The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 332-353

基於聽覺感知模型之類神經網路及其在語者識別上之應用

Two-stage attentional auditory model inspired neural network and its application to speaker identification

羅玉雯[a], 廖元甫[b], 冀泰石[a]
[a]國立交通大學電機工程學系
[b]國立台北科技大學電子工程系
yuwenlo0320@gmail.com, yfliao@ntut.edu.tw, tschi@mail.nctu.edu.tw

摘要

根據神經生理學研究，耳朵會針對聲音的各個頻率進行分頻，並產生出聽覺頻譜，研究人員根據專注聽覺現象和生物聽覺實驗，也發現了大腦聽覺皮質上神經作用的模式。於本論文中，我們運用類神經網路，建構出一種模擬人類聽覺的類神經網路模型，並在語者識別這個應用上進行討論，期望能成功連結神經生理學的知識與工程的技術。而我們所設計的模型，是利用兩層不同維度的卷積神經網路(Convolutional Neural Network)，分別模擬初期耳蝸階段及大腦皮質階段，透過設計卷積核初始值，即耳蝸階段多組一維分頻濾波器和大腦皮質階段同時解析時頻資訊的二維濾波器，以使模型能夠快速地達到收斂狀態。而透過模型訓練，根據目的與環境變因的不同，模型會自動調整其中參數，使輸入資料映射至目標的型態。同時我們也針對所提出的模型架構，進行了多種形態的比較，進而發現在給定初始值的狀況下，即使訓練不夠充分，也能產生不錯的結果。

1. 研究背景

語者識別的目標為有效地準確地辨別目前的說話者，而發展至今已有很多成熟的方法。在本論文中，我們所設計的語者識別系統，是以神經網路學習中的卷積神經網路(CNN)實現，並以模擬人耳聽覺感知為目標。近年來隨著神經網路技術的普及，研究學者發展出了許多以神經網路為核心的語者識別演算法，但這些系統基本上僅利用對原始訊號進行特徵抽取(例如MFCC)，再透過類神經網路進行個別與者的模型訓練[1][2][3]。但是在計算特徵的同時，可能會遺失掉原本語音訊號的其他重要資訊，因此我們師法人的耳蝸功能，不抽取特定特徵而任由卷積神經網路對原始時域訊號進行濾波[4][5][6]，來進行語者識別。

然而，聽覺神經學的學者透過實際動物實驗發現，哺乳類動物的聽覺形成主要經過兩階段，分別為初期耳蝸階段以及大腦皮質階段。在初期耳蝸階段中，聲音訊號進入到耳朵後，耳蝸會針對聲音頻率進行解析，並且會根據頻率的高低而有不同的解析度，其解析中心頻率與頻寬的比值呈現一個常數 Q 的關係，也就是對低頻聲音有著較為精細的解析；而對高頻聲音則進行較為廣泛的頻率解析。而我們可以透過這個關係，將原始聲音訊號轉換成二維的聽覺頻譜圖，與傅立葉頻譜圖的不同處在於聽覺頻譜圖更能表現出耳朵對聲音所解析出的時頻特徵[7][8]。

之後，將耳蝸階段解析出來的聽覺頻譜送往下個階段，也就是大腦皮質(A1)階段，其神經元會針對聽覺頻譜的時域調變及頻域調變同時進行解析[9][10]，亦即聽覺感知是一對二維時-頻訊息的綜合反應，當頻譜上兩頻率通道資訊互換或者時間軸上兩時間點資訊對調，皆會對聲音的解讀產生困難。基於動物實驗而得到的 A1 神經元紀錄，美國 NSL 實驗室提出了一聽覺感知模型 [11]，而這個模型，能解析出語音能量頻譜中所隱含的多種重要資訊，像是音高(pitch)、諧波成份(harmonic)、振幅調變(AM)、頻率調變(FM)、語音起始(onset)與終止處(offset)等資訊。近年來已成功的應用在許多語音與音訊處理的研究議題上，如評估語音清晰度[12]、語者識別[13]及從背景音樂中進行聲源分離[14][15]等等。

而由動物實驗中，研究人員亦發現聽覺皮質層的神經元會因為認知的目的不同，自我調整出一個專注的機制來選擇提取相對重要的訊息[16][17][18]，換句話說人類常在聽到聲音時表現出注意行為，而這些注意行為是由較高層次的認知功能所引起的，而在聽覺中，這個專注行為可以幫助我們在吵雜的環境中更有效地辨別目標的聲音。

近年來，具強大功能的類神經網路解決了許多困難的工程問題，並且也有許多將其成功應用於語音方面的例子，如語音辨識[19][20]、音源分離 [21]、情緒辨識[22]，而卷積神經網路(convolutional neural network, CNN)，擁有能夠提取二維特徵資訊的功能，除了廣泛的應用在圖像辨識[23][24]，也能成功的應用於語音辨識[25][26]上。

我們根據上述這些理論，在本論文中提出了基於類神經網路的兩階段聽覺感知模型，此模型主要概念為模擬耳蝸對於原始訊號的分頻以及大腦聽覺皮層區的神經元針對輸入封包會有不同的時域及頻域選擇性。也就是說，在經過耳蝸的分頻機制後，我們可以將原始的音源訊號轉換成二維聽覺頻譜圖；而在大腦皮質階段，會透過卷積神經網路建構出其神經元的專注機制，對轉換而成的二維聽覺頻譜圖進行解析。

我們參考的兩階段感知聽覺模型模擬了初期耳蝸階段以及聽覺皮質(A1)區對聲音的解析，但沒有包含聽覺感知更上層的神經細胞的反應，因此，為了模擬更完整的聽覺路徑上神經整合資訊的過程，我們對於 A1 區之後的神經細胞的作用，以標準的神經網路學習演算法來近似。此方法的最大優點是：資料導向(data-oriented)及非模型導向(model-oriented)，亦即在我們不知道大腦運作的任何數學模型下，也能夠利用神經網路演算法自主學習的能力，模擬更深層神經元之間的運算，以完整模擬人類聽覺路徑上所有神經元的作用。

在本論文中，我們採用卷積神經網路，來實現所參考聽覺模型的初期耳蝸階段以及大腦皮質階段，我們以一維的卷積層並透過設定基於耳蝸分頻機制的多組濾波器，來模擬聽覺感知模型的初期耳蝸階段；再利用二維卷積層搭配多個二維時頻調變濾波器，來模擬在大腦皮質階段對於聽覺頻譜的專注機制所進行的特徵提取，並透過特徵映射層來模擬大腦更深層的資訊連結。因此，我們預期所提出的演算法相當接近人類耳朵及大腦處理語音資訊的實際情況，進而在語者識別上能達成不錯的效果。

2. 感知訊號處理

2.1 生理聽覺現象與特性

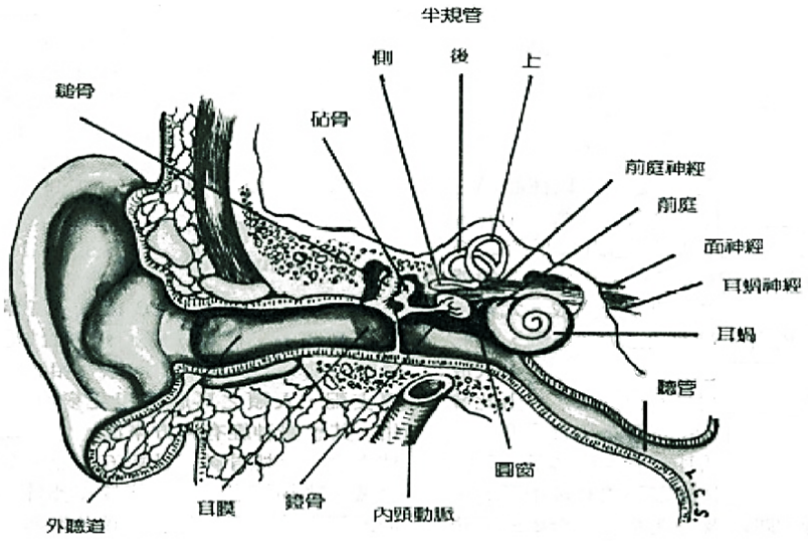

圖 2.1：耳朵基本構造 [7]

人耳的基本構造，由外而內主要分為外耳、中耳、以及內耳三個部份，如上圖(2.1)所示。外耳包含耳殼及外聽道；中耳包含三小聽骨(鎚骨、砧骨及鐙骨)；而內耳則是由主司聽覺的耳蝸以及主司平衡的前庭與半規管所組成。外界的聲波，經由外耳、中耳、內耳的順序依序傳遞，將聲波轉換成最後的電訊號，使我們能聽到聲音。首先，聲音訊號由耳殼及外聽道接收後撞擊耳膜，耳膜震動，進而帶動中耳的三小聽骨以槓桿原理運動推動卵圓窗，此時聲波已被轉換成機械能傳遞。之後由於卵圓窗受到推擠，能量進而傳入充滿組織液的內耳，由機械能再轉為動能，帶動內耳內組織液的流動，並於基底膜(basilar membrane)上產生行進波。

由於基底膜上的質地和寬度差異，靠近膜底部(前端，base)的質地較硬寬度較窄；而靠近頂部(後端，apex)較寬軟。這樣的結構使得不同頻率的訊號，在基底膜上所產生的行進波會在不同的位置產生最大振幅。因此，基底膜可視為一系列的分頻濾波器，較低頻的訊號會傳至較遠處才產生共振；而頻率較高的訊號，在靠近基底膜底部的位置就會產生共振，而可接受的波頻率範圍大約為 20 Hz 到 20000 Hz，即正常人類聽覺範圍。

基底膜上行進波的運動會拉動基底膜內柯替氏器(organ of Corti;圖 2.2)上，附著的數以千計的內毛細胞和外毛細胞，使之產生一連串的電化學變化，引發神經脈衝以電訊號刺激聽神經，再傳遞至大腦進行分析與整合。

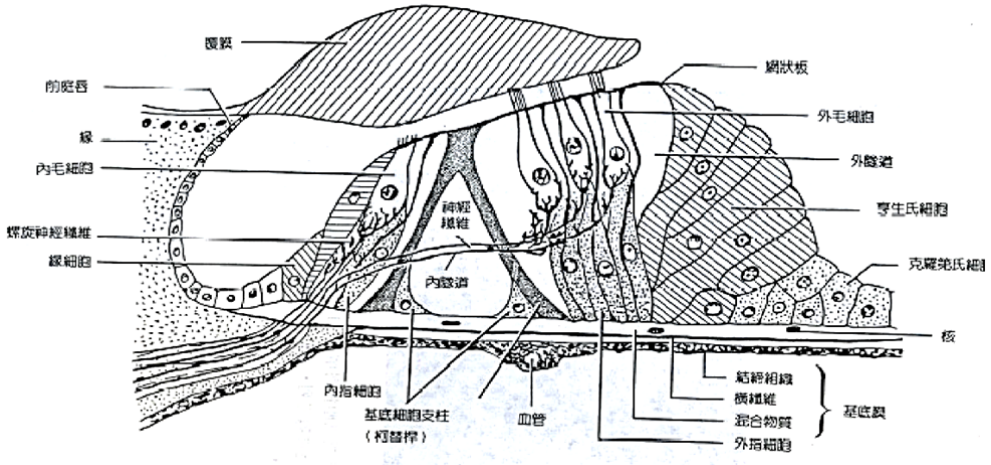

圖 2.2：基底膜內柯氏器 [7]

2.2 聽覺感知模型

此模型是由 NSL(Neural Systems Laboratory)實驗室所提出，藉由實際進行動物聽覺實驗，進而建構出符合哺乳類動物的聽覺系統模型,它模擬了聲音訊號從耳蝸到大腦皮質 A1 區的過程。這個模型包含了兩個主要的部分：初期耳蝸階段(early cochlear stage)及大腦皮質階段(cortical stage)。前者為聲音訊號被內耳耳蝸上的內外毛細胞所解析的過程，即預估聲音的聽覺頻譜階段；後者在於模擬大腦皮質(A1)區對其聽覺頻譜的解析，由多組的時頻域二維調變濾波器所實現。

2.3 聽覺感知模型-初期耳蝸階段

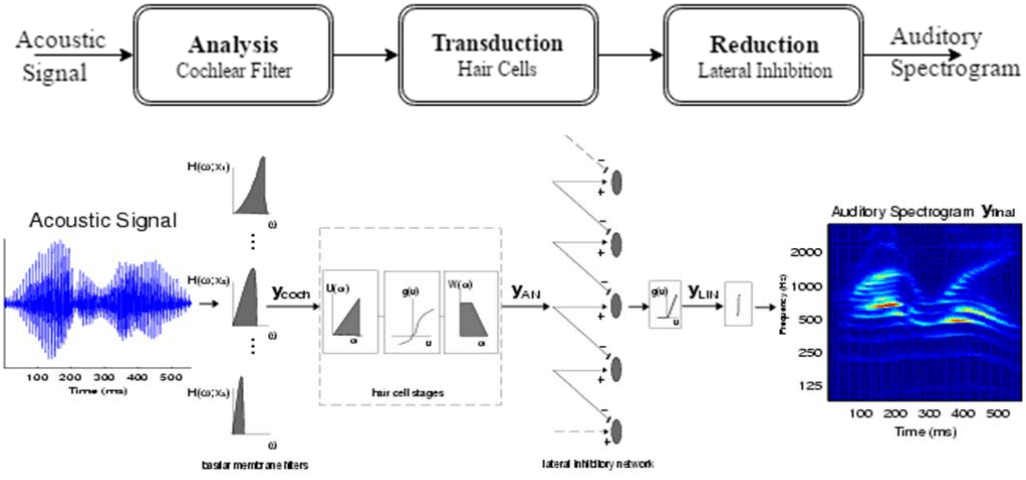

圖 2.3：初期耳蝸階段之訊號處理流程圖 [7]

初期耳蝸階主要的流程如圖(2.3)所示，每個聲音訊號進來後都會經過三個階段：分析(analysis)、傳導(transduction)以及壓抑(reduction)，而我們將利用以下數學式來完成其模擬：

$$y_{coch}(t,x) = s(t) \otimes_t h(t;x) \tag{2.1}$$

$$y_{AN}(t,x) = g(\partial_t y_{coch}(t,x)) \otimes_t w(t) \tag{2.2}$$

$$y_{LIN}(t,x) = \max(\partial_x y_{AN}(t,x), 0) \tag{2.3}$$

$$y_{final}(t,x) = y_{LIN}(t,x) \otimes_t u(t;\tau) \tag{2.4}$$

式(2,1)為分析部分，用來模擬聲音 $s(t)$ 傳至耳蝸後，在基底膜上依照其本身不同的共振頻率，被不同的位置上被解析出來。$h(t,x)$ 代表基底膜上位置為 x 的共振響應，x 即為基底膜上距離耳蝸底部的距離，而模型中使用 128 個具不同中心頻率及頻寬的帶通濾波器組(band pass filter bank)來模擬各位置的共振響應，其中中心頻率和頻寬成常數 Q(constant Q)的關係，如式(2.5)

$$\text{中心頻率}(center\ frequency) \Big/ \text{頻寬}(bandwidth) = \text{常數}Q \tag{2.5}$$

中心頻率在對數軸上是均勻分布的，接著，每個濾波器的輸出將被傳送到一非線性壓縮階段，對應到式(2.2)。這個非線性壓縮是用來模擬耳蝸基底膜的震動轉化成內毛細胞的電位，而內毛細胞的飽和現象。接著相近的內毛細胞彼此之間會有一階側抑制作用(lateral inhibitory network ,LIN)，如式(3.3)，也模擬了聽覺上鄰近頻率的遮蔽效應。

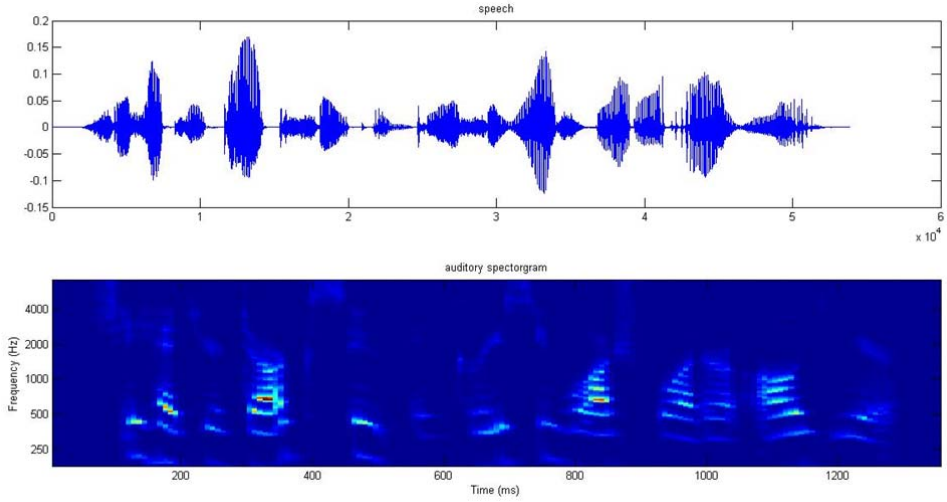

圖 2.4：語音訊號經在聽覺模型中經過初期耳蝸階段所產生之聽覺頻譜

最後，訊號會通過一封包擷取器，如式(2.4)。而其積分窗函式則寫成式子(2.6)：

$$u(t;\tau) = e^{-\frac{t}{\tau}} \times \mu(t) \tag{2.6}$$

經過以上的處理，我們可以得到y_{final}，也就是時頻域的聽覺頻譜圖(auditory spectrogram)。和一般的短時傅立葉轉換頻譜圖(STFT spectrogram)不同之處為，此頻譜的頻率軸是以對數呈現，如圖(2.4)所示，接著在第二階段的大腦皮質分析將針對此聽覺頻譜作進一步的分析。

2.4 聽覺感知模型-大腦皮質階段

第二階段是在模擬大腦皮質 A1 區的神經元對於時頻的選擇性。在聽覺模型中，y_{final}是經過初期耳蝸階段所得到的聽覺頻譜圖。大腦皮質 A1 區的神經可以被視為一系列具有不同特徵參數的二維時頻調變濾波器 (spectro-temporal modulation filters , STMFs)，可以用來解析所得到的聽覺頻譜。換句話說，A1 模型將原本的聽覺頻譜根據不同的時頻調變進行解析，我們假設，在 A1 後的神經元可以收集並整合許多經 STMF 解析後的具體資訊，進而建構出更高階的大腦認知功能。

生成 STMF 的參數包含了 rate ω_c (Hz) 、scale Ω_c (cycle/octave) 以及方向性。rate 捕捉了聽覺頻譜沿著時間軸的變化速度，而 scale 則是捕捉了其沿著頻率軸的能量分布狀況，此外，rate 的符號代表了 STMF 的方向性(正/負 符號代表 向下/向上 的方向)，而 STMF 的頻率響應可以寫成 (2.7)及(2.8)：

$$STMF_{+w_c,\Omega_c}(\omega,\Omega) = \begin{cases} |F\{h_{rate}(t;\omega_c)\} \otimes F\{h_{scale}(f;\Omega_c)\}|, 0 \le \omega;\Omega \le \pi \\ 0, \ otherwise \end{cases} \tag{2.7}$$

$$STMF_{-w_c,\Omega_c}(\omega,\Omega) = \begin{cases} |F\{h_{rate}(t;\omega_c)\} \otimes F\{h_{scale}(f;\Omega_c)\}|, -\pi \le \omega \le 0;0 \le \Omega \le \pi \\ 0, \ otherwise \end{cases} \tag{2.8}$$

而F代表一維的傅立葉轉換，$\otimes$ 是外積。rate (ω, 以 Hz 為單位) 和 scale (Ω, 以 ms 為單位) 分別是時間的頻域軸以及頻率的頻域軸。而 h_{rate} 和 h_{scale}則是代表利用珈瑪形狀濾波器(Gammatone filters)所得到的以ω_c 及Ω_c為中心的一維時間及頻率脈衝響應，如(2.9)。

$$\begin{cases} h_{rate}(t;w_c) = t^4 e^{-2\pi BW_{rate}t}\cos(2\pi w_c t) \\ h_{scale}(f;\Omega_c) = f^4 e^{-2\pi BW_{scale}f}\cos(2\pi \Omega_c t) \end{cases} \tag{2.9}$$

而頻寬BW_{rate} 和 BW_{scale}會根據中心頻率ω_c 和Ω_c而增加。

圖(2.5)則代表 24 個二維的 STMF 的脈衝響應，其參數分別為 ω_c = { 4, 8, 16, 32 } Hz, Ω_c = { 0.5, 1, 2 } cyc/otc，且其方向性為雙向。而圖(2.6)則是某例句經過初期耳蝸階段後所得到的聽覺頻譜再經過 8 種不同的 STMF 濾波後之結果。

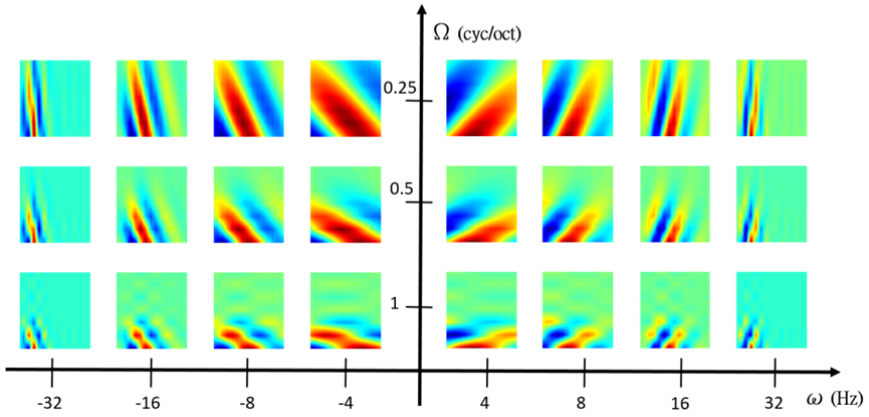

圖 2.5：STMFs 二維的脈衝響應範例

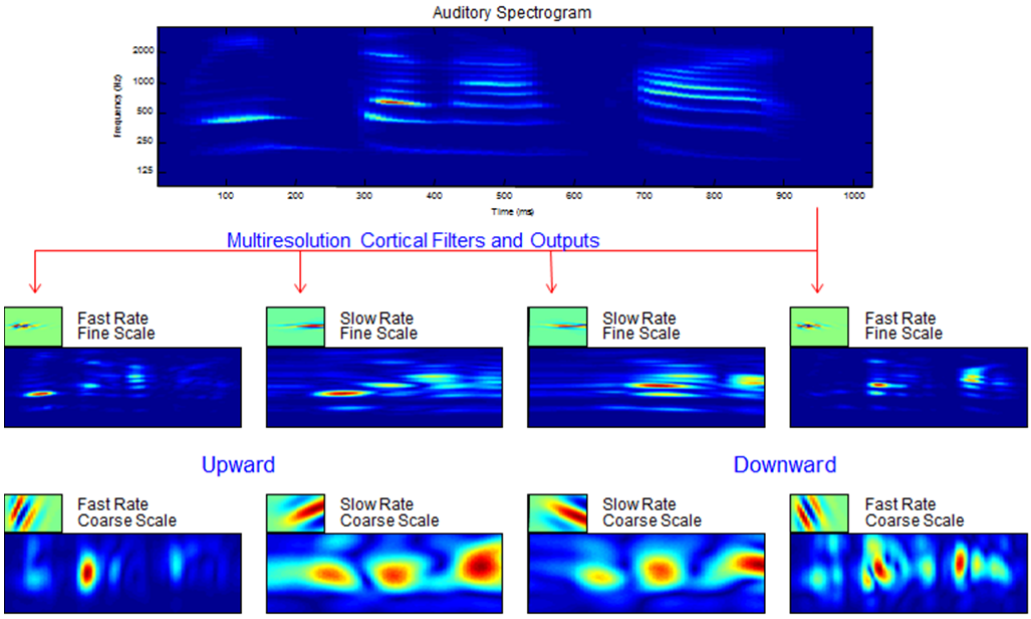

圖 2.6：經過初期耳蝸階段後所得到的聽覺頻譜再經過 8 種不同的 STMF 濾波後之結果

3. 類神經網路系統架構與參數設定

3.1 卷積神經網路簡介

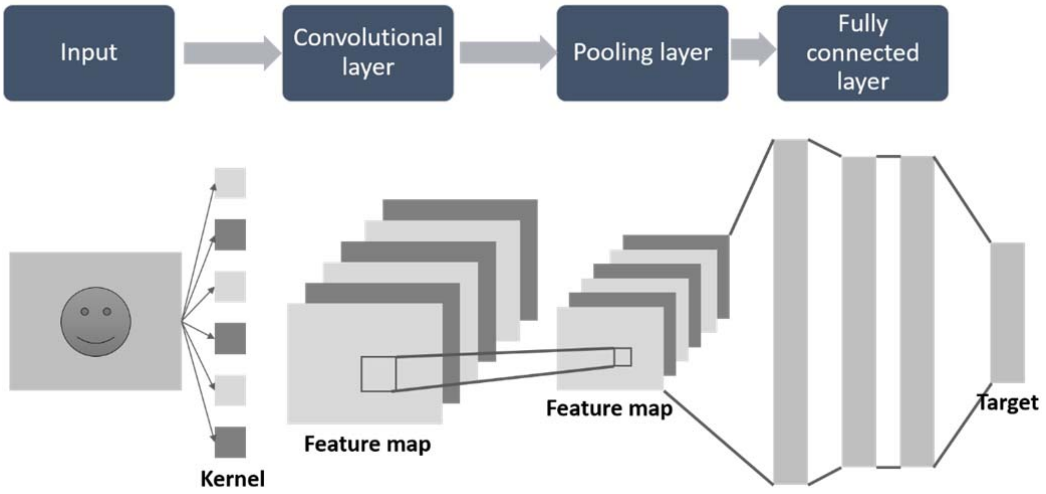

圖 3.1：卷積神經網路架構示意圖

卷積神經網路(Convolutional Neural Network, CNN)為神經網路的變形，於近代發展起來，並備受重視，已廣泛的被應用到解決各種關於辨識與分類問題上。其由來為 20 世紀 60 年代，Hubel 和 Wiesel 在研究貓大腦皮質層中對局部方向選擇敏感的神經元時發現其獨特的結構可以有效地降低反饋神經網路的複雜性，既而提出。

一般的卷積神經網路包含三層：卷基層(convolutional layer)、池化層(pooling layer)以及特徵整合層(fully connected layer)。圖(3.1)為標準的卷積神經網路架構之範例。以圖片分類為例，我們的輸入可以是一張二維的原始圖像，在卷積層中經過與卷積核(kernel)的運算後，可以提取到其相對應的特徵圖(feature map)，每個卷積核所得到的特徵圖皆為一獨立平面，且其平面上所有神經元之權值相等，此步驟於物理意義上為提取與目標相關之特徵，以利我們之後的計算。圖(3.2)為卷積層數學運算之範例。

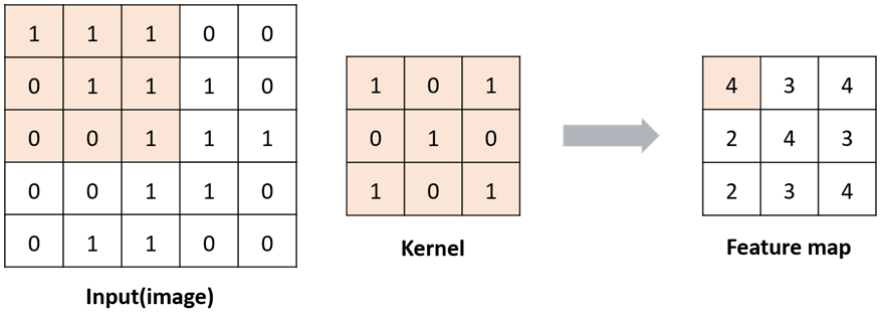

圖 3.2：卷積核大小為 3x3 之卷積層範例

當透過卷積層得到特徵圖之後，我們希望能利用這些特徵來做分類，但是對於一個太大的特徵輸入分類器來說，需要過於龐大的計算量而且很容易出現過擬合(over-fitting)的情形。因此我們希望得到的特徵圖具有平移不變性，並透過這個特性將對於不同位置的特徵值進行聚合統計，一般來說就是計算某個特定區域的最大值或平均值，而這種聚合統計的過程就稱為池化(pooling)，圖(3.3)顯示一最大池化的例子。

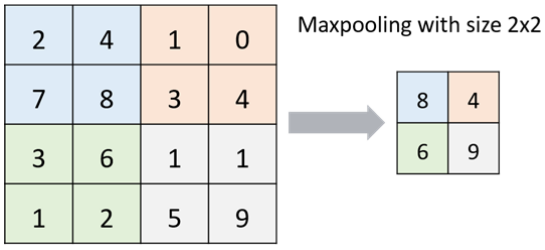

圖 3.3：大小為 2x2 之最大池化範例

而特徵整合層，為卷積神經網路最後一個階段，此層的運算方法和傳統神經網路相同，即透過輸入神經元和輸出神經元間互相連結而成，可把前面提取之參數用於分類(classification)或回歸(regression)的議題上。

3.2 模型架構

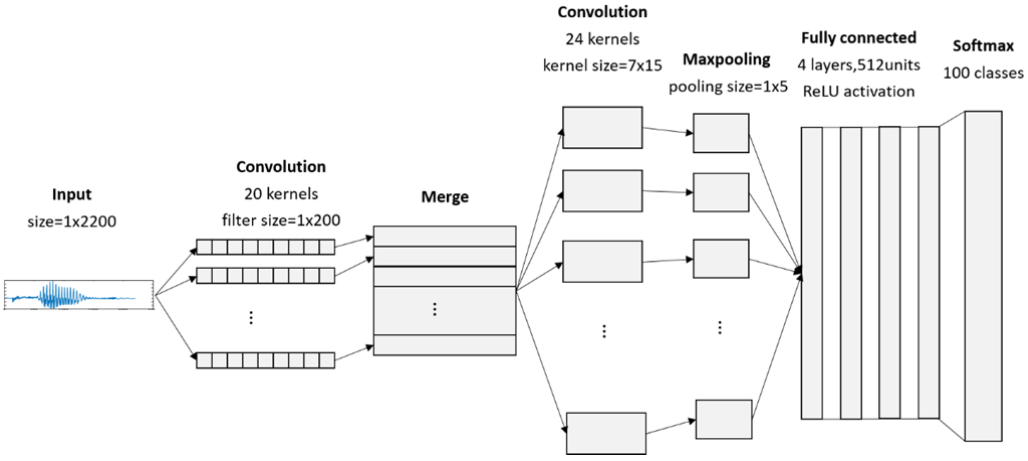

圖 3.4：所提出的模型架構

藉由聽覺感知模型的啟發，我們提出了一個基於卷積神經網路的語者識別系統。我們所提出的類神經模型，包含了輸入層、一維的卷積層、二維的卷積層、池化層，以及四層特徵整合層，如圖(3.4)所示。其中，為了要完整的模擬聽覺感知模型，我們輸入層的是未經過任何處理的一維原始音檔。

在一維的卷積層時，我們利用了卷積核權值共享的特性，我們認為在對原始音檔做卷積時，相當於對其做了不同頻率的濾波。因此我們根據耳蝸對於不同中心頻率以及頻寬的常數 Q 關

係，選擇使用 20 個卷積核進行濾波，並將每個濾波器所得到的結果進行排列，可以得到基於聽覺感知模型初期耳蝸階段的聽覺頻譜。

得到聽覺頻譜後，在二維的卷積層時，我們選用了 24 個 7x15 的卷積核，來模擬大腦皮質階段時，會對聽覺頻譜作一個二維調變資訊的擷取動作。而池化層則是將我們所得到的結果，保留重要資訊並進行降維，來降低我們整體的運算量。而特徵整合層則是將我們所得到的資訊進行統整、分析，藉以模擬大腦更高階層的資訊統整動作。

3.3 實驗語料

本論文中，我們使用 2008 NIST SRE (Speaker Recognition Evaluation)資料庫中的語音訊號，此資料庫是由語言數據聯盟(Linguistic Data Consortium, LDC)及美國國家標準技術研究所(National Institute of Standards and Technology, NIST)所提出。

我們所用的資料為 training set 中 short2 的電話語料，每個音檔約 5 分鐘，左右聲道分別為不同的語者。我們隨機抽取 100 人，並將靜音的部分先行移除、合併，再將其切成 24 份 5 秒的音檔，並加入了由 NOISEX-92 資料庫取得的背景雜訊，訊雜比會在第四章詳述實驗結果時進行說明。而為了確保測試音檔的可信及穩定度，我們從 24 份音檔之中，挑選出 2 個能量最強，也就是語音資訊最豐富的音檔當作測試用資料，而剩餘的 22 份音檔則當作語者模型之訓練資料。

3.4 語音處理背景知識與參數設定

本論文中我們模型的輸入為 275ms 的片段語音，所有音檔的取樣頻率定在 8k Hz，此設定既能保有語音中的重要資訊又能有效的降低輸入維度(即輸入維度為 2200 點)。在一維的卷積層時，為了能完整的表現其濾波器的大小能夠涵蓋各種頻率，因此選擇卷積核大小為 25ms (200 點)，此設定可以模擬中心頻率為 80Hz ~ 4000Hz 的帶通濾波器的脈衝響應，並且以每 10ms (80 點)為音框彼此間的間隔。藉此來模仿在做頻譜分析時，原始訊號時間軸上的處理方式。

經過一維卷積層後，我們將 20 個濾波器結果排成一張頻譜圖(大小為：20 kernel * n frame)，考慮到二維卷積核的物理意義，在時間軸上，判斷一個音素最少須 50ms 的時間，我們設計的卷積核 y 軸大小為 15，根據一維卷積核 10ms 為一間格，我們可以得知，y 軸大小為 15 的狀況下，能夠包含 150ms 的資訊；而在頻率軸上，我們選擇 x 軸大小為 7 的卷積核，是因為這個大小能夠包含兩個八度音，以人平常在講話為例子，即可以包含能量較為明顯的第一共振峰，因此在卷積的過程中，我們可以有效的擷取出較有意義的能量區塊中的隱藏資訊。

3.5 一維卷積核初始化

在聽覺感知模型中，聲音傳至耳蝸後，會在基底膜上依照其本身不同的頻率，而被不同的聽神經元解析出來。而我們所提出的模型中，一維卷積層即是要模擬耳蝸分頻的動作，也就是利用 20 組帶通濾波器，針對各個位置的共振響應，來對原始訊號進行濾波。因為基底膜對於不同位置的聲音響應過程相當於一個濾波過程，而珈瑪形狀濾波器(gammatone filter)結合了人耳的聽覺特性，也就是對中心頻率呈對數分布來模擬基底膜的特性，其數學式如下：

$$g(t) = at^{n-1}e^{-2\pi bt}\ cos(2\pi ft + \phi) \tag{3.1}$$

其中，f (Hz) 是中心頻率，ϕ 是載波相位，a 是振幅，n 是濾波器的順序，b (Hz) 是濾波器的頻寬，t 是時間。這是一個以珈瑪分布(gamma distribution) 函數來調變一單音的函式。

因此，在一維的卷積層時，我們希望所濾出來的波型的中心頻率，能根據其頻寬成常數 Q (constant Q)關係，故我們利用珈瑪形狀濾波器來產生具希望之頻率響應之濾波器組。下圖(3.5)為根據實驗設定所得到的 20 個濾波器，再分別經過 400 點的傅立葉轉換所得到的頻率振幅響應，並依照其中心頻率之高低排列(x 軸，filter index)後的結果。

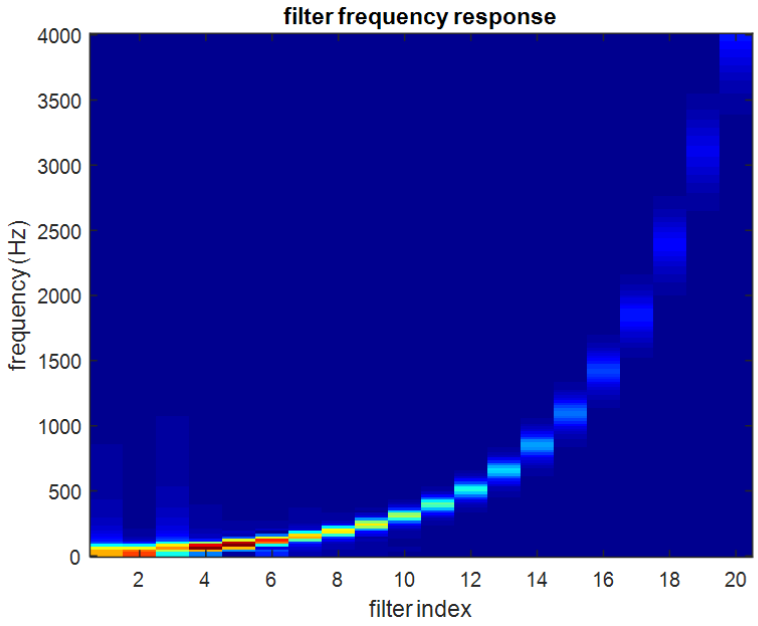

圖 3.5：20 個一維卷積核經過傅立葉轉換所得之頻率振幅響應

3.6 二維卷積核初始化

在所提出的模型中，二維的卷積層中我們使用了 24 個卷積核。我們利用 24 個 STMF 的脈衝響應，擷取比較強烈的部分，也就是 7x15 的大小，來當作我們的初始值，而我們所用到的參數分別為 rate = { 4, 8, 16, 32 } Hz, scale = { 0.25, 0.5, 1 } cyc/otc，及雙方向[28]，如圖(3.6)所示。

在頻率軸上，我們選擇大小為 7 的卷積核，是因為 7 能夠涵蓋兩個八度音(octave)，從圖(3.5)簡單來看，編號第 14 到編號第 20 個一維卷積核所涵蓋的頻率範圍，即是 1000Hz~4000Hz，也就是兩個八度音。

在時間軸上，因為 50ms 大約是人能夠理解語音中的最小單位的時間，但我們又希望能夠分析到較小的 rate 所包含的語音長時資訊，因此我們選定 150ms 為我們卷積核 x 軸的大小，而其倒數，也就是 6.7Hz，是分析所得到的聽覺頻譜的語音封包變化最小單位，但在大小為 150ms 的音框上，我們可以看到波長為 250ms 的一半以上的波型，故這個時間軸的大小，可大約分析 rate 最低至 4Hz 的語音封包變化情形。

再加上一維卷積核是以每 10ms(也就是 100Hz)為音框彼此間的間隔，以取樣定理我們可以得知，最高觀察 50Hz 的變化量。綜合以上兩點，在時間軸的分析上，我們可以觀察到 rate 為 4~50Hz 的時域調變變化情形。

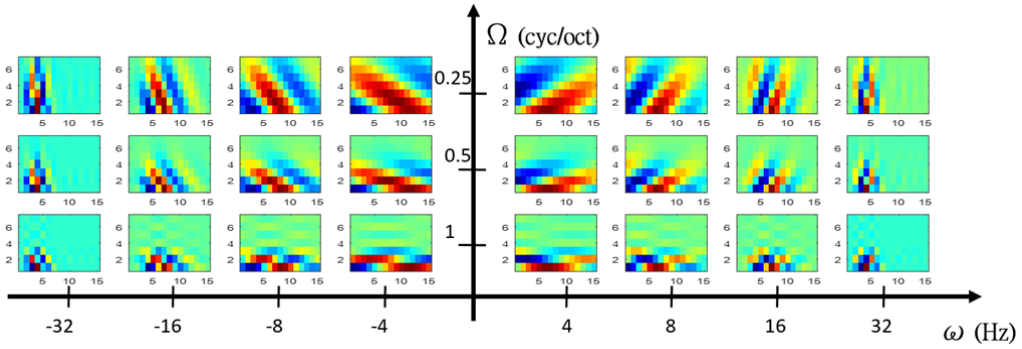

圖 3.6：24 個根據不同的 rate-scale 參數所圈出來的二維卷積核初始值

4. 實驗與討論

4.1 比較系統介紹

根據聽覺感知模型的特性，我們將考慮五大類模型來進行實驗，如下表(4.1)所示，以下針對各類模型的設定做說明：

Model		
1D CNN kernel type	2D CNN kernel type	Referred to
Gammatone Fix	A1 initial	Gammafix_A1init
	A1 random	Gammafix_A1rand
Gammatone Initial	A1 initial	Gammainit_A1init
	A1 random	Gammainit_A1rand
Both random		Bothrand

表 4.1：五大類比較之模型

Gammafix_A1init：一維卷積核固定為 20 個 gammatone 濾波器之結果，訓練時無法對此一維卷積核組進行修正；而二維卷積核的初始形狀給定計算出來之 24 個 STMF，後來透過前饋以及反向傳播演算法(feed-forward and back-propagation)進行訓練。這個模型的假設是耳蝸階段的分頻，是沒有辦法依照應用目的的不同而進行調整的；而大腦皮質 A1 區可以根據應用目的進行調整。此設定與動物神經實驗所觀察到的現象類似。

Gammafix_A1rand：一維卷積核固定為 20 個 gammatone 濾波器之結果，訓練時無法對此一維卷積核組進行修正；而二維卷積核不給特定的初始值，直接透過前饋以及反向傳播演算法進行訓練。這個模型的假設與第一類型(**Gammafix_A1init**)相似，不同的地方是二維卷積核給的是隨機初始值。我們最後會對所訓練出的二維卷積核進行分析與討論。

Gammainit_A1init：一維卷積核的初始形狀給定為 20 個 gammatone 濾波器之結果、二維卷積核的初始形狀給定為計算出來之 24 個 STMF，後來透過前饋以及反向傳播演算法對兩階段的卷積核進行調整。這個模型的假設是，聽覺感知模型的兩個階段的神經反應皆可以針對應用目的的不同而進行調整。

Gammainit_A1rand：一維卷積核的初始形狀給定為 20 個 gammatone 濾波器之結果；而二維卷積核不給特定的初始值，直接透過前饋以及反向傳播演算法進行訓練。這個模型的假設與第三類型(**Gammainit_A1init**)相似，不同的地方是二維卷積核給的是隨機初始值。我們最後會對所訓練出的二維卷積核進行分析與討論。

Bothrand：一維與二維卷積核，皆不給定特定初始值，直接透過前饋以及反向傳播演算法進行訓練。我們想藉由不給定任何初始值的狀況，來觀察在此架構下訓練調整完後一維及二維卷積

核的形狀，並檢視此模型架構下與前數種模型之效益比較。

這裡所有參與比較的模型均有同樣的架構，亦即一維卷積層包含 20 個 1x200 的卷積核，並以每 80 點做一次平移相乘；而二維卷積層包含 24 個 7x15 的卷積核；後面連接大小為 1x5 的最大池化層，並在之後接上 4 層節點數為 512 的特徵整合層。如圖(3.4)所示。

4.2 實驗結果

我們利用所提出的類神經網路模型，來模擬聽覺模型中，初期耳蝸階段對於聲音訊號的分頻；以及大腦皮質 A1 區對於聽覺頻譜的時頻選擇性。因此在這個章節中，我們除了將比較五種模型對正確率的影響，同時也會針對我們所提出的類神經網路模型經過訓練後，與傳統的聽覺感知模型的相關性及意義進行討論。

在這個實驗裡我們將兩種不同的背景雜訊分別以訊雜比-5、0、5dB 與語音相混，一共產生六種不同情境下的語句同時對模型進行訓練。我們在這次的實驗中選定兩種背景雜訊，分別為 buccaneer 及 factory。

下表(4.2)為此次實驗的實驗結果，從中可以發現，在有參考 gammatone 濾波器的模型，無論是否固定其一維的卷積核，在語者識別上的校能都會比一維、二維都隨機給定初始值的模型來的好。在此我們將針對以下幾點進行討論：

I　前四種模型，對一維卷積核的形狀進行討論。
II　前四種模型，對二維卷積核形狀進行討論。
III　第五種模型 Bothrand 的結果討論。

Model		SNR(dB)		
1D CNN kernel	2D CNN kernel	-5	0	5
Gammatone Fix	A1 initial	59.50%	77.25%	95.00%
	A1 random	63.50%	73.25%	93.75%
Gammatone Initial	A1 initial	67.00%	77.50%	92.00%
	A1 random	69.25%	76.50%	92.75%
Both random		56.00%	65.75%	87.00%

表 4.2：各模型在多訊雜比與多雜訊種類條件下的語者識別正確率

I　前四種模型，對一維卷積核的形狀進行討論

首先，我們先針對前四種模型，也就是有參考 gammatone 濾波器的模型，分成固定其值

以及透過訓練去修正其值兩類，其結果如圖(4.1)所示。因為固定 gammatone 濾波器的結果並不會因為是否給定二維卷積核初始值而有所差異，故圖(4.1)左邊，代表著模型 gammafix_A1init 及 gammafix_A1rand 所固定的一維卷積核頻率振幅響應。

我們可以從圖(4.1)中右邊兩圖發現，能夠透過訓練而修正一維卷積核的模型，其卷積核大致上仍保留著 gammatone 濾波器的頻率選擇特性，但是對於不同頻帶，卻有著不同強度的增益。以高頻的卷積核來說，其明顯比左圖中原始的高頻卷積核，能量來的強。因此我們可以推論，模型經過訓練後，的確會根據應用目的的不同，或者背景雜訊的不同，來調整在該目的之下重要頻帶資訊的權重。

而我們也可以透過表(4.2)的結果發現，在低訊雜比之下，固定 gammatone 濾波器的模型明顯表現較差，故我們可以合理的推論，因為在低訊雜比下，原始訊號被破壞的較為嚴重，因此需要比較能夠凸顯某些較不受噪音影響的特定頻帶的濾波器，而透過重要濾波器所得的聽覺頻譜圖，在模型後面的階段，也就是大腦皮質階段擷取語音重要資訊時，也能較有幫助。

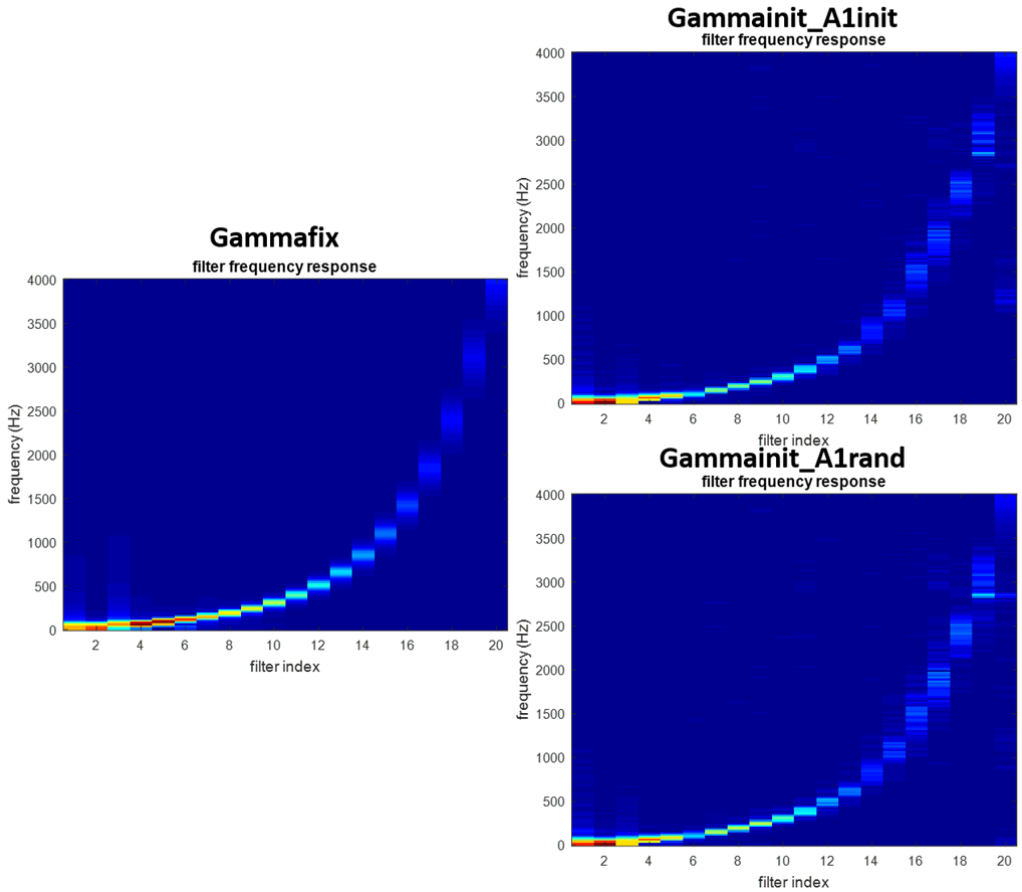

圖 4.1：多訊雜比及多雜訊種類條件下，模型訓練後之一維卷積核頻率振幅響應圖

II　前四種模型，對二維卷積核形狀進行討論

由上述討論我們可以知道，前四種模型無論是否透過訓練進行修正，一維卷積核都大致上

仍保留著 gammatone 濾波器的頻率選擇特性，因此，在這個階段我們將針對經過一維卷積核所得到的聽覺頻譜圖，經過第二階段，也就是仿大腦皮質階段的二維卷積核進行討論。

從表(4.2)中我們可以發現，在有給定 gammatone 濾波器結果之前四種模型，都有著差不多的表現，即使固定 gammatone 濾波器的模型，在 -5dB 訊雜比的狀況下有著稍微較差的表現，但其在 0dB 及 5dB 上仍有相似的表現。因此我們推論，這四種模型都有著類似功能的二維卷積核。而其結果如圖(4.2)所示。

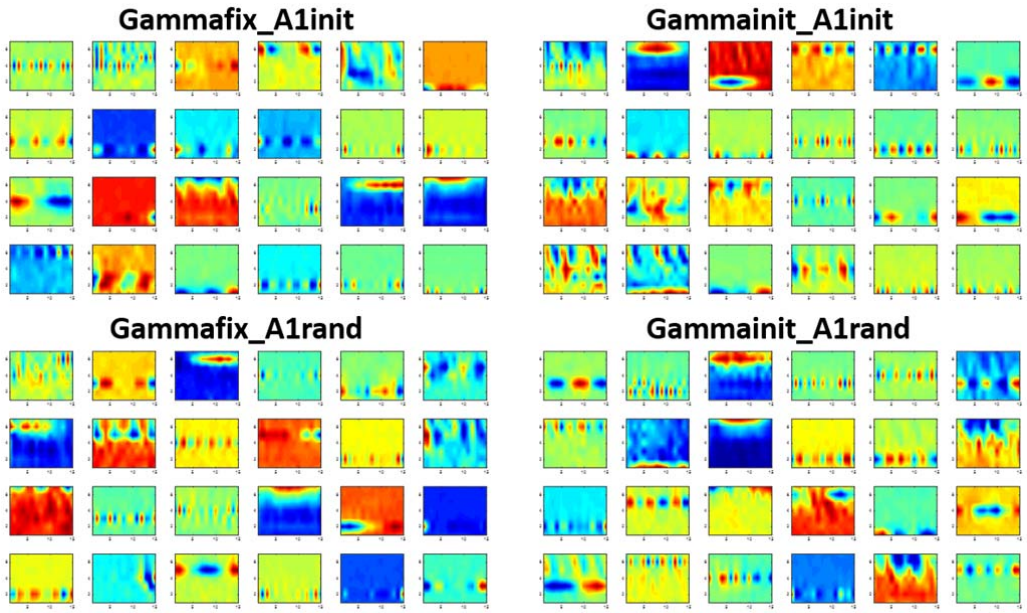

圖 4.2：多訊雜比及多雜訊種類條件下，各模型訓練後之二維卷積核形狀結果圖

而在圖(4.3)中，我們將一些重複於多個模型中功能類似的卷積核圈出，我們可以發現無論在何種模型中，都存在著類似功能的卷積核，這說明了無論二維卷積核是否有給定初始值，經過大資料的訓練後都會演化生成出類似的卷積核，而導致這些模型的最終結果差距不大。

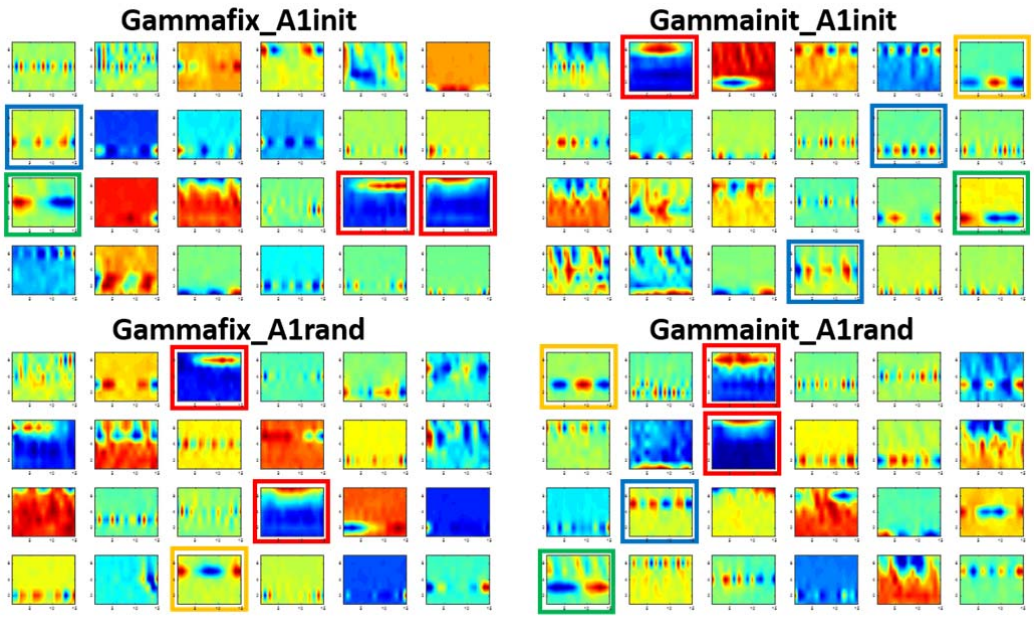

圖 4.3：多訊雜比及多雜訊種類條件下，各模型訓練後功能相同之二維卷積核形狀結果圖

而我們將針對特定卷積核進行討論：以下圖(4.4)為例，圖中我們可以看到該卷積核約包含 0.6 個波長，而我們的卷積核設定為可以涵蓋 150ms 的資訊，因此透過計算，我們可以得到其波型為波長 250ms 也就是頻率變化為 4Hz 的卷積核。

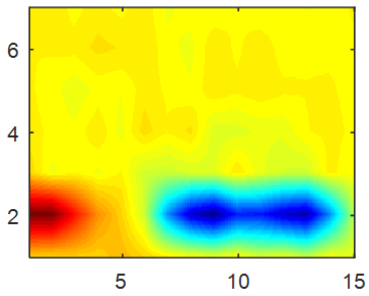

圖 4.4：擷取調變頻率變化為 4Hz 的卷積核

同樣的我們也可以從卷積核找出其他調變頻率變化，如 8、16、32Hz 的波型，如下圖(4.5)所示。當然，所有卷積核代表的調變頻率變化不單單只有這些，因此我們推斷，在語者辨識這個議題上調變頻率變化是一項重要的資訊。

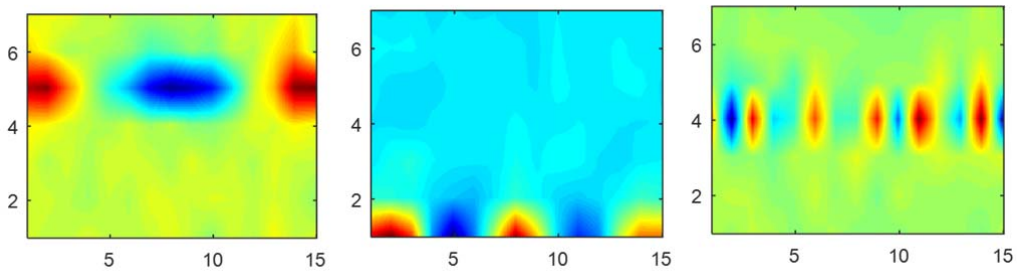

圖 4.5：由左至右為擷取調變頻率變化為 8、16、32Hz 的卷積核

然而，除了從調變頻率變化上來觀察訓練後得到的卷積核之外，我們也可以發現有些卷積核的區域能量特別的強，如下圖(4.6)所示，這表示此卷積核除了包含 3.5 個波型，也就是代表擷取調變頻率變化 23.3Hz 的語音資訊外，其能量呈現的方式則是代表著擷取語音資訊能量較大的共振峰部分。

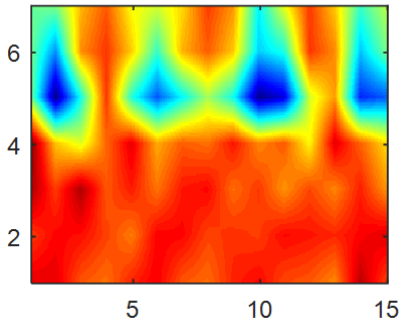

圖 4.6：擷取調變頻率變化為 23.3Hz 以及語音共振峰的卷積核

III　第五種模型 Bothrand 的結果討論

根據表(4.2)，我們可以發現 Bothrand 模型的表現不如其他有給定初始形狀的模型，我們猜想可能原因是 Bothrand 模型的卷積核可能還需要較長的時間或較多的資料才能訓練出更有效果的形狀，在此，我們僅就現階段的結果進行說明。

下圖(4.7)我們可以看到 Bothrand 模型的一維卷積核頻率振幅響應圖，不同於先前的比較模型，Bothrand 因為是隨機給定初始值而直接進行訓練，因此其並沒有像我們先前用來給定初始值的 gammatone 濾波器有依照中心頻率來排定大小順序，因而呈現出一組沒有規則的濾波器組。但我們可以從該頻率響應圖中發現，其對於不同頻率仍會有不同的解析效果，就如同 gammatone 濾波器在低頻時解析較為精細，而高頻時解析則較差。

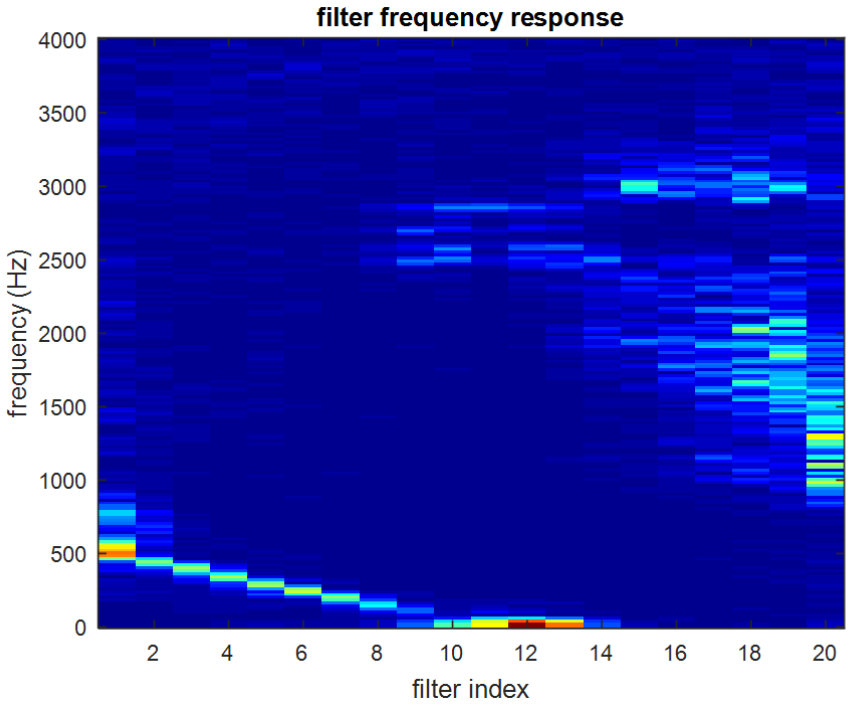

圖 4.7： Bothrand 模型的一維卷積核頻率振幅響應圖

並且因為其一維卷積核並沒有依照中心頻率高低順序而排列，故所得到的結果並非我們所理解的聽覺頻譜圖，因此從圖(4.8)中我們可以看到，Bothrand 所產生的二維卷積核內容顯得較為雜亂，所以要從中判讀出任何有關語音意義的資訊是非常困難的。

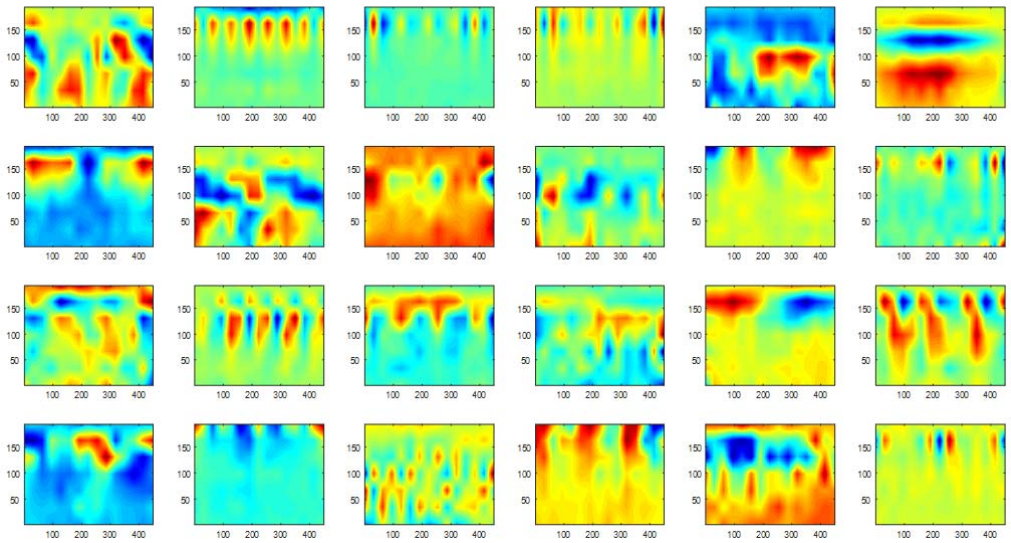

圖 4.8： Bothrand 模型的二維卷積核形狀結果圖

5. 結論與未來展望

在本論文裡，我們提出了一個基於兩階段之聽覺感知模型之類神經網路的模型，並將其應

用來辨識語者。我們透過給予具有其物理意義的兩階段卷積層之卷積核初始值，再利用類神經網路前饋以及反向傳播演算法(feed-forward and back-propagation)進行訓練，並根據語者識別的目標來改善模型的性能表現。而我們也可以透過經過訓練後而調整的卷積核發現，無論是在第一階段的耳蝸分頻亦或是第二階段的大腦皮質階段，我們皆可以透過其訓練調整後的卷積核形狀，進行判讀與分析。這種透過輸入原始訊號(raw data)的架構理念，也許可以和以聽覺科學作為基礎的參數系統做比較，同樣的，我們也可以透過初始化卷積核，來使模型在相同的時間條件中或者較少的資料量下，其表現優於不給予任何初始值的模型，這代表著即使在較為嚴苛情況下，我們也可以透過給予卷積核初始值，使其朝著這個方向進行微調修正，來達到較好的收斂結果。

人類的聽覺感知系統，並非只用於單一一種目標，而近年來，有許多透過卷積神經網路(CNN)成功地應用於自動語音識別(automatic speech recognition , ASR) [4][29][30]等等議題上的例子， 因此我們希望，未來能發展一套基於感知聽覺模型並且同時應用於多種目標的架構，例如：同時應用於語音辨識及語音增強。而在此架構底下，該模型能夠隨著目標的改變進行本身參數的微調，來達到相對於其應用之較好的狀態。

6. 參考資料

[1] Khan Suhail Ahmad, Anil S. Thosar, Jagannath H. Nirmal, and Vinay S. Pande, "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network," in *Proc. of Advances in Pattern Recognition (ICAPR)*, pp. 1-6, 2015.

[2] Yi Wang, and Bob Lawlor, "Speaker recognition based on MFCC and BP neural networks," in *Proc. of Signals and Systems Conference (ISSC)*, pp. 1-4, 2017.

[3] Xiaojia Zhao, Yuxuan Wang, and DeLiang Wang, "Deep neural networks for cochannel speaker identification," in *Proc. of ICASSP*, pp. 4824-4828, 2015.

[4] Yedid Hoshen, Ron J. Weiss, and Kevin W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Proc. of ICASSP*, pp. 4624-4628, 2015.

[5] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *Proc. of ICASSP*, pp. 421-425, 2017.

[6] Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. of INTERSPEECH,* pp. 1-5, 2015.

[7] 張斌. *耳鼻喉科學*, 正中書局, 台北 (1996).

[8] Andrew Morris, Jean-Luc Schwartz, and Pierre Escudier, "An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram," *Computer Speech & Language* 7.2: 121-136, 1993

[9] Larry E. Humes, and Lisa Roberts, "Speech-recognition difficulties of the hearing-impaired elderly: The contributions of audibility," *Journal of Speech, Language, and Hearing Research,* 33.4: 726-735, 1990.

[10] Brian C. J. Moore, "Perceptual consequences of cochlear hearing loss and their implications for the

design of hearing aids," *Ear and hearing,* 17.2: 133-161, 1996.
[11] T. Chi, P. Ru, and S. A. Shamma, “Multi-resolution spectro- temporal analysis of complex sounds,” *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.
[12] M. Elhilali, T. Chi, and S. Shamma, “A spectro-temporal modulation index (stmi) for assessment of speech intelligibility,” *Speech Communication*, pp. 331–348, 2003.
[13] T.-S. Chi, T.-H. Lin, and C.-C. Hsu, “Spectro-temporal modulation energy based mask for robust speaker identification,” *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. EL368–EL374, 2012.
[14] T. E. Lin, C. C. Hsu, Y. C. Chen, J. H. Chen, and T. S. Chi, “Spectro-temporal modulation based singing detection combined with pitch-based grouping for singing voice separation,” in *Proc. of INTERSPEECH.*, pp. 2920–2923, 2013.
[15] F. Yen, Y.-J. Luo, and T.-S. Chi, “Singing voice separation using spectro-temporal modulation features,” in *Proc. of Annual Conference of International Society for Music Information Retrieval (ISMIR)*, pp. 617–622, 2014.
[16] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, “Auditory attention-focusing the searchlight on sound,” *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437– 455, 2007.
[17] E. R. Hafter, A. Sarampalis, and P. Loui, “Auditory attention and filters,” *Auditory perception of sound sources*, Springer US, pp. 115–142, 2008.
[18] M. Elhilali, J. Fritz, T. Chi, and S. Shamma, “Auditory cortical receptive fields: Stable entities with plastic abilities,” *J. Neuroscience*, vol. 27, no. 39, pp. 10 372– 10 382, 2007.
[19] Z. Q. Wang and D. Wang, “A joint training framework for robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.
[20] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, “Learning acoustic frame labeling for speech recognition with recurrent neural networks,” in *Proc. of ICASSP*, pp. 4280–4284, 2015.
[21] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.
[22] L.-Y. Yeh and T.-S. Chi, “Spectro-temporal modulations for robust speech emotion recognition,” in *Proc. of INTERSPEECH*, pp. 789–792, 2010.
[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, pp. 1097–1105, 2012.
[24] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” In *Proc. of International Conference on Artificial Neural Networks*, pp. 52–59, 2011.
[25] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks.” in *Proc. of ICASSP*, pp. 4580–4584, 2015.
[26] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition.” in *Proc. of ICASSP*, pp. 4277–4280, 2012.
[27] Jing Chen, Thomas Baer, and Brian CJ Moore. "Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired." *J. Acoust. Soc. Am.,* 131.4: 2987-2998, 2012
[28] Tai-Shih Chi, and Chung-Chien Hsu. "Multiband analysis and synthesis of spectro-temporal modulations

of Fourier spectrogram." *J. Acoust. Soc. Am.,* 129.5: EL190-EL196, 2011.
[29] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent, Y. Bengio, and A. Courville, “Towards end-to-end speech recognition with deep convolutional neural networks.” in *Proc. of INTERSPEECH*, pp. 410– 414, 2016.
[30] Z.-Q. Wang and D. Wang., “Robust speech recognition from ratio masks.” in *Proc. of ICASSP*, pp. 5720–5724, 2016.