1)式並轉成Z-domain並將(2.4.2)式帶入：

$$\boldsymbol{Y}(z) = \boldsymbol{X}(z) \frac{1}{1 - \sum_{k=1}^{p} z^{-k} a_k} + \boldsymbol{E}(z) \tag{2.4.3}$$

假設空間中的雜訊遠小於激勵源訊號並移項整理得出：

$$\text{Assume } \boldsymbol{X}(z) >> \boldsymbol{E}(z),$$
$$\boldsymbol{Y}(z)\left(1 - \sum_{k=1}^{p} z^{-k} a_k\right) = \boldsymbol{X}(z) \tag{2.4.4}$$

轉回time-domain，就可以看到AR Model的標準形式：

$$y(n) - \tilde{\boldsymbol{y}}(n)^T \boldsymbol{a} = x(n) \qquad , \tilde{\boldsymbol{y}}(n) = [y(n-1) \cdots y(n-p)]^T$$
$$\boldsymbol{a} = [a_1 \cdots a_p]^T \tag{2.4.5}$$

則預測的均方誤差(Mean Square Error，MSE)為：

$$E = \sum_{n=-\infty}^{\infty} x(n)^2 = \sum_{n=-\infty}^{\infty} \left(y(n) - \tilde{\boldsymbol{y}}(n)^T \boldsymbol{a} \right)^2 \tag{2.4.6}$$

此時的激勵訊號被看為預測誤差，解釋為此方法希望盡量用前面取樣的資料去預測目前取樣的資料，所以當前面的資料無法預測目前的資料時也就代表有新的激勵訊號進來。所以預測誤差代表無法預測的資料，也就是激勵訊號。

為了得到最小均方誤差的線性預測係數，對(2.4.6)式取 $\boldsymbol{a}$ 的偏微分並令為0：

$$\begin{aligned} \frac{\partial E}{\partial \boldsymbol{a}} &= \sum_{n=-\infty}^{\infty} \left(-2y(n)\tilde{\boldsymbol{y}}(n) + 2\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)^T \boldsymbol{a} \right) = 0 \\ &\sum_{n=-\infty}^{\infty} \left(y(n)\tilde{\boldsymbol{y}}(n) - \tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)^T \boldsymbol{a} \right) = 0 \\ &\sum_{n=-\infty}^{\infty} \left(y(n)\tilde{\boldsymbol{y}}(n) \right) = \sum_{n=-\infty}^{\infty} \left(\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)^T \right) \boldsymbol{a} \end{aligned} \tag{2.4.7}$$

寫成矩陣的形式：

$$\begin{bmatrix} R_{yy}(1) \\ R_{yy}(2) \\ \vdots \\ R_{yy}(p) \end{bmatrix} = \begin{bmatrix} R_{yy}(0) & R_{yy}(1) & \cdots & R_{yy}(p-1) \\ R_{yy}(1) & R_{yy}(0) & \cdots & R_{yy}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{yy}(p-1) & R_{yy}(p-2) & \cdots & R_{yy}(0) \end{bmatrix} \boldsymbol{a} \tag{2.4.8}$$

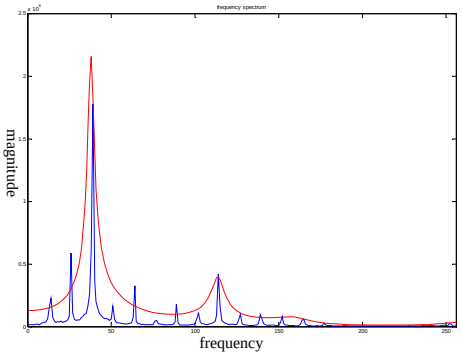
其中 $R_{yy}(\tau) = \sum_{n=-\infty}^{\infty} \left(y(n)y(n-\tau) \right)$，稱為自相關函數(Autocorrelation Function)。

最後定義兩個參數：

$$\mathbf{r}_{yy} = \begin{bmatrix} R_{yy}(1) \\ R_{yy}(2) \\ \vdots \\ R_{yy}(p) \end{bmatrix}, \mathbf{R}_{yy} = \begin{bmatrix} R_{yy}(0) & R_{yy}(1) & \cdots & R_{yy}(p-1) \\ R_{yy}(1) & R_{yy}(0) & \cdots & R_{yy}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{yy}(p-1) & R_{yy}(p-2) & \cdots & R_{yy}(0) \end{bmatrix} \tag{2.4.9}$$

就可以得出一個簡單的解：

$$\boldsymbol{a} = \mathbf{R}_{yy}^{-1} \mathbf{r}_{yy} \tag{2.4.10}$$



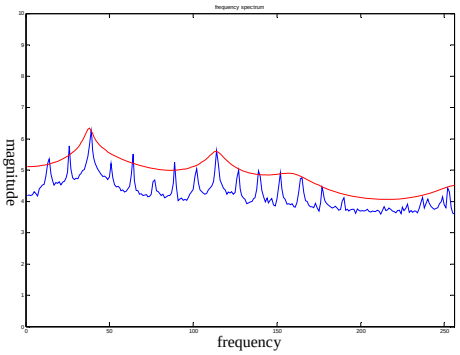


圖2.14　LPC模型之頻率響應　　圖2.15　LPC模型之頻率響應(dB表示)

最後得出LPC模型的頻率響應就為共鳴曲線，也被稱為頻譜包絡線(Spectral Envelope)，如圖2.14與圖2.15。在LPC模型預測的階數，則依照取樣頻率(Sampling Rate)的範圍，共振峰可能出現的數目而定，通常在語音訊號中大都是在10-16 階。

另外在實作上會以(2.4.12)式去近似自相關函數，N 代表計算自相關函數的資料數(frame size)：

$$R_{yy}(\tau,N)=\sum_{n=0}^{N-1}\left(y(n)y(n-\tau)\right) \qquad (2.4.12)$$

最後LPC模型除了使用在共鳴曲線的估測，也可以作為語音合成器的設計方式(Speech Synthesized)或應用於語音訊號壓縮(Speech Compression)。

2.5 Bayes Risk

貝氏風險(Bayes Risk)是一種總括性的決策法則。給定兩個已知的機率分布(Probability Distribution)，藉由設定各種決策結果的代價(Cost)，決定出最佳的門檻值(Threshold)以使得整體代價最小[2-10]。

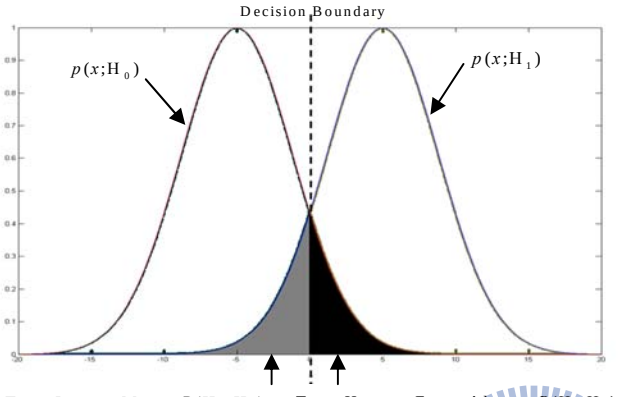


圖 2.16 兩個機率分布圖

	H_0 is Ture	H_1 is True
Decided as H_0	C_{00}	C_{01}
Decided as H_1	C_{10}	C_{11}

表 2.2 不同決策結果的代價表

給定兩個機率分布如圖 2.16，定義 Hypothesis：

H_0: μ=0　　null hypothesis

H_1: μ=1　　alternative hypothesis

並定義 C_{ij} 為表 2.2，則總共的代價為：

$$\begin{aligned} E(C) &= C_{00}P(H_0;H_0)+C_{01}P(H_0;H_1)+C_{10}P(H_1;H_0)+C_{11}P(H_1;H_1) \\ &= C_{00}P(H_0\mid H_0)P(H_0)+C_{01}P(H_0\mid H_1)P(H_1) \\ &\quad +C_{10}P(H_1\mid H_0)P(H_0)+C_{11}P(H_1\mid H_1)P(H_1) \end{aligned} \tag{2.5.1}$$

令 R_1 為判斷成 H_1 的區域，R_0 為判斷成 H_0 的區域，改寫(2.5.1)式：

$$\begin{aligned} E(C) &= C_{00}P(H_0)\int_{R_0} p(x\mid H_0)\,dx + C_{01}P(H_1)\int_{R_0} p(x\mid H_1)\,dx \\ &\quad + C_{10}P(H_0)\int_{R_1} p(x\mid H_0)\,dx + C_{11}P(H_1)\int_{R_1} p(x\mid H_1)\,dx \end{aligned} \tag{2.5.2}$$

根據機率的理論得知：

$$\int_{R_1} p(x\mid H_i)\,dx = 1-\int_{R_0} p(x\mid H_i)\,dx\ ,\ i\in\{0,1\} \tag{2.5.3}$$

把(2.5.3)式帶入(2.5.2)中整理得到：

$$\begin{aligned}\mathrm{E}(C) &= C_{00}P(\mathrm{H}_0)+C_{01}P(\mathrm{H}_1)\\ &+\int_{\mathrm{R}_1}(C_{10}P(\mathrm{H}_0)-C_{00}P(\mathrm{H}_0))p(x\,|\,\mathrm{H}_0)\,dx\\ &+\int_{\mathrm{R}_1}(C_{11}P(\mathrm{H}_1)-C_{01}P(\mathrm{H}_1))p(x\,|\,\mathrm{H}_1)\,dx\end{aligned} \tag{2.5.4}$$

觀察(2.5.4)式，為了使$\mathrm{E}(C)$最小，可以發現：

$$(C_{10}-C_{00})P(\mathrm{H}_0)p(x\,|\,\mathrm{H}_0)=(C_{11}-C_{01})P(\mathrm{H}_1)p(x\,|\,\mathrm{H}_1) \tag{2.5.5}$$

也就是說當(2.5.6)式成立時，決策為H_1：

$$(C_{10}-C_{00})P(\mathrm{H}_0)p(x\,|\,\mathrm{H}_0)<(C_{11}-C_{01})P(\mathrm{H}_1)p(x\,|\,\mathrm{H}_1) \tag{2.5.6}$$

最後可以化簡為一個簡單的判斷式：

$$\frac{p(x\,|\,\mathrm{H}_1)}{p(x\,|\,\mathrm{H}_0)}>\frac{(C_{10}-C_{00})P(\mathrm{H}_0)}{(C_{01}-C_{11})P(\mathrm{H}_1)}=\gamma \tag{2.5.7}$$

通常假設$C_{10}>C_{00}$，$C_{01}>C_{11}$。

所以在此方法中，只要預先設定好各種決策結果的代價，就可以簡單的判斷兩個可能性函數(Likelihood Function)的比值是否大過γ值來決定為H_0還是H_1。

值得一提，如果設定$C_{00}=C_{11}=0$、$C_{01}=C_{10}=1$，則貝氏風險就會同等於最大化事後機率偵測(Maximum a Posteriori Probability Detector，MAP)，如(2.5.8)式。

$$\frac{p(x\,|\,\mathrm{H}_1)}{p(x\,|\,\mathrm{H}_0)}<\frac{P(\mathrm{H}_0)}{P(\mathrm{H}_1)}=\gamma \tag{2.5.8}$$

如果再加上設定$P(\mathrm{H}_0)=P(\mathrm{H}_1)$，則貝氏風險就會同等於最大化可能性偵測(Maximum Likelihood Detector，ML)，如(2.5.9)式。

$$\frac{p(x\,|\,\mathrm{H}_1)}{p(x\,|\,\mathrm{H}_0)}<1 \tag{2.5.9}$$

2.6 Cascade Detector

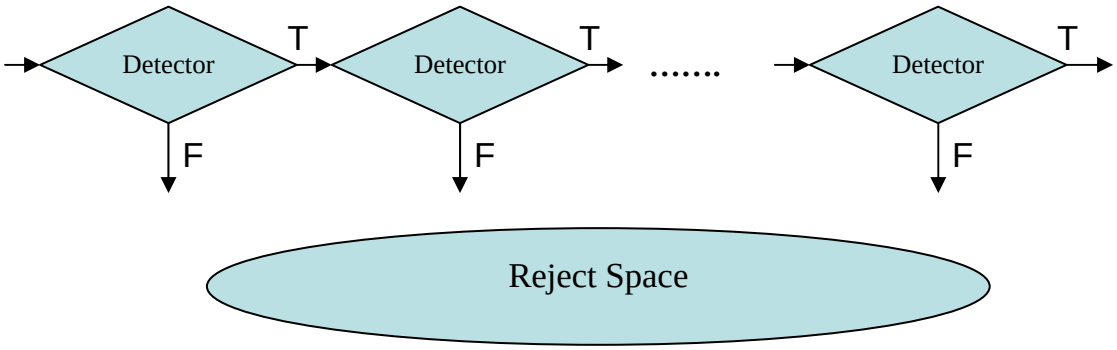


圖 2.17 串聯式偵測器架構

串聯式偵測器的概念，是使用多個弱偵測器(Weak Detector)的串接來完成一個強偵測器(Strong Detector)。此方法的作法是，依據部分特徵互相獨立、互不相關的特性，為各自獨立的特徵空間設計各自的偵測器，並串接起來。

其優點是可以降低只有單一偵測器時的複雜度，並可以專注且獨立的針對各自的特徵空間去簡化設計，使得各階偵測器可以很簡易並有效率的篩選掉錯誤的情況。另外還擁有容易擴充的特性，當發現額外的特徵可以加入判別時，不用考慮到前面原本各階的設計狀況，只需針對目前的特徵作設計。注意到各階都是獨立的，交換排序不會有結果上的分別。

例如在人臉偵測上，第一階可能設定為擷取有膚色的地方，第二階只把輪廓為橢圓形的部份保留，第三階為再去掉沒有偵測到眼睛的地方。可以想像，如果要同時把上述的特徵空間全部混合在一起，並設計出一個偵測器去檢測，則偵測器會相當複雜且難以設計。事實上目前在人臉偵測(Face Detection)中，Adaboost 和 Cascade Detector 的結合有很好的效用[2-11]。

第三章　利用空間域特徵空間一致性及共鳴曲線相似性之喚醒關鍵字偵測演算法

傳統在關鍵字偵測的方法上，為擷取當下語音的頻譜特徵，並對這些特徵作辨識分析判斷是否為關鍵字，簡單來說，即為單一關鍵字的辨識問題。而本論文所提出的方法，除了延續判斷頻譜特徵－共鳴曲線相似性之外，提出了新的特徵－空間性特徵空間一致性，作為關鍵字偵測的依據。

所提出新的方法概念是，在不同的語音，發聲者所講的不同字，各有特定的特徵頻率代表了這個字，並決定了這個字聽起來的樣子。當這個字在環境中發聲的時候，聲源到達方位估測(DOA)會偵測出這些特徵的頻率共同來自於某個方位。反過來思考，藉由預先設定偵測器一個關鍵字(如機器人的名字)，並依照語音特徵的分析，擷取出關鍵字的特徵頻率，就可以透過持續觀測這些特徵頻率是否來自於同一方位，也就是空間性特徵空間的一致性，來判斷這個字有沒有在這個方位發聲。而此想法的空間特徵資訊獲取則必須使用麥克風陣列。

精確來說，先設定一個機器人喚醒關鍵字，搜集大量語料後使用 LPC 預先擷取關鍵字的共鳴曲線與共振峰，並基於貝氏風險理論作為上述兩個獨立偵測器門檻值(Threshold)判別。運作時持續偵測共鳴曲線的相似度是否通過偵測器門檻值，並同時用共振峰作為特徵頻率持續估測其特徵空間的一致性是否通過門檻值，再將兩個獨立的偵測器串接作為整體偵測的機制。

3.1 演算法流程架構

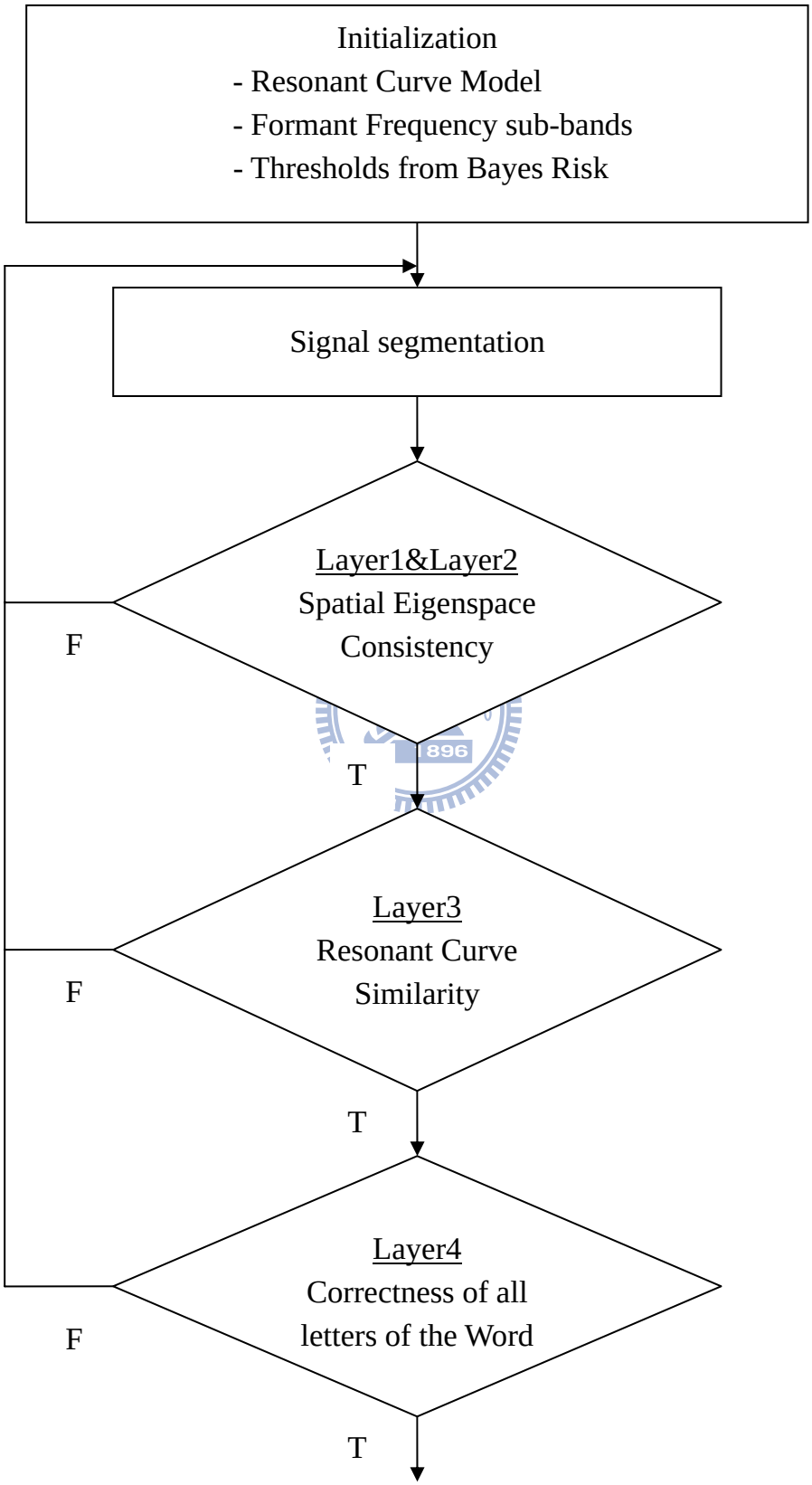


圖 3.1 喚醒關鍵字偵測演算法流程圖

喚醒關鍵字偵測演算法解說：

1. 根據選定好的喚醒關鍵字，蒐集大量的語音樣本，以 10 階的 LPC 估測出每個語音樣本的共鳴曲線，並建立一個共鳴曲線的模型(Resonant Curve Model)。依照共鳴曲線模型的峰值處選定前兩個共振峰的頻帶(Formant Frequency Sub-band)。並依據收集到大量的語音樣本，使用貝氏風險決定所有檢測器會使用到的門檻值(Threshold)。

2. 把語音訊號中每一個字元(Letter)分割出來。目前在本論文中假設語音訊號已經做好分割，先不考慮語音分割的問題。

3. 依照選定的前兩個共振峰頻帶，判斷分割好的字元在特徵頻帶內其空間性特徵空間的一致性是否大於事先依貝氏風險所決定的門檻值。為串聯式偵測器的第一部份。

4. 判斷切割好的字元的共鳴曲線相似度是否大於事先依貝氏風險所決定的門檻值。為串聯式偵測器的第二部份。

5.判斷連三個切割好的字元其空間性特徵空間一致性與共鳴曲線相似性是否都正確，若有一個字元在串聯式偵測器中的任一階不正確，則判斷不為關鍵字。為串聯式偵測器的第三部份。

3.2 空間域特徵空間一致性偵測

回顧章節 2.2，在各個頻率下對麥克風陣列做主要成分分析，得出各個頻率下的空間性特徵空間，包含了訊號子空間與雜訊子空間。回顧章節 2.3，共振峰為語音特徵之一，並且能量較大的諧波會出現共振峰頻率附近處，所以選定共振峰及附近的頻帶為特徵頻帶。

在這章節完整的作法為，在設定完一個機器人的喚醒關鍵字，並搜集大量語料後擷取關鍵字的共鳴曲線與共振峰，對選定的共振峰頻帶下的特徵頻率，持續觀測其估測出的空間性特徵空間特徵的一致性，並基於貝氏風險理論門檻值的判別，判斷目前已經分割出的字有沒有發聲。

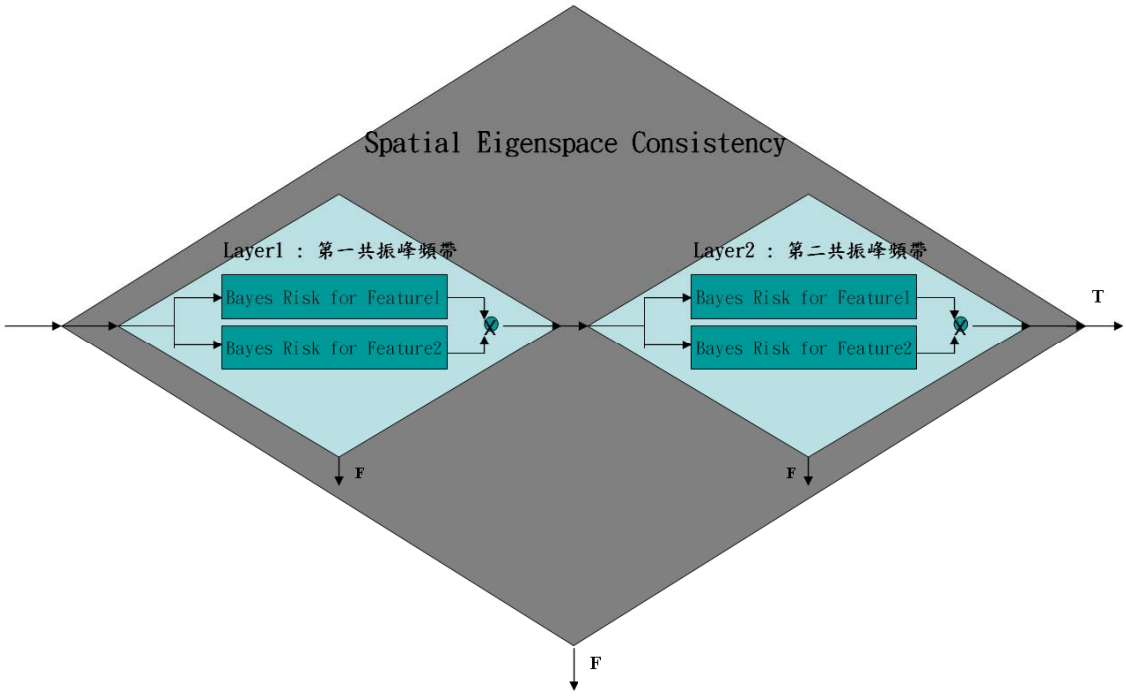


圖 3.2 Spatial Eigenspace Consistency 的整體架構圖

3.2.1 擷取語音特徵

蒐集大量的語音樣本，並使用 LPC 估測出共鳴曲線，再選取第一個及第二個共振峰附近的頻帶，如圖 3.3。紅色線為共鳴曲線的模型，藍色線為能量頻譜，黑色線為取出的共振峰頻帶。

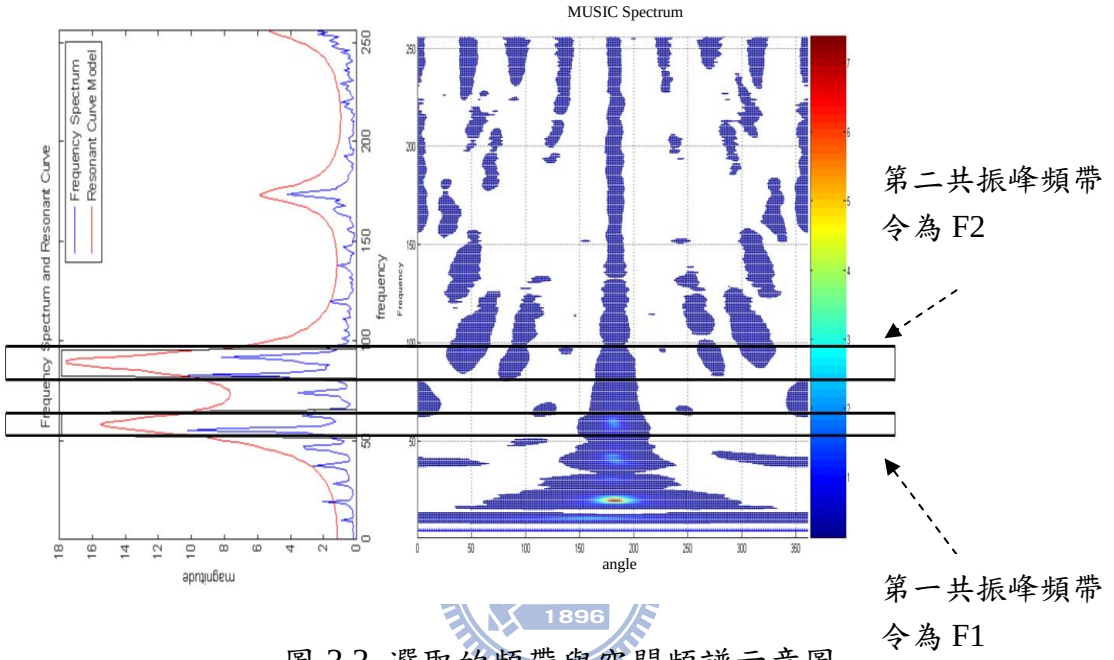


圖 3.3 選取的頻帶與空間頻譜示意圖

從圖 3.4 可以清楚看到，如果是關鍵字，則共振峰頻帶會選擇到能量較高的諧波成分，所以頻帶內的 Spatial Eigenspace 將比較有一致性，因為空間資訊會保持在訊噪比較高的頻帶上。

關鍵字 **非關鍵字**

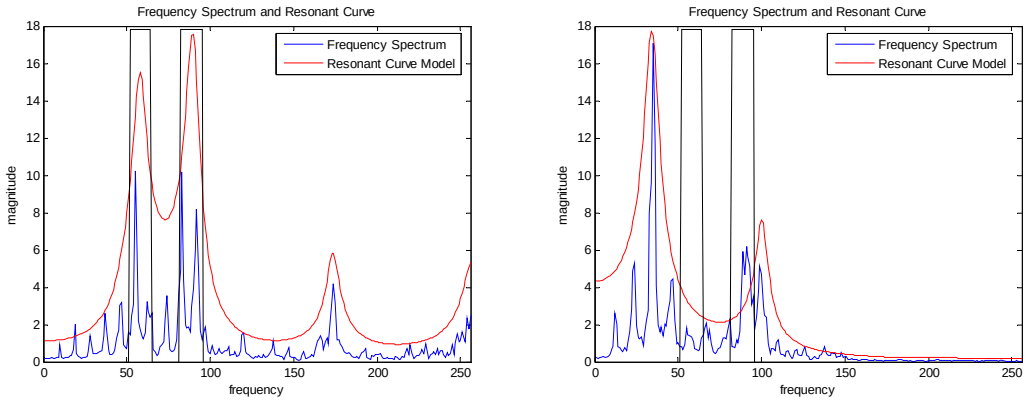


圖 3.4 兩個字的能量頻譜與共振曲線

3.2.2 建立空間域特徵空間一致性的特徵參數

把第一共振峰頻帶中的雜訊子空間取出，並建立一個投影到雜訊子空間的投影矩陣 $\mathbf{P}_N$ ，參考第 2.2 節：

$$\mathbf{P}_N(\omega_f) = \sum_{i=D+1}^{M} \boldsymbol{V}(\omega_f)_i \boldsymbol{V}(\omega_f)_i^H = \mathbf{P}_A^{\perp}(\omega_f)\,,\ f \in F1 \tag{3.2.1}$$

計算空間頻譜：

$$\boldsymbol{S}_{MUSIC}(\theta,\omega_f) = \frac{1}{\boldsymbol{a}^H(\theta,\omega_f)\mathbf{P}_N(\omega_f)\boldsymbol{a}(\theta,\omega_f)}\,,\ f \in F1 \tag{3.2.2}$$

則角度的估測為：

$$\hat{\theta}(\omega_f) = \arg\max_{\theta} \boldsymbol{S}_{MUSIC}(\theta,\omega_f)\,,\ f \in F1 \tag{3.2.3}$$

本論文對空間性特徵空間一致性(Spatial Eigenspace Consistency)建立兩個特徵(第二共振峰頻帶作法也相同)：

1. 角度估測量值

$$x_1 = \max_{\theta}\left(\frac{\sum_{f\in F1} \boldsymbol{S}_{MUSIC}(\theta,\omega_f)}{D}\right),\ D\text{:normailed factor} \tag{3.2.5}$$

2. 角度估測變異數

$$x_2 = \mathrm{var}(\hat{\theta}(\omega_f))\,,\ f \in F1 \tag{3.2.4}$$

觀察這兩個特徵的特性，當特徵空間比較有一致性的時候，角度估測量值會比較大，相對的角度估測變異數會比較小。下圖 3.5 為第一個特徵－角度估測量值不同字的實際結果。圖 3.6 為第二個特徵－角度估測變異數不同字的實際結果。

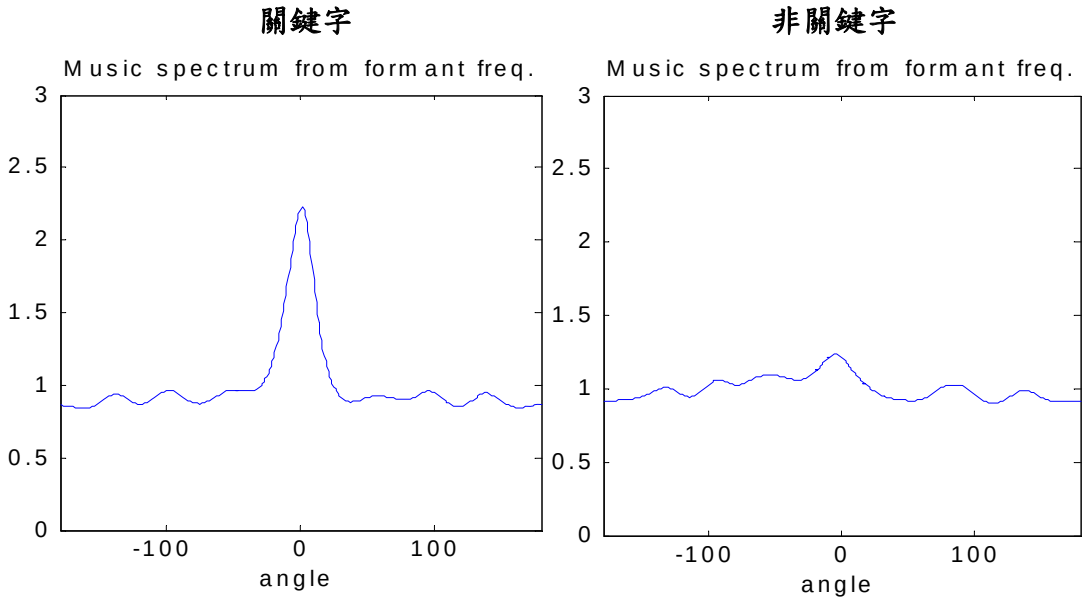


圖 3.5 $\frac{\sum_{f \in F1} S_{MUSIC}(\theta, \omega_f)}{D}$，圖形峰值處即為角度估測量值

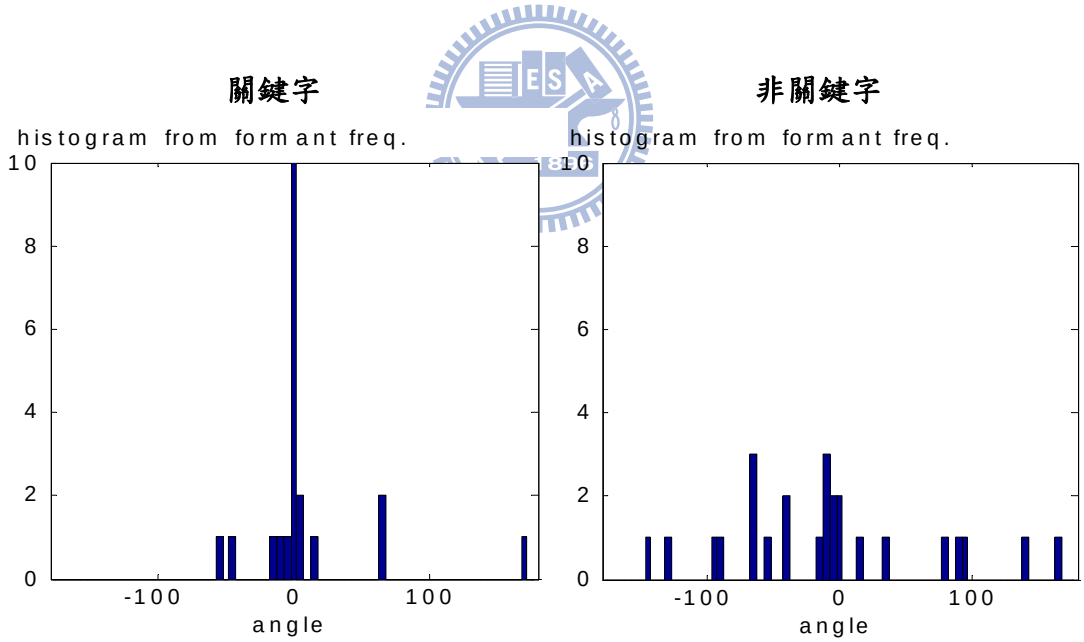


圖 3.6 $\hat{\theta}(\omega_f)$ 的統計圖，離散程度即為角度估測變異數

圖 3.5 與圖 3.6 驗證出特徵頻率其空間性特徵空間一致性，的確可以當作區別不同字的特徵之一。偵測字為關鍵字時，角度估測量值比較大，角度估測變異數比較小。

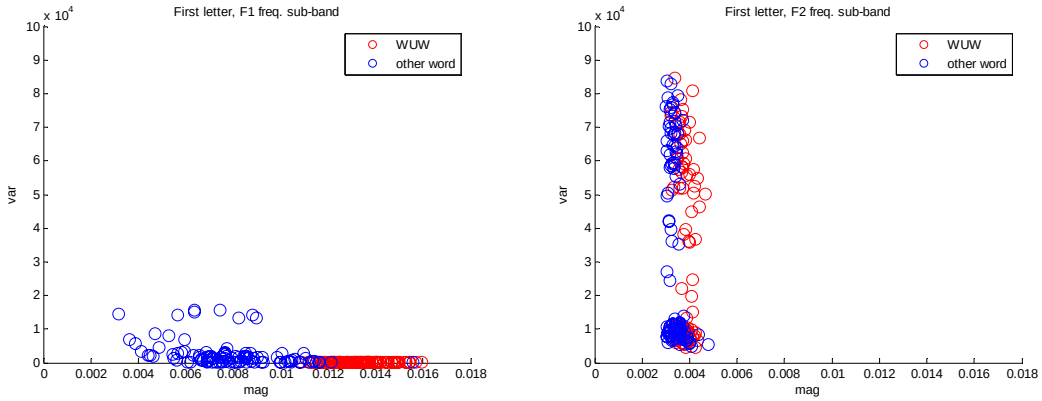


圖 3.7 語音特徵分布圖

圖 3.7 為第一共振峰及第二共振峰，累積多筆資料後對兩個特徵所做出的散佈圖(紅色為關鍵字、藍色為非關鍵字)。第一個觀察發現偵測字為關鍵字時，角度估測量值比較大或角度估測變異數比較小。第二個觀察發現兩個特徵算是低相關性(low correlation)，因而可以各自設計各自的偵測器，再串接起來。

3.3 共鳴曲線相似度偵測

回顧章節 2.3，共鳴曲線為語音完整的特徵。先前的論文步驟建立好了關鍵字的共鳴曲線模型，在這章節要計算目前的偵測字其共鳴曲線和關鍵字的共鳴曲線的相似度，並基於貝氏風險理論門檻值的判別，判斷關鍵字有沒有發聲。如圖 3.8，比較兩條不同共鳴曲線的相似度(紅色為共鳴曲線、藍色為能量頻譜)。

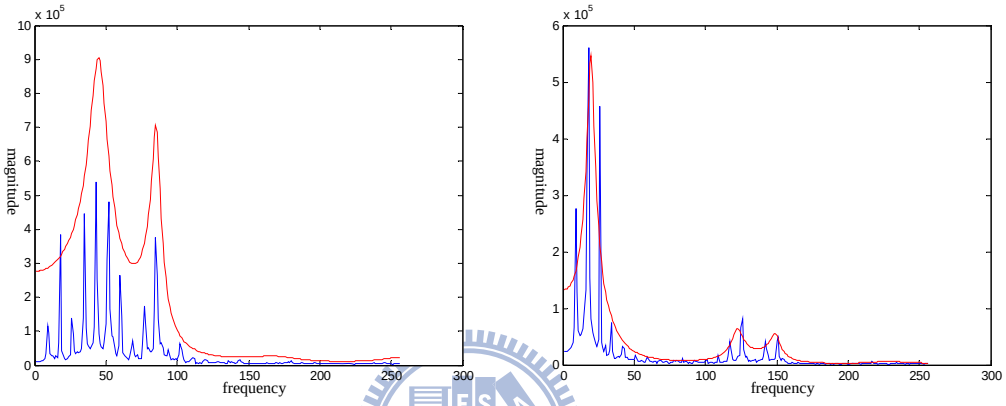


圖 3.8　兩條不同字的共鳴曲線

此論文所用的相似度定義為Bhattacharyya Distance[3-1]

$$BC(p,q)=\int\sqrt{p(x)q(x)}\,dx \tag{3.3.1}$$

圖 3.9 為累積多筆資料後對兩個特徵所做出的直方圖(紅色為關鍵字、藍色為非關鍵字)。

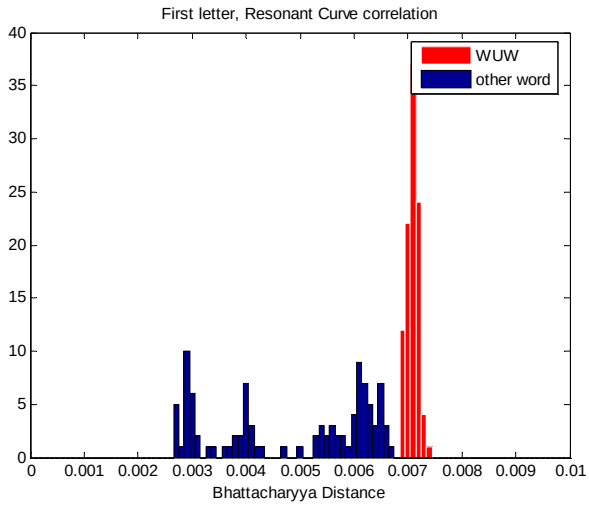


圖 3.9 語音特徵分布圖

第四章　實驗結果與分析

本章節將介紹本論文的方法於同一發聲者與不同發聲者在不同 SNR 下的測試結果，並分析不同的 SNR 下特徵的變化及對於門檻值的設定，最後測試固定門檻值下不同 SNR 的整體測試結果。

圖 4.1 為錄音環境的實際照片，環境中有一個喇叭作為人聲的播放，並於麥克風陣列平台的不同方位(0 度、90 度、180、270 度)及距離(2M、4M)錄音，如圖 4.3。另外語料庫依照同一發聲者與不同發聲者分為兩部份，細節如表 4.1 及表 4.2。不同 SNR 的訊號則用雜訊去合成，雜訊為 Babble noise，以喇叭播放並放置在錄音環境的 4 個角落以建立出 Diffuse noise。SNR 的估算由於語料庫字數眾多，只能大概估出部分字的 segmental SNR，所以約有 3dB 左右的誤差。

圖 4.1 錄音環境實際照片

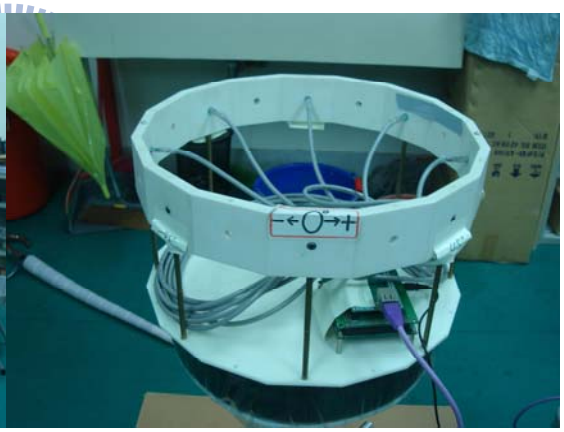


圖 4.2 麥克風陣列平台

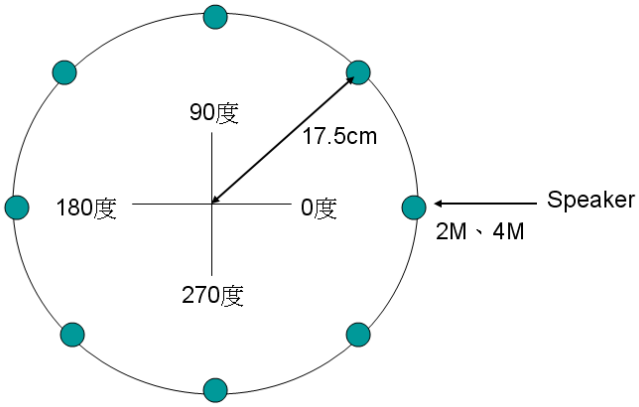


圖 4.3 錄音環境平面關係的俯視圖

關鍵字	‘阿凡達’	**發聲者人數**	1 人
陣列架構	環型麥克風陣列	**發聲者角度**	0 度
陣列半徑	17.5cm	**發聲者距離**	2M
取樣頻率	8K	**關鍵字次數**	200 次
音框大小	512	**關鍵字總數**	200 個字
Overlap	256	**非關鍵字數目**	20 個字
		非關鍵字次數	10 次
		非關鍵字總數	200 個字
		總數	400 個字

表 4.1 同一發聲者語料庫

關鍵字	‘阿凡達’	**發聲者人數**	8 人
陣列架構	環型麥克風陣列	**發聲者角度**	0 度、90 度、180 、270 度
陣列半徑	17.5cm	**發聲者距離**	2M、4M
取樣頻率	8K	**關鍵字次數**	5 次
音框大小	512	**關鍵字總數**	320 個字
Overlap	256	**非關鍵字數目**	20 個字
		非關鍵字次數	1 次
		非關鍵字總數	1280 個字
		總數	1600 個字

表 4.2 不同發聲者語料庫

本實驗指標參數的定義如下：

	Wake-Up-Word	**Other Word**
Decided as Wake-Up-Word	True Positive	False Positive
Decide Other Word	False Negative	True Negative

表 4.3 本實驗指標參數的定義表

1.偵測率(Detection Rate)：True Positive/(True Positive+ False Negative)

在給定關鍵字下，被判斷為關鍵字的機率。

2.誤報率(False Positive Rate)：False Positive/(False Positive+ True Negative)

在給定不是關鍵字下，被誤判為關鍵字的機率。

3.漏報率(False Negative Rate)：1- Detection Rate

在給定關鍵字下，被誤判為不是關鍵字的機率。

4.等錯誤率(Equal Error Rate)：False Negative Rate = False Positive Rate

調整偵測器的參數，使得誤報率剛好等於漏報率時的機率，作為測量偵測器整體效能的參數。

4.1 同一發聲者之實驗結果與分析

在此章節中實驗結果與分析分為三大項，總共有六個測試。

第一項：本論文採用串聯式偵測器，設計重點需測試 Detection Rate 為 100% 下最低的 False Positive Rate，作為表示以此論文方法架構為前級的檢測器可以初步篩選掉多少錯誤的情況。如實驗 A。

第二項：分析不同 SNR 下本論文提出方法的差異。如實驗 B、C、D、E。

第三項：在實際的聲場環境，使用者音量與環境中的吵雜程度是不可預知的，因而需要測試在同一組門檻值下，整體條件的辨識率。如實驗 F。

總共六個測試如以下：

A. 在偵測率 100%下測試最低的 False Positive Rate

B. 各個 SNR 下的 Equal Error Rate

C. 各個 SNR 下字元與字組的分析

D. 不同 SNR 下特徵的分析

E. 不同 SNR 下門檻值的分析

F. 固定門檻值下不同 SNR 的整體測試結果

4.1.A 在偵測率 100%下測試最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	0	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0%

表 4.4 SNR=14.15 時偵測率為 100%時最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	0	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0%

表 4.5 SNR=7.3 時偵測率為 100%時最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	0	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0%

表 4.6 SNR=-0.3 時偵測率為 100%時最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	0	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0%

表 4.7 SNR=-2.24 時偵測率為 100%時最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	0	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0%

表 4.8 SNR=-3.82 時偵測率為 100%時最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	0	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0%

表 4.9 SNR=-6.32 時偵測率為 100%時最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	1	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0.5%

表 4.10 SNR=-8.26 時偵測率為 100%時最低的 False Positive Rate

	WUW (200)	Other word (200)	
判斷為 WUW	200	12	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 6%

表 4.11 SNR=-11.78 時偵測率為 100%時最低的 False Positive Rate

整體測試結果如表 4.12，可以看到在同一個發聲者的偵測結果很好，因為同一人的共鳴曲線與共振峰的位置在正常狀況下變動不大，因而會有不錯的偵測結果。

SNR	Detection rate	False positive rate
14.15 dB	100%	0 %
7.3 dB	100%	0 %
-0.3 dB	100%	0 %
-2.24 dB	100%	0 %
-3.82 dB	100%	0 %
-6.32 dB	100%	0 %
-8.26 dB	100%	0.5 %
-11.78 dB	100%	6 %

表 4.12 各個 SNR 下的 False positive rate

4.1.B 各個 SNR 下的 Equal Error Rate

SNR	EER	1- EER
14.15 dB	0 %	100 %
7.3 dB	0 %	100 %
-0.3 dB	0 %	100 %
-2.24 dB	0 %	100 %
-3.82 dB	0 %	100 %
-6.32 dB	0 %	100 %
-8.26 dB	0.5 %	99.5 %
-11.78 dB	5 %	95 %

表 4.13 各個 SNR 下的 EER

4.1.C 各個 SNR 下字元與字組的分析

圖 4.4 與圖 4.5 為在乾淨的語音時(SNR=14.15)，各個字元(letter)與最後字組(word)的偵測結果。前面 200 筆為關鍵字，後面 200 筆為非關鍵字。

字元偵測結果為'1'代表偵測器判斷此字元是正確的字元，也就是通過了本論文中的 Layer1、Layer2、Layer3 與 Layer4。可以注意到第一個字元的索引 201~230 和索引 241~260 也為'阿'，所以都會通過偵測器。

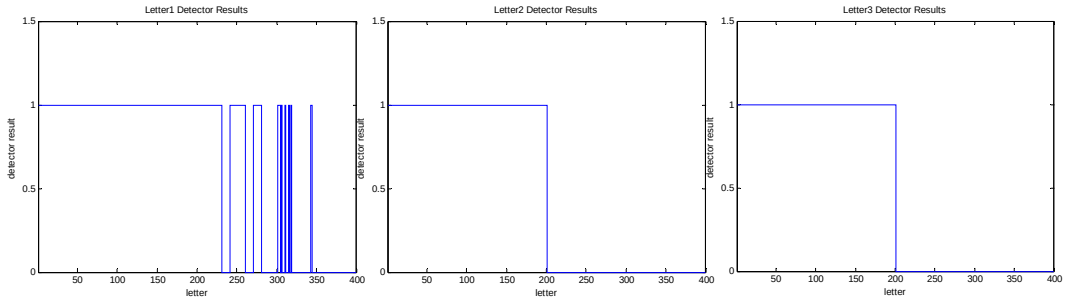


圖 4.4　三個字元各自的偵測結果(SNR=14.15 dB)

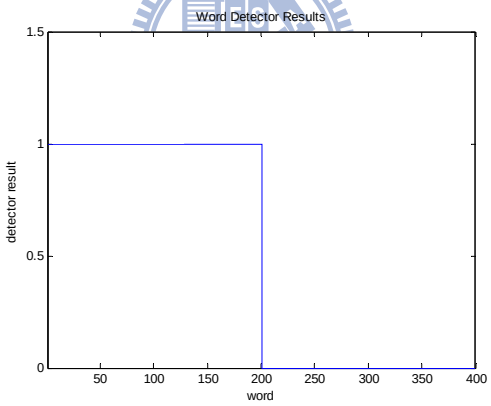


圖 4.5　字組的偵測結果(SNR=14.15 dB)

	WUW (200)	**Other word (200)**	
判斷為 WUW	200	0	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 0%

表 4.14 SNR=14.15 時偵測率為 100%時最低的 False Positive Rate

圖 4.6 與圖 4.7 為在非常惡劣環境下的語音(SNR=-11.78)時，時各個字元和最後字組的偵測結果。從結果上可以看到，一個字元的偵測常常會出現誤判，把不是關鍵字的字元判斷成關鍵字的字元。但是經由其他字元的篩選，可以得出不錯的結果。

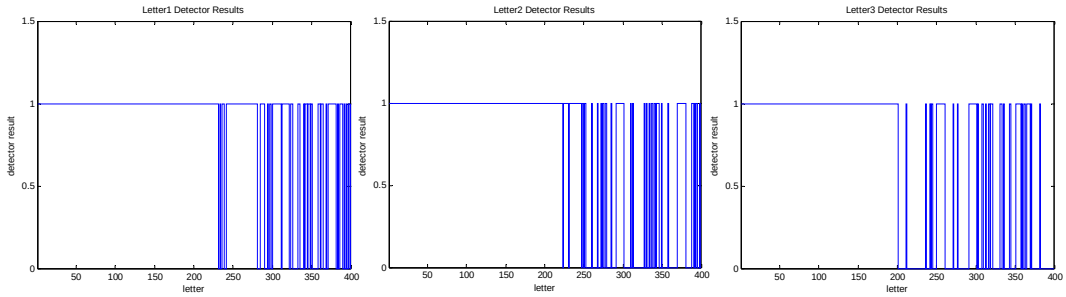


圖 4.6 三個字元各自的偵測結果(SNR=-11.78)

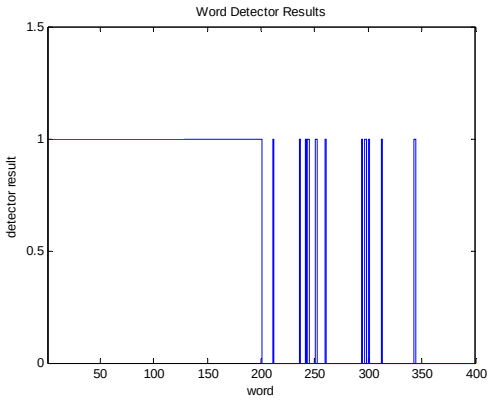


圖 4.7 字組的偵測結果(SNR=-11.78)

	WUW (200)	**Other word (200)**	
判斷為 WUW	200	12	Detection rate = 100%
判斷不為 WUW	0	200	False positive rate = 6%

表 4.15 SNR=-11.78 時偵測率為 100%時最低的 False Positive Rate

4.1.D 不同 SNR 下特徵的分析

在這節的分析中，選定某個字元，觀察在最高的 SNR 與最低的 SNR 下，Layer1、Layer2、Layer3 的特徵差異。

先觀察 Layer1 及 Layer2，此階層是判斷空間性特徵空間一致性。隨著 SNR 的下降，角度估測變異數(variance)會變大且角度估測量值(magnitude)會減少，也就是空間性的特徵空間一致性降低，所以特徵的分布也會跟著重疊起來。但是還是可以看到正確的字有比較小的變異數和比較大的量值，如圖 4.8 與圖 4.9。另外從特徵的分布也可以看到 Layer1 及 Layer2 的兩個特徵算是低相關性(low correlation)，因而可以各自設計各自的偵測器，效用就像兩條直線分割出的二維分類器。

SNR = 14.15 **SNR = -11.78**

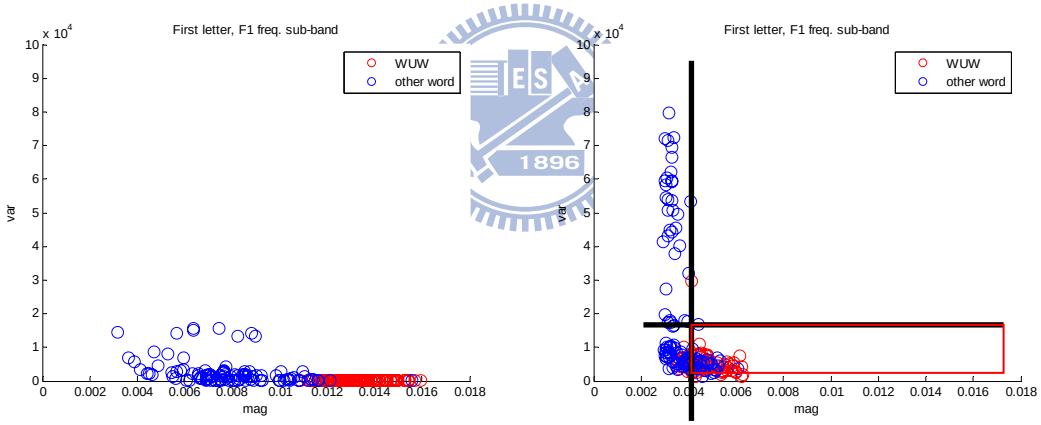


圖 4.8 Layer1 語音特徵分布圖

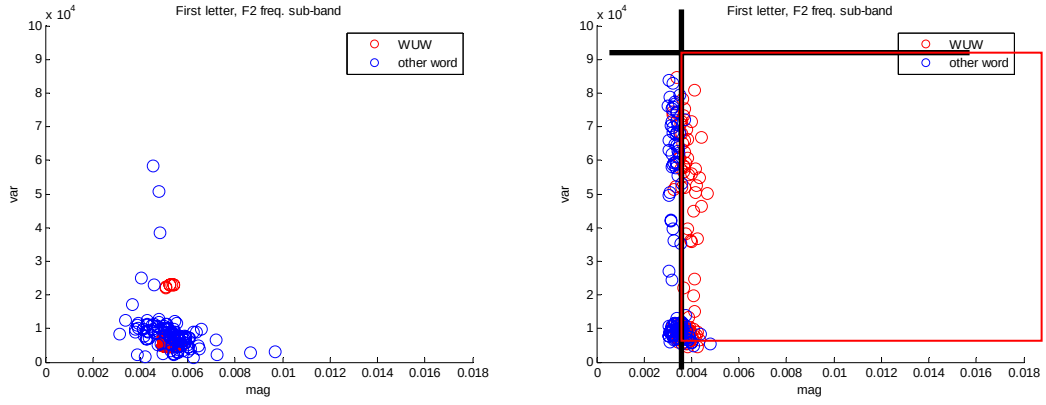


圖 4.9 Layer2 語音特徵分布圖

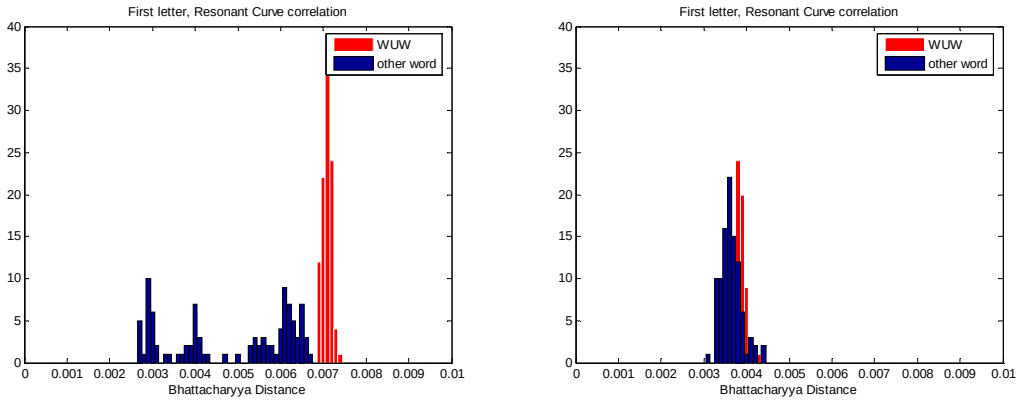


圖 4.10 Layer3 語音特徵分布圖

再觀察 Layer3，此階層是判斷共鳴曲線的相似度，相當於語音辨識的標準作法。可以預期到當 SNR 很高時，特徵會分離的很明顯，因為頻譜很乾淨未受到汙染，比對共鳴曲線相似度時效果會不錯。但是當 SNR 下降時，頻譜因為雜訊而失真，特徵也因此重疊在一起，變成主要都是靠 Layer1 和 Layer2 去篩選，這也是擷取空間資訊比較抗雜的優點之一。

4.1.E 不同 SNR 下門檻值的分析

在這節分析中，選定某個字元，觀察在 8 個不同的 SNR 下，偵測率剛好為 100%時，階層分類器的 Layer1、Layer2、Layer3 總共五個門檻值的差異。

下圖 4.11 從左到右為五張圖為:

1. Layer1 角度估測量值(magnitude)
2. Layer1 角度估測變異數(variance)
3. Layer2 角度估測量值(magnitude)
4. Layer2 角度估測變異數(variance)
5. Layer3 共鳴曲線相似度(similatiry)

八個 SNR 由左到右為:

[14.15(dB) 7.3(dB) -0.3(dB) -2.24(dB) -3.82(dB) -6.32(dB) -8.26(dB)]

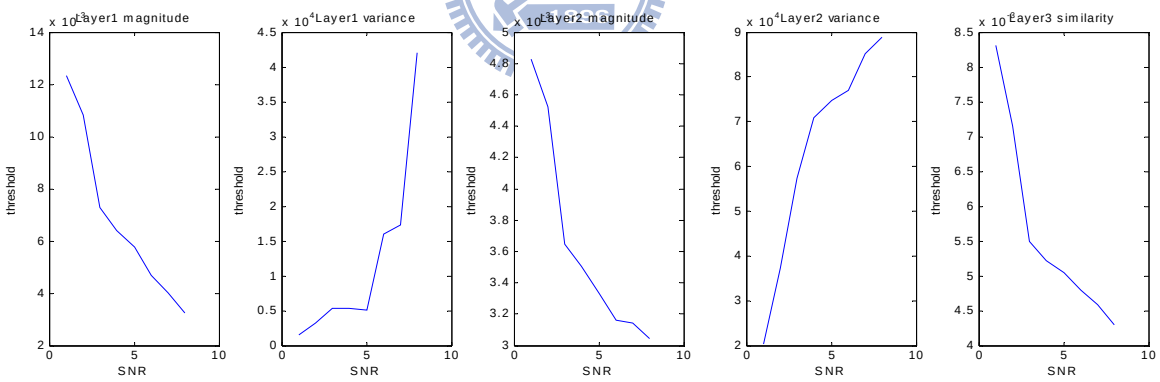


圖 4.11 不同 SNR 下使偵測率剛好為 100%時門檻值的變化

觀察若要維持剛好偵測率為 100%，則不同的 SNR 下的門檻值將不同，角度估測量值的門檻值會因為 SNR 減小而減小，角度估測變異數的門檻值會隨著 SNR 減小而增加，而共鳴曲線相似度的門檻值也會因為 SNR 減小而減小。這是很合理的變化。

4.1.F 固定門檻值下不同 SNR 的整體測試結果

在這節分析中，把不同 SNR 下的資料混合，測試在發聲者與環境雜訊變動很大時整體偵測的效果，並測試其 Equal Error Rate。8 個 SNR 與其資料數如表 4.16。

觀察表 4.17，若不去考慮過低 SNR 的狀況下，如 SNR1~SNR5，設定一組門檻值是可行的，共有 2000 筆資料，只會有 54 筆被誤判。但是考慮最差的情況，雖然最低的 EER 仍有 9.06%，但其 3200 筆資料約有 290 筆被誤判，其中大部分都是 SNR8 的狀況，所以若要在非常低的 SNR 下保有辨識率，則必須為特別低 SNR 的狀況建立一組門檻值。

	SNR	資料數
SNR1	14.15 dB	400 筆
SNR2	7.3 dB	400 筆
SNR3	-0.3 dB	400 筆
SNR4	-2.24 dB	400 筆
SNR5	-3.82 dB	400 筆
SNR6	-6.32 dB	400 筆
SNR7	-8.26 dB	400 筆
SNR8	-11.78 dB	400 筆

表 4.16 8 個不同 SNR 與其資料數

	EER	1-EER
只看 SNR1~SNR5　共 2000 筆資料	2.7 %	97.30 %
只看 SNR1~SNR6　共 2400 筆資料	6.42 %	93.58 %
只看 SNR1~SNR7　共 2800 筆資料	6.79 %	93.21 %
只看 SNR1~SNR8　共 3200 筆資料	9.06 %	90.94 %

表 4.17 固定門檻值下不同 SNR 的整體測試結果

4.2 不同發聲者之實驗結果與分析

在此章節中實驗結果與分析分為三大項，總共有五個測試。

第一項：本論文採用串聯式偵測器，設計重點需測試 Detection Rate 為 100% 下最低的 False Positive Rate，作為表示以此論文方法架構為前級的檢測器可以初步篩選掉多少錯誤的情況。如實驗 A。

第二項：分析不同 SNR 下本論文提出方法的差異。如實驗 B、C、D。

第三項：在實際的聲場環境，使用者音量與環境中的吵雜程度是不可預知的，因而需要測試在同一組門檻值下，整體條件的辨識率。如實驗 E。

總共五個測試如以下：

A. 在偵測率 100%下測試最低的 False Positive Rate

B. 各個 SNR 下的 Equal Error Rate

C. 各個 SNR 下字元與字組的分析

D. 不同 SNR 下門檻值的分析

E. 固定門檻值下不同 SNR 的整體測試結果

4.2.A 在偵測率 100%下測試最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	68	Detection rate = 100%
判斷不為 WUW	0	1212	False positive rate = 5.31%

表 4.18 SNR=14.15 時偵測率為 100%時最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	67	Detection rate = 100%
判斷不為 WUW	0	1213	False positive rate = 5.23%

表 4.19 SNR=7.3 時偵測率為 100%時最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	88	Detection rate = 100%
判斷不為 WUW	0	1192	False positive rate = 6.87%

表 4.20 SNR=-0.3 時偵測率為 100%時最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	106	Detection rate = 100%
判斷不為 WUW	0	1174	False positive rate = 8.28%

表 4.21 SNR=-2.24 時偵測率為 100%時最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	132	Detection rate = 100%
判斷不為 WUW	0	1148	False positive rate=10.31%

表 4.22 SNR=-3.82 時偵測率為 100%時最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	256	Detection rate = 100%
判斷不為 WUW	0	1024	False positive rate = 20%

表 4.23 SNR=-6.32 時偵測率為 100%時最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	432	Detection rate = 100%
判斷不為 WUW	0	848	False positive rate=33.75%

表 4.24 SNR=-8.26 時偵測率為 100%時最低的 False Positive Rate

	WUW (320)	Other word (1280)	
判斷為 WUW	320	929	Detection rate = 100%
判斷不為 WUW	0	351	False positive rate=72.58%

表 4.25 SNR=-11.78 時偵測率為 100%時最低的 False Positive Rate

整體測試結果如表 4.26，可以看到不同發聲者的效果相對於同一發聲者的效果降低不少，尤其在低 SNR 的狀況下。原因是同一人的共鳴曲線與共振峰的位置在正常狀況下變動不大，而不同人的差異相對來說就比較大。

SNR	Detection rate	False positive rate
14.15 dB	100 %	5.31 %
7.3 dB	100 %	5.23 %
-0.3 dB	100 %	6.87 %
-2.24 dB	100 %	8.28 %
-3.82 dB	100 %	10.31 %
-6.32 dB	100 %	20 %
-8.26 dB	100 %	33.75 %
-11.78 dB	100 %	72.58 %

表 4.26 各個 SNR 下的 False positive rate

4.2.B 各個 SNR 下的 Equal Error Rate

SNR	EER	1- EER
14.15 dB	5.08 %	94.92 %
7.3 dB	5 %	95.00 %
-0.3 dB	5.28 %	94.72 %
-2.24 dB	7.11 %	92.89 %
-3.82 dB	9.44 %	90.56 %
-6.32 dB	11.1 %	88.90 %
-8.26 dB	13.9 %	86.10 %
-11.78 dB	21.63 %	78.37 %

表 4.27 各個 SNR 下的 EER

4.2.C 各個 SNR 下字元與字組的分析

圖 4.12 左圖為在乾淨的語音時(SNR=14.15)，字組偵測的結果。前面 320 筆為關鍵字，後面 1280 筆為非關鍵字。字組偵測結果為'1'代表偵測器判斷此字組是關鍵字。

觀察被誤判成關鍵字的字組可以發現，大部分的字組都是'阿拉拉'。原因是'阿拉拉'和'阿凡達'三個字中母音的共鳴曲線都相當相似，這是只針對母音所設計的偵測器會有缺陷的地方。

令'阿拉拉'的結果都為 0，觀察除了此字組外其他被誤判為關鍵字的有多少，如圖 4.12 右圖與表 4.28。可以看到原本 68 個 false positives 中有 62 個為'阿拉拉'，也就是總共 64 個'阿拉拉'有 62 個被判斷成關鍵字。

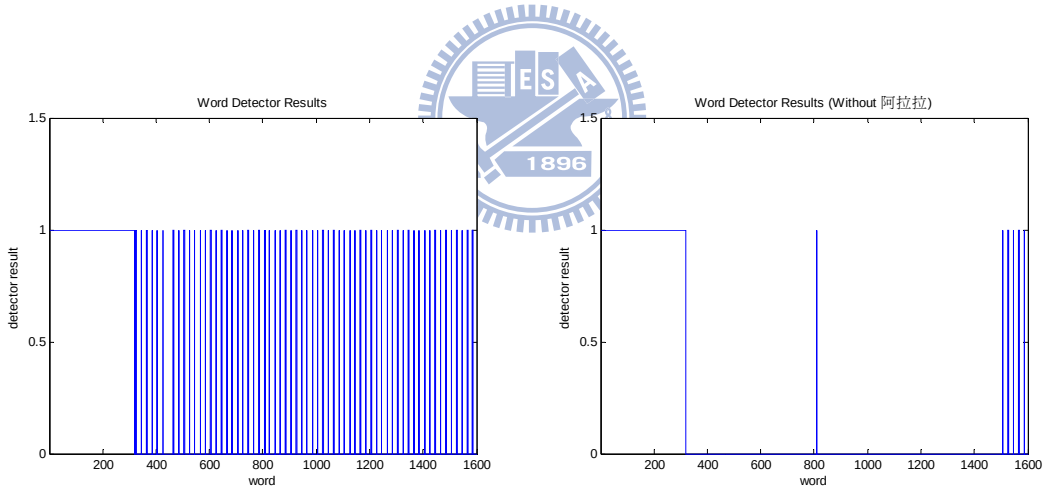


圖 4.12 排除'阿拉拉'前後字組的偵測結果(SNR=14.15 dB)

	Other word	Number of Force positive
去掉'阿拉拉' 前	1280	68
去掉'阿拉拉' 後	1216	6

表 4.28 排除'阿拉拉'前後 False Positive 的比較

在正常的講話速度下，一個字長度通常不超過 200ms，依照不同人與對話當時的情況而異。而在取樣頻率=8k、音框大小=512、overlap=256 下，一個字大約只有 5 個音框。

特徵空間(Eigenspace)需要多個音框來估測，所以沒辦法偵測到稍縱即逝的子音，因而在空間性特徵空間的一致性中變成一種取捨，若要特徵空間估測的準確則無法偵測到子音的變化。

不過在共鳴曲線相似性中是可以克服的。所以將來可能解決的方法為修正本論文中的 Layer3，或直接以語音辨識的作法替代 Layer3，換句話說就是用 Layer1、Layer2 空間性特徵空間的一致性來當作語音辨識器的前級，初步篩選掉錯誤的狀況。

4.2.D 不同 SNR 下門檻值的分析

在這節分析中，選定某個字元，觀察在 8 個不同的 SNR 下，偵測率剛好為 100%時，階層分類器的 Layer1、Layer2、Layer3 總共五個門檻值的差異。

下圖 4.13 從左到右為五張圖為:

1. Layer1 角度估測量值(magnitude)
2. Layer1 角度估測變異數(variance)
3. Layer2 角度估測量值(magnitude)
4. Layer2 角度估測變異數(variance)
5. Layer3 共鳴曲線相似度(similatiry)

八個 SNR 由左到右為:

[14.15(dB) 7.3(dB) -0.3(dB) -2.24(dB) -3.82(dB) -6.32(dB) -8.26(dB)]

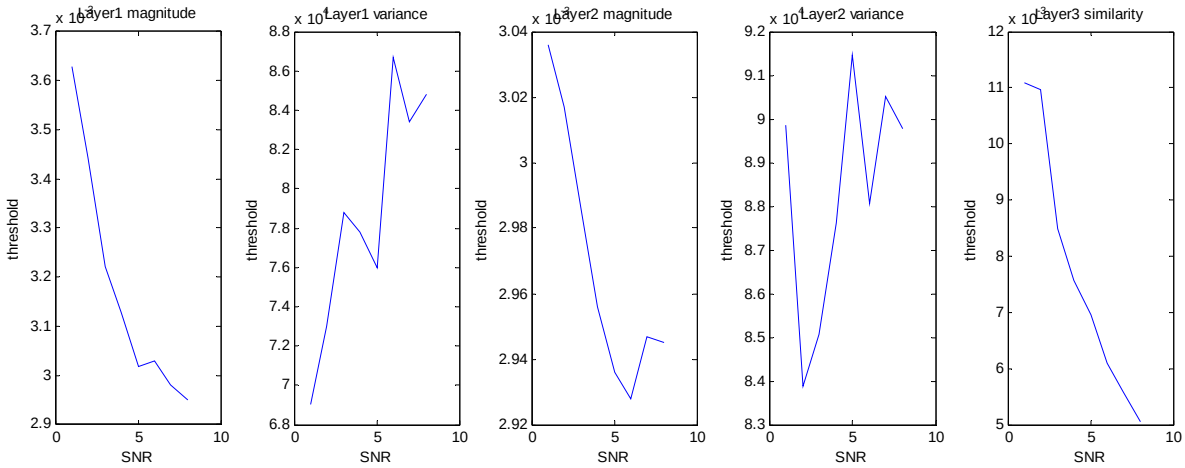


圖 4.13 不同 SNR 下使偵測率剛好為 100%時門檻值的變化

觀察若要維持剛好偵測率為 100%，則不同的 SNR 下的門檻值將不同，角度估測量值的門檻值大體來說會因為 SNR 減小而減小，角度估測變異數的門檻值會隨著 SNR 減小而增加，而共鳴曲線相似度的門檻值也會因為 SNR 減小而減小。這是很合理的變化。

但是相對於同一發聲者，不同發聲者門檻值變化變得比較不規則。在 Layer2 的門檻值就不是單調的特性，尤其是其中變異數門檻值的變化，這樣的結果代表第二個共振峰附近頻帶角度估測不是很穩定，有可能是不同人的第二個共振峰位置變異較大。

4.2.E 固定門檻值下不同 SNR 的整體測試結果

在這節分析中，把不同 SNR 的資料混合，測試在發聲者與環境雜訊變動很大時整體偵測的效果如何，並測試其 Equal Error Rate。8 個 SNR 與其資料數如表 4.29。

觀察表 4.30，跟同一發聲者的情況一樣，雖然最低的 EER 只有 13.2%，但其 12800 筆資料約有 1689 筆被誤判，其中大部分都是 SNR8 的狀況；而在 SNR1~SNR7 中 11200 筆約有 1440 筆被誤判，大部份也是 SNR7 的狀況。所以若要在各個 SNR 下保有辨識率，為不同 SNR 的狀況建立一組門檻值將會有效提升辨識率，但代價則是需要估測環境的 SNR。

	SNR	**資料數**
SNR1	14.15 dB	1600 筆
SNR2	7.3 dB	1600 筆
SNR3	-0.3 dB	1600 筆
SNR4	-2.24 dB	1600 筆
SNR5	-3.82 dB	1600 筆
SNR6	-6.32 dB	1600 筆
SNR7	-8.26 dB	1600 筆
SNR8	-11.78 dB	1600 筆

表 4.29 8 個不同 SNR 與其資料數

	EER	**1-EER**
只看 SNR1~SNR5　共 8000 筆 data	8.75 %	91.25 %
只看 SNR1~SNR6　共 9600 筆 data	9.69 %	90.31 %
只看 SNR1~SNR7　共 11200 筆 data	12.86 %	87.14 %
只看 SNR1~SNR8　共 12800 筆 data	13.2 %	86.80 %

表 4.30 固定門檻值下不同 SNR 的整體測試結果

第五章　結論

5.1 研究成果

本論文提出了一個在低訊噪比下偵測關鍵字的方法。利用關鍵字中特徵頻率其空間性特徵空間一致性並與共鳴曲線相似度來作篩選，使得可以在低訊噪比(SNR)下有相當的辨識率，因而可以適用在遠距關鍵字語音偵測或者在嘈雜的環境下作為關鍵字語音喚醒機制，並且還能同時估測出關鍵字的聲源方向。基於貝氏風險(Bayes Risk)理論的門檻值(Threshold)判別以及利用串接多個偵測器的組合，使本方法得以在極低的訊噪比之下仍保有非常高的辨識率。此方法在低訊噪比下有相當的辨識率，因而可以適用在遠距關鍵字喚醒或者在惡劣的環境下作為關鍵字喚醒機制。

經大量的語料測試，本方法在同一發聲者可以在-11.78dB 的訊噪比之下達成 100%的 detection rate 以及 6%的 false positive rate 錯誤偵測率。在不同發聲者可以在-3.82dB 的訊噪比之下達成 100%的 detection rate 以及 10.32%的 false positive rate。同時，本研究串聯式偵測器保有串接其他偵測器的能力，在有額外的語音特徵或空間特徵可以加入時，能夠簡易的設計新的偵測器，串接到原本的架構上以持續增進辨識率。

5.2 未來展望

在空間性特徵空間一致性中，除了角度估測量值與角度估測變異數兩個特徵外，也許可以再找到其他的特徵以展開偵測器的維度以增加辨識率。而在共鳴曲線相似性中可以套用語音辨識器的模型，改善目前方法只能區別出母音的缺點。在 4.2.E 實驗中提到，若建立估測環境的訊噪比的機制，並為不同訊噪比的狀況各自建立一組門檻值將會有效提升辨識率。另外本論文目前假設訊號已經被分割好，之後在實作上必須要先完成訊號分割的機制。

Reference

[1-1] J. Makhoul, "*Linear prediction: A tutorial review*", Proceedings of the IEEE, 1975.

[1-2] Jyh-Shing Roger Jang , *Audio Signal Processing and Recognition*, class note.

[1-3] H. Hermansky, "*Perceptual linear predictive (PLP) analysis of speech*", Journal of the Acoustical Society of America, 1990.

[1-4] B.H. Juang, L.R. Rabiner, *"Hidden Markov models for speech recognition*", Technometrics, 1991.

[1-5] L.R. Rabiner, "*A tutorial on hidden Markov models and selected applications inspeech recognition*", Proceedings of the IEEE, 1989.

[1-6] Y. Gong, "*Speech recognition in noisy environments: A survey*", Speech communication, 1995.

[1-7] P.J. Moreno, "*Speech recognition in noisy environments*", Doctoral thesis, Carnegie Mellon University, 1996.

[1-8] D. Pearce, H.G. Hirsch, "*The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions*",ICSA ITRW ASR2000, September 2000.

[1-9] P. R. Cohen, S. L. Oviatt, "*The Role of Voice Input for Human-Machine Communication*", Proceedings of the National Academy of Sciences, 1995.

[1-10] J.F. Allen, D.K. Byron, M. Dzikovska, G. Ferguson, L. Galescu,A. Stent. "*Towards ConversationalHuman-Computer Interaction*", AI Magazine, 22(4), 27-35, 2001.

[1-11] V.Z. Kepuska, T.B. Klein, "*Nonlinear Analysis: Theory, Methods & Applications*", vol.71, issue 12, pp. e2772-e2789, 2009.

[1-12] J. Junkawitsch, L. Neubauer, H. Hoge, G. Ruske, "*A new keyword spotting algorithm with pre-calculated optimal thresholds*", Fourth International Conference on Spoken Language Proceedings, ICSLP 96. Volume 4, 3-6, 1996.

[1-13] A. Stiles, B. Schmitt, F. Gertz, T. Klein, V. Kepuska, "*Testing and Improvement of the Triple Scoring Method for Applications of Wake-up Word Technology*", 2007.

[1-14] V.Z. Kepuska, T.B. Klein, "A novel Wake-Up-Word speech recognition system, Wake-Up-Word recognition task, technology and evaluation", 2009.

[2-1] J. Chen, J. Benesty, Y. Huang**, "***Time delay estimation in room acoustic environments: an overview***",** EURASIP Journal on applied signal ,vol 2006, pp.170-188, 2006.

[2-2] C.P. Mathews, D. Zoltowski, "*Eigenstructure techniques for 2-D angle*

estimation with uniform circular arrays", IEEE Transactions on signal processing, 1994.
[2-3] R.O. Schmidt, "*Multiple Emitter Location and Signal Parameter Estimation*", IEEE Trans. Antennas and Propag., vol. AP-34, no. 3, pp.276-280,March 1986.
[2-4] D. Johnson and D. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.
[2-5] A.S. Spanias, "*Speech Coding: A Tutorial Review*," Proceedings of the IEEE,
vol. 82, no. 10, pp. 1541-82, October 1994.
[2-6] K.N. Stevens, A.S. House, "*An acoustical theory of vowel production and some of its implications*", J Speech Hear Res, 1961.
[2-7] 來源網站：http://www.sciscape.org/articles/overtone/
[2-8] G. Fant, Acoustic theory of speech production, Mouton, 1960.
[2-9] K.H. Davis, R. Biddulph, S. Balashek, "*Automatic Recognition of Spoken Digits*", Journal of the Acoustical Society of America Vol. 24 No. 6, November 1952
[2-10] S.J. Wang, *Detection and Estimation*, class note.
[2-11] P. Viola, M.J. Jones, "*Robust real-time face detection*", International Journal of Computer Vision, 2004.
[3-1] A. Bhattacharyya, "*On a measure of divergence between two statistical populations defined by their probability distributions*", Bulletin of the Calcutta Mathematical Society 35: 99–109. , 1943.