國 立 交 通 大 學

電信工程研究所

碩 士 論 文

基於深度學習之降噪演算法應用於

可保留聲響聽覺之人工電子耳

Deep Learning Based Noise Reduction

for Acoustic Hearing Preserved Cochlear Implant

研 究 生：吳宗振　　Student: Tsung-Chen Wu

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

中華民國一百零六年八月

基於深度學習之降噪演算法應用於
可保留聲響聽覺之人工電子耳

Deep Learning Based Noise Reduction
for Acoustic Hearing Preserved Cochlear Implant

研 究 生：吳宗振　　Student: Tsung-Chen Wu

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

國立交通大學

電信工程研究所

碩士論文

A Thesis
Submitted to Institute 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
August 2017
Hsin-Chu, Taiwan, Republic of China

中華民國一百零六年八月

基於深度學習之降噪演算法應用於
可保留聲響聽覺之人工電子耳

學生：吳宗振　　指導教授：冀泰石 博士

國立交通大學電信工程研究所

感知訊號處理實驗室

摘要

聲碼器模擬是普遍被用來模擬人工耳蝸所產生的電訊號，聲碼器處理過後的信號讓正常聽力者聆聽時，可以臨摹聽損患者所感受到的聲音。我們計劃團隊正在研發一項聽覺聽力保留的高頻四電極點人工電子耳系統，為了要事先評估此系統的功能性，本論文利用了實驗室先前開發模擬聽損人士的個人化聽損模型，結合四通道的高頻聲碼器，分別模擬聽覺聽力保留的聲響聽覺和人工耳蝸信號處理的電聽覺，讓正常聽力者聽到聽覺保留之高頻聲碼模擬器模擬的聲音。透過中文聽辨度的心理聲學實驗，我們找尋到此人工耳蝸的最佳頻率分布方式。經過不管在乾淨語音或噪音背景下的實驗，結果顯示中文的聲母比韻母更加難以被聽損患者辨識，增加了此系統的電子聽覺，中文的聽辨度被顯著地提升了，尤其是聲母更為明顯。接著結合深度學習方式，對於生活環境中常見的噪音進行降噪演算法，討論在模型沒訓練過的噪音之下降噪，進而增加整體的中文聽辨度，並且用客觀評分標準與主觀聽辨度測量來驗證此模型。

Deep Learning Based Noise Reduction for Acoustic Hearing Preserved Cochlear Implant

Student: Tsung-Chen Wu Advisor: Dr. Tai-Shih Chi

Institute of Communication Engineering

National Chiao-Tung University

Perception Signal Processing Laboratory

Abstract

Vocoder simulations are generally adopted to simulate the electrical hearing induced by the cochlear implant (CI). Our research group is developing a new four-electrode CI microsystem that induces high-frequency electrical hearing while preserving low-frequency acoustic hearing. To assess the functionality of this CI, a previously developed hearing-impaired (HI) hearing model is combined with a 4-channel vocoder in this thesis to respectively mimic the perceived acoustic hearing and electrical hearing. Psychoacoustic experiments are conducted on Mandarin speech recognition for determining spectral coverages of electrodes for this CI. Simulation results show that initial consonants of Mandarin are more difficult to recognize than final vowels of Mandarin via acoustic hearing of HI patients. After electrical hearing being induced through logarithmic-frequency distributed electrodes, speech intelligibility of HI patients is boosted for all Mandarin phonemes, especially for initial consonants. Similar results are consistently observed in clean and noisy test conditions. Next, we combine a deep neural network based noise reduction algorithm with the proposed CI system in the hope to improve the Mandarin speech intelligibility for seen and unseen noise types. Ultimately, we use objective evaluation and subjective evaluation scores to verify this model, hence, to provide the proof of concept of this combinational system.

誌謝

研究所這兩年裡，對我來說既是時光飛逝，抑是度日如年，在實驗室中的生活我真的像是在社會中得到了預習，許多人生的體驗都在此發生，然而這都是造就了現在的我，我非常感謝在這實驗室裡面發生的點點滴滴。

若要感謝，我一定第一個想到冀泰石老師，至上無比的敬意和謝意給冀泰石老師，老師不僅花費很多耐心在每次的研究討論中，也在二年級參與計畫時提供很好的建議，對於我的研究是一盞耀眼的明燈，口試時也會盡所有可能的幫我補充，填補我的失誤；另外，老師也會很關心我的生活，給予我很大的自由，並且督促我，讓我有機會去國外參加會議，對於我的人生這一定是一項相當棒的經驗，冀泰石老師，是我在交大的六年裡面，最感謝的一位老師。

在實驗室中的各位夥伴們，黃群、玉雯、周歆，這一句「畢業之路有你有我」，在我焦頭爛額的最後幾個月裡，對於我心中給予我很大的鼓勵，也有相當大的感動，知道實驗室我不是孤單的。學長姐們對於研究的開導也有很大的幫助，尤其是劉兆倫學長、陳祖昊學長、許凱鈞學長、黃茂彰學長和蔡佩均學姊，有你們耐心的指點，我才可以走到今天的這一步，對了還有賴貞延學姊。學弟妹們，別說啦，有你們陪我打屁抓寶打嚕談心，真的是很開心，你們來向我詢問很多事情，其實我本來就是一位喜歡教導的人，我回答你們時很愉快，最後還是點名一下好了，文成、曜雲、明佐、姿羽、翊瑄和星瑋。

絕不能忘記讓我擁有最強後背的家人，有你們真的很好，家族裡就我最晚開始賺錢，但你們也願意給我很充裕的經濟支柱，讓我無後顧之憂的在研究上面，而你們在研究所這段期間對我的恩情，這只是冰山一角，永遠的感謝，永遠。最後，我的女朋友，函均，若是沒有妳的陪伴沒有妳的鼓勵，很多時間我根本撐不下去，甚至想放棄，你的純真確確實實得，提醒我在這社會上走的道路是正確的，我們的路，將會繼續一直走下去。

七一一實驗室，我要離開了，各位再見，不過我還是會回來的，我新竹人很近，公司也很近哈哈。

目　　錄

中文摘要 i
英文摘要 ii
誌　　謝 ii
目　　錄 iv
表 目 錄 vi
圖 目 錄 vii
第一章 緒論 1
1.1 研究背景 1
1.2 研究方法 2
1.3 章節大綱 2
第二章 感知訊號處理基礎 3
2.1 生理聽覺現象與特性 3
2.1.1 聽覺的產生 3
2.1.2 響度 5
2.2 聽損現象簡介 7
2.2.1 最小可聽水平與響度聚集 7
2.2.2 分頻解析度降低 8
2.3 短時間傅立葉轉換 9
第三章 個人化聽損模型與聲碼器 12
3.1 響度模型 13
3.2 頻譜模糊化模型 13
3.2.1 濾波器變寬程度計算 14
3.2.2 模糊化計算以及各頻帶附載上額外生成成分 15
3.2.3 載波計算 16
3.3 等響度曲線增益 17
3.4 個人化聽損模型 17
3.5 聲碼器 18
第四章 基於個人化聽損模型的聲碼器模擬以測量主觀聽辨度 20
4.1 系統架構圖 20
4.2 個人化聽損模型參數選取 21
4.3 心理聲學實驗方法 22
4.3.1 中文語音資料庫 23
4.3.2 心理聲學實驗受試者 24
4.3.3 語料計分方式 24
4.4 主觀中文聽辨度測量(一)：帶通濾波器的頻率覆蓋範圍 24

4.4.1 實驗條件 25
4.4.2 實驗結果與討論 26
4.5 主觀中文理解度測量(二)：噪音下聲母與韻母的聽辨度提升 28
4.5.1 實驗條件 28
4.5.2 實驗結果與討論 29
第五章 基於深度學習降噪演算法提升語音聽辨度 33
5.1 深度神經網路學習 33
5.1.1 背景 33
5.1.2 神經網路架構系統 33
5.2 心理聲學實驗方法 35
5.2.1 心理聲學實驗受試者 35
5.2.2 中文語音資料庫 35
5.2.3 語料計分方式 37
5.3 客觀中文聽辨度測量 37
5.3.1 短時客觀聽辨度(STOI) 37
5.3.2 實驗條件 38
5.3.3 實驗測量結果 39
5.4 主觀中文聽辨度測量 41
5.4.1 實驗條件 41
5.4.2 實驗測量結果與討論 41
第六章 結論與未來展望 47
參考文獻 49

表 目 錄

表 2-1：聽力受損程度分級 7
表 4-1：受試聽損病患的各頻率最小可聽水平 21
表 4-2：利用 NOTCH-NOISE METHOD 所測出來的變寬因子 22
表 4-3：使用 SUB2 之帶通濾波器頻率覆蓋範圍的變異數分析結果 26
表 4-4：使用 SUB8 之帶通濾波器頻率覆蓋範圍的變異數分析結果 27
表 4-5：使用 SUB2 之語音形狀噪音下聽辨度的三因子變異數分析結果 29
表 4-6：使用 SUB2 之雙語者噪音下聽辨度的三因子變異數分析結果 29
表 4-7：使用 SUB8 之語音形狀噪音下聽辨度的三因子變異數分析結果 29
表 4-8：使用 SUB8 之雙語者噪音下聽辨度的三因子變異數分析結果 30
表 4-9：使用 SUB2 之噪音下聽辨度的變異數分析結果 32
表 4-10：使用 SUB2 之噪音下聽辨度的變異數分析結果 32
表 5-1：基於深度學習降噪演算法提升語音聽辨度實驗的聽損模型參數 41
表 5-2：使用 SUB8 聽損模型參數下語音聽辨度結果 42

圖 目 錄

圖 2-1：耳朵基本構造 3
圖 2-2：中耳及耳蝸 4
圖 2-3：柯替式器 4
圖 2-4：倍頻單音在基底膜的共振位置示意圖 5
圖 2-5：1KHZ 單音的強度-響度對應關係 6
圖 2-6：以 1KHZ 為 基準的等響度曲線圖 6
圖 2-7：正常聽力與受損聽力者的聽覺強度-響度比較圖 8
圖 2-8：受損患者與正常人的聽覺濾波器比較 8
圖 2-9：時域訊號與其 FFT 圖 9
圖 2-10：矩形視窗與其頻率響應 10
圖 2-11：HANN 視窗與其頻率響應 11
圖 3-1：頻譜模糊化演算法流程圖 14
圖 3-2：NOTCHED－NOISE METHOD 實驗所使用的刺激與遮蔽物 15
圖 3-3：原始語音、使用白雜訊載波模糊化後的語音的聲譜圖 16
圖 3-4：人耳可聽到不同頻率及音強的聲音 17
圖 3-5：混合模型流程圖 18
圖 3-6：聲碼器流程圖 19
圖 4-1：基於個人化聽損模型的聲碼器模擬流程圖 20
圖 4-2：心理聲學實驗流程圖 22
圖 4-3：合理反和弦轉換 22
圖 4-4：音素平衡之中文單字聽力測試語料表 23
圖 4-5：(A)對數分布的帶通濾波器頻率響應(B)線性等寬分布的帶通濾波器頻率響應 .25
圖 4-6：低頻對應高頻分布的封包萃取與製造載波之帶通濾波器頻率響應 25
圖 4-7：使用 SUB2 聽損模型參數的平均辨識正確率 26
圖 4-8：使用 SUB8 聽損模型參數的平均辨識正確率 27
圖 4-9：中文字「上，尸尢丶」的語音頻譜圖 28
圖 4-10：使用 SUB2 之噪音下的平均辨識正確率 31
圖 4-11：使用 SUB8 之噪音下的平均辨識正確率 31
圖 5-1：以深層神經網路降噪的流程圖 34
圖 5-2：深層神經網路(DNN)的模型架構 34
圖 5-3：結合降噪演算法之實驗流程 35
圖 5-4：所有 240 句測試句子音調比例分布圖 36
圖 5-5：所有句子組別在安靜及吵雜環境的 RTS 平均值和標準差 36
圖 5-6：雞尾酒宴會噪音下的客觀語音聽辨度 39
圖 5-7：雞尾酒宴會噪音 SNR -3 DB 下降噪前後的對數頻譜圖 39

圖 5- 8：語音形狀噪音下的客觀語音聽辨度 40
圖 5- 9：語音形狀噪音 SNR -3 dB 下降噪前後的對數頻譜圖 40
圖 5- 10：經過 Sub 8 個人化聽損模型的對數頻譜 42
圖 5- 11：使用假設患者聽損模型參數的雞尾酒宴會噪音底下之平均辨識正確率 43
圖 5- 12：使用假設患者聽損模型參數的語音形狀噪音底下之平均辨識正確率 44
圖 5- 13：十種噪音訓練語料 DNN 模型下的 SSN 客觀聽辨度結果 45
圖 5- 14：SSN 情況經過兩種不同模型降噪後的降噪語音對數頻譜圖 45
圖 5- 15：十種噪音訓練語料 DNN 模型下的 SSN -5 dB 平均辨識正確率 46

第一章 緒論

1.1 研究背景

耳朵是人類重要的感知器官，正常的聽力讓我們能無礙地接受身邊的訊息，察覺環境中的突發狀況，最重要的是也能使我們有效的和人們溝通，幫助語言的學習，融入人類的社會。隨著高齡社會的來臨，老年人聽力退化，或長期暴露於高噪音環境之下、受到突發性巨大聲響等等因素而造成聽力受損，不僅會帶來環境訊息接收上的困難，在人際溝通上，要求重複的口述，或是誤聽訊息更會造成人際關係的疲勞和隔閡，因此聽力受損問題已備受重視。

人工電子耳是現今常見的侵入式醫療器材，利用電極點刺激聽神經，已經成功地幫助嚴重聽力受損的病患，然而傳統的人工電子耳是將電極點刺穿軟圓窗進入耳蝸內，此作法經醫學證實會有失去原有聽覺聽力和罹患腦膜炎的風險[1]，因此就有發展可保留聲響聽覺之高頻人工電子耳的想法，為將四個電極點放置耳蝸外骨的表面上，因軟圓窗沒有被刺穿，所以可以在保留原有的聲響聽覺條件下提供電聽覺。此人工電子耳可適用於高頻受損的聽損者和聽力衰退的老年人。

常見預測人工電子耳對聽損病患語音辨識度效果的方式，是利用聲碼器模擬(vocoder-based simulation)來臨摹病患所產生的電聽覺[7-9]，正常聽力者會聆聽聲碼器輸出的合成聲音，並且辨識語音來進行評測效果，模擬的結果能在真正做病患實驗之前，提供醫生重要的預先評估資訊。其中，噪音聲碼器(noise vocoder)常常被使用來模擬病患的電聽覺[11-12]，我們在此也將使用噪音聲碼器來評估此「聲響聽覺保留之高頻人工電子耳」的效能。為了能同時模擬聽損患者原本的聲響聽覺，我們將結合實驗室先前開發的「模擬聽損病患的個人化聽損模型」[10]以模擬患者原來的聽覺。

在人工電子耳議題中，為了達到更高的聽辨度，常見的一項功能為消除雜訊(又稱降噪，noise reduction)。在近年來，用深度學習(deep learning)的方式消除雜訊有許多很好的成果[13-14]，所以我們將用此模型結合深層學習的降噪演算法，進行人工電子耳的心理聲學實驗，評測對其聽辨度的助益。

1.2 研究方法

本論文為了評估此新型的可保留聲響聽覺之高頻人工電子耳的可行性，我們進行心理聲學實驗，利用四通道高頻噪音聲碼器，結合實驗室先前開發模擬聽損病患的個人化聽損模型，建立可保留聲響聽覺之高頻聲碼器(acoustic hearing preserved high-frequency vocoder)，簡稱 AHPHFV，進而臨摹新型人工電子耳提供給聽損患者的聲響聽覺(acoustic hearing)和電聽覺(electrical hearing)，並將將聲響聽覺保留之高頻聲碼器輸出的合成音檔給正常聽力者聆聽，最後用中文聽辨度來評測。

其中為了決定電極點放置在耳蝸外的位置，我們利用此模型測試三種不同的頻率分布方式。得到聽辨度最好的頻率分布方式後，為了符合現實生活中的情況，將進行不同雜訊和訊雜比下的中文聽辨度評測，以此來驗證聽覺保留之高頻人工電子耳的想法。

第二部分，我們將利用客觀評量來分析不同深度模型的降噪效果，將效果最佳的深度模型與聲響聽覺保留之高頻聲碼器結合來進行心理聲學實驗，來了解降噪對使用此新式人工電子耳病患的幫助。

1.3 章節大綱

本論文的內容章節如下：

第一章　緒論：用來介紹研究背景、聽損現象簡介、研究方法及各章節概要。

第二章　感知訊號處理基礎：介紹聽覺生理現象、人耳如何解析聲音以及常見聽損現象，故會介紹常見的訊號處理方法。

第三章　個人化聽損模型與聲碼器：介紹聲響聽覺保留之高頻聲碼器架構中的兩大模組。

第四章　聲響聽覺保留之高頻聲碼器的主觀中文聽辨度量測：介紹提出的模型，進行心理聲學實驗，統計分析結果，評估此模型對於提升聽辨度效果。

第五章　利用深度學習降噪演算法提升聽辨度：結合深度學習降噪演算法，提升可保留聲響聽覺人工電子耳使用者在噪音環境底下的聽辨度。

第六章　結論與未來展望。

第二章 感知訊號處理基礎

2.1 生理聽覺現象與特性

2.1.1 聽覺的產生

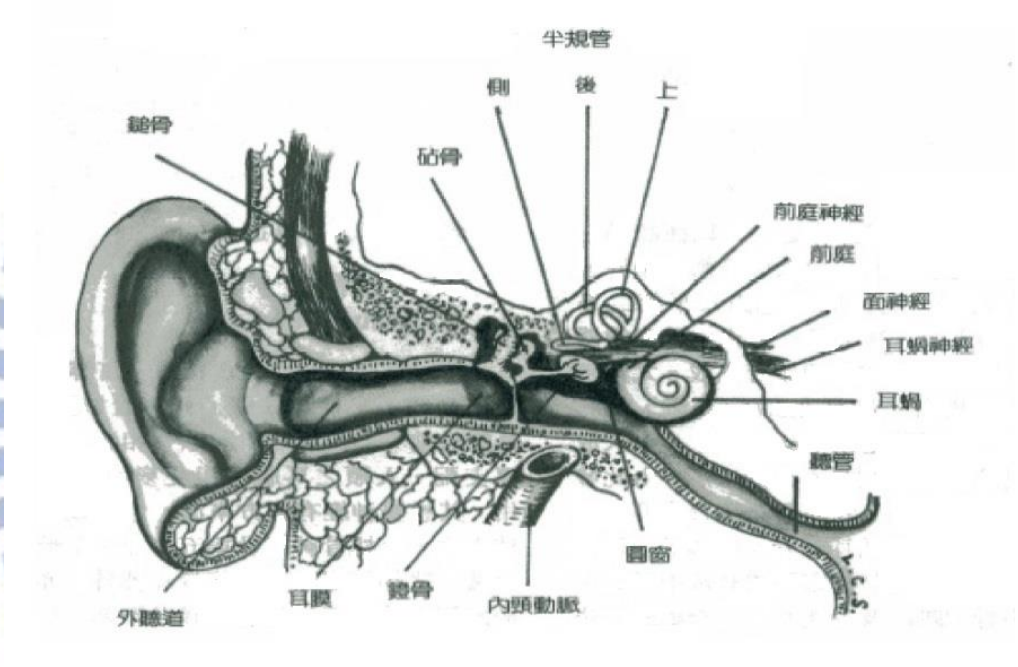


圖 2-1：耳朵基本構造 [16]

人耳的構造包含外耳、中耳及內耳三個部分，如圖 2-1 所示。外耳包含耳殼及外聽道，具有保護、放大、收音和定位的功能，外耳殼負責接收外界的聲波傳遞至外聽道內；中耳則由三小聽骨(鎚骨、砧骨及鐙骨)，將耳膜所接收到的聲波震動，經三小聽骨以槓桿原理放大，推動耳蝸上的卵圓窗；內耳包含前庭、半規管及耳蝸，前兩者主司平衡，耳蝸主司聽覺，在耳蝸上有卵圓窗(oval window)和圓窗(round window)兩個窗口，窗口內

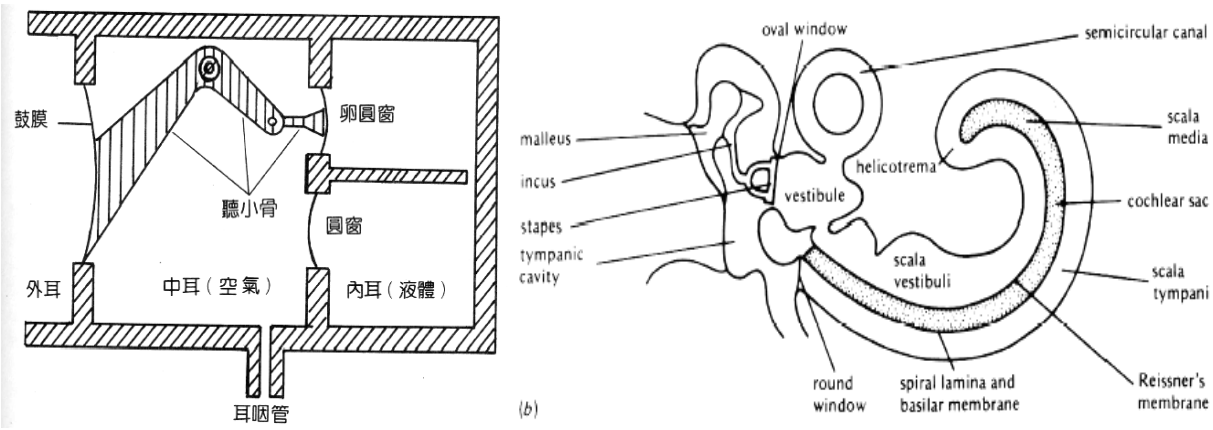


圖 2-2：中耳及耳蝸，左圖為簡易構造圖，右圖為關係模式圖 [16]

淋巴液充滿耳蝸空腔組織，空腔組織中間由基底膜(basilar membrance)隔開，如圖 2-2 所示。當卵圓窗被推動時，使得淋巴液有所流動而於基底膜上會產生行進波(travelling wave)，機械能轉換為液態能，基底膜上的柯替氏器(organ of Corti)上分布數著以千計的內毛細胞(inner hair cell)與上萬個外毛細胞(outer hair cell)，如圖 2-3，行進波擠壓毛細胞而造成離子濃度差異而進出神經細胞，當電離子電位大於某個臨界值後開始放電，此過程可將行進波的訊號轉為電訊號而傳給聽覺神經。

基底膜的共振範圍大約為 20Hz~20kHz，即為正常人的聽力範圍。越高頻的聲音在基底膜共振的位置越靠近基底部(base)，因此高頻的行進波會在近 base 處產生較大的振幅，反之低頻所產生共振的位置就越靠近深處頂部(apex)，因此低頻的行進波會在近 apex 處產生較大的振幅，如圖 2-4 所示。負責解析高頻的耳蝸基底部分，因為最靠近內耳的門戶-卵圓窗，一旦內耳受損，基底首當其衝，導致高頻的毛髮細胞(hair cells)容易受損這也就是為什麼退化型的聽力受損會從高頻先開始。

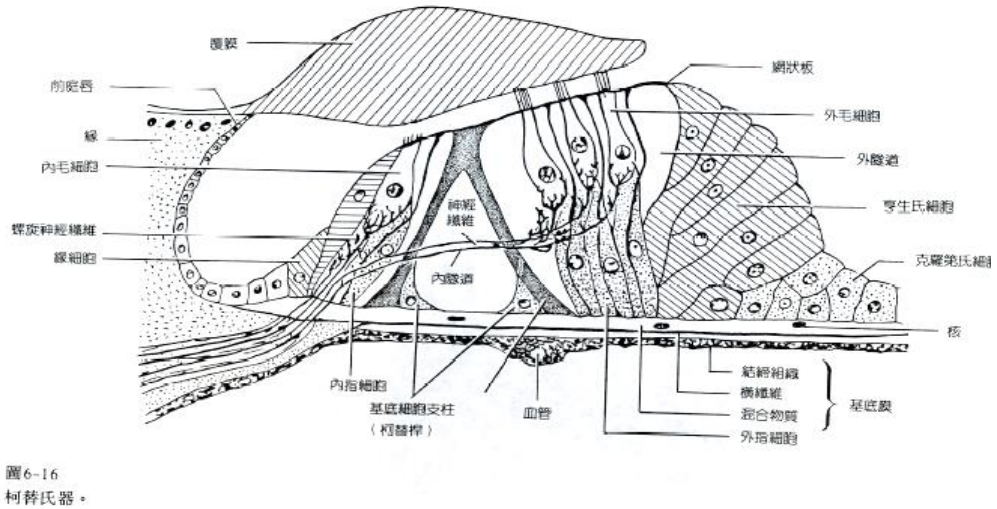


圖6-16
柯替氏器。

圖 2-3：柯替式器 [16]

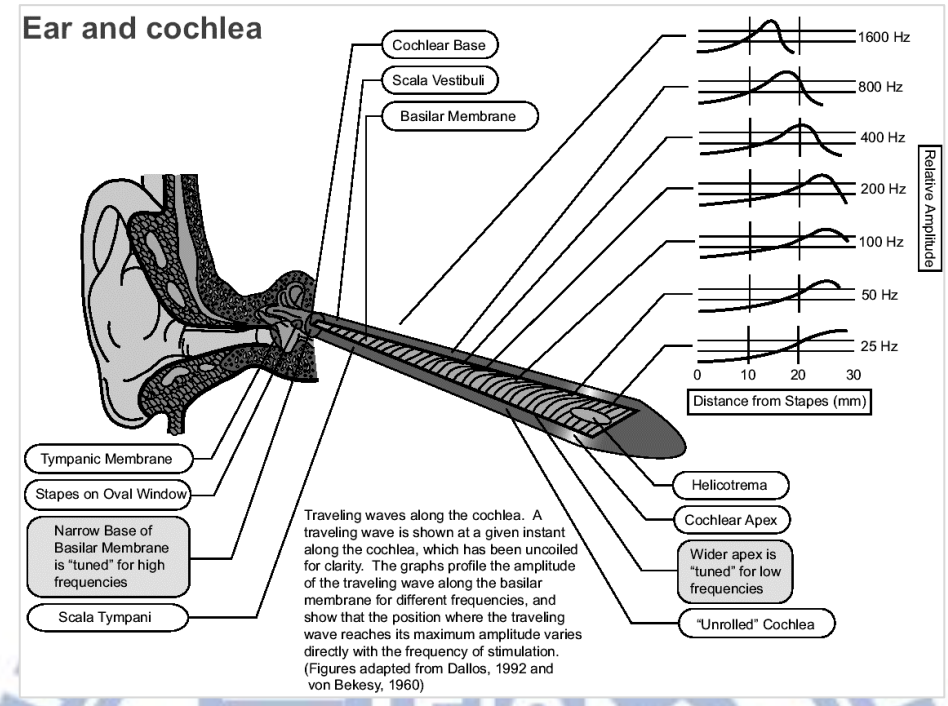


圖 2-4：倍頻單音在基底膜的共振位置示意圖 [16]

2.1.2 響度

1. 響度(loudness)與強度(intensity)

聲音強度(sound intensity)指的是每單位時間單位面積中經過的聲音能量，常用來表示聲音的大小，衡量聲音強度的數值稱音壓值(sound pressure level)，其單位使用對數形式的分貝(dB)。響度(loudness)指的是在主觀下表示聲音大小，與頻率有關係，響度值(sound level)有兩種：方(phon)與宋(sone)。方等同於單頻 1 kHz 聲音時的音壓值，每上升 10 方，能量會上升兩倍；宋可以表示測量響度的數量關係，1 宋等於頻率為 1 kHz、音壓值為 40dB 的純音的響度。意旨每相差 10 方相當於相差 2 倍宋，例如響度為 2 宋的聲音比 1 宋的聲音響兩倍。彼此的換算為：

$$\text{phon} = 40 + 10\log_2(\ \text{sone}\ ) \qquad (2\text{-}1)$$

強度與響度之間的關係可由下式近似[15]：.

$$\text{loudness in sone} \propto (\ \text{sound intensity in dB}\ )^{1/3} \qquad (2\text{-}2)$$

將式子 2-1 畫出來，即可得到圖 2-5，圖中縱軸的響度單位呈現對數分布，而強度大小亦呈現線性分布(dB)，故響度的對數變化與聲音強度成正比關係。

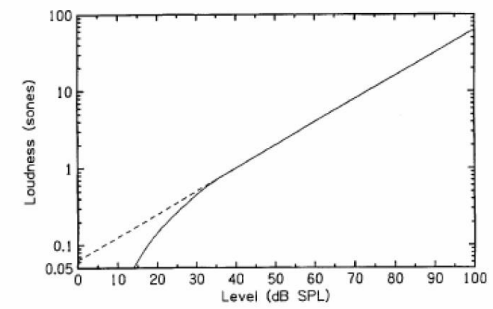


圖 2-5　1kHz 單音的強度-響度對應關係[15]：實線為實際測量，虛線為公式(2-1)

2. 最小可聽水平和最大可聽水平

最小可聽水平(minimum audible level, MAL)是指人耳能聽到不同頻率的單頻聲音所需的最小聲音強度；反之，最大可聽水平（maximum audible level）則是人耳對於不同頻率的單頻聲音能承受的最大強度，超過此強度後聽力會受損。在不同頻率的最小可聽水平所連接而來的臨界曲線稱作聽力閾值(hearing threshold)，如圖 2-7 所示。

3. 等響度曲線(equal loudness curve：ELC)

以 1kHz 的單頻率聲音為參考基準，在不同頻率下，聽覺上的響度與 1kHz 時的響度一樣時，對應的音壓值連成一曲線，這就是等響度曲線(equal-loudness contours)，如圖 2-6 所示。

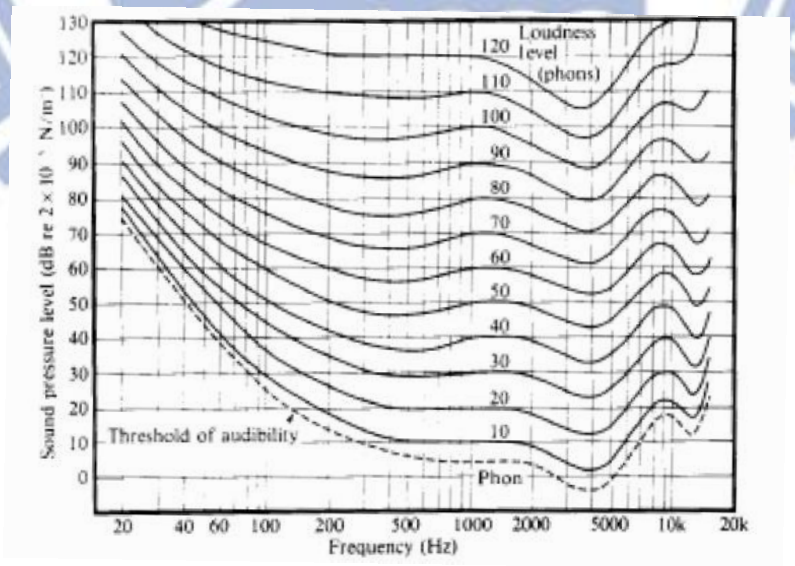


圖 2-6 以 1kHz 為基準的等響度曲線圖 [15]

2.2 聽損現象簡介

2.2.1 最小可聽水平與響度聚集

聽障者常見的問題就是最小可聽水平的提升，常見的聽力檢查會測量受試者 250、500、1k、2k、4k、8k Hz 的單頻聽力閾值(pure tone threshold)，利用平均純音聽力閾值(pure tone threshold average, PTA)：

$$\frac{0.5k\text{的閾值}+1k\text{的閾值}+2k\ Hz\text{的閾值}}{3} \tag{2-3}$$

式子(2-3)的計算能得知受試者整體的聽力受損狀況，平均聽力閾值的受損程度可歸納如表 2-1。

聽力受損程度	閾值(dB)
輕度	26-40
中度	41-55
中重度	56-70
重度	71-90
極重度	>90

表 2-1　聽力受損程度分級

資料來源：[2]

然而最小可聽水平的提升，時常伴隨者響度聚集(Loudness recruitment)的發生。因為聽損患者與正常聽力者的最小可聽水平會對應到相同的響度(Loudness)，最大可聽水平亦然，亦指聽力區間變小，結果聽障患者的「聽覺響度－物理強度」圖的斜率會比正常人的斜率來的大，如圖 2-8 所示，此即為響度聚集現象。

圖 2-7 的範例中，下方曲線和上方曲線為聽損患者和正常聽力者。根據下方曲線，當強度為 30dB 時響度為 0.001Sones(最小可聽水平)，當物理強度上升，會越靠近正常聽力者的曲線。然而斜率越大，造成的聲音響度變化會放大，造成聽損患者的不適和解讀語句的困難[3-4]。

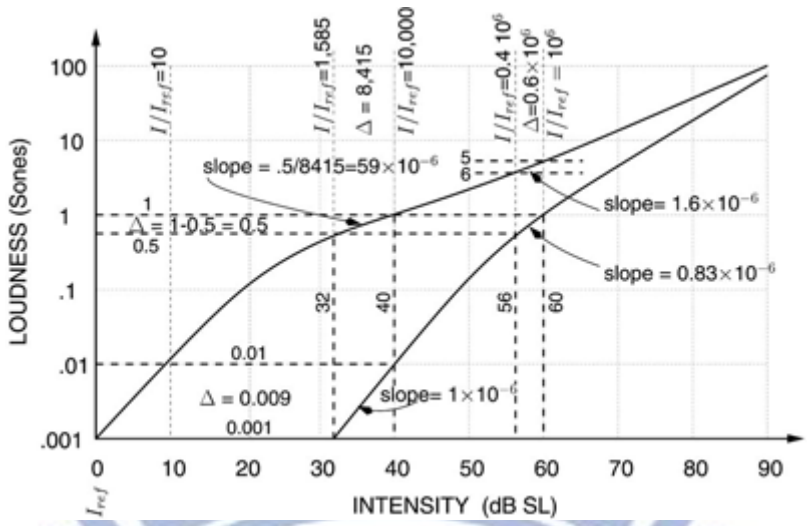


圖 2-7 正常聽力與受損聽力者的聽覺強度-響度比較圖

資料來源：MITCogNet

2.2.2 頻率析度降低

聲音在耳蝸內傳遞，基底膜會產生對應的行進波，不同頻率的聲音會讓不同地方的內毛細胞(inner hair cell)產生反應，換句話說，我們可以將耳蝸近似為一組帶通濾波器(band pass filter banks)。但是當耳蝸老化、受損，內毛細胞對於頻率的敏感度降低，聽覺上對於頻率的解析度也就下降，舉例來說，某聲音對正常聽力者會影響 800Hz 至 1200Hz 的內毛細胞，因為聽損患者頻率敏感度下降的關係，可能會影響到 500Hz 至 1500Hz 的內毛細胞，也就是聽覺濾波器的頻寬變寬，如圖 2-8 所示。如此會造成聲音的音色改變，聽起來的聲音會變得模糊。

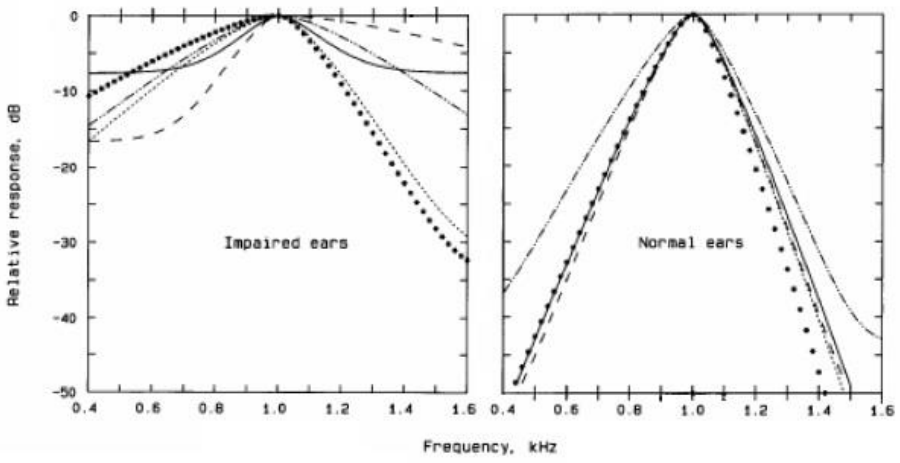


圖 2-8 受損患者與正常人的聽覺濾波器比較 [5]

2.3 短時間傅立葉轉換

在語音訊號處理中，傅立葉轉換(Fourier Transform：FT)是常見的分析訊號的方式，可將訊號的時域轉換到頻域，觀察訊號的頻率成分分布，利用聲音擁有不同的頻譜特性進行分類、擷取、增強等動作。然而，大部分我們想分析的訊號卻不能直接使用傅立葉轉換來進行分析，例如語音訊號，因為使用傅立葉轉換的基本假設是訊號為非時變訊號，但是現實中大多數的訊號都是會隨著時間改變的時變訊號，因此傅立葉轉換是一項不適合的工具。

以圖 2-9 為例：(a)為一非時變訊號，由 5 Hz、15 Hz、20 Hz 弦波所組合時間長度為 1.5 秒，(b)為此訊號傅立葉轉換後的頻譜圖，在 5 Hz、15 Hz、20 Hz 的頻率位置上均有能量產生，代表清楚地解析出這三個頻率成分。(c)為一時變訊號，前 0.5 秒為 5 Hz 弦波、0.5 至 1 秒為 15 Hz 弦波、1 至 1.5 秒為 20 Hz 的弦波，(d)為此訊號做傅立葉轉換所得之頻譜圖，依然可在 5 Hz、15 Hz、20 Hz 的頻率上看到較大的能量值，但無法看到訊號隨著時間，每過 0.5 秒而改變的現象。由此例子得知，利用傅立葉轉換來分析訊號會因為缺少時間變化的資訊而無法一窺全貌。

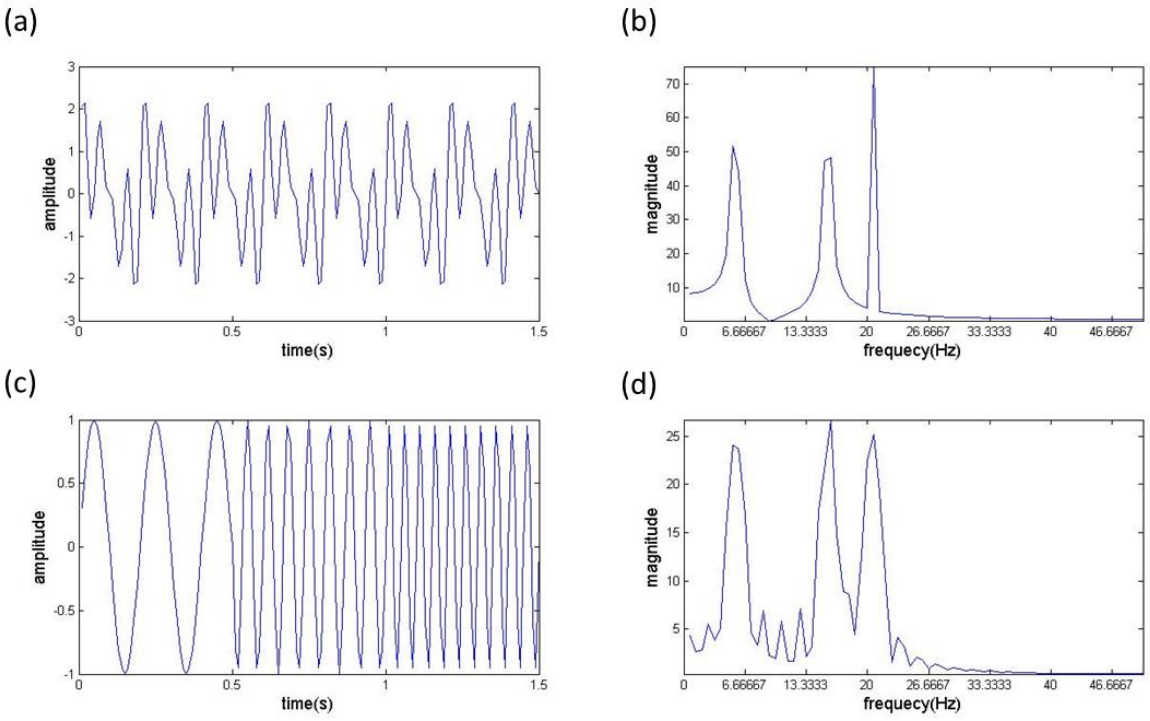


圖 2-9 時域訊號與其 FFT 圖

以此例子說明了在日常生活中的語音為時變訊號，在不同的時間區域有不同的周期和振幅大小，若利用傅立葉轉換分析長時間的語音，只能求出整段時間的平均頻率組合成分，也無法得知頻率組成成分隨著時間變化的內容。因為時間資訊的遺失，也會造成我們無法做傅立葉逆轉換還原得到原本的時域訊號，這也是另外一個我們無法用傅立葉轉換對時變訊號進行分析原因。

所以，在分析像是語音的時變訊號時，為了保留瞬時頻譜的資訊，我們會對訊號進行短時間傅立葉轉換(short-term Fourier transform：STFT)，其作法是將一長時段的語音訊號，用固定長度的視窗(window)，將此語音訊號分拆成若干個音框(frame)，並且假設在這短時音框內的訊號為非時變訊號，我們就能在每個音框內進行個別分析，若是音框太長，會無法捕捉語音隨時間變化的特性，而音框太短，又難以正確捕捉語音頻率的特性，一般而言，音框長度大約為 20~30 ms，足以抓出語音的頻率特徵，為了減少運算量，我們會對每個音框使用快速傅立葉轉換(fast Fourier transform：FFT)。

視窗的選取是重要的議題，若我們直接進行切割音框，等同於在時域上將原訊號乘上矩形視窗(rectangular window，圖 2-10)，則在頻域上原訊號頻譜會與 sinc 函數做摺積，因為 sinc 函數的主瓣(main lobe)雖然很集中，但旁瓣(side lobe)也很高，做摺積的結果將原訊號頻譜模糊化，並於高頻產生細微雜訊。最理想的視窗就是其時域和頻域上的解析度都要高，但事實上這兩者在傅立葉分析架構下是相互抵觸，故使用者只能就其想要的功能選擇視窗。

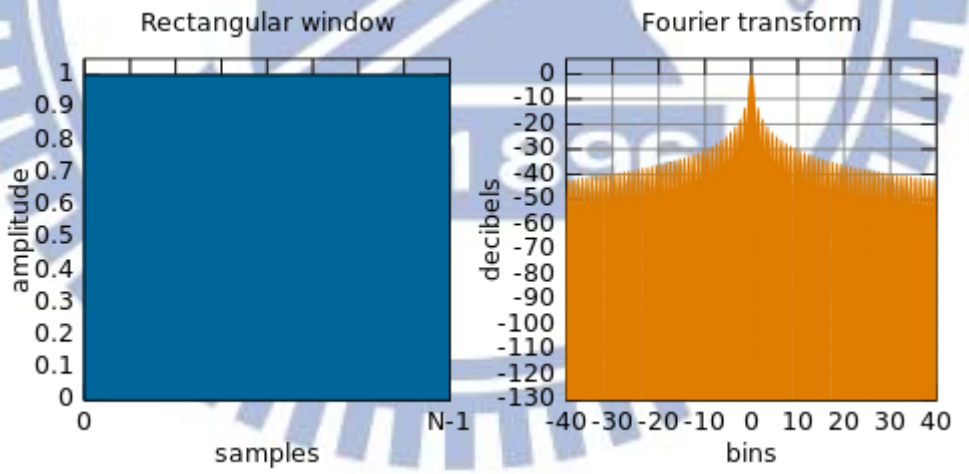


圖 2-10 矩形視窗與其頻率響應 [17]

另外，考慮前一個音框的後幾點與後一個音框的前幾點彼此有關連性和延續性，通常我們會選用 Hann 窗(Hann window，圖 2-11)來進行音框拆解，Hann 窗是利用升餘弦函數(raised cosine function)，兩邊端點會降至零，為了完美的還原短時間傅立葉轉換的訊號，音框間有區塊重疊(overlap)，並且重疊時間為音框的 1/2，如此各個時間平移的 Hann 窗加總起來會得到常數 1。Hann 窗其好處為頻率響應的主瓣(main lobe)寬度頗

窄，副瓣(side lobe)高度很低，所以可消除音框兩邊的不連續且降低音框內訊號頻譜模糊化的程度。

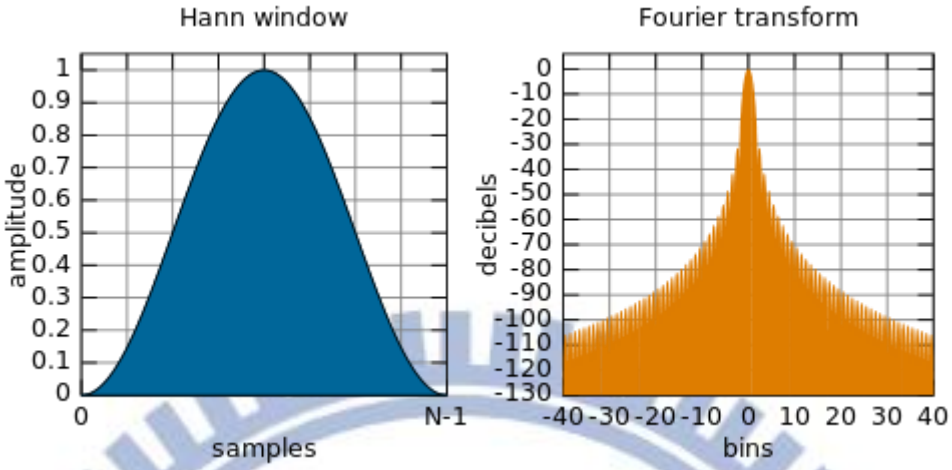


圖 2- 11 Hann 視窗與其頻率響應 [17]

第三章 個人化聽損模型與聲碼器

當要研究開發針對聽損者的語音增強演算法時，基於每位聽損者的情況大不相同，所以歸納出基本的發展步驟應為：

1. 針對某類型聽損狀況開發出演算法，提出理論上可達到的效果並且保留許多可調參數，開發過程中必須不斷實際測試於病患，測試出對群體真正有用的演算法參數。
2. 因應每位聽損者的狀況不相同，事前必須縝密量測患者的聽損情況，再細部調整出最適合患者個人的演算法有效參數組合，來達到語音增強的效果。

這些步驟看似不多且簡單，但是實際操作上會遇到很大的問題，就是對病患來說會需要大量的時間與精力，導致病患參與實測的意願度下降許多，試想為了減輕聽損者的負擔，是否能由正常聽力者代替來受試，因此本實驗室之前開發出個人聽損模型[18-19]，想法在於：「正常語音經過個人聽損模型處理過後的聲音，使正常聽力者聆聽，相當於聽損患者聆聽正常語音而感知到的聲音。」如此，將此模型處理過的聲音讓正常聽力者聆聽，我們就可以減少研究演算法所需要的時間，有助於更有效率的開發演算法。

在本論文中，個人化聽損模型是很重要的關鍵，用來模擬「聽覺保留之高頻人工電子耳」中聽力保留的狀況，關於此個人化聽損模型，我們考慮聽力閾值上升導致響度聚集，以及因聽覺濾波器變寬所造成語音模糊化兩個聽損情形，以下針對響度模型、頻譜模糊化模型以及混合模型做介紹。另外在本章也會介紹聲碼器的原理，以作為人工電子耳的訊號處理模擬。

3.1 響度模型

從在章節 2.2.1 中提到聽損者響度聚集的現象之中，我們知道當聲音響度很高的時候，正常人與聽損患者對應的響度是一樣的，我們假設響度聚集指發生在聲音物理強度 100 dB 以下。根據 Moore 在 1985 年的研究[20]，為達到同一響度，正常人所需的聲音強度與聽損者所需的聲音強度做對應，會發現近乎為一斜率為 N 的直線，th 是病患在該頻率的聽力閾值，病患響度聚集程度由聽力閾值與斜率 N 所決定。將上述的生理情形以數學形式呈現，假設 L_U 是原聲音訊號的強度 dB 值，L_P 是經過處理、模擬聽損患者所聽到的聲音強度 dB 值，我們可以列出下列式子：

$$L_P = NL_U + K \tag{3-1}$$

K 為某一個常數，為直線與縱軸的截距。若量測到了病患的聽力閾值，就可以在這張圖上表示此條直線與橫軸的交叉點，也可以決定 K 值截距以及直線斜率 N，強度大於 100 dB 的聲音就以斜率 N=1 計算其對應值，意旨沒有響度聚集的現象；若強度低於閾值的聲音則輸出為負的 dB 值，代表患者聽不到強度低於聽力閾值的聲音。根據以上的敘述，將函數值依各能量的對應方式拆解如下：

$$\begin{cases} L_P = 0 + (L_U - th), & L_U < th \\ L_P = 0 + N(L_U - th), & th \le L_U < th + 100/N \\ L_P = 100 + [L_U - (th + 100/N)], & th + 100/N \le L_U \end{cases} \tag{3-2}$$

其中 N 為響度聚集的程度、th 為最小可聽水平(threshold)、100/N 則是響度聚集的範圍大小。

3.2 頻譜模糊化模型

頻譜模糊化模型即是模擬章節 2.2.2 提到的耳蝸頻率解析度降低的聽損現象。我們根據某特定聽覺濾波器變寬的程度，記算出所有聽覺濾波器所需的個別增益為近似此變寬的濾波器的頻率響應，並利用這些個別增益計算出通過變寬濾波器的封包，最後加載到變寬前的載波上，成為一個具模糊化封包的訊號，模型流程如圖 3-1 所示。

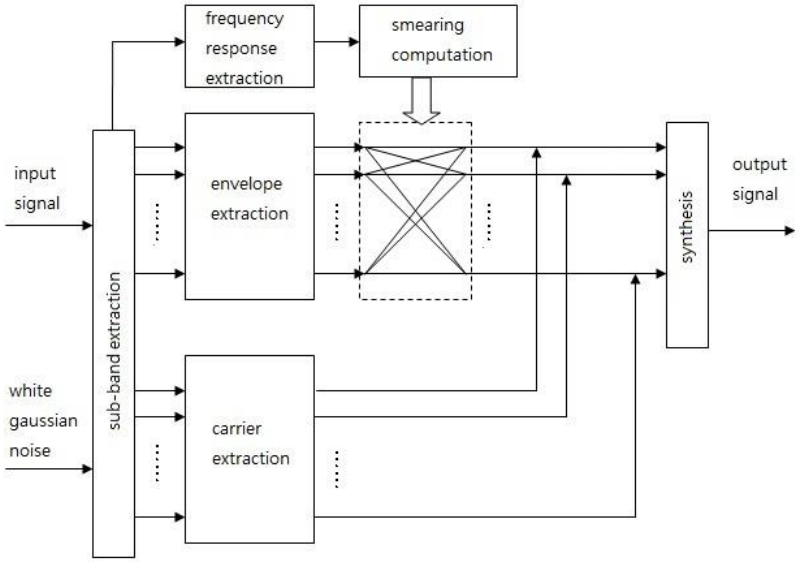


圖 3-1 頻譜模糊化演算法流程圖 [18]

在此流程圖中，上半部為外界聲音作為輸入訊號在不同子頻帶萃取出封包(envelope)，下半部由白雜訊產生出各個頻帶上的載波(carrier)。先將外界聲音經過不同子頻帶的濾波器分析成不同頻率成分的時間訊號，並且在各子頻帶上互相影響，對各個子頻帶產生互相的增益，造成模糊化的效果，最終將處理過的封包乘上載波合成出有模糊化效果的輸出訊號。其中各步驟的詳述如下：

3.2.1 濾波器變寬程度計算

針對個人化聽損模型，我們必須先量測聽損病患的耳蝸濾波器在各頻率的變寬程度，將此作為模型參數來計算頻率模糊化的個別增益參數。

為了量測受試者的聽覺濾波器形狀，我們會根據 notched-noise method 實驗來進行[21]：先假設某特定中心頻率的聽覺濾波器，如圖 3-2 所示，以中心頻率當作單頻刺激音，兩側的遮蔽聲音使用白雜訊，進行量測受試者聽到此單頻聲音的最小分貝數(dB SPL)，實驗過程中，將相對中心頻率的偏移量 Δf 由小至大，依序量測不同 Δf 情況的雜訊干擾下，所需單頻不同的聲音大小，而雜訊頻寬為中心頻率的 0.2 倍；在對受試者的人體實驗結束後，將數據輸入 Filter fitting 程式[22]，即可得到受試者於此中心頻率聽覺濾波器的等效矩陣頻寬(equivalent rectangular bandwidth，ERB)。

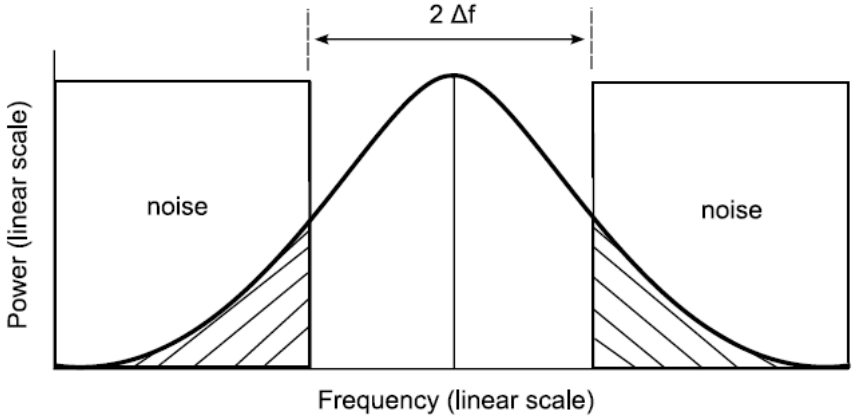


圖 3-2 notched–noise method 實驗所使用的刺激與遮蔽物 [23]

當量得病患各頻率聽覺濾波器的 ERB 值之後，再拿去跟正常人測出的各頻率聽覺濾波器 ERB 的平均值相除，便得到病患聽覺濾波器的變寬倍數 w。過去的研究挑選中心頻率 500、1000 和 2000Hz 的單頻聲音進行實驗，以此三個頻率的變寬程度為基準，內插及外插出其他中心頻率聽覺濾波器的變寬程度。

3.2.2 模糊化計算以及各頻帶附載上額外生成成分

得到變寬程度 w 之後，利用 Moore 團隊的 roex filter(rounded-exponential filter)來描述聽覺濾波器組的形狀[7][19]，其數學形式如下：

$$W(g) = \left(1 + \left(\frac{p}{w}\right)g\right)e^{-\frac{p}{w}g}, \qquad g = \frac{|f - f_c|}{f_c} \tag{3-3}$$

其中 fc 為中心頻率、g 為目標頻率相對於該頻帶中心頻率正規化後的差值，W(g)為此濾波器在特定 g 值的增益，而 p 則為控制濾波器寬窄的係數，p 值與變寬參數 w 刻畫出變寬的濾波器形狀。ERB 變寬代表聽覺濾波器所涵蓋的範圍包含原先濾波器以外的頻寬，意即包含其他聽覺濾波器的頻率響應。我們將此變寬濾波器的頻率響應想成是由原先許多正常濾波器的頻率響應線性組合而來的，所以我們要解決的問題變成「尋求一組最佳權重組合以組合出變寬的濾波器」，最佳化的公式(3-4)如下：

$$\min_x |Ax - z|^2,$$

$$\text{subject to } 0 \le x \le 1$$

$$A \in R^{8000*128}\ ,\ x \in R^{128*1}\ ,\ z \in R^{8000*1} \tag{3-4}$$

其中 A 為 128 個正常人耳聽覺濾波器的頻率響應，總共涵蓋頻寬由 0 至 7999Hz；z 是變寬濾波器的頻率響應；x 是每個聽覺濾波器對此變寬濾波器所貢獻的增益權重，即是經最佳化運算所求得之最佳權重組合。上述 3.2.1 與 3.2.2 部分即為圖 3-1 中模糊化計算

(smear computation)，以 matlab 內建的 quadratic programming 求解後發現，將 128 個變寬濾波器分別經過以上的最佳化運算之後，我們得到一個 128*128 的矩陣，代表 128 個變寬的濾波器的線性組成參數。

3.2.3 載波計算

在得到各頻帶濾波器的時域封包之後，我們將載入到各頻帶上的載波來調變，將各頻帶的輸出合成後才可以合成出模糊化的語音。若我們使用原始語音來產生載波，調變出來的訊號不僅擁有原始的基頻和諧頻，也保留原本的聲譜結構，導致正常人耳也輕易地可以聽出原始語音。分析此問題的原因，是因為原始音檔萃取出載波與封包均留下個別頻帶的語音封包及相對應的相位結構，所以將變寬處理後的封包載上原始的語音相位結構，並不會使諧頻變寬變模糊。因此，我們使用白雜訊經過個別子頻帶後取其載波作為相位資訊，如此一來合成聲音的諧頻就有變寬變模糊的效果，而正常人聽起來確實有模糊化的感受，如圖 3-3 所示。以白雜訊來產生個別頻帶的相位資訊，是噪音聲碼器的標準做法。

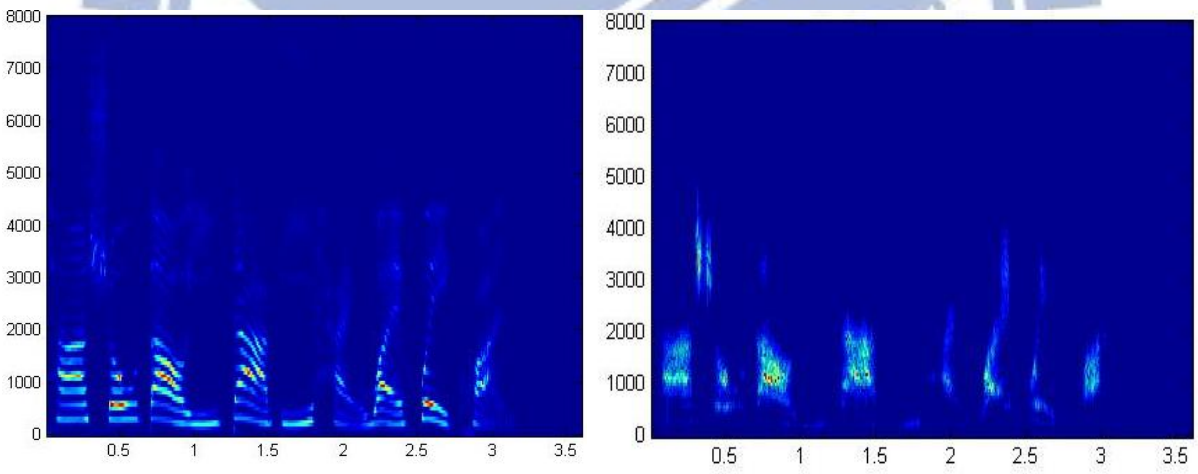


圖 3-3 原始語音、使用白雜訊載波模糊化後的語音的聲譜圖

3.3 等響度曲線增益

在結合兩個模型時，考慮了 ELC correction (equal loudness curve correction)的效應，這是用來模擬外耳與中耳對不同頻率聲音所產生的增益：許多研究指出，當頻率低於 1kHz 時，內部雜訊會變得很強[24-25]，使的最小可聽水平在 1kHz 以下會急遽上升[26-27]；而人耳對於頻率 3 kHz 附近的聲音較為敏感，如圖 3-5 所示。Moore 團隊為了要模擬外耳與中耳對不同頻率聲音的放大效果，他們認為強度大的等響度曲線較能代表外耳及中耳的增益，即選用 100dB 的 1 kHz 單音的等響度曲線為基準，給予每一個子頻帶訊號其濾波器中心頻率相對應的增益值，以將訊號由能量更正至響度軸再進行處理。

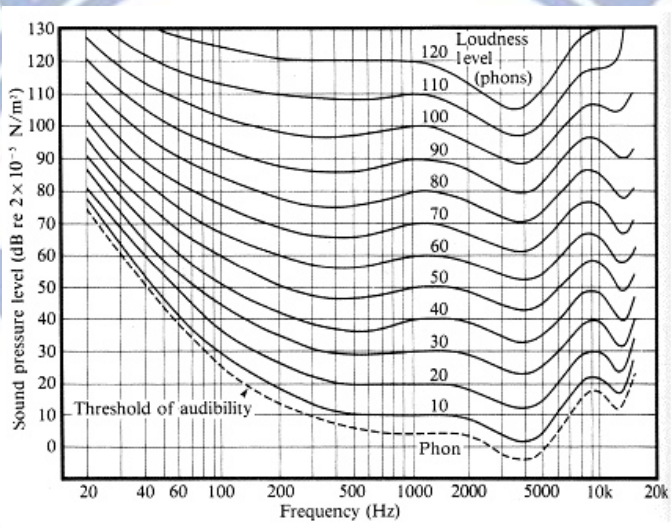


圖 3-4 人耳可聽到不同頻率及音強的聲音 [15]

3.4 個人化聽損模型

本實驗室先前所提出的個人化聽損模型即是結合響度模型與頻譜模糊化模型，並考量 ELC 增益。關於模型順序問題，若先考慮響度聚集再進行模糊化的話，會將閾值以下所裁斷的語音成分經模糊化而生成出來，使響度聚集的效果消失，所以結合兩模型的訊號處理步驟為先進行語音模糊化再考慮響度聚集。因為此模型處理過後的聲音是要給正常人聽，所以我們必須將 ELC 增益預先強調的地方解除回來，即後端 inverse ELC 增益部分。個人化聽損模型流程圖如下圖 3-6 所示：

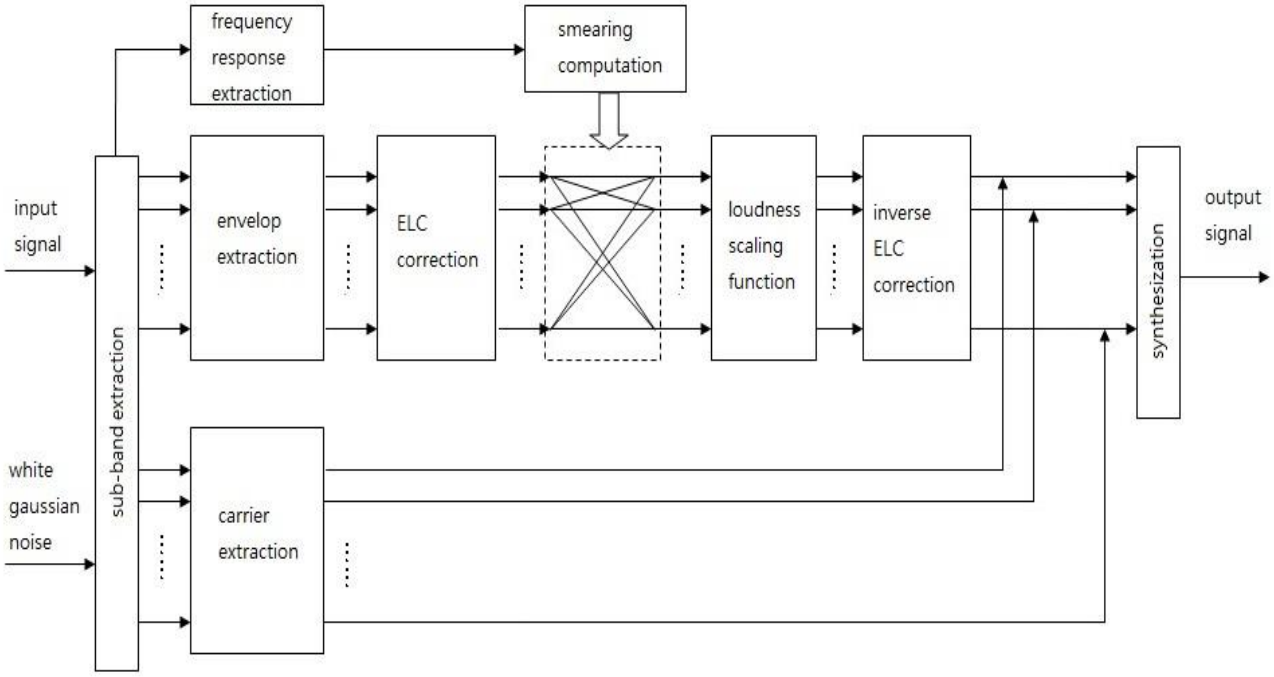


圖 3-5 混合模型流程圖

3.5 聲碼器

聲碼器(Vocoder)是由聲音操作編碼器(Voice Operated reCOrDER)簡稱而來，是個可以分析與合成人聲訊號的人聲處理系統，被廣泛運用於音訊壓縮、人聲加密與傳輸和修改人聲等應用。發展至今，聲碼器對人工電子耳的發展有深刻的影響，基於聲碼器的人聲模擬已被廣泛的運用來預測人工電子耳使用者的語音辨識效果[28-29]。

人聲是由喉頭的聲帶開關聲門所產生，其中包括了音高的波形(waveform)與許多諧波(harmonic)組成，這些週期波可視為基本的聲源信號。換句話說，可以想成喉嚨製造了語音中的高頻精細結構(fine structure)，也就是各頻帶的載波，這些諧波經過口腔與鼻腔變動所產生的複雜共振響應而形成語音。我們可以將語音分解成許多不同頻帶上固定的載波(carrier)，載上此頻帶的時間封包(temporal envelope)，若是頻帶數目越多，會得到更精確的拆解，最後將所有頻帶的聲音合成起來，就可以合成出原本的語音，此種將語音分解為封包與載波，就是一種語音編碼的方法。

我們實作聲碼器的方法是根據[9]，如圖 3-7 所示，上半部分析部分為封包萃取，先將輸入語音經過不同頻帶的帶通濾波器，各頻帶的時域封包會被封包萃取器(envelope detector)所取出，其中封包萃取器是由半波或全波整流器與低通濾波器所組成的；下半

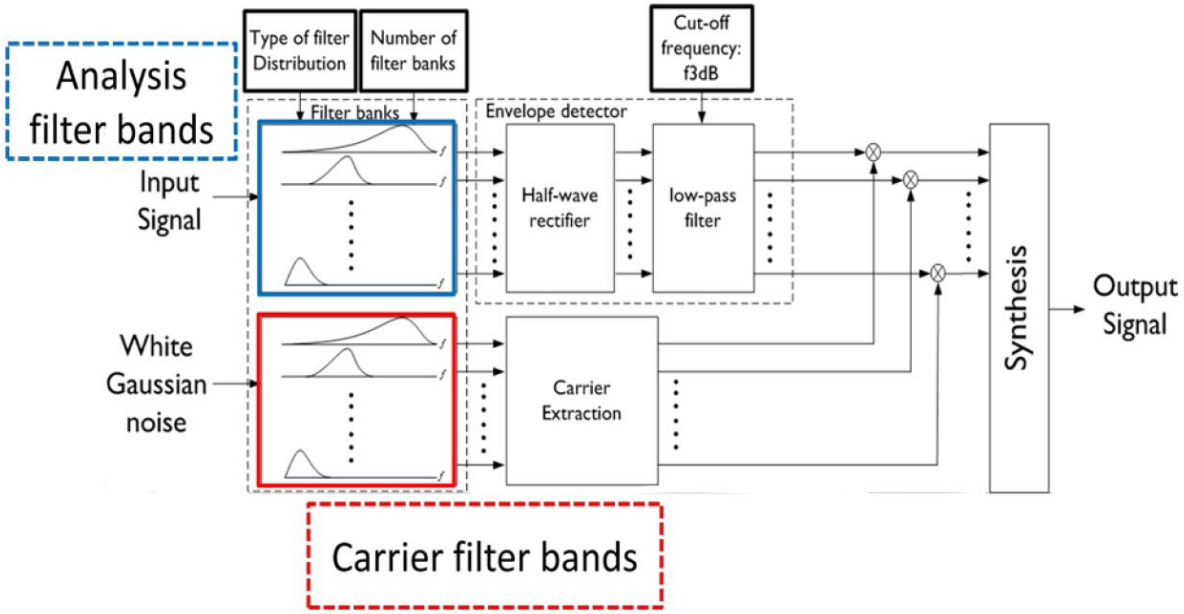


圖 3-6 聲碼器流程圖

部為載波部分，常見的載波有單頻載波(頻率為各帶通濾波器的中心頻率)，稱 tone vocoder；或通過各子頻帶的白雜訊，稱 noise vocoder [31]。從過去的研究中發現兩種聲碼器在聽辨度方面沒有顯著性的差別。最後將各頻帶時間封包載到載波上的訊號加總起來，就可以得到聲碼器的合成語音。最終將此合成語音給予正常聽力者聆聽，即可模擬人工電子耳使用者的語音感知結果。

第四章 基於個人化聽損模型的聲碼器模擬以測量主觀聽辨度

4.1 系統架構圖

根據我們團隊提出的可保留聲響聽覺的高頻四電極點人工電子耳系統，若育在聽損病患上直接評估本系統的效能，會承受極大的風險與成本，因此我們在開發時，進行模擬預測效果是很必要的。本論文提出聲響聽覺保留之高頻聲碼器，其架構如圖 4-1 所示。

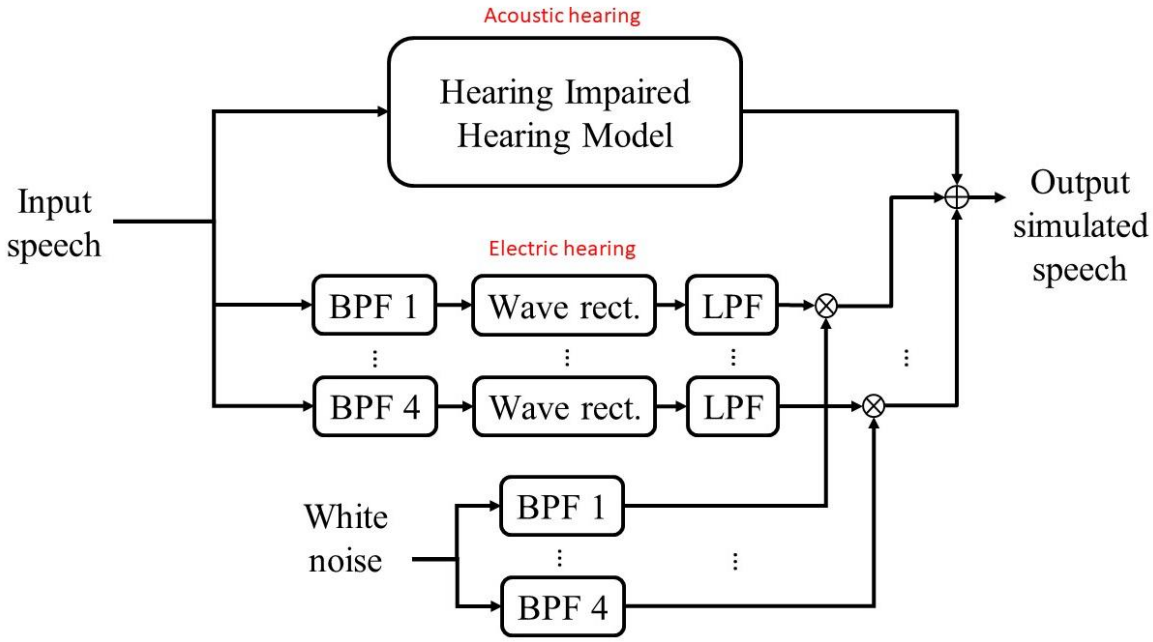


圖 4-1 基於個人化聽損模型的聲碼器模擬流程圖

在並排的兩個系統中，上半部是章節 3.4 中提及的個人化聽損模型，模擬病患之聲響聽覺，影響聽損模型的參數包括聽力閾值、響度聚集程度和耳蝸濾波器變寬程度，其中參數細節將會在下章節介紹。

下半部是章節 3.5 中提及的聲碼器，模擬四通道的人工電子耳所提供的電聽覺[9]，依照過去的經驗，載波的部分，我們選用白雜訊載波，在各子頻帶中，我們選用全波整流器和 8-order Butterworth 的低通濾波器來擷取封包，其截止頻率是在 400 Hz。我們將所需用高頻電聽覺補償的頻率範圍設定在 4000 至 7999Hz 內，然而各子頻帶的範圍選取

是一個重要的議題，我們將會在實驗中討論，最後我們也會討論此系統對提升聽辨度的效果。

我們將兩並排系統的訊號加總起來，為最後輸出的模擬訊號，以提供給正常聽力者聆聽進行主觀的中文聽辨度測量，此測量結果可以給醫師未實際病患植入此人工電子耳系統前提供重要的評估依據。

4.2 個人化聽損模型參數選取

在過去的個人化聽損模型研究中[10] [32]，本實驗室測量了六位有效的聽損患者聽損模型參數，分別代表他們不同的聽損狀況，影響聽損模型參數的現象包括聽力閾值、響度聚集程度和耳蝸濾波器變寬程度，其中聽力閾值和響度聚集程度可用一組在各頻段的最小可聽水平(簡稱 MAL)參數所表示，而耳蝸濾波器變寬程度可用聽覺濾波器的變寬因子(broaden factor，簡稱 BF)參數表示，表格 4-1 與表格 4-2 分別為六位病患的 MAL 和 BF 參數[32]。

在本次實驗中，我們選取兩位聽損病患的參數來進行測試，選取的標準在於需符合此人工電子耳系統的使用者特性，意即聽力在高頻受損較嚴重者。我們尋找最小可聽水平在 4000Hz 很高的病患，即分別是 2 號病患的 70 dB 與 8 號病患的 90 dB，並且在過去的主觀中文聽辨度測試研究中，尤其是 8 號病患在乾淨或噪音環境下的聽辨度不高，有很大的進步幅度空間，因此 2 號與 8 號病患的參數即被我們使用來進行實驗測試。(在[10]中，兩位受試者分別表示為 P1 與 P3。)

	Age	Better ear	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
Sub 2	**27**	**L**	**35**	**40**	**60**	**70**	**70**
Sub 6	72	L	55	55	55	55	55
Sub 8	**36**	**R**	**50**	**50**	**40**	**65**	**90**
Sub 9	unknown	R	35	40	60	65	70
Sub11	56	L	30	35	40	55	45
Sub12	unknown	L	65	65	65	65	60

表 4-1 受試聽損病患的各頻率最小可聽水平 MALs (dB)　[32]

	0.5k	1k	2k
Sub2	**2.85**	**3.29**	**3.19**
Sub6	2.6	4.62	2.57
Sub8	**3.2**	**2.95**	**4.66**

Sub9	1	2.55	1.7
Sub11	3.41	1.27	3.25
Sub12	2.85	2.48	1.23

表 4-2 利用 Notch-noise method 所測出來的變寬因子 [32]

4.3 心理聲學實驗方法

實驗流程為將中文單字的正常音檔(包括乾淨環境與噪音環境)輸入聲響聽覺保留之高頻聲碼器得到聽損音檔或聽損與高頻補償音檔，將處理後的音檔給正常聽力受試者聆聽，正常聽力者將所聽到的中文字寫下來，最後計算辨識正確率來評定效果，如圖 4-2 所示。

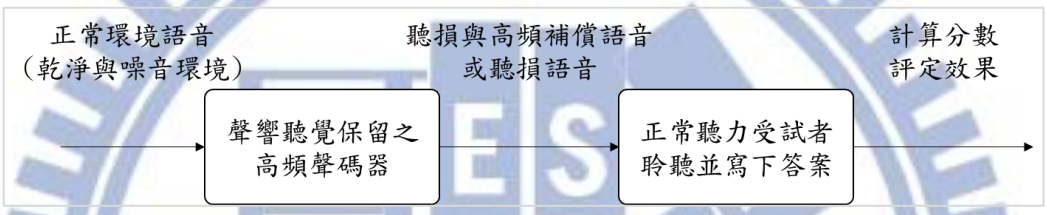


圖 4-2 心理聲學實驗流程圖

對於辨識正確率的實驗結果，我們需要利用統計學常用到的「變異數分析(Analysis of Variance，ANOVA)方法」來分析實驗結果數據，進而討論數據間是否有顯著的差異，其中表示顯著性(significance)的數據為「p 值」，p 值越低則顯著性越高，顯著性分為三個層級： $p > 0.05$ 為不顯著、 $0.01 < p < 0.05$ 為顯著、 $p < 0.01$ 為極顯著。

另外，為了進行統計分析，我們將結果的平均正確率使用「合理反和弦轉換(Rational arcsine transform)」轉換為合理反和弦元(Rational arcsine units)再進行 ANOVA 分析，其目的是減少分數的天花板效應(ceiling effect)，換句話說，就是使趨近滿分與零分分數的區別度增加[34]，其合理反和弦轉換函數如圖 4-3 所示。

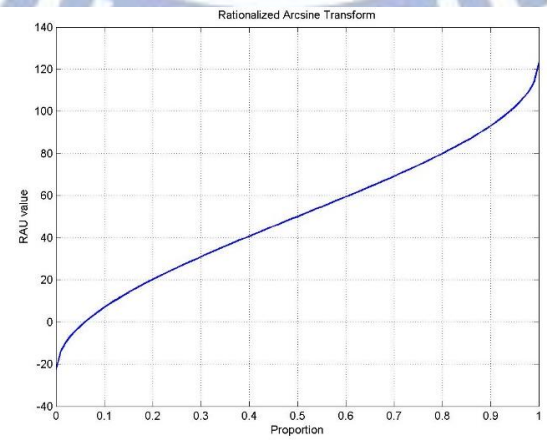


圖 4-3 合理反和弦轉換

4.3.1 中文語音資料庫

本實驗是採用[33]中所發展的中文測試語料表，來進行中文音素(phoneme)的主觀聽辨度聽力測量實驗。根據[33]，此單字表的來源是從日常生活中挑出 700 個最常使用的單字，再利用實驗從中找出 384 個單字，再組出 6 組具有相近數量的音調(tone)、聲母(consonant)和韻母(vowel)的測試語料，每組共有 25 個中文字，如圖 4-4 所示。

根據此中文測試語料表，我們從公開資源「教育部中文學習網站」下載相對應的音檔(http://stroke-order.learningweb.moe.edu.tw/home.do?rd=72)，音檔取樣頻率為 16 kHz，所有音檔也正規化至同一能量。

	List A1		List A2		List A3		List B1		List B2		List B3	
No.	Item	Word	Item	Word	Item	Word	Item	Word	Item	Word	Item	Word
1	zhi1	知	di4	地	zhi1	知	jing1	經	shi4	是	yi3	以
2	shang4	上	yu2	魚	he2	河	yuan2	元	cheng2	成	cheng2	成
3	hou4	後	jian4	建	yu2	魚	yu3	雨	ji4	計	ru2	如
4	neng2	能	zi4	字	ji3	擠	qi2	其	yu3	雨	si4	四
5	jin4	進	shuo1	說	ying4	硬	xia4	下	bian4	便	du4	度
6	qu4	去	xiao3	小	wu2	吳	jiang4	降	du4	度	cong2	從
7	mu4	木	kan4	看	jie2	結	bian4	便	fei1	飛	ling4	另
8	xian1	先	mu4	木	ban4	半	si4	四	ying3	影	lü4	綠
9	jie3	姐	tou2	頭	chu4	觸	fei1	飛	xi2	習	jian3	檢
10	diao4	掉	jie3	姐	zhan3	展	kao3	考	ru4	入	gei3	給
11	ying2	營	feng1	風	qi1	七	guang1	光	jiao1	交	jun1	軍
12	chu2	除	zhan3	展	diao4	掉	liao4	料	qiang2	強	zhuan1	專
13	li2	梨	gai1	該	xin4	信	tan2	談	liu4	六	yang2	陽
14	wei3	偉	chu2	除	pian4	片	huan1	歡	kuang4	礦	tui1	推
15	ge1	歌	a1	阿	zeng1	增	you1	優	mian3	免	xi3	洗
16	ban3	板	yin3	引	tuo1	脫	cu4	促	tan1	貪	kuang4	礦
17	fan2	凡	zhen4	鎮	long2	龍	re4	熱	zhui1	追	mao4	貿
18	pi2	皮	sui2	隨	gang3	港	xian2	賢	guan4	冠	bang4	棒
19	ao4	傲	ni2	尼	wei1	威	chu3	楚	lun2	輪	pian1	偏
20	kong3	孔	bang1	幫	fa2	罰	zhui1	追	pao3	跑	shun4	順
21	za2	雜	pi2	皮	kao4	靠	shun4	順	song1	松	fan1	翻
22	zhuo1	桌	long2	龍	shou2	熟	dong3	懂	cang2	藏	yao2	搖
23	ai1	哀	ao4	傲	nai4	耐	meng2	蒙	xuan2	玄	qiu1	秋
24	ti4	替	he1	喝	mi3	米	lang2	郎	o1	喔	xuan2	玄
25	sen1	森	qing4	慶	sen1	森	pi4	僻	he4	賀	he4	賀

Note: The items of the word lists are represented by traditional Chinese characters, and their Romanization is in the Hanyu Pinyin system.

圖 4-4 音素平衡之中文單字聽力測試語料表

資料來源：[33]

4.3.2 心理聲學實驗受試者

為測試我們提出的聲響聽覺保留之高頻聲碼器的效能，我們對於正常聽力受試者進行心理聲學實驗，找了 9 位台灣在地的正常聽力中文使用者，年齡介於 20 至 24 歲，其中有 6 位男生與 3 位女生，每位受試者將寫下他們所聽到的中文字，音檔不可重複播放。實驗過程中將會讓每位受試者帶上 AKG 型號 k702 的耳機，音量調整在 65 dB 至 75 dB SPL 的舒適範圍，在半無響室進行(唯獨地板無吸音海綿)。

4.3.3 語料計分方式

計算分數時，我們會將字猜解成音調(tone)、聲母(consonant)和韻母(vowel)三個部分來分別計算，其中音調的部分，在之前的研究[10]中提到，變寬因子在 6 以上才會失去音調輪廓，而在我們實驗中的病患參數皆在 6 以內，對於音調的辨認並不會有太大的影響，所以我們只計算聲母和韻母的分數。

計算的方式為正確的聲母或韻母，如：

1. 「喘，ㄔㄨㄢˇ」若聽成「闖，ㄔㄨㄤˇ」，如此得聲母 1 分，韻母 0 分。
2. 「棒，ㄅㄤˋ」若聽成「方，ㄈㄤ」，如此得聲母 0 分，韻母 1 分。

最後，分別將每組單字測驗的聲母與韻母的總分除上總個數，再換算成百分比，即為受試者對此組單字的中文單字聽辨度。

4.4 主觀中文聽辨度測量(一)：帶通濾波器的頻率覆蓋範圍

我們欲探討會影響中文聽辨度的因素－子頻帶在頻率上的分布，也就是聲碼器中，四個帶通濾波器的頻率覆蓋範圍。在此實驗中，我們將探討三種情形：對數分布(Logarithmic space，簡稱 LOG)、線性等寬分布(Linear space，簡稱 LIN)和低頻對應高頻分布(Low-map-high-frequency arrangement，簡稱 LMH)。

對於對數分布，我們將四個 Butterworth 濾波器的中心頻率設置在 4367、5199、6182 和 7343 Hz 上，並且將頻寬設定為 1/4 倍頻(octave) ，如圖 4-5 (a)所示；對於線性等寬分布，我們將四個 Butterworth 濾波器的中心頻率設定在 4500、5500、6500 和 7500 Hz 上，頻寬設定為 1000 Hz，如圖 4-5 (b)所示。

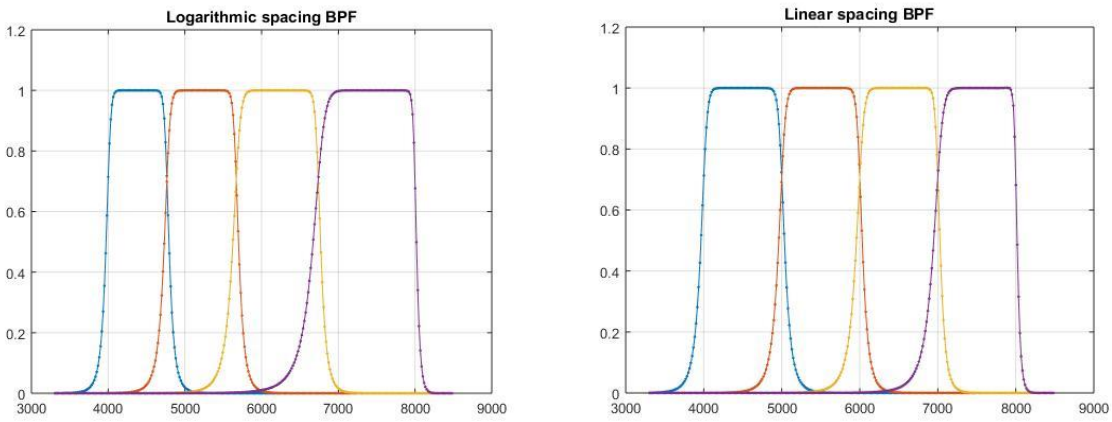


圖 4-5 (a)對數分布的帶通濾波器頻率響應 (b)線性等寬分布的帶通濾波器頻率響應

以上兩種分布情況中，封包萃取和製造子頻帶白雜訊載波的帶通濾波器頻寬分布都是相同的，對於低頻對應高頻分布，我們將探討頻帶不對應相同的情況，將封包萃取部分的帶通濾波器中心頻率設定在 1000、3000、5000 和 7000 Hz 上，頻寬為 2000 Hz，而製造載波的帶通濾波器頻率分布與線性等寬分布一樣，如圖 4-6 所示，此作法的想法是探討全頻段的語音資訊都包含在封包裡面，是否對於聽辨度會有提升的效果。

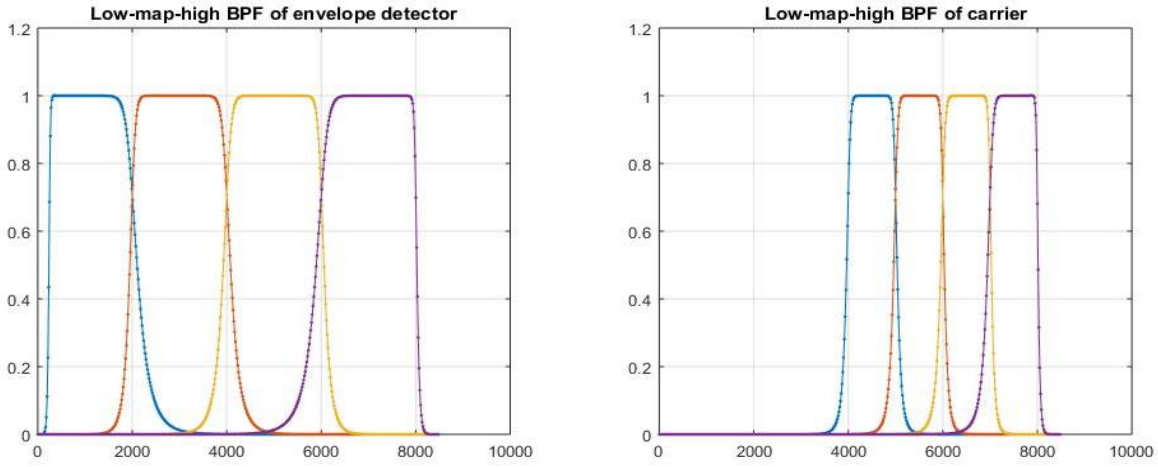


圖 4-6 低頻對應高頻分布的封包萃取與製造載波之帶通濾波器頻率響應

4.4.1 實驗條件

此實驗是對於帶通濾波器的頻率範圍來做討論，對於每位測試者，共進行 8 種實驗，分別是：

「2 位聽損病患之個人化模型 *(無高頻聲碼器 +
對數分布之聲碼器 + 線性等寬分布之聲碼器 + 低頻對應高頻分布之聲碼器)」

在每種實驗中，都沒有加入噪音，並且隨機選用[33]中的某一項 25 字中文測試語料列表，再將 8 種實驗總共 200 個字中，以 5 個同條件無意義順序的字為一組隨機排列後，使測試者進行中文聽辨度的心理聲學實驗。

4.4.2 實驗結果與討論

在選擇帶通濾波器的頻率覆蓋範圍實驗中，我們討論音素(聲母與韻母總和)的平均辨識正確率實驗結果，根據單因子變異數分析(one-way ANOVA)，改變「濾波器的頻率覆蓋範圍」這項因素在統計上對於辨識正確率有顯著($F[3,32] = 9.33,\ p = 0.001$)的影響。兩組聽損模型個人化參數配上各四種條件的平均辨識正確率即顯示在圖 4-7 與圖 4-8，每項條件上都有標示 ± 1 的標準差，我們將三種聲碼器條件的平均分數與無聲碼器條件的平均分數做比較，而紅色星號「*」表示其分數有顯著地($p < 0.05$)高於無聲碼器的分數，其中詳細的 ANOVA 統計分析數據在表 4-3 與表 4-4 表示。

根據圖 4-7 與表 4-3，我們可以看到對於使用 Sub2 病患的參數來說，加上對數分布和線性等寬分布的聲碼器皆能極顯著地($p = 0.0003$、$p = 0.0024$)提升聽辨度 10%左右，雖然前者比後者的平均辨識率略高一點，但是兩者彼此之間卻沒有顯著的差別；另外，加上低頻對應高頻分布的聲碼器並無法顯著提升聽辨度($p = 0.2604$)。

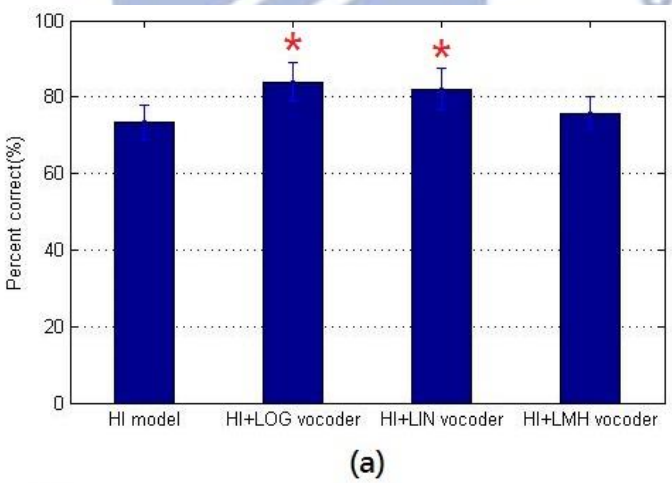


圖 4-7 使用 Sub2 聽損模型參數的平均辨識正確率

Sub2 聽損模型受試者			
LOG 分布與無聲碼器	F(1,16)=21.23	**p=0.0003 ***	極顯著
LIN 分布與無聲碼器	F(1,16)=12.91	**p=0.0024 ***	極顯著
LMH 分布與無聲碼器	F(1,16)=1.36	p=0.2604	不顯著

表 4-3 使用 Sub2 之帶通濾波器頻率覆蓋範圍的變異數分析結果

而使用 Sub8 的聽損模型參數時，圖 4-8 與表 4-4 可顯示出，唯獨加上對數分布的聲碼器可顯著地($p = 0.0185$) 提升聽辨度，然而在此組參數下，加上線性等寬分布的聲碼器卻無法顯著提升聽辨度($p = 0.5530$)，而且低頻對應高頻分布的聲碼器與無聲碼器條件也沒有顯著的差異($p = 0.1935$)，甚至其辨識正確率的平均更是略低於無聲碼器條件下的平均值。

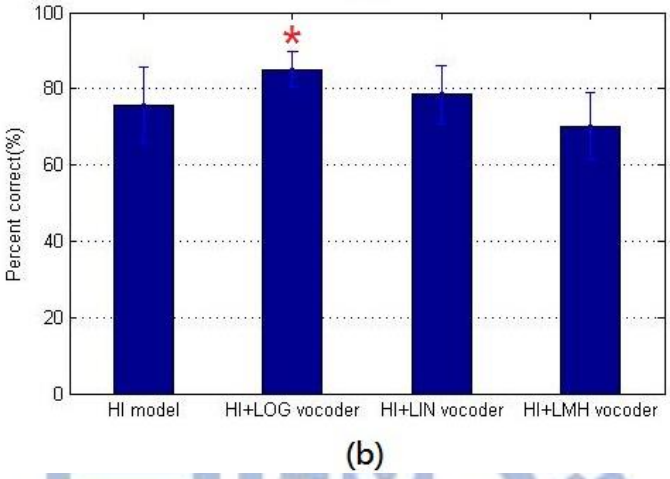


圖 4-8 使用 Sub8 聽損模型參數的平均辨識正確率

Sub8 聽損模型受試者			
LOG 分布與無聲碼器	F(1,16)=6.87	**p=0.0185 ***	顯著
LIN 分布與無聲碼器	F(1,16)=0.37	p=0.5530	不顯著
LMH 分布與無聲碼器	F(1,16)=1.84	p=0.1935	不顯著

表 4-4 使用 Sub8 之帶通濾波器頻率覆蓋範圍的變異數分析結果

根據上述的結果，我們可知道「濾波器的頻率覆蓋範圍」是對於中文的辨識率有影響的，而且對數分布的影響是最為顯著的，甚至對於高頻聽損最嚴重的測試者 Sub8 來說，加上對數分布的聲碼器更是唯一可顯著提升中文聽辨度的條件。

然而，討論到低頻對應高頻分布，儘管萃取語音封包時的頻寬範圍最廣，包含最多的語音資訊，但是中文語音的辨識率結果卻是最差的，這結果可能是因為存在非對應的頻率範圍。對於聽者來說，這做法是會破壞平常的語音結構，或許聽者需要更多的時間針對此種語音進行訓練，或可助於對於此分布條件下的辨識度。

總而言之，根據此實驗的結果，我們將會使用最有顯著效果的對數分布聲碼器模型，進行接下來的「噪音下聲母與韻母的聽辨度提升」與「利用深度學習降噪演算法提升聽辨度」實驗。

4.5 主觀中文聽辨度測量(二)：噪音下聲母與韻母的聽辨度提升

在 4.4 的實驗中，我們是依據乾淨語音來討論帶通濾波器的頻率覆蓋範圍，然而，在正常的環境中皆會有噪音的存在，在這項實驗中，我們針對常見的噪音下，使用聲響聽覺保留之對數分布聲碼器來討論此電子耳系統是否可提升聽辨度。

中文字包含聲母與韻母兩個部分，前者相對於後者在頻譜分布中是沒有諧波，就像是英文中的子音，常使用塞音(像是[p]、[t]、[k])、擦音(像是[s])等等。例如中文字「上，ㄕㄤˋ，/Shàng/」，如圖 4-9 的頻譜所示，韻母「ㄤ」的時間範圍在 0.4 至 0.7 秒，有明顯的諧波，能量多分布於低頻，而高頻也有部分能量分布；而聲母「ㄕ」的時間範圍是在 0.2 秒至 0.38 秒左右，能量分布皆在頻率 3500 至 6000 Hz 之中。此項我們提出的電子耳系統的使用對象為高頻聽損病患，我們預計高頻聲碼器對於聲母的聽辨度會有比較大的幫助，所以此實驗中我們將聲母與韻母的聽辨度分開討論。

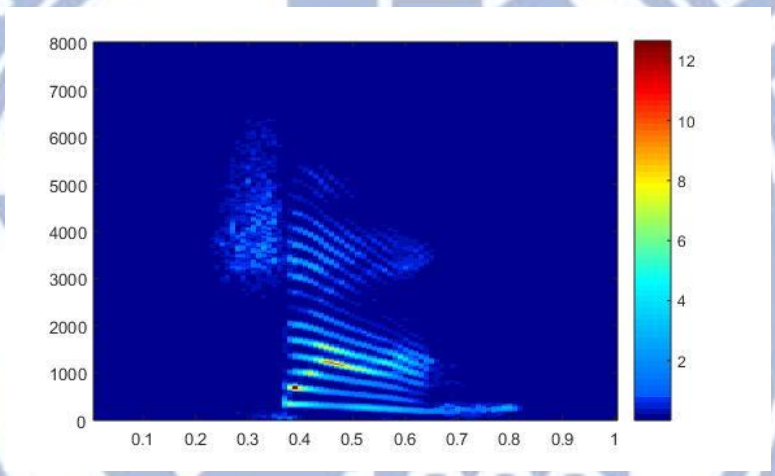


圖 4-9 中文字「上，ㄕㄤˋ」的語音頻譜圖

我們測試五種條件：加上 0 dB 與 4 dB SNR 的語音形狀噪音(speech shaped noise，SSN)、加上 3 dB 與 7 dB SNR 的雙語者噪音(two-talker speech，TTS)和乾淨語音(clean)。

4.5.1 實驗條件

此實驗是對於噪音下聲母與韻母的聽辨度來做討論，對於每位測試者，共進行 16 種實驗，分別是：　「2 位聽損病患之個人化模型數 * 2 種噪音

* 2 種 SNR * 2 種情況 (無聲碼器與有對數分布聲碼器) 」

在每種實驗中，隨機選用[33]中的某一項 25 字中文測試語料列表，再將 16 種實驗總共 400 個字中，以 5 個同條件無意義順序的字為一組隨機排列後，使測試者進行中文聽辨度的心理聲學實驗。

4.5.2 實驗結果與討論

本實驗中，將兩種雜訊分開測試三因子變異數分析(three-way ANOVA)，討論訊雜比(SNR)、因素種類(Phoneme type)和聲碼器存在與否(Vocoder presence)這三個因素對於辨識正確率的影響。

我們先測試 Sub2 的聽損模型參數，對於語音形狀噪音，我們將實驗 4.4 中的對數分布與無聲碼器兩種情況的結果當作乾淨語音，表示為訊雜比無限大的條件，並且將其併入語音形狀噪音 4 dB 與 0 dB，共三種訊雜比條件，做三因子變異數分析，其結果如表 4-5 所示。其中三個因子在統計上皆對於此實驗有極顯著的影響。

Sub2 SSN Three-way ANOVA			
SNR	F(2,98) = 6.07	**p<0.005 ***	極顯著
Phoneme type	F(1,98) = 36.16	**p<0.0005 ***	極顯著
Vocoder presence	F(1,98) = 264.99	**p<0.0005 ***	極顯著
SNR & Phoneme type	F(2,98) = 33.5	p=0.7343	不顯著
SNR & Vocoder presence	F(2,98) = 2.23	p=0.1134	不顯著
Phoneme type & Vocoder presence	F(1,98) = 33.57	**p<0.005 ***	極顯著

表 4-5 使用 Sub2 之語音形狀噪音下聽辨度的三因子變異數分析結果

對於雙語者噪音(包含 3 dB 與 7 dB 兩種訊雜比條件)，我們做三因子變異數分析，其結果如表 4-6 所示，與上述的語音形狀噪音所得到的結果相同，亦即三個因子在統計上皆對於此實驗有極顯著的影響。

Sub2 TTS Three-way ANOVA			
SNR	F(1,65) = 8.4	**p<0.05 ***	極顯著
Phoneme type	F(1,65) = 36.75	**p<0.0005 ***	極顯著
Vocoder presence	F(1,65) = 171.24	**p<0.0005 ***	極顯著
SNR & Phoneme type	F(1,65) = 1.1	p=0.2983	不顯著
SNR & Vocoder presence	F(1,65) = 0.04	p=0.8351	不顯著
Phoneme type & Vocoder presence	F(1,65) = 34.75	**p<0.005 ***	極顯著

表 4-6 使用 Sub2 之雙語者噪音下聽辨度的三因子變異數分析結果

接下來，我們測試 Sub8 的聽損模型參數，對於語音形狀噪音，我們依然將實驗 4.4 中的對數分布與無聲碼器兩種情況的結果當作乾淨語音，表示為訊雜比無限大的條件，並且將其併入語音形狀噪音 4 dB 與 0 dB，共三種訊雜比條件，做三因子變異數分析，其結果如表 4-7 所示。其中三個因子在統計上皆對於此實驗有極顯著的影響。

Sub8 SSN Three-way ANOVA			
SNR	F(2,98) = 10.41	**p<0.005 ***	極顯著
Phoneme type	F(1,98) = 21.97	**p<0.0005 ***	極顯著
Vocoder presence	F(1,98) = 209.34	**p<0.0005 ***	極顯著
SNR & Phoneme type	F(2,98) = 1.53	p=0.2212	不顯著
SNR & Vocoder presence	F(2,98) = 6.89	**p<0.005 ***	極顯著
Phoneme type & Vocoder presence	F(1,98) = 1.98	p=0.1627	不顯著

表 4-7 使用 Sub8 之語音形狀噪音下聽辨度的三因子變異數分析結果

對於雙語者噪音(包含 3 dB 與 7 dB 兩種訊雜比條件)，我們做三因子變異數分析，其結果如表 4-8 所示，與上述的語音形狀噪音所得到的結果相同，亦即三個因子在統計上皆對於此實驗有極顯著的影響。

Sub8 TTS Three-way ANOVA			
SNR	F(1,65) = 7.36	**p<0.05 ***	極顯著
Phoneme type	F(1,65) = 20.81	**p<0.0005 ***	極顯著
Vocoder presence	F(1,65) = 115.04	**p<0.0005 ***	極顯著
SNR & Phoneme type	F(1,65) = 1.28	p=0.2619	不顯著
SNR & Vocoder presence	F(1,65) = 1.72	p=0.1939	不顯著
Phoneme type & Vocoder presence	F(1,65) = 0.05	p=0.8317	不顯著

表 4-8 使用 Sub8 之雙語者噪音下聽辨度的三因子變異數分析結果

在經過上述的三因子變異數分析當中，我們最想知道的是「聲碼器存在與否」是否會產生影響。因此，我們將使用 Post hoc 成對分析來討論有無聲碼器在每種條件下，對於語音的辨識度是否會有影響。

圖 4-10 與 4-11 分別表示了 Sub 2 和 Sub 8 在噪音下之平均辨識正確率，每項條件上都有標示 ±1 的標準差，我們將三種聲碼器條件的平均分數與無聲碼器條件的平均分數做比較，而紅色星號「*」表示其分數有顯著地($p < 0.05$)高於無聲碼器條件下的分數，其中詳細的 ANOVA 統計分析數據在表 4-9 與表 4-10 表示。

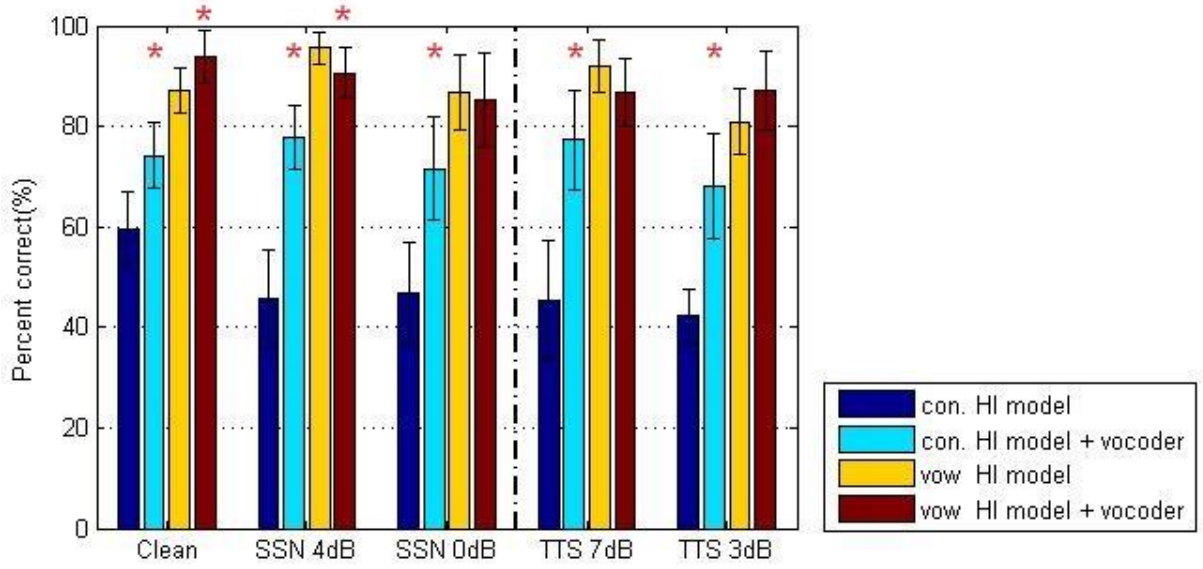


圖 4-10 使用 Sub2 之噪音下的平均辨識正確率
(其中 con.表示聲母，vow.表示韻母)

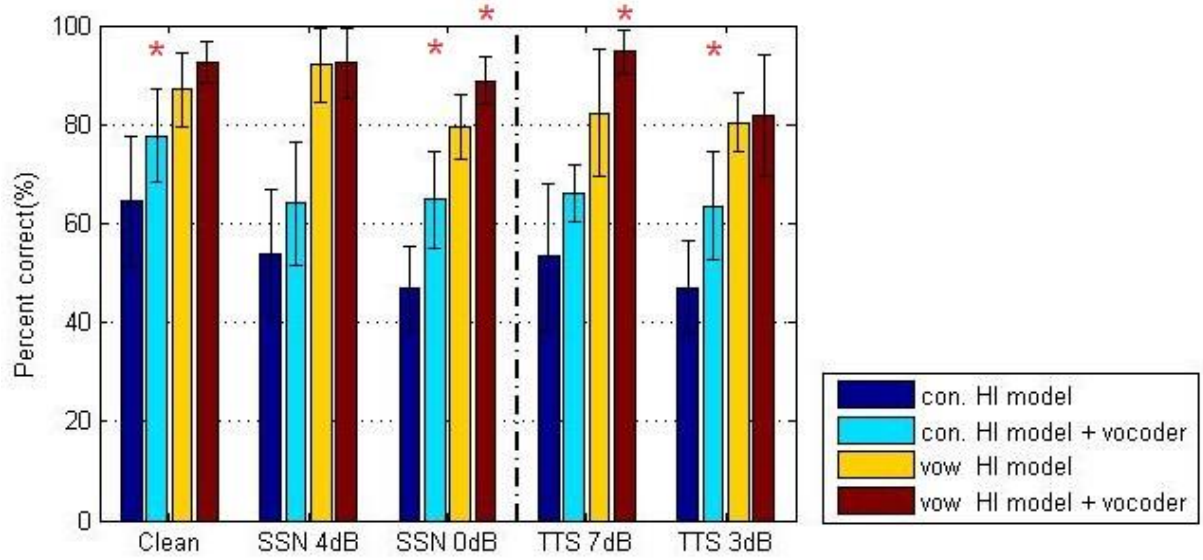


圖 4-11 使用 Sub8 之噪音下的平均辨識正確率
(其中 con.表示聲母，vow.表示韻母)

從圖 4-10 所見，對於 Sub 2 參數，我們可以發現在每種條件之下，聲碼器都能顯著地提升聲母辨識正確率，甚至在噪音情況下，平均的提升率比乾淨語音情況來的多，而韻母的識別正確率卻沒有在全部條件之下顯著地被提升，可能原因是原本韻母的辨識正確率就已經很高了。然而從圖 4-11 中，我們看到對於 Sub 8 參數，加上聲碼器之後，雖然某些條件在統計上沒有顯著性，不過不管是聲母或者是韻母的平均辨識正確率皆有提升。

綜合本實驗以上的結論，我們的系統可以對於噪音下的聽辨度有相當的幫助，尤其是對於聲母的幫助最大，而這些模擬對於人工電子耳系統的評估有相當大的幫助。

Sub2 聲母 (Consonant)			
Clean	F(1,16) = 18.94	**p=0.0005 ***	極顯著
SSN 4dB	F(1,16) = 26.58	**p=0.00009 ***	極顯著
SSN 0dB	F(1,16) = 69.35	**p=0.0000003 ***	極顯著
TTS 3dB	F(1,16) = 43.83	**p=0.0000058 ***	極顯著
TTS 7dB	F(1,16) = 38.4	**p=0.0000128 ***	極顯著
Sub2 韻母 (Vowel)			
Clean	F(1,16) = 8.41	**p=0.0104 ***	顯著
SSN 0dB	F(1,16) = 0.11	p=0.7432	不顯著
SSN 4dB	F(1,16) = 6.37	**p=0.0226 ***	顯著
TTS 3dB	F(1,16) = 3.41	p=0.0834	不顯著
TTS 7dB	F(1,16) = 3.56	p=0.0776	不顯著

表 4-9 使用 Sub2 之噪音下聽辨度的變異數分析結果

Sub8 聲母 (Consonant)			
Clean	F(1,16) = 6.1	**p=0.0251 ***	顯著
SSN 0dB	F(1,16) = 17.11	**p=0.0008 ***	極顯著
SSN 4dB	F(1,16) = 2.86	p=0.11	不顯著
TTS 3dB	F(1,16) = 11.88	**p=0.0033 ***	顯著
TTS 7dB	F(1,16) = 6.02	**p=0.026 ***	顯著
Sub8 韻母 (Vowel)			
Clean	F(1,16) = 3.51	p=0.0793	不顯著
SSN 0dB	F(1,16) = 12.08	**p=0.0031 ***	顯著
SSN 4dB	F(1,16) = 0.02	p=0.9885	不顯著
TTS 3dB	F(1,16) = 0.09	p=0.7732	不顯著
TTS 7dB	F(1,16) = 7.56	**p=0.0143 ***	顯著

表 4-10 使用 Sub2 之噪音下聽辨度的變異數分析結果

第五章 基於深度學習降噪演算法提升語音聽辨度

5.1 深度神經網路學習

5.1.1 背景

深度學習(deep learning)是屬於機器學習(machine learning)領域的一支，從人工神經網路(artificial neural network)的概念演化而來，模擬生物中樞神經系統結構與運作方式的參數化模型，它的複雜結構包含多重非線性轉換構成的多個隱藏層，來對資料進行高層抽象的特徵擷取。要將自然語音交給機器學習處理，首先要經過數學化來表示，在機器學習中的基礎就是使用分散表示(distributed representation)方法，語音在機器學習的過程中經過不同因子互相作用生成要學習目標，每項因子代表不同抽象概念，整體而言，在分散表示此基礎上，深度學習更將這一互相作用拆解為多個層次，不同的層數和規模可擷取不同程度的抽象意義。深度學習利用分層抽象代表的概念，將更高層次的概念會由低層次的概念學習而得，此分層的結構常使用誤差倒傳遞演算法(error backpropagation)來更新參數，並從中選取有效特徵。

5.1.2 神經網路架構系統

為了在噪音環境下提升更高的聽辨度，降噪(noise reduction)就是一個可能的方向，近年來，深度學習成功地應用於效於語音的降噪[35]，也有用於人工電子耳的模擬器上[13]，在此我們將參考於[13]中所提出深層神經網路(deep neural nerwork，DNN)模型的基本架構，在我們所提出的新型電子耳進行降噪，來觀察其對語音聽辨度的影響。降噪流程架構如圖 5-1 所示，我們將吵雜語音(noisy speech)以 16 毫秒為一音框，做短時間傅立葉轉換，音框覆蓋範圍為 1/2 個音框，頻譜在第 w 個時間點可表示為：

$$y(m) = [\mathrm{Y}(1,m), \mathrm{Y}(2,m), \dots, \mathrm{Y}(129,m)]^{\mathrm{T}} \qquad (5\text{-}1)$$

其中$\mathrm{Y}(k,m)$為頻譜上第 m 個時間上且第 k 個頻率的點，再將轉換後的絕對值取對數變成 $log(abs(y(m)))$，這裡有個小細節，是會在$abs(y(m))$後加入小偏差 0.01，目的是防止log函數的輸出成無意義或縮小輸出值的範圍。因為語音的連續特性很強，所以 l 為包含前後音框的數量，將第 $m\pm l$ 個音框同時當作深層神經網路的輸入。因硬體運算效能的限制，所以每個音框只取的 256 點的快速傅立葉轉換，因實數傅立葉轉換後的振幅為偶函數，所以我們只取 256 點的一半和 DC 值，共 129 個點作為深層神經網路模型的輸入。

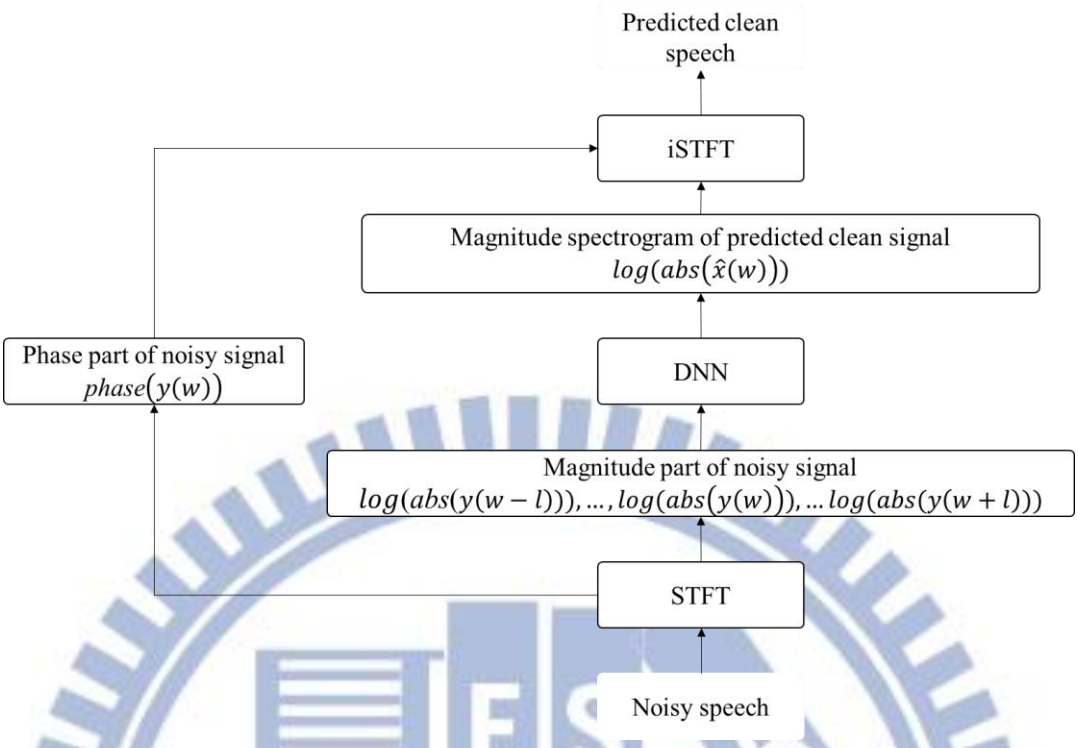


圖 5-1 以深層神經網路降噪的流程圖

其中的深層神經網路，如圖 5-2 所示，模型架構共有五層，三層隱藏層中各有 500 個神經元，並且加入 batch normalization 來防止收斂速度過慢和梯度爆炸的問題，因為是回歸模型，所以最後一層的激活函數用「linear」，而其他層皆用「Sigmoid function」。我們首先對前後音框的數量 l 對於客觀評估的影響進行了解，經實驗發現越高的 l 就會有越高的語音聽辨度，因為硬體的關係我們最後選擇 l=5，前後包含共 11 個音框。經過深層神經網路之後，估計出來乾淨語音的振幅頻譜對數量值，再將此估計值與吵雜語音頻譜的相位結合，做逆短時間傅立葉轉換，合成回時域的估計乾淨語音(predicted clean speech)。

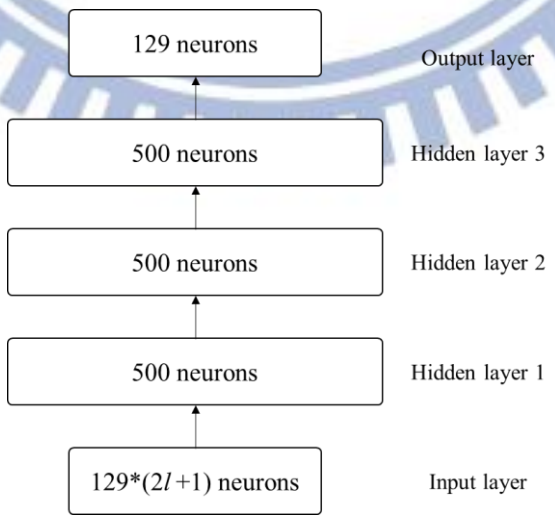


圖 5-2 深層神經網路(DNN)的模型架構

5.2 心理聲學實驗方法

本實驗中，我們將以客觀評估的方式選取降噪演算法的最佳參數，再將聽覺保留之高頻聲碼模擬器結合此基於深度學習的最佳降噪演算法，進行心理聲學實驗，做主觀中文聽辨度的測量。圖 5-3 所示為實驗流程。對於辨識正確率的實驗結果，我們依然使用變異數分析方法來分析實驗數據，並且在使用變異數分析前經過合理反和弦轉換來減少天花板效應。

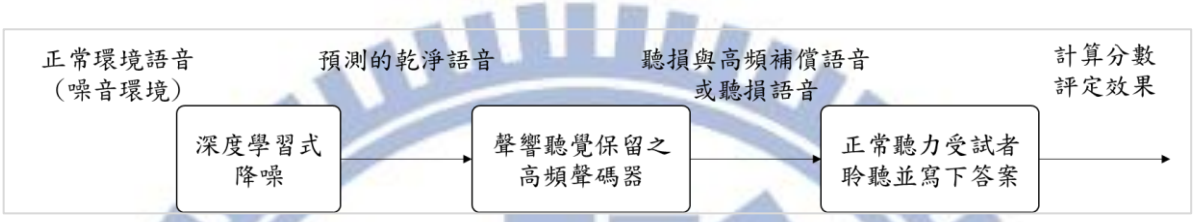


圖 5-3 結合降噪演算法之實驗流程

5.2.1 心理聲學實驗受試者

在對於正常聽力受試者的心理聲學實驗中，我們找了 9 位台灣在地的正常聽力中文使用者，年齡介於 20 至 24 歲，其中有 6 位男生與 3 位女生，每位受試者將寫下他們所認為聽到的中文句子，音檔會重複播放兩次。實驗過程在半無響室進行(唯獨地板無吸音海綿)，每位受試者會帶上 AKG 型號 k702 的耳機，音量會調整在 65 dB 至 75 dB SPL 的舒適範圍，。

5.2.2 中文語音資料庫

跟第四章實驗不同的地方是，中文句子為此語音聽辨測試的語料，句子的特性為包含前後文語意和結構性，比較貼近日常生活中的使用狀況，因此我們採用的是針對台灣地區的普通話在噪音底下測試(Taiwan Mandarin hearing in noise test，TMHINT)語音聽辨度所發展出來的中文句子語料[30]。每句句子符合以下四點規範：(1)每句話需包含十個字；(2)每句話必須連小學一年級的學生都聽得懂；(3)每句話必須是台灣地區很普遍的日常生活用語；(4)句子裡面不能含有大眾口號。此語料總共發展出 16 組(包含 12 組測試句子和 4 組練習句子)，每組包含 20 句的中文句子測試語料。所有測試句子的音調統計圖如圖 5-4 所示。

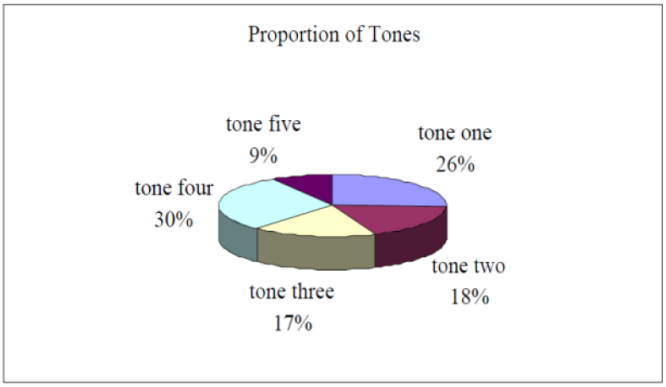


圖 5-4 所有 240 句測試句子音調比例分布圖

而每組測試句子在安靜及吵雜環境底下句子的接受閾值(Reception Thresholds of Sentences，RTSs)為圖 5-5 所示，可以看出來每組句子的差異都維持在正負 1dB 之內。

所有中文測試語料皆在一無響室(3x2x3 m3)裡面錄製。我們聘請一名接受過中文演講訓練的女性擔任錄音員，錄音器材為 A SHURE SM58 麥克風 ALESIS iO2 的 USB 錄音介面及一台筆記型電腦，並以 MATLAB 發展一錄音 GUI 操作介面。語音取樣頻率為 16k Hz，儲存格式為 16-bit PCM，且每句話或字的能量都正規化至同一水平。

在本章中，本語料同時使用於客觀聽辨度測量和主觀聽辨度量測的實驗。

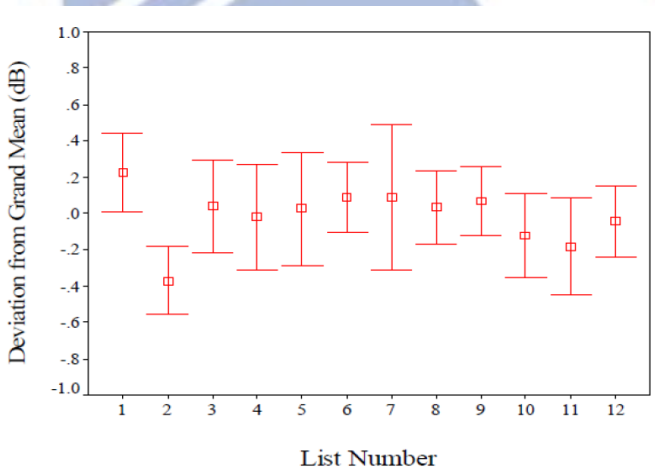


圖 5-5 所有句子組別在安靜及吵雜環境的 RTS 平均值和標準差[30]

5.2.3 語料計分方式

受試者會針對每組測試句子用電腦填寫答案並儲存。但因為可能有同音不同字或是字沒有對齊的狀況發生，我們會以人工的方式核對答案。答對一個字給一分，將每種條件中，統計答對的字數除上此條件中的總字數乘上百分比(ex:某條件共 5 句，50 個字)，來當作最後句子的語音聽辨度。

5.3 客觀中文聽辨度測量

由於主觀測量會耗費大量的時間成本，因此在主觀的心理聲學實驗測量之前，我們先要用客觀聽辨度量測來找到降噪演算法的最佳參數，選用的評估方式是短時客觀聽辨度(STOI)，我們使用表現最佳的降噪演算法模型來進行接下來第 5.4 章節的主觀中文語音聽辨度測量實驗。

5.3.1 短時客觀聽辨度(STOI)

利用短時客觀聽辨度(short-term objective intelligibility)，我們可以計算出受測語音(degraded speech)對於參考語音(clean speech)的語音聽辨度。首先我們將兩者語音降至 10kHz，並且去除靜音(silence)在時間上對齊，用漢寧窗(Hann-window)切割成長度 256 個樣本點、1/2 重疊切割的音框，並且補零至 512 個樣本點。接下來將每個音框分頻成 15 個 1/3 倍頻帶(octave band)，最低中心頻率為 150 Hz，最高為 4.3 kHz，時頻點(TF unit)表示為$X_j(m)$，計算方式為：

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} \left|\left|\hat{X}(k,m)\right|\right|^2} \qquad (5\text{-}2)$$

其中$\hat{X}(k,m)$表示在第 m 個音框經離散傅立葉轉換(DFT)後的第 k 個頻率值，$k_1(j)$~$k_2(j)$為第 j 個 1/3 倍頻帶濾波器組對應的頻寬範圍。也將受測語音的單位時頻點利用同方式計算為$Y_j(m)$。再取鄰近 N 個音框為短時間封包，如下式所示：

$$x_{j,m} = \left[X_j(m-N+1), X_j(m-N+2), \ldots, X_j(m)\right] \qquad (5\text{-}3)$$

其中依據文獻[14]中的 N 取 30，對應音框時間為 384 毫秒，以相同的方式得到受試語音的短時間封包$y_{j,m}$，為了不讓能量大小影響語音聽辨度的判斷，再進行比較前，受測語

音的短時間封包必須經過能量正規化，使其與參考語音的能量相同。接著需要設定語音受損的上限值，代表最嚴重的語音受損狀況，超過此值代表此語音完全無法聽辨，由下式表示：

$$\tilde{y}_{j,m}(n) = \min\left(\frac{||x_{j,m}||}{||\tilde{y}_{j,m}||} y_{j,m}(n), \left(1 + 10^{-\frac{\beta}{20}}\right) \times x_{j,m}(n)\right) \quad (5\text{-}4)$$

其中β為訊損比(signal-to-distortion, SDR)，在文獻[14]中設定-15 為下限值，以下式表示：

$$\text{SDR} = 10\log_{10}\left(\frac{x_{j,m}(n)^2}{\left(\tilde{y}_{j,m}(n) - x_{j,m}(n)\right)^2}\right) \geq \beta \quad (5\text{-}5)$$

計算參考語音與受測語音在每個單位時頻點的短時封包向量的相關係數，並取其平均值即為短時客觀聽辨度，如下式所示：

$$\text{STOI} = \frac{1}{MJ}\sum_{j,m} correlation(\tilde{y}_{j,m}(n), x_{j,m}(n)) \quad (5\text{-}6)$$

短時客觀聽辨度的結果值在 0~1 之間，越大代表聽辨度越高。

5.3.2 實驗條件

根據[13]，深層神經網路模型的訓練語料為將 TMHINT 的前 280 句(每句約 3.6 秒)結合 9 種隨機擷取的同時間長度噪音噪音類型，包含：男生語音 1、男生語音 2、女生語音 1、女生語音 2、群眾歡呼、客機艙噪音、粉紅噪音、汽車噪音和餐廳噪音，並且結合 5 種訊雜比(-10, -5, 0, 5 和 10 dB)，總共 12600 句，共 12.6 個小時。

而測試語料為 TNHINT 的後 40 句(長度約一樣)，結合 2 種不同的噪音型式：雞尾酒宴會噪音和語音形狀噪音(希望對於使用不同噪音類型的測試語料也能帶來與訓練語料一樣的降噪效果)，並且結合 9 種訊雜比(-12, -9, … ,12)，總共 720 句。

最後我們將 40 句的測試語料，不管是吵雜語音或是經過深層神經網路模型的降噪語音，都經過個人化聽損模型和聽覺保留之高頻聲碼模擬器，因此會有六種條件，將每種條件的 40 句輸出音檔計算客觀語音聽辨度後取平均。

5.3.3 實驗測量結果

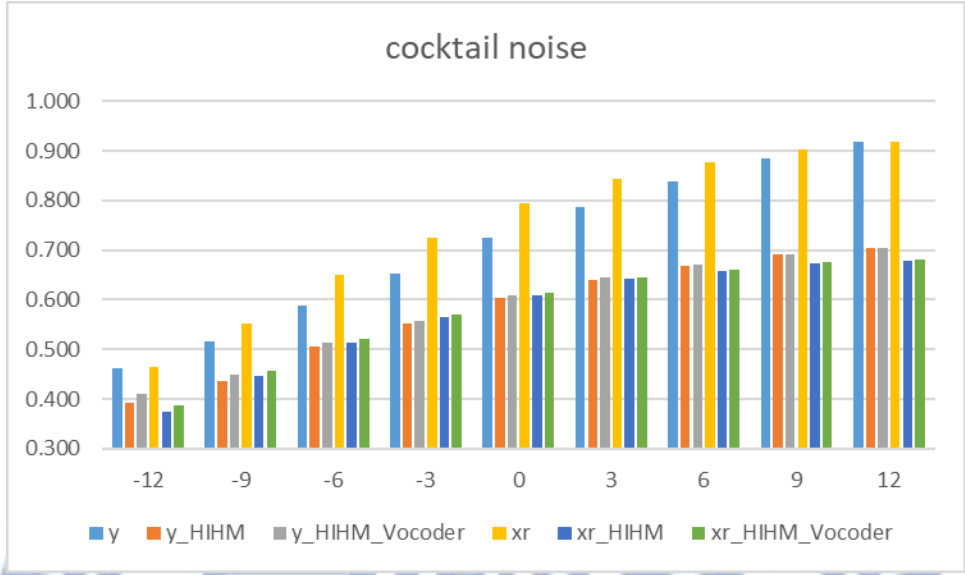


圖 5-6 雞尾酒宴會噪音下的客觀語音聽辨度

由圖 5-6 中，可以發現經過降噪語音(以 xr 表示)的 STOI 分數在所有的 SNR 條件下都比吵雜語音(以 y 表示)來的高，證明此深層神經網路模型在雞尾酒宴會噪音下的降噪效果客觀上是有效的；若觀察經過個人化聽損模型(以 HIHM 表示)和經過聽覺保留之高頻聲碼模擬器(以 HIHM_Vocoder 表示)的降噪語音，我們可以發現其分數在 SNR -9 dB 至 3 dB 皆是有提升，然而在 SNR -12 dB 和 3 dB 以上時分數會下降，可能的原因為在低 SNR 及高 SNR 條件下，降噪演算法對語音所造成的失真會在個人化聽損模型中被放大，產生聽辨度降低的問題。但是整體來說，當 SNR 高的情況下，語音聽辨度本來就很高，即使掉落一點，相信還是有一定的聽辨程度。

圖 5-7 為雞尾酒宴會噪音 SNR -3 dB 的對數頻譜圖，取對數的原因是讓語音輪廓

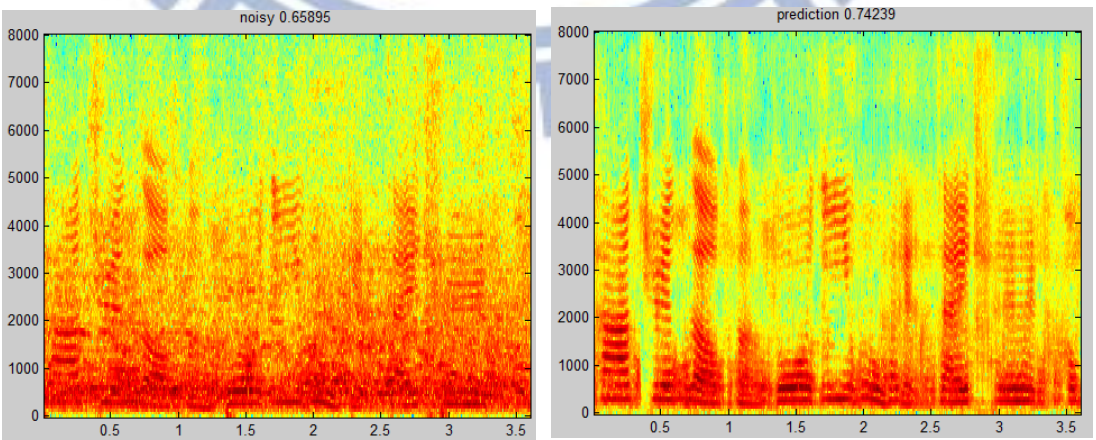


圖 5-7 雞尾酒宴會噪音 SNR -3 dB 下降噪前後的對數頻譜圖

左圖為吵雜語音 右圖為降噪語音

(harmonic contour)更明顯，中文語句為「他捐了很多衣物給災區」，上方數字為此句的STOI 分數。比較降噪前後，發現語音輪廓不管在低頻還是高頻的地方皆有被強化出來，靜音的地方也清楚地降低了能量。

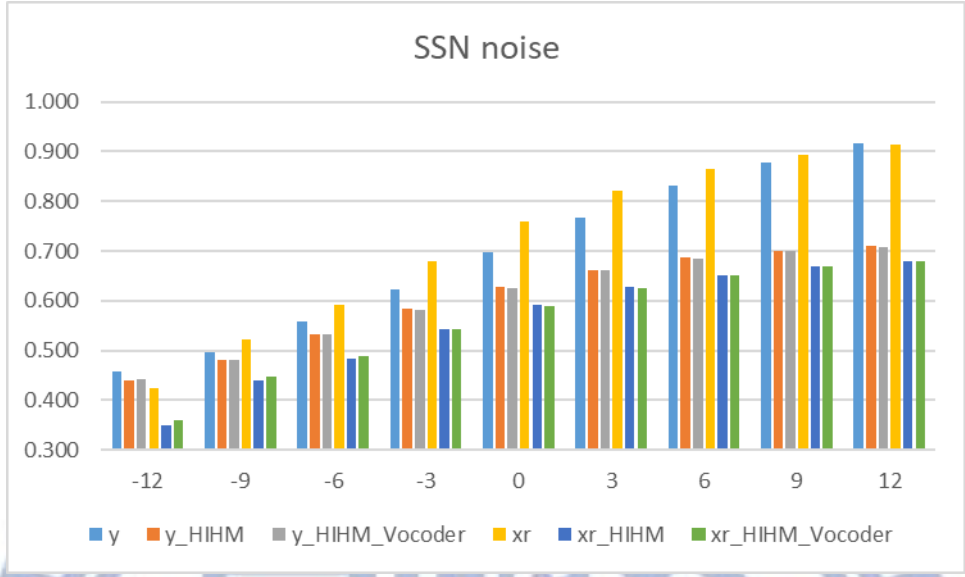


圖 5-8 語音形狀噪音下的客觀語音聽辨度

圖 5-8 為語音形狀噪音下的 STOI 分數，可以看出降噪語音只有在 SNR -12 dB 時其 STOI 值比吵雜語音 STOI 值來的低，可能原因是語音能量被此種噪音覆蓋的太嚴重，使得降噪模型無法辨識出語音的成分，除此之外，降噪語音的 STOI 值皆比吵雜語音的 STOI 值來的高。

然而，在經過個人化聽損模型和經過聽覺保留之高頻聲碼模擬器後，語音聽辨度完不升反降，此結果出乎預期，我們看到圖 5-9 的 SNR -3 dB 下降噪前後的對數頻譜圖，發現縱使語音輪廓有突顯，但是噪音成分還是遍佈在所有頻譜之中，就連靜音的地方也是還有些許的能量殘留。但是這是客觀聽辨度評估，或許語音輪廓的突顯是成功的提升主觀辨識率。

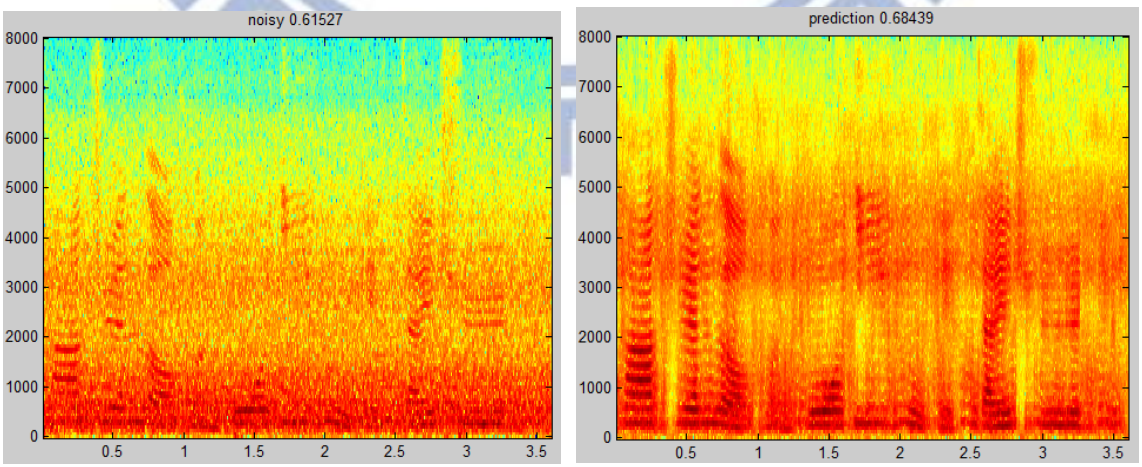


圖 5-9 語音形狀噪音 SNR -3 dB 下降噪前後的對數頻譜圖
左圖為吵雜語音 右圖為降噪語音

5.4 主觀中文聽辨度測量

經過客觀的聽辨度測量，我們將效果最佳的降噪模型(包含前後 5 個音框)，結合聽覺保留之高頻聲碼模擬器，期望在主觀的心理聲學實驗中，也能得到降低噪音可提升中文語音聽辨度的結果。在本主觀實驗中使用的語料是句子，跟第四章所使用的單字不同，我們會用聲碼器模擬來探討聽覺保留之人工電子耳是否有助於中文語句聽辨度的提升。

5.4.1 實驗條件

本次實驗在聽覺保留之高頻聲碼模擬器中，選用高頻聽損最嚴重的 Sub8 聽損模型參數，除此之外，還討論一位頻率解析度更低導致變寬因子更大嚴重的假設性聽損患者，藉此討論降噪演算法合適的使用者條件。此假設患者的最小可聽水平和 Sub8 一樣，三個聽覺濾波器的變寬因子調高至 6.0，如表 5-1 所示。

MALs	Age	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
Sub 8	36	50	50	40	65	90
Hypothesis	36	50	50	40	65	90
BFs	0.5k Hz		1k Hz		2k Hz	
Sub 8	3.2		2.95		4.66	
Hypothesis	**6.0**		**6.0**		**6.0**	

表 5-1 基於深度學習降噪演算法提升語音聽辨度實驗的聽損模型參數

實驗要探討兩種因素：加入人工電子耳是否有助於語句的聽辨和加入降噪演算法是否有助於語句的聽辨。因此產生四種條件，包含(1)聽損的吵雜語音、(2)聽損結合人工電子耳的吵雜語音、(3)聽損的降噪語音和(4)聽損結合人工電子耳的降噪語音。在上述四種條件之下，我們也會討論四種噪音情況：雞尾酒宴會噪音-6 dB 和-3 dB、語音形狀噪音-6 dB 和-3 dB。選擇此兩訊雜比的原因是，在更高的訊雜比情況下，本來就會有很好的語音聽辨度，如此一來識別度就無法區分。

結合探討因素和訊雜比之下，總共產生 16 種測試條件，每種條件使用不重複的 5 句 10 字的句子，共 80 句隨機排序後使測試者進行中文聽辨度的心理聲學實驗。

5.4.2 實驗測量結果與討論

1. Sub 8 聽損模型參數

在使用 Sub8 聽損模型參數的情況下，降噪演算法是否可提升聽辨度？實驗結果如表 5-2 所示。

	cocktail -9 dB	cocktail -6 dB	SSN -9 dB	SSN -6 dB
聽損的吵雜語音	8%	42%	21%	35%
聽損的降噪語音	18%	12%	5%	17%

表 5-2 使用 Sub8 聽損模型參數下語音聽辨度結果

結果顯示對聽損的吵雜語音，聽辨度本身就不低，訊雜比在-6 dB 以上的聽辨度更是達到至七八成，若是經過本論文提出的降噪演算法，反而語音失真的部分會導致聽辨度降低，若將訊雜比調降至-9 dB，在雞尾酒噪音環境中，降噪演算法才使宇與者聽辨度有些微的提升。對於這結果的可能原因，我們觀察吵雜語音和降噪語音經過個人化聽損模型的對數頻譜，如圖 5-10 所示。

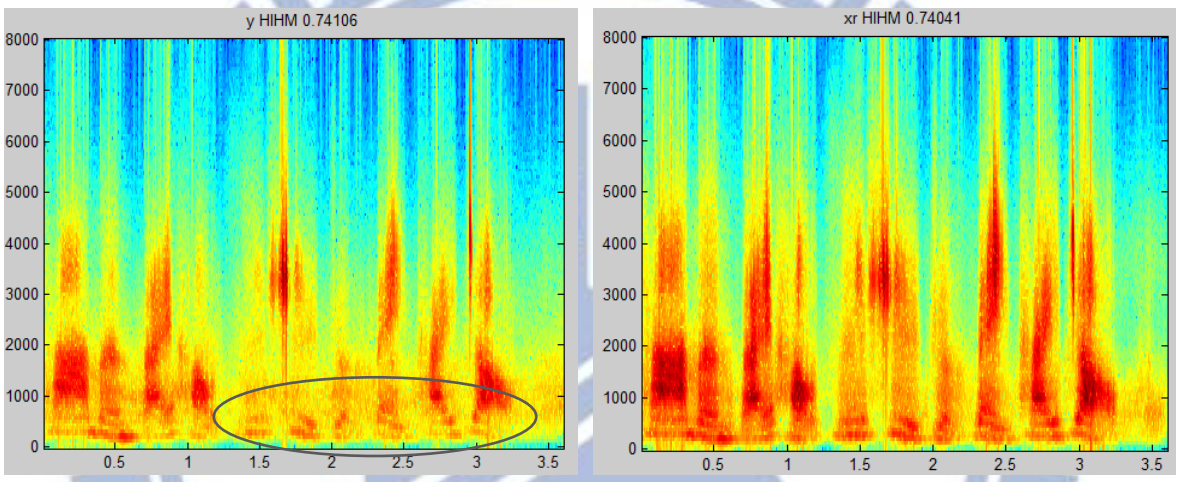


圖 5-10 經過 Sub 8 個人化聽損模型的對數頻譜，左圖為吵雜語音，右圖降噪語音

從圖 5-10 中發現語音經 Sub 8 聽損模型後的語音輪廓幾乎還可辨識，吵雜語音沒有被個人化聽損模型所模糊，而且語句和單字的差別就在於有前後文的連貫性，導致聽取前後文就能推得語言資訊而推測出中間的字詞，以致聽辨度被保留了下來，而降噪演算法造成的失真反而破壞了聽辨度。我們過去在開發個人化聽損模型時[10]，所量測的聽損患者皆為助聽器的使用者，即使最小可聽水平提升，但還有一定程度的頻率解析度，所以當音量調大後，依然對語音有不錯的聽辨度。從過去的實驗中[10]發現，影響頻率解析度的變寬因子 BF 提升至 6 以上才會使語音輪廓模糊化，因此，我們針對某假設性的聽損患者其所有頻率的 BF 參數皆為 6，重複進行實驗。

2. 假設患者聽損模型參數 (All BF = 6)

在經過 9 位正常聽力測試者的主觀測試之後，將他們的分數計算統計後取平均的辨識正確率即顯示在圖 5-11、5-12，四種語音條件分別為：NH 表示聽損的吵雜語音(Noisy speech through Hearing-impaired hearing model)，NHV 表示聽損結合人工電子耳的吵雜語音(Noisy speech through Hearing-impaired hearing model with Vocoder)，DH 表示聽損的降

噪語音(De-noised speech through Hearing-impaired hearing model)，DHV 表示聽損結合人工電子耳的降噪語音(De-noised speech through Hearing-impaired hearing model with Vocoder)。

在此要討論兩個問題：(1)加入人工電子耳是否有助於語音的聽辨程度；(2)加入降噪演算法是否有助於語音的聽辨程度。我們將每種噪音情況下的四個分數，依據這兩個問題的有效性，分成四組兩兩比較：

第一組 在吵雜語音加入人工電子耳的有效性 (比較 NH 和 NHV)
第二組 在降噪語音加入人工電子耳的有效性 (比較 DH 和 DHV)
第三組 在無加入人工電子耳情況下降噪的有效性 (比較 NH 和 DH)
第四組 在加入人工電子耳情況下降噪的有效性 (比較 NHV 和 DHV)

依照這四個組別，分別經過合理反和弦轉換再做變異數分析，討論是否有顯著的差異，在圖中的紅色星號表示，組別比較之下其中一者顯著($p<0.05$)高於另外一者。

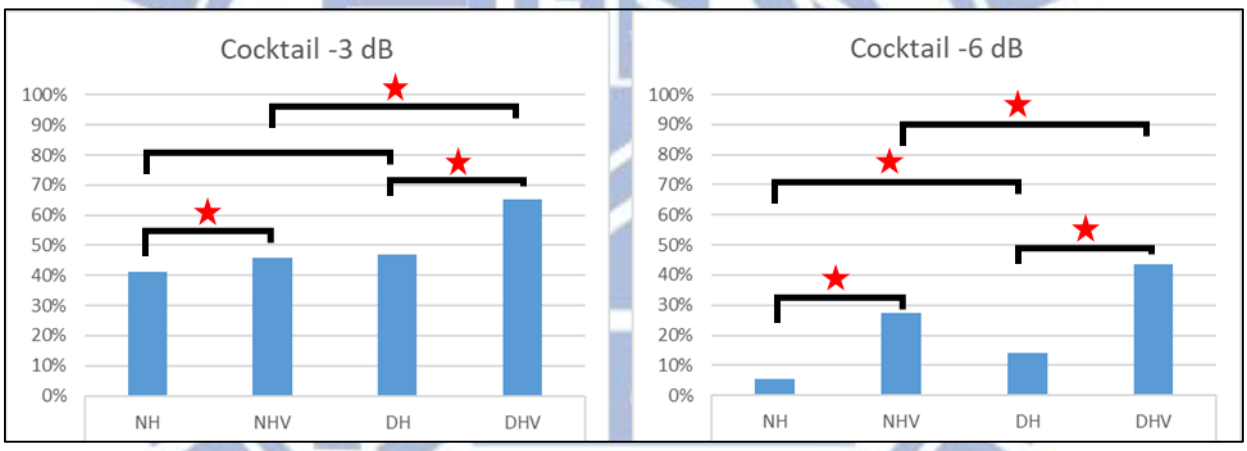


圖 5-11 使用假設患者聽損模型參數的雞尾酒宴會噪音底下之平均辨識正確率

針對雞尾酒宴會的噪音，各比較結果如圖 5-11 中。在 SNR-3 dB 的條件下，第一組別的比較($F(1,16)=33.67$，$p<0.0001$)和第二組別的比較($F(1,16)=4.74$，$p=0.0448$)結果顯著，加入人工電子耳可有效提升語音的聽辨度，在 SNR -6 dB 的條件下，第一組別的比較($F(1,16)=374.88$，$p<0.0001$)和第二組別的比較($F(1,16)=69.42$，$p<0.0001$)也有相同的結論。而討論到降噪的有效性，除了在 SNR -3 dB 的情況下，儘管第三組別的比較($F(1,16)=0.59$，$p=0.4532$)沒有顯著的差異，但 DH 的平均值還是稍微高於 NH 的，其他的條件像是 SNR -3 dB 的第四組別($F(1,16)=6.76$，$p=0.0194$)、SNR -6 dB 的第三組別($F(1,16)=106.26$，$p<0.0001$)和第四組別($F(1,16)=23.95$，$p=0.0002$)的比較都可以發現降噪之後的聽辨度是顯著高於沒有降噪時的聽辨度。

在此噪音環境之下的整體討論，發現加入人工電子耳和加入降噪演算法都是有助於

語音的聽辨，尤其是降低了噪音之後，再加入人工電子耳就能把語音聽辨度提升地更高，因此，直接比較 NH 和 DHV 的情況，發現語音聽辨度分別在 SNR -3 dB 中提高了 24%，在 SNR -6 dB 中提升了將近 40%，可見此兩項方法的有效性非常顯著。

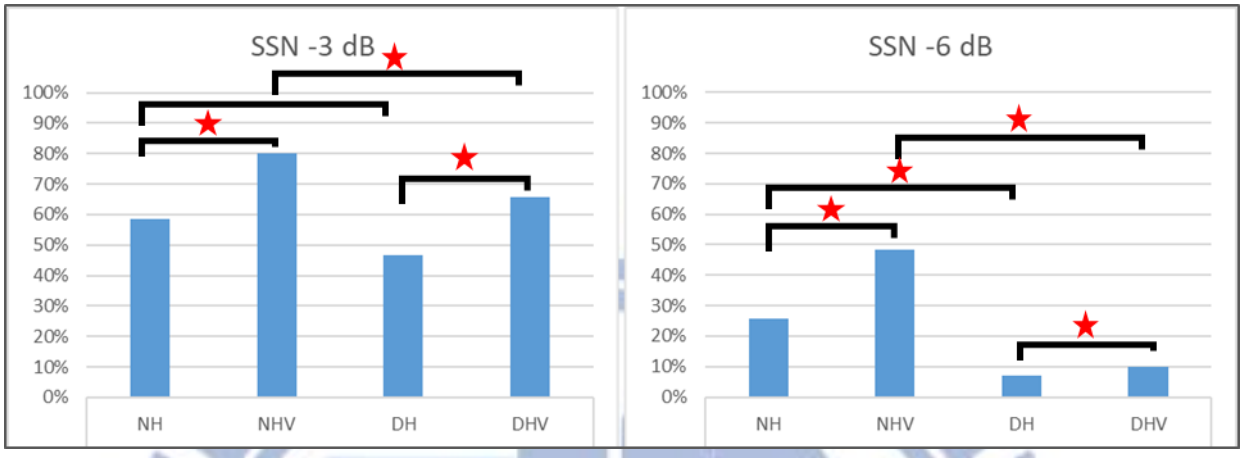


圖 5-12 使用假設患者聽損模型參數的語音形狀噪音底下之平均辨識正確率

針對語音形狀噪音，各比較結果如圖 5-12 中。在 SNR -3 dB 的條件下，第一組別的比較(F(1,16)=8.6，p=0.0098)和第二組別的比較(F(1,16)=10.68，p=0.0048)，和在 SNR -6 dB 的條件下，第一組別的比較(F(1,16)=81.8，p<0.0001)和第二組別的比較(F(1,16)=130.93，p<0.0001)，四種比較結果皆和在雞尾酒宴會的噪音環境的結果相同，加入人工電子耳可顯著提升語音的聽辨度。然而，討論噪音之下的降噪的有效性，SNR -3 dB 的情況儘管第三組別的比較(F(1,16)=2.93，p=0.1064)沒有顯著的差異，但是在 SNR -3 dB 的第四組別的比較(F(1,16)=7.86，p=0.0127)、SNR -6 dB 的第三組別的比較(F(1,16)=11.39，p=0.0039)和第四組別的比較(F(1,16)=5.34，p=0.0345)都很巧妙地發現降噪後的聽辨度會顯著的低於沒有經過降噪演算法的聽辨度，這是一項很奇特的現象，對於此結果竟然與章節 5.3 實驗的客觀聽辨度實驗結果吻合。

因此為什麼對於語音形狀噪音進行降噪之後，會提升語音聽辨度，但是再經過個人化聽損模型反而會使聽辨度降低的這個問題，我們進行更深的討論。

3. 語音形狀噪音下聽辨度降低的問題討論

我們先從兩種噪音差別來討論，雞尾酒宴會噪音為 Babble noise，或許跟訓練語料中的餐廳噪音是相像的，而語音形狀噪音類似於白雜訊，可能在訓練語料中沒有類似的噪音，導致降噪效果不佳，儘管降噪是有效客觀地提升語音聽辨度，但是殘留的噪音和語音失真卻在經過個人化聽損模型後模糊了原本的語音成分。

所以我們將不同時間段的語音形狀噪音加入到訓練語料之中，將總共十種噪音的語料，在其他條件皆與原本相同的條件下，重新訓練深層神經網路模型，得到的語音形狀雜訊客觀聽辨度結果如圖 5-13 所示。

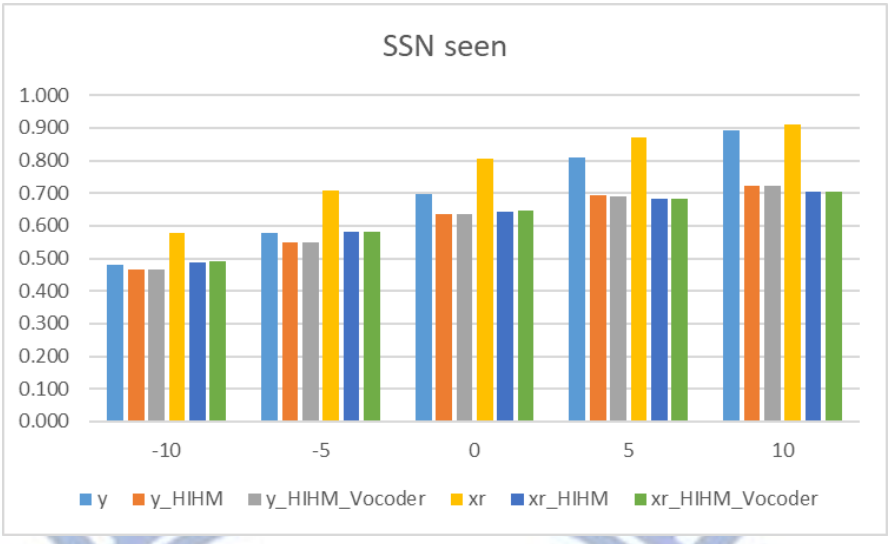


圖 5-13 十種噪音訓練語料 DNN 模型下的 SSN 客觀聽辨度結果

從圖 5-13 中發現，在語音形狀噪音 SNR -10 至 SNR -5 的條件下，經過個人化聽損模型和聽覺保留之高頻聲碼模擬器之後，聽辨度確實提升了，再比較如圖 5-14 所示，九種噪音和十種噪音深層神經網路模型的輸出的對數頻譜圖，比較之後發現，模型訓練過語音形狀噪音之後，會得到比較好的降噪效果，儘管在有語音時段內，所有的頻率上在語音輪廓中的噪音沒有很好的消除，但是有成功的降低靜音時的能量，能把語音斷開有助於語音聽辨。

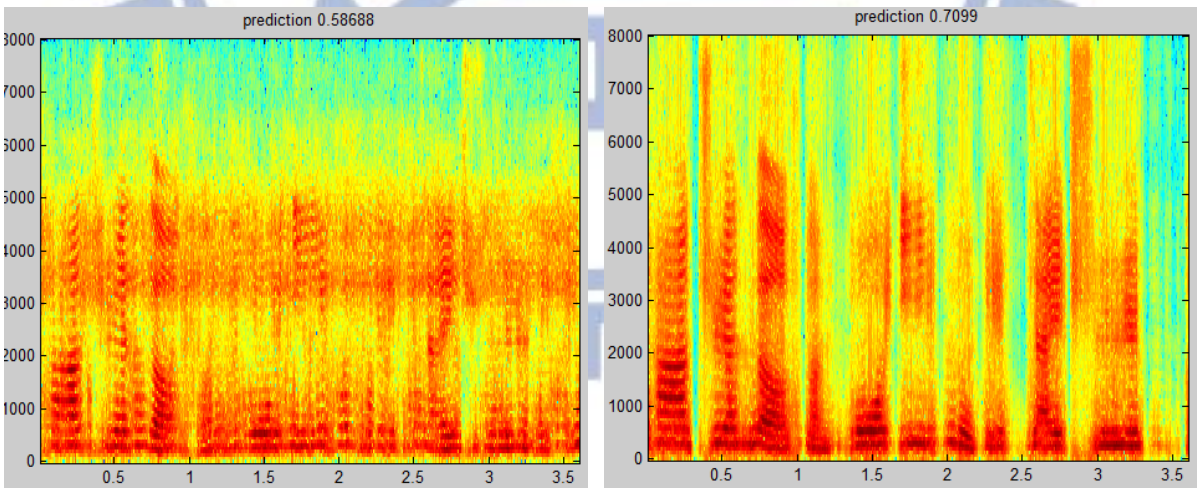


圖 5-14 SSN 情況經過兩種不同模型降噪後的降噪語音對數頻譜圖
左圖為九種噪音訓練語料的 DNN，右圖十種噪音訓練語料的 DNN

因此，從上述第 2 點的假設患者聽損模型參數主觀中文聽辨度測量實驗中，發現結果會跟客觀聽辨度測量的結論相符的結論上，我們可推論在十種噪音深層神經網路模型的訓練結果之下，也能在主觀測量得到相符合的結果。

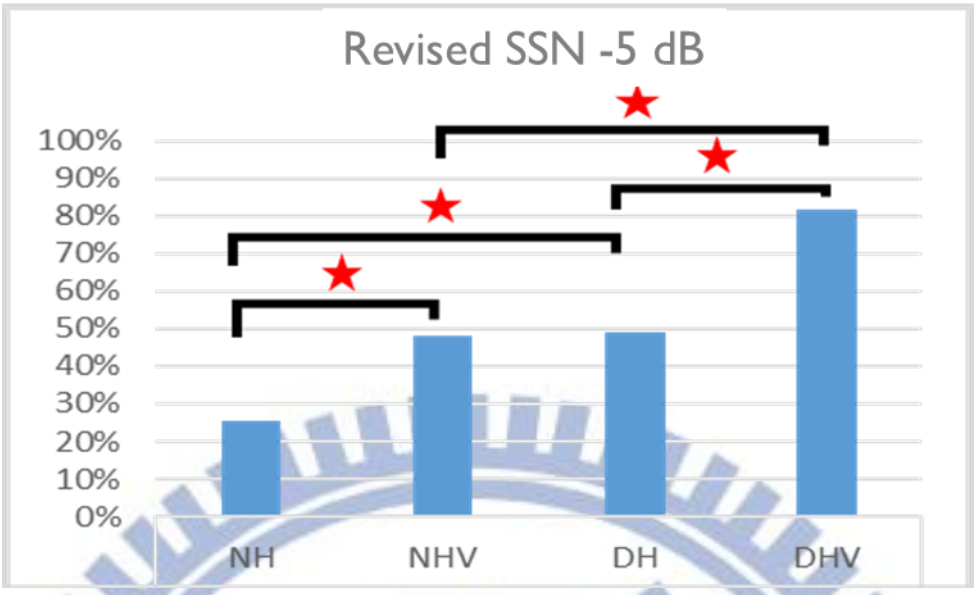


圖 5-15 十種噪音訓練語料 DNN 模型下的 SSN -5 dB 平均辨識正確率

接著，使用主觀聽辨度測量實驗來驗證十種噪音深層神經網路模型的效果，如圖 5-15 中顯示，在 SNR -5 dB 的條件下，第一組別的比較($F(1,16)=8.55$，$p=0.0099$)和第二組別的比較($F(1,16)=32$，$p<0.0001$)結果顯著，加入人工電子耳可有效提升語音的聽辨度。而討論到降噪的有效性，DH 的正確率顯著高於 NH 的正確率($F(1,16)=9.67$，$p=0.0067$)，DHV 的正確率也顯著高於 NHV 的正確率($F(1,16)=28.64$，$p=0.0001$)。因此，在十種噪音訓練語料深度神經網路模型下的語音形狀噪音卻時被消除，使得聽辨度提高，發現完全符合客觀聽辨度實驗的推論。

最終，結合了深度降噪模型的可保留聲響聽覺之高頻聲碼器被證實了有效地提升語音聽辨度，對於聽損患者是一項很大的幫助。

第六章 結論與未來展望

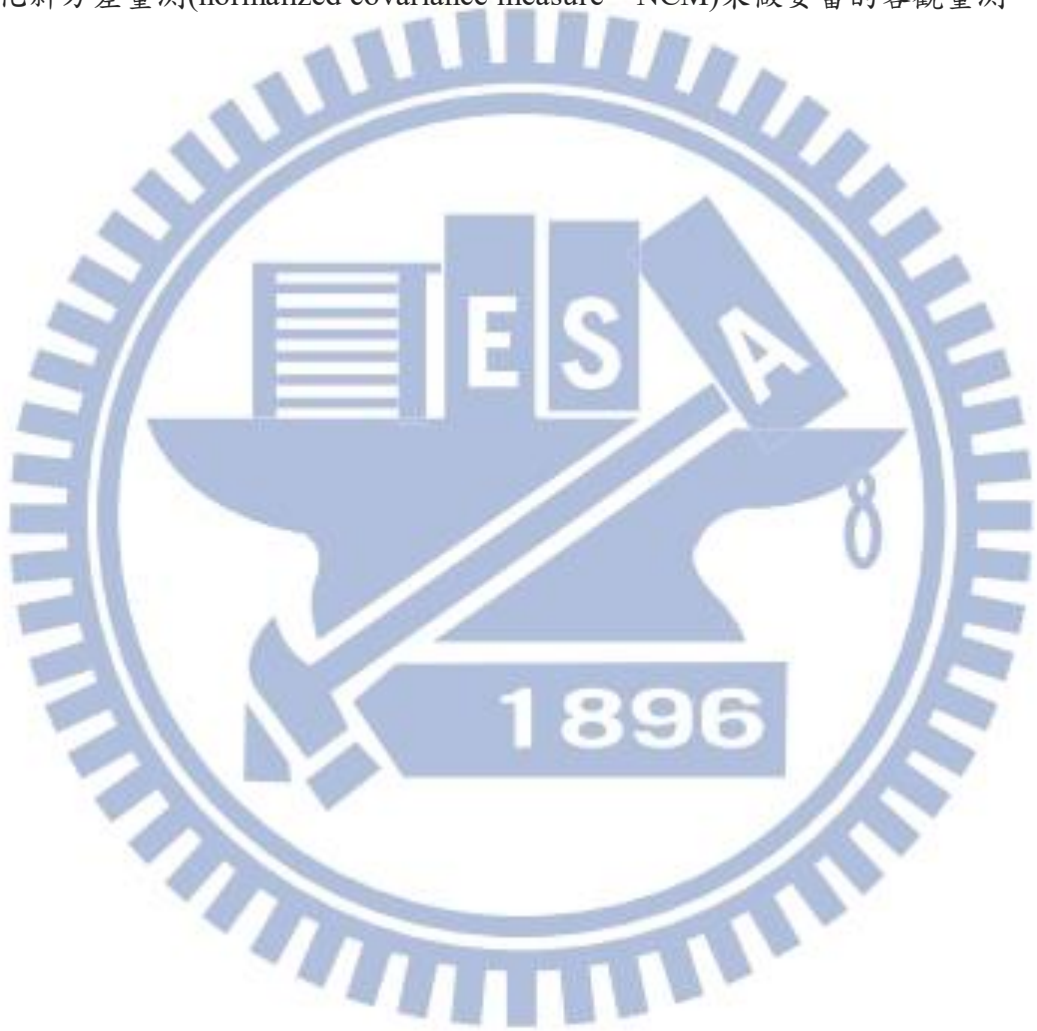
本篇論文的最初目的是在建立一個可以驗證「新型聽覺聽力保留的高頻四電極點人工電子耳系統」的模型，再基於這模型，進而研究演算法以提升語音聽辨度，因此，最終開發了結合深度學習降噪演算法的聽覺保留之高頻聲碼模擬器。此項模擬器能以正常聽力人做心理聲學實驗，加速人工電子耳演算法的開發，提升語音聽辨度，對於聽損病患而言是非常重要的開發。

為了開發這項模擬器，過去實驗室開發的個人化聽損模型是實現保留聲響聽覺很重要的關鍵，再結合聲碼器模擬人工電子耳所提供的電聽覺。在第四章的實驗，成功的驗證在正常環境常見的噪音中，加入人工電子耳的電聽覺是有效的提升中文單字的語音聽辨度，尤其更能提升對中文字的韻母聽辨率。在第五章的實驗，我們驗證此人工電子耳對於中文語句的聽辨度也有實質的幫助，並且利用本模擬器開發人工電子耳演算法速度快速的特性，結合深度學習降噪演算法，實現降噪的效果，對於此章節實驗做出以下結論：

1. 對於不同噪音的抗噪效果，會隨著該種噪音是否有被包含在訓練資料中而有很大的差別，想當然，訓練時加入越多類型的資料，一定對於深度學習的演算法有幫助。
2. 依據實驗結果，我們發現客觀聽辨度測量和主觀聽辨度測量的趨勢會有一致的結果，所以客觀聽辨度也會是很好的開發人工電子耳演算法工具。

有了以上的所有實驗結果，我們確實驗證了此新型人工電子耳對於聽損患者的語音聽辨度有所幫助，然而在未來，為了能夠模擬更多類型的聽損患者，希望能開發更多聽損模型的參數類型，或是測試更多聽損患者的個人化參數，進而幫助個人化聽損模型的

完整度；也為了真正能實作在助聽器或人工電子耳上，希望能發展出一套架構小並且可以降噪的深層神經網路；演算法開發上，雖然降噪對於本實驗模型有效果，但是此降噪的語音聽起來不連續的問題，希望提出更好合乎人工電子耳的降噪演算法，像是遞迴神經網路等來提升語音品質，並且，除了能降噪也要能抗迴響，更加應付現實生活中的情況；另外，我們使用到聲碼器來模擬人工電子耳，在評估客觀語音聽辨度上，可以使用標準化斜方差量測(normalized covariance measure，NCM)來做妥當的客觀量測。



參考文獻

[1] J. Reefhuis, et al. “Risk of bacterial meningitis in children with cochlear implants,” New England Journal of Medicine, vol. 349, no. 5, pp. 435–445, 2003.

[2] National Taiwan University Hospital, ENT, report. https://www.ntuh.gov.tw/ENT/DocLib9/HSUCJ20110829.pdf

[3] P. M. Sellick, et al. “Measurement of basilar membrane motion in the guinea pig using the Mössbauer technique,” The journal of the acoustical society of America, vol.72, no. 1, pp.131-141, 1982.

[4] M. A. Ruggero, and N. C. Rich, “Furosemide alters organ of Corti mechanics: Evidence for feedback of outer hair cells upon the basilar membrane,” Journal of Neuroscience, vol. 11, no. 4, pp. 1057-1067, 1991.

[5] L. E. Humes, and L. Roberts, “Speech-recognition difficulties of the hearing-impaired elderly: The contributions of audibility,” Journal of Speech, Language, and Hearing Research. vol. 33, no. 4, pp. 726–735, 1990.

[6] X.-H. Qian et al., “A bone-guided cochlear implant CMOS microsystem preserving acoustic hearing,” in Proc. Int. Symposium on VLSI Circuits. IEEE, pp. C46–C47, 2017.

[7] R. D. Patterson, and I. Nimmo-Smith. “Off-frequency listening and auditory-filter asymmetry,” The Journal of the Acoustical Society of America, vol. 67, no. 1, pp. 229-245, 1980.

[8] R. V. Shannon, F.-G. Zeng, and J. Wygonski, “Speech recognition with altered spectral distribution of envelope cues,” The Journal of the Acoustical Society of America, vol. 104, no. 4, pp. 2467–2476, 1998.

[9] R. V. Shannon, et al. “Speech recognition with primarily temporal cues,” Science, vol. 270, no. 5234, p. 303, 1995.

[10] P.-C. Tsai, S.-T. Lin, W.-C. Lee, C.-C. Hsu, T.-S. Chi, and C.-F. Lee, “A hearing model to estimate mandarin speech intelligibility for the hearing impaired patients,” In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference, pp. 5848–5852, 2015 .

[11] T. Green, A. Faulkner, and S. Rosen, “Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants,” The Journal of the Acoustical Society of America, vol. 112, no. 5, pp. 2155–2164, 2002.

[12] B. Roberts, R. J. Summers, and P. J. Bailey, “The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes,” Proc. of the Royal Society of London B: Biological Sciences, vol. 278, no. 1711, pp. 1595–1600, 2011.

[13] Y. H. Lai, et al. “A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation,” IEEE Transactions on Biomedical Engineering, vol. 64, no. 7, pp. 1568–1578, 2017.

[14] C. H. Taal, R. C. Hendriks, et al. “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125-2136, 2011.

[15] T. S. Chi, class notes of Auditory and Acoustical Information Processing, Department of Communication Engineering, National Chiao-Tung University, Taiwan, 2016.

[16] J. Rouat, “Computational auditory scene analysis: Principles, algorithms, and applications (wang, d. and brown, gj, eds.; 2006)[book review],” IEEE Transactions on Neural Networks, vol. 19, no. 1,pp. 199-199, 2008.

[17] G. Heinzel, A. Rüdiger, R. Schilling, “Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new flat-top windows,” 2002.

[18] W.-C. Lee, and T.-S. Chi, “A Construction of Hearing Impaired Cochlea Model about

Loudness and Frequency Selectivity" A Thesis for the Degree of Master of Science in Communication Engineering, National Chiao-Tung University, 2011.

[19] R. D. Patterson, et al. "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," The Journal of the Acoustical Society of America, vol. 72, no. 6, pp. 1788-1803, 1982.

[20] B. C. J. Moore, B. R. Glasberg, R. F. Hess, and J .P. Birchall, "Effects of flanking noise bands on the rate of growth of loudness of tones in normal and recruiting ears," The Journal of the Acoustical Society of America, vol. 77, no. 4, pp. 1505-1515, 1985

[21] B. R. Glasberg, B. C. Moore, "Auditory filter shapes in subjects with unilateral and bilateral cochlear," The Journal of the Acoustical Society of America. vol. 79, no. 4, pp. 1020-1033, 1986.

[22] Auditory Perception Group University of Cambridge provides Auditory demonstrations and useful software.

[23] C. Jurado, D. Robledano, "Auditory filters at low frequencies: ERB and filter shape," Technical report, Aalborg University, Spring 2007

[24] D. R. Soderquist, and J. W. Lindsey, "Physiological noise as a masker of low frequencies: the cardiac cycle," The Journal of the Acoustical Society of America, vol. 52, no. 4B, pp. 1216-1220, 1972.

[25] V. Nedzelnitsky, "Sound pressures in the basal turn of the cat cochlea," The Journal of the Acoustical Society of America, vol. 68, no. 6, pp. 1676-1689, 1980.

[26] T.J. Lynch, V. Nedzelnitsky, and W.T. Peake, "Input impedance of the cochlea in cat," The Journal of the Acoustical Society of America, vol. 72, no. 1, pp. 108-130, 1982.

[27] J.J. Zwislocki, "The role of the external and middle ear in sound transmission," The Nervous System, vol. 3, pp. 45-55, 1975.

[28] P. C. Loizou, "Introduction to cochlear implants," IEEE Engineering in Medicine and Biology Magazine, vol. 18, no. 1, pp. 32-42, 1999.

[29] F. Chen, and A. H. Lau. "Effect of vocoder type to Mandarin speech recognition in cochlear implant simulation," Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE, Sep. 2014.

[30] L. L. Wong, S. D. Soli, S. Liu, et al. "Development of the Mandarin Hearing in Noise Test (MHINT) ," Ear and hearing, vol. 28, no. 2, pp. 70–74, 2007.

[31] N. A. Whitmal III, et al. "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," The Journal of the Acoustical Society of America, vol. 122, no. 4, pp. 2376-2388, 2007.

[32] P.-C. Tsai, "A Study of Perceptual Effects of Spectral Sharpening on the Hearing-impaired (Unpublished master's thesis)," Institute of Communications Engineering, National Chiao Tung University, Taiwan. 2014.

[33] K. S. Tsai, L. H. Tseng, C. J. Wu, S. T. Young. "Development of a Mandarin Monosyllable Recognition Test," Ear and Hearing, vol. 30, no. 1, pp. 90-99. 2009.

[34] G. A. Studebaker. "A "rationalized" arcsine transform," Journal of Speech and Hearing Research, vol. 28, no. 3, pp. 455-462, 1985.

[35] X. Lu, et al. "Speech enhancement based on deep denoising autoencoder," In Interspeech. pp. 436-440. Aug. 2013.