國 立 交 通 大 學

電信工程研究所

碩 士 論 文

時域封包上的雜訊消除

Noise Reduction in Temporal Modulation Domain

研 究 生：紀雅文

指導教授：冀泰石

中 華 民 國 一百零一 年 十 月 十六 日

時域封包上的雜訊消除

Noise Reduction in Temporal Modulation

研 究 生：紀雅文　　Student：Ya-Wen Chi

指導教授：冀泰石　　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

October 2012

Hsinchu, Taiwan, Republic of China

中華民國一Ｏ一年十月

時域封包上的雜訊消除

學生：紀雅文　　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

本論文所提出之方法是在單通道下對每個子頻帶在時域上的調變封包作雜訊消除，對輸入的語音訊號作了兩層的遮蔽，第一層是根據語音與雜訊能量大小的差異設計出了每個頻帶的第一個臨界值，將判定為非語音的音框作遮蔽，接著第二層則是將語音封包作了快速傅立葉轉換，根據語音與雜訊在調變振幅上的差異設計了每個頻帶的第二個臨界值，將上一層所誤判實際上則為非語音的部分再加以壓抑，為了使效果更加顯著，我們在系統前端加入在 spectral 上進行去雜訊的 Wiener 濾波器作為預強濾波器，因此整個系統為分別作在 spectral 與 temporal 上最後再結合的方法，其計算複雜度低，所需時間較少，未來可能應用於助聽器上。

在後面的實驗評估裡，我們選了只作在 spectral 上的 Wiener 濾波器、本篇所提出的只作在 temporal 上的封包調變消噪法以及在 spectro 與 temporal 上同時處理的 Joint spectro-temporal subband Wiener filter 這三種方法作比較，採用的是客觀評分，分別為 PESQ 與 IS distance，並且也將四種方法計算所需的時間也一併比較，所得之結果在客觀評分裡，Joint spectro-temporal subband Wiener filter 的效能最好，其次是本篇所提出的封包調變消噪法與 Wiener 濾波器結合的系統，而 Wiener 濾波器與封包調變消噪法相比之下，在高斯白雜訊的環境裡，當訊雜比較高，Wiener 濾波器的分數比較高，然而當訊雜比越差時，封包調變消噪法的分數會越來越接近 Wiener 濾波器，甚至在 0 dB 的情況時優於前者。

Noise reduction in temporal modulation domain

Student: Ya-Wen Chi Advisor: Dr. Tai-Shih Chi

Institute of Communication Engineering

National Chiao-Tung University

Perception Signal Processing Laboratory

Abstract

We propose a single-channel two-stage masking algorithm based on temporal modulations for noise reduction. The first masking stage is based on the temporal modulation energy and the second stage is based on the amplitude modulation of the input signal to distinguish speech from non-speech segments. The algorithm is developed under a filter bank structure with a frame-by-frame analysis paradigm. The pure temporal noise reduction algorithm is then combined with a conventional Wiener filter for further enhancement of speech. The whole system conducts noise reduction in spectral and temporal domain separately and it may be applied on the digital hearing-aid in the future since the computation complexity is low comparing with the complexity of the joint spectro-temporal subband Wiener filter.

As for the performance comparison, we evaluate four systems in this thesis. They are: (1) the proposed pure temporal noise reduction algorithm, (2) a conventional Wiener filter, (3) a joint spectro-temporal subband Wiener filter and (4) the proposed temporal algorithm combined with the conventional Wiener filter. Objective measures of PESQ and IS distance are used in our evaluations. The system (4) outperforms system (1) and (2) and has slightly lower performance than the system (3). However system (4) can achieve the request of real-time process compare to the joint spectro-temporal subband Wiener filter.

誌　　謝

這篇論文的完成有好多好多人要感謝。

首先是我的指導老師，冀泰石教授。剛開始對環境的不熟悉老師總是很關切，到後來不只是課業、研究甚至是日常生活一些自己的迷糊導致的麻煩，老師都給了我相當多的指引與幫助。在研究上毫無頭緒時，老師也總是很有耐心的給予指導，讓我不致於在研究這條路上迷失方向，到現在跌跌撞撞也總算是完成了這一本論文，最感謝的是老師，最幸運的是我，能成為老師的學生。

再來是實驗室的學長姐們，同學們，感謝阿郎、大師、大樹、勝哥、文中、華山、靖雯、雞排、張彰、家銘、炮哥、小何、勛正以及暐桓在課業、研究還有日常生活中給了我相當多的幫助，也常常會一起聊天開玩笑一起去打球不致讓研究生活太過苦悶，能成為感知訊號實驗室的一份子我真的很開心，也為將來的學弟妹們加油。感謝消息理論實驗室的朋友們常常邀我去聊天喝咖啡，三不五時還會找我一起運動，讓我不致於怠惰發胖。感謝我的朋友們。這期間你們的加油和鼓勵從沒少過，謝謝你們給了我很多幫助。

最後要感謝我的父母和家人，含辛茹苦拉拔我到大，讓我無後顧之憂的念書，給予我最多的鼓勵和關心並包容我的一切。

目　　錄

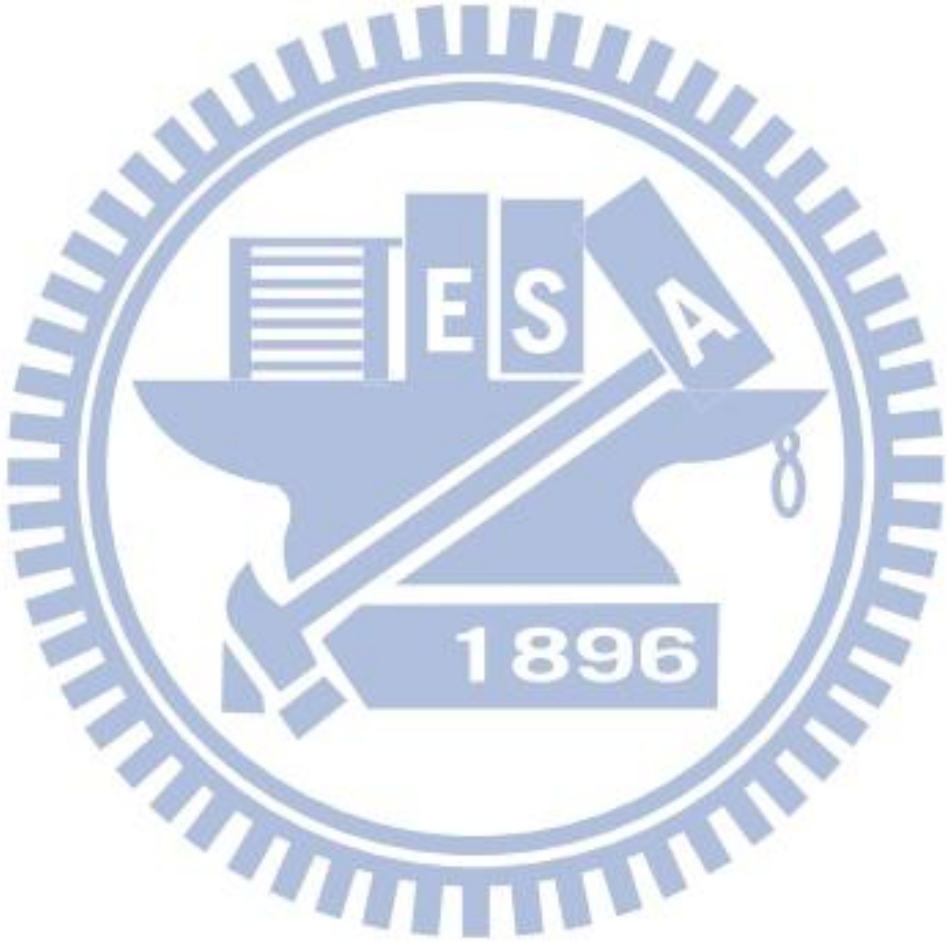


表 目 錄

圖 目 錄

一、 緒論

1.1 研究背景

在現今科技發達的社會，手持通訊產品、語音辨識器乃至於助聽器等等對於語音的傳輸都有相當大的需求，然而生活中存在著各種特性不同的背景雜訊訊號，在使用的儀器設備裡也無法避免的會產生一些人為雜訊，這些都會導致語音辨識率以及語音品質大幅降低，雜訊對語音的破壞對於聽損的患者造成的影響更甚，因此消噪的技術在助聽器上顯得更為重要。

在過去幾十年裡，已有相當多語音增強的研究，有在時域上 (temporal)對語音作降噪，也有在頻域上 (spectral)作，也有兩者皆有的系統，例如: spectral subtraction [8]、Wiener filter [8]、minimum mean square error (MMSE) [8]、Karhunen-Loeve transform (KLT) [8]、Kalman filter [38] 以及 phase spectrum compensate (PSC) [28]。

近幾年內，已有多位學者基於聽覺感知的原理對語音訊號作處理的技術提出，如 [1]。由於人耳對於不同的語音之間以及語音和背景雜訊之間的分辨相當精準，這些基於聽覺感知訊號處理技術的特色就在於模擬人耳對於接收到的訊號作分析、處理以及合成，以求能達到跟人耳一樣的效果，在 [1]有針對聽覺感知系統的詳細解說。隨後也有研究致力於在對語音作時域軸和頻域軸上的感知分析，最後再把處理過後的訊號合成回語音，而這些基於聽覺感知的語音處理技術在結果上優於傳統技術，如 [4]、[12]、[15]以及 [16]。

本篇論文所提出的方法其系統流程是基於 AMS (Analysis-Modification-Synthesis) 架構下所作的訊號處理，如圖 1：

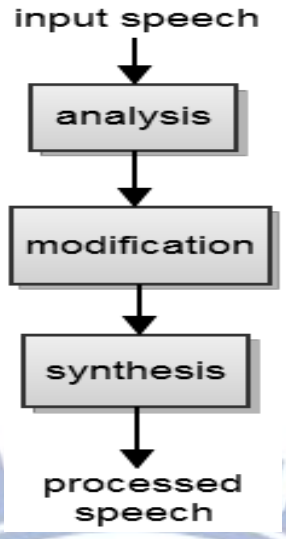


圖 1　　AMS 架構

現今有相當多系統皆在此架構下進行語音增強，如[7]、[26]以及 [29]等等。

1.2　研究動機

在這個資訊傳輸量龐大的世代，人與人之間最直接的語音溝通仍然佔有舉足輕重的地位，但由於日常生活裡各種背景雜訊的干擾，會導致接收端那方在收聽語音時無法得到乾淨且正確的資訊，更甚者，對於聽損患者來說，過多過大的干擾源對他們而言會讓語音的接收更加吃力。

考慮到未來應用在助聽器上之可行性，本研究除了著重於語音的雜訊消除與聽覺感知的特性結合以外也考量到助聽器的即時處理需求，亦即計算量有一定的限制，如 [2]就是為了將 MMSE 應用在助聽器上，必須將此方法中的反矩陣計算步驟以即時且運算速度較為快速的迭代方式取而代之。

而本篇論文所提出的方法是 Wiener 濾波器 [8]後加上一個時域調變後強濾波器 (post-filter) 把剩餘雜訊再加以消除，此方法將輸入的語音分成若干音框，若當下此音框被判定為非語音，則作遮蔽令此音框為極小值；若被判定為語音，則僅作極輕微的遮蔽，此作法可將 Wiener 濾波器在訊雜比 (signal to noise ratio) 低時處理後留下的過多剩餘雜訊再作進一步的壓抑，若應用在助聽器上，可加強其處理後的語音清晰度，且經實驗測試後 perceptual evaluation of speech quality (PESQ) [9]與 Itakura-Saito distance (IS dist.) [10]在四種不同訊雜比兩種不同雜訊環

境的情況下其結果分數較傳統 Wiener 為佳，並與本篇提及時域封包消噪系統比較，其結果分數亦較優，同時也跟 Joint spectro-temporal Wiener filter [7]做比較，雖然結果略差，但計算複雜度較低，因此計算速度也較 Joint spectro-temporal Wiener filter 為快，詳細的內容將於後面解說。

1.3 章節大綱

接下來要介紹的各章內容如下：

第二章為基礎理論的簡介，包括了聽覺的行成、接收到的訊號在時頻域上的分析、進入大腦皮質層所作的更高維度分析以及各種生理限制所造成的聽覺現象；

第三章介紹了維納 (Wiener) 濾波器、本篇論文所提出的方法以及此方法所依據的訊號分析特性；

第四章則設計了實驗參數與測試方法，與另外三種不同的系統在各環境下比較其語音品質的高低，以此來作為系統性能評比的依據；

第五章提出結論並對於現階段的系統提出可能提升效能的方法。

二、基礎理論介紹

2.1 訊號之初期分析

2.1.1 聽覺的產生

圖 2 為人耳的構造：

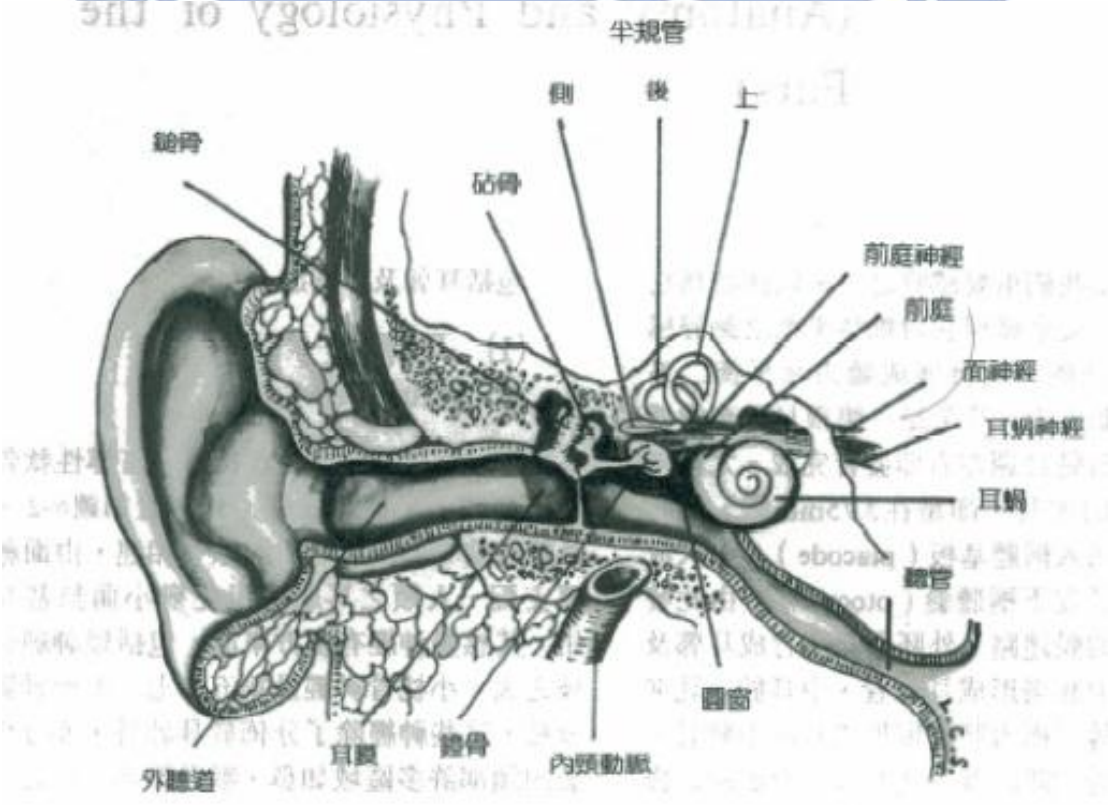


圖 2 人耳構造示意圖

資料來源：[3]

在外耳的部分，聲波在經過外耳的耳殼收集之後，由外聽道傳送至鼓膜產生振動。連接著也振動著中耳部分的三小聽骨（砧骨、鎚骨以及鐙骨），藉由三小聽骨將聲波轉換為機械能，並傳送至三小聽骨連接的另一端也就是內耳的圓窗，藉由三小聽骨擠壓圓窗內的液體產生行進波，藉由耳蝸裡的基底膜（basilar membrane）來傳遞行進波。

不同頻率的訊號所產生的行進波會在基底膜上不同的部位產生最大的振福，主要是因為基底膜的橫向寬度遞增且越接近圓窗基底膜的質地會越硬，也因為此兩種特性，越高頻的訊號在越接近圓窗的部位（base）會有最大的振幅，而越低頻的訊號則是在基底膜的越深處（apex）有最大振幅，因此基底膜可視為是一組由高頻分布至低頻的濾波器組，其特徵頻率大約是20~20,000 Hz。圖 3為示意圖：

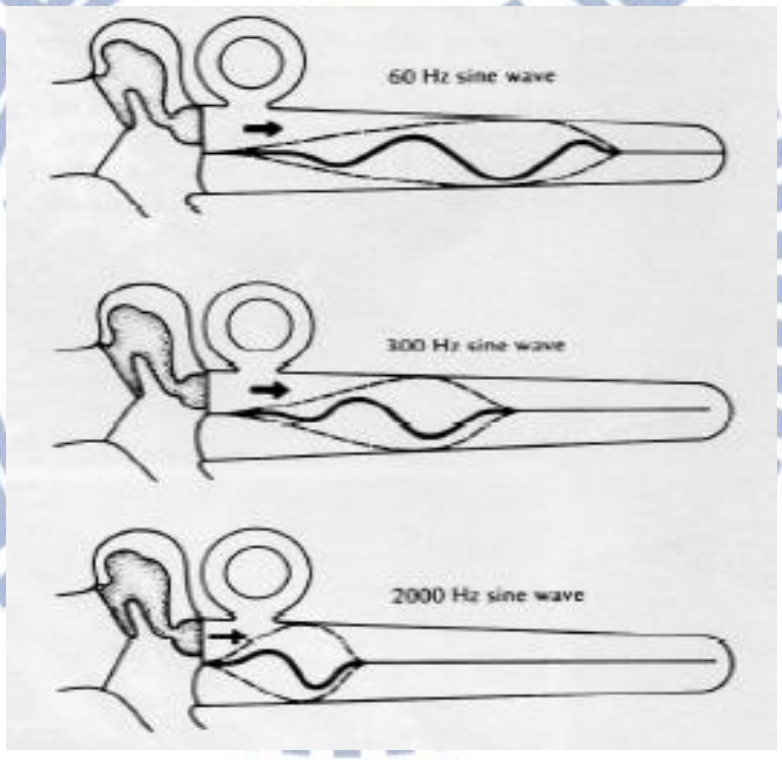


圖 3　　不同頻率的行進波在基底膜不同位置上的振動情形
資料來源：[3]

而基底膜上佈有柯替式器（organ of Corti），如圖 4，主要由毛細胞（hair cell）組成，毛細胞主要是將行進波轉換為神經傳導電能，藉由毛細胞的放電、表面覆膜的拉扯以及內耳組織液的流動產生電位差，因此將行進波轉換為電能，由連接的聽覺神經將聲音傳送至大腦，進行更高階的處理。毛細胞依照分布位置與功能又可細分為外毛細胞（outer hair cell）與內毛細胞（inner hair cell），外毛細胞主要是增強聽覺神經在頻率上的選擇性、放大電位能以及當輸入訊號強度過強達到

臨界值時加以壓抑避免產生過大電能傷害大腦的保護動作。內毛細胞則是與許多的聽覺神經連接，將己身所感受到的機械能大小轉化成與自己相連的聽覺神經之高低電位，主要的能量轉換是由內毛細胞完成。以上過程可參照下圖 5：

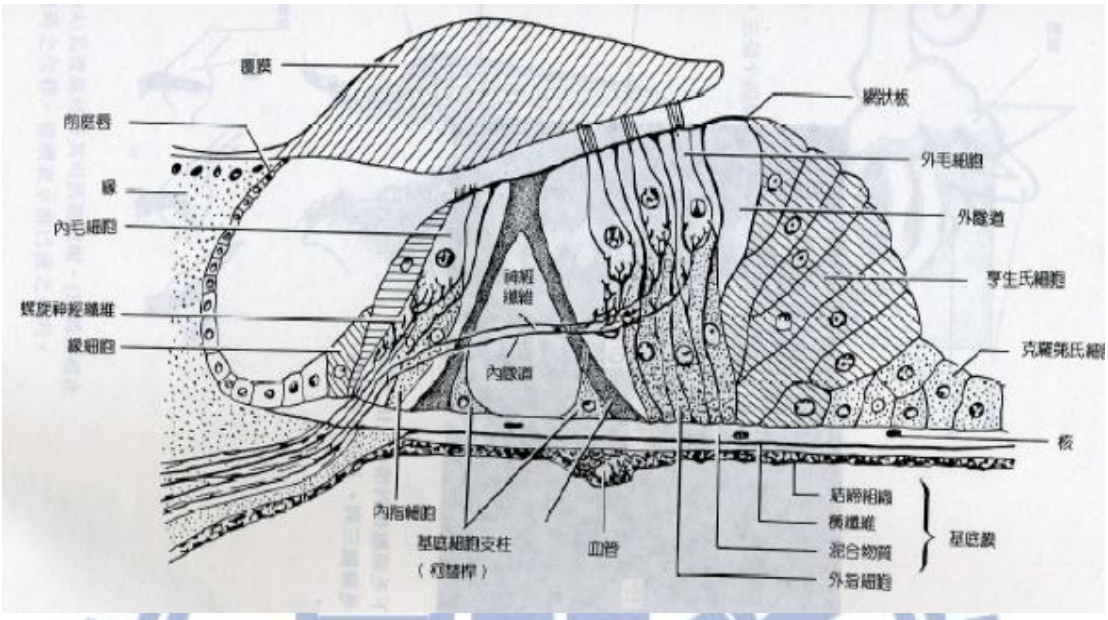


圖 4　　柯替式器 (organ of Corti)
資料來源：[3]

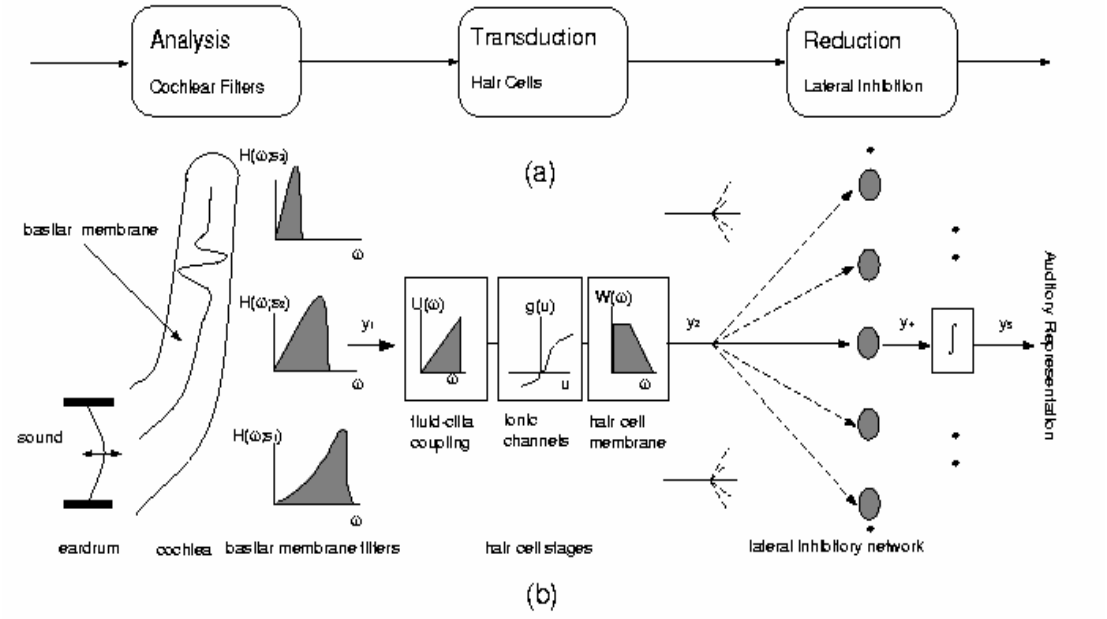


圖 5　　行進波進入耳蝸內轉換為由低頻至高頻的電訊號示意圖
資料來源：[3]

2.1.2 生理的聽覺現象

觀察上圖 5 我們可發現基底膜上的另一特性，那就是不同位置的脈衝響應 (impulse response) 會隨著中心頻率 (center frequency) 越高，其脈衝響應的頻寬也會越寬，此一特性在後面的模擬中會加以詳述。

圖 6 為行進波各頻率在基底膜上產生響應的位置：

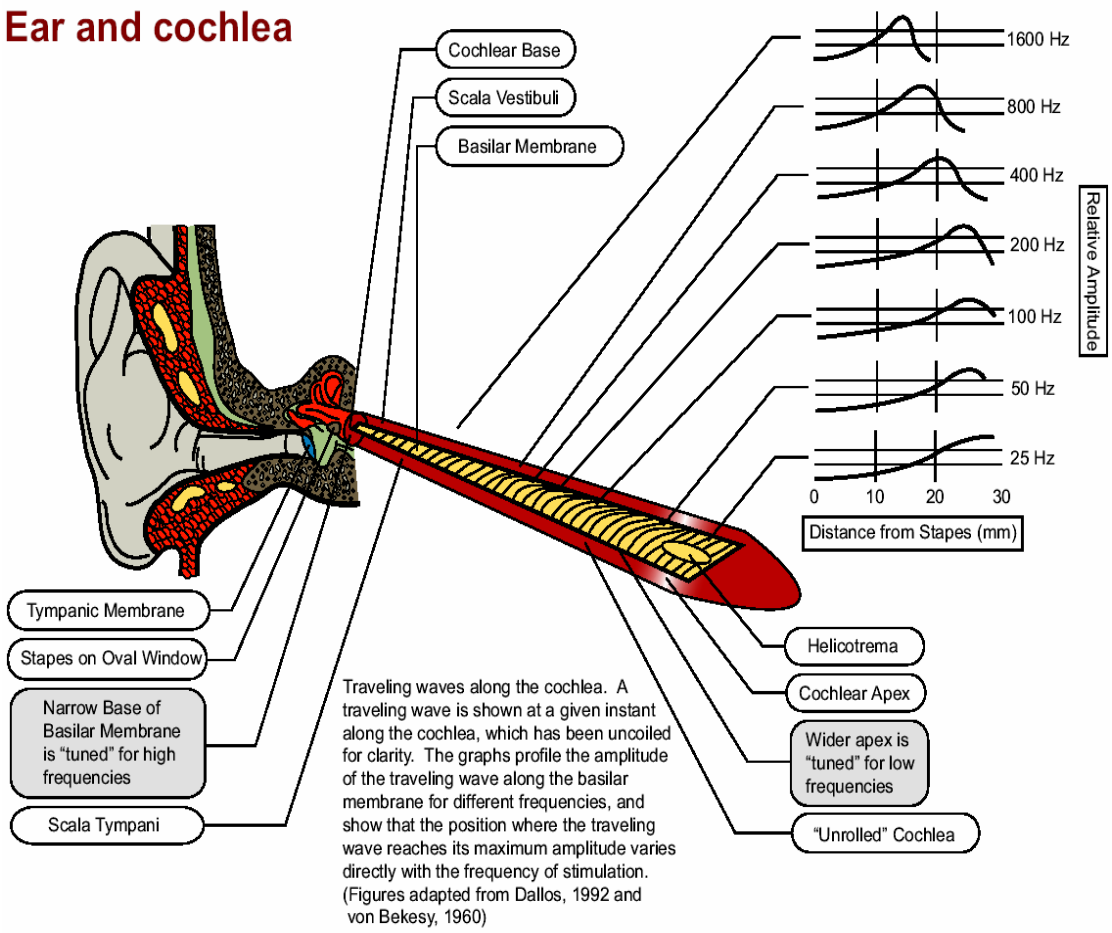


圖 6　　行進波上相差倍頻之成分於基底膜的最大振幅位置示意圖
資料來源：[3]

從圖 6 裡可發現，每個相差兩倍頻率的行進波之間，在基底膜上的間距都是等距，這可以解釋基底膜上各個準濾波器的中心頻率是呈現對數分佈而非一般常用的線性分布，而此現象會在後面的聽覺模型佔有重要的地位。

圖 7 為行進波在基底膜上的振動示意圖：

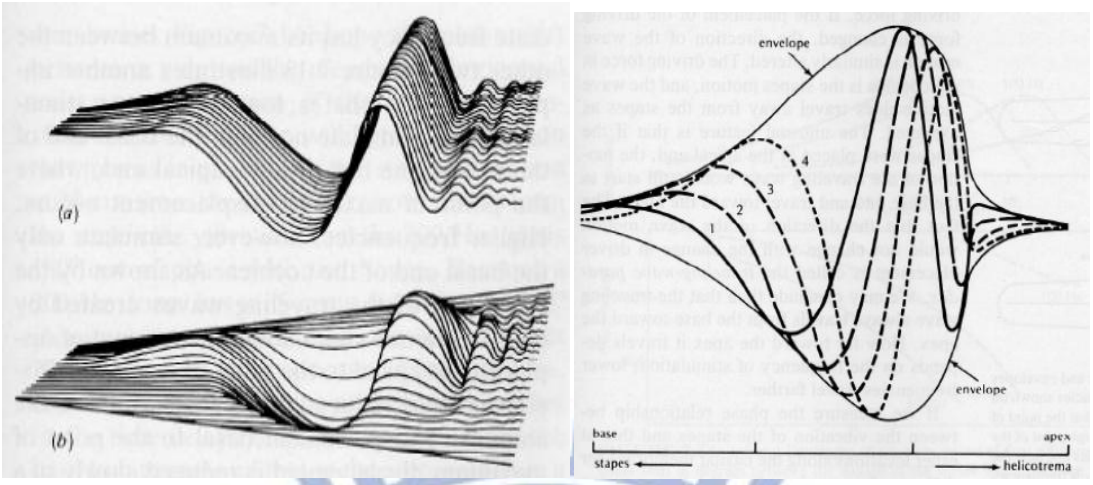
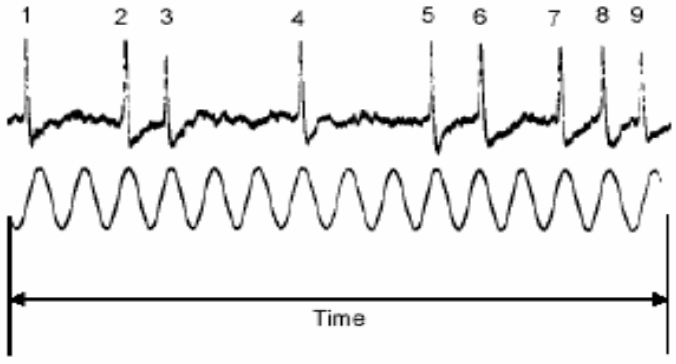


圖 7　　基底膜上行進波的振動示意圖

資料來源：[3]

在圖 7 的行進波示意圖中可以觀察到，以基底膜產生最強響應的地方為原點 (也就是封包最強振幅處)，往左右兩邊均有衰減現象，往高頻處的衰減比較慢，往低頻衰減的速度則比較快。當一個聲音包含兩個鄰近頻率的單音，其中一個單音的振動模式幾乎被另一個單音的振動模式所包含時，就會產生遮蔽效應 (masking effect)。

由於內毛細胞將行進波轉換成與連接神經之間的高低電位差，訊息得以能透過神經繼續傳遞至大腦，但因為神經元在發射出動作電位後，會進入靜止電位，使得神經發射速率無法跟上較高頻的訊號，內毛細胞最高的神經發射速率大約在 4000~5000 Hz，示意圖如圖 8：



Firing Rate = Number of Spikes/Time = 9 spikes/50 ms = 180 spikes/s

圖 8　　聽覺神經細胞的發射速率與對應到的輸入單音示意圖

資料來源：[3]

2.1.3 聽覺模型的模擬

此章節將會介紹聽覺模型是如何去模擬出來的，圖 9 為流程圖：

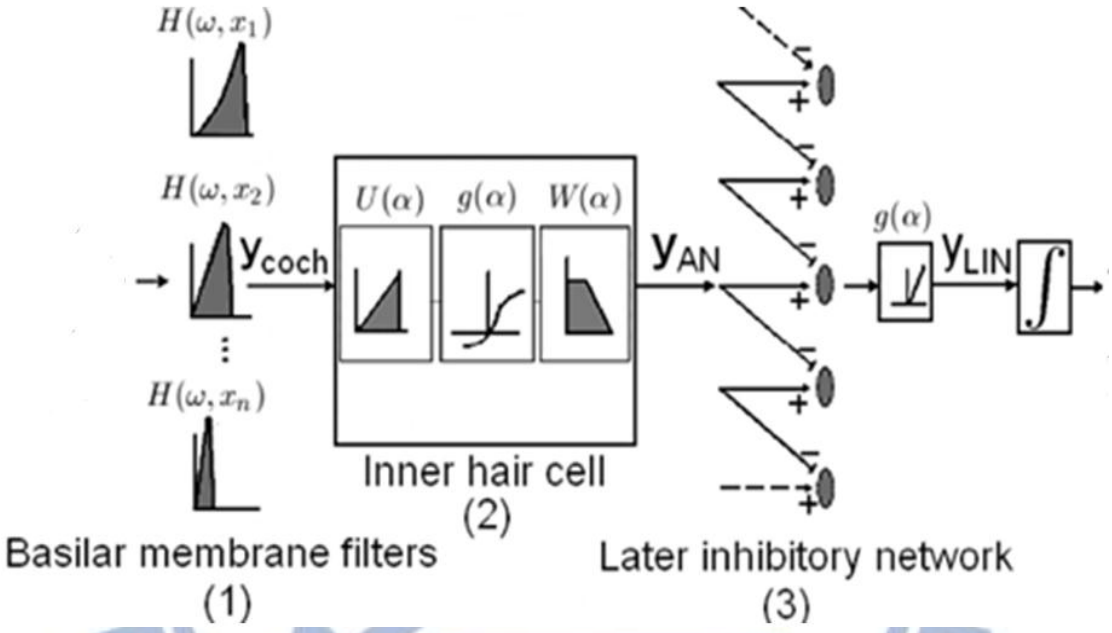


圖 9　　聽覺模型的模擬架構

資料來源：[3]

$y_{coch}(t,f) = s(t) *_t h(t; f)$　　(2-1)

$y_{AN}(t,f) = g(\partial_t y_{coch}(t,f)) *_t w(t)$　　(2-2)

$y_{LIN}(t,f) = \max(\partial_f y_{AN}(t,f), 0)$　　(2-3)

$y_{final}(t,f) = y_{LIN}(t,f) *_t \mu(t; \tau)$　　(2-4)

式子 (2-1) 等同於將輸入訊號通過一組濾波器，模擬訊號在基底膜上被分解的反應，$*_t$ 表示在時域上的摺積。

式子 (2-2) 內對時間的偏微則是在模擬內毛細胞將行進波轉換為與連接神經間高低電位，至於函數 g 為：

$$g(u) = 1/(1+e^{-u}) \qquad (2\text{-}5)$$

主要是在模擬外毛細胞的輸入訊號過大時加以壓抑的自我保護機制，而低通濾波器 w(t) 則是計算聽覺神經的漏電流。

式子 (2-3) 模擬的現象則是鄰近頻率間的遮蔽效應並且對訊號做了半波整流，這裡所模擬的遮蔽效應只考慮了與鄰近較高頻互相遮蔽的現象，但實際上聽覺的遮蔽效應是不只有鄰近一個的高頻會影響，特別一提的是遮蔽效應也會發生在不同的音源之間，近幾年也有相當多的研究是利用不同的音源間之遮蔽效應達到對干擾源之壓抑以作語音增強，如[12]、[13]以及[15]等等。
此處也模擬了基底膜上這一組濾波器其中心頻率越高頻寬會越寬的特性，如下圖 10：

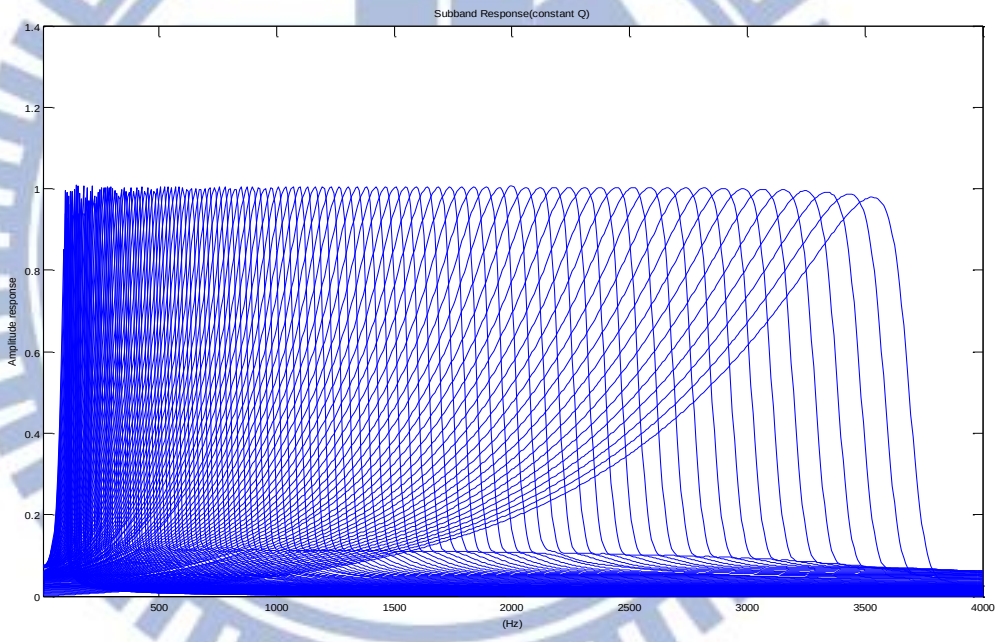


圖 10　基底膜上濾波器組之模擬圖

圖 10 中共有 129 個濾波器，由於鄰近頻率的遮蔽效應，輸出後的訊號兩兩相減，使得輸出只有 128 個訊號，而這 129 個濾波器的頻寬都遵守 constant Q 的原則：

$$f_{center}/\text{bandwidth} = Q \qquad (2\text{-}6)$$

式子 (2-4) 的 μ(t; τ)函數為：

$$\mu(t;\tau) = e^{-t/\tau} * u(t) \qquad (2\text{-}7)$$

τ 為一時間常數，主要是模擬訊號傳至中腦時所面臨的時域上的動態縮減現象。

圖 11 為輸入訊號以及輸出訊號的結果圖：

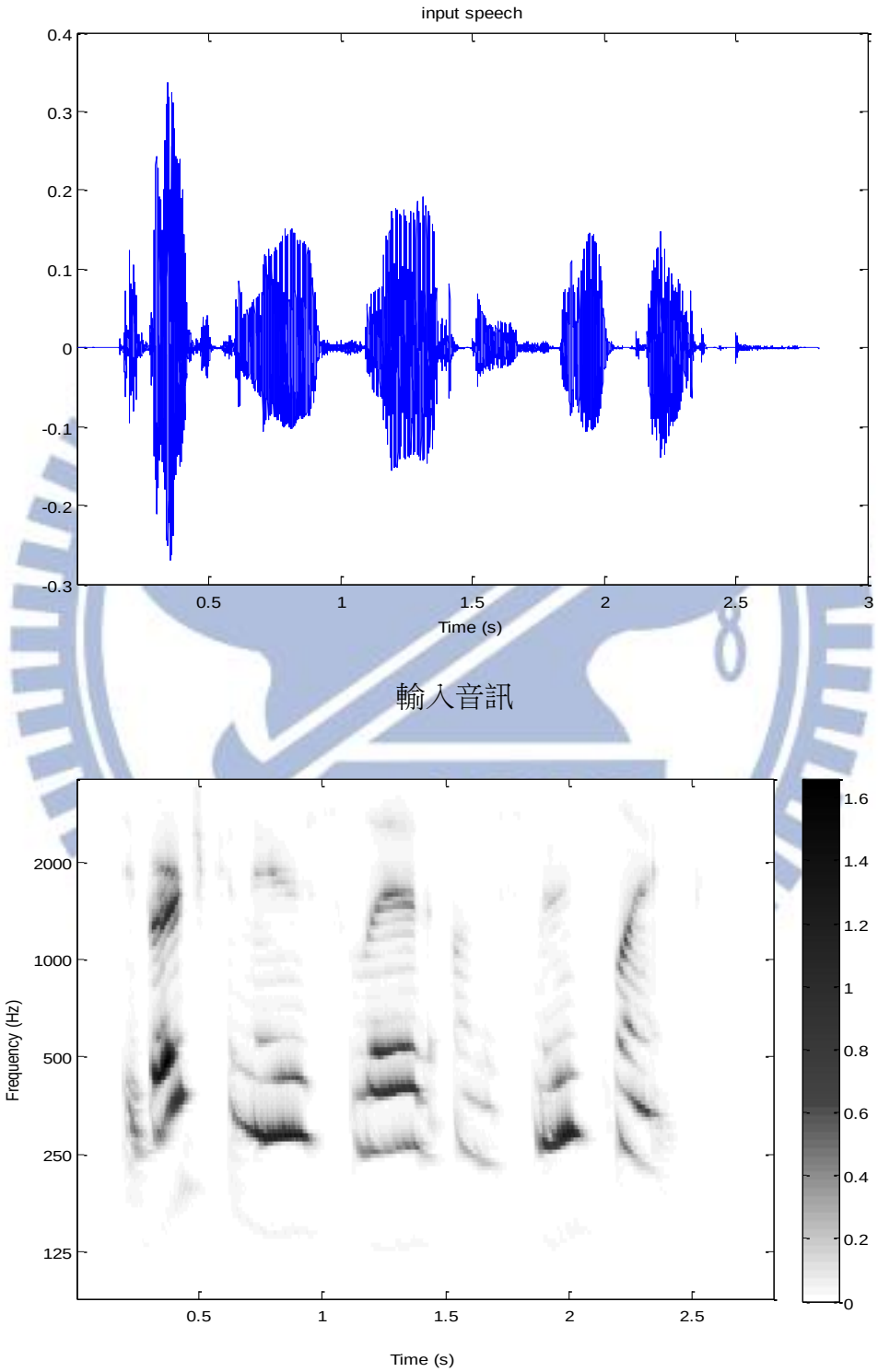


圖 11　輸出的二維頻譜圖，使用語句 ”The birch canoe slid on the smooth planks”

2.2 高維度之多頻帶訊號分析

人耳在中腦階段會將接收到的聲音處理成圖 11 的輸出頻譜圖，接著進入大腦皮質層作更高維度的分析，所使用的是一組同時包含頻域-時域 (spectro-temporal) 的二維調變濾波器來模擬皮質層的功能。聲音裡包含的音色以及基頻這些詳細的資料皆能顯示在此分析結果裡，主要是因為此分析能針對頻譜圖的局部頻寬與非對稱性作有效地解析。

在生理學實驗當中，研究人員會使用單一頻率來測試大腦對於個頻率的反應為何，在此，因為進入大腦皮質層前的訊號是屬於二維的，我們設計出移動波紋刺激源 (moving ripple stimulus)，如下圖 12，來作為二維的基本訊號基底。

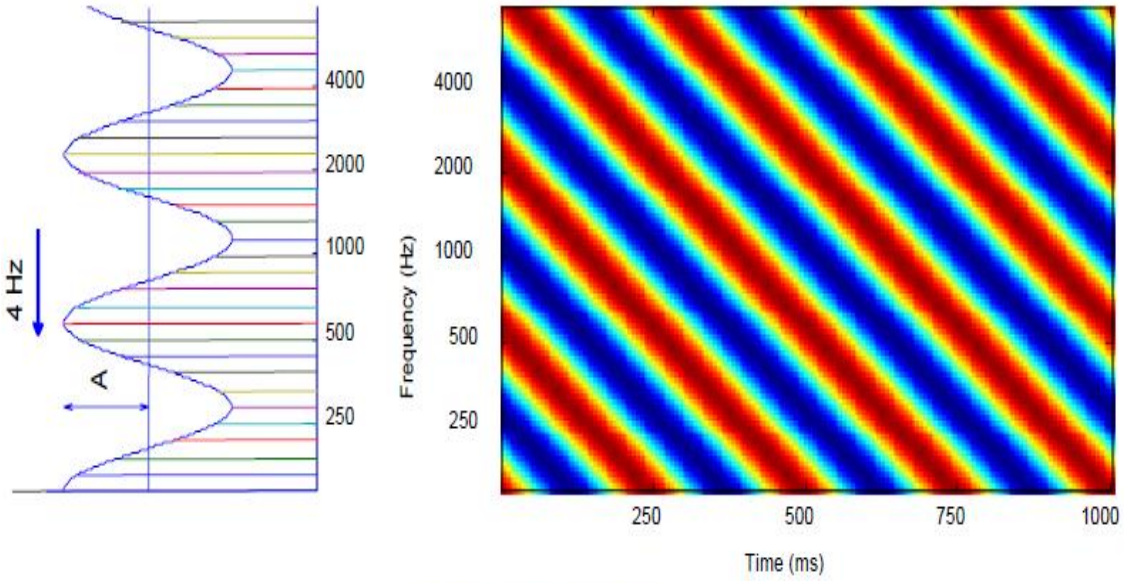


圖 12 移動波紋刺激源 (moving ripple stimulus)
資料來源：[3]

時域軸上的變化率被定義為 rate，單位是 Hz，從圖 12 我們可看出 rate 值為 4 Hz，此二維基底訊號在時域上以 250 ms 為一週期；而在頻域軸上的變化率則是定義為 scale，單位是 cycle/octave，從圖 12 可看出此訊號的 scale 為 0.5 cyc/oct，週期則包含了兩個倍頻。從生理學實驗中可發現大腦皮質層上不同位置對於不同的 rate 與 scale 會有最大響應，所以可將大腦皮質層視為二維的濾波器組，對不同的 rate 與 scale 的訊號會有最大的響應，見式子 (2-8)：

$$r(t, f, \omega, \Omega) = y_{final}(t, f) *_{tf} SRTF(t, f; \omega, \Omega) \quad (2\text{-}8)$$

此外，大腦皮質層對於調頻 (FM) 的上升與下降也會有不同的反應，因此 rate 也定義了正值，表示其對調頻的下降 (downward) 有反應；負值則是對調頻的上升 (upward) 有反應，例如圖 12 裡為下降，即為正的 rate。

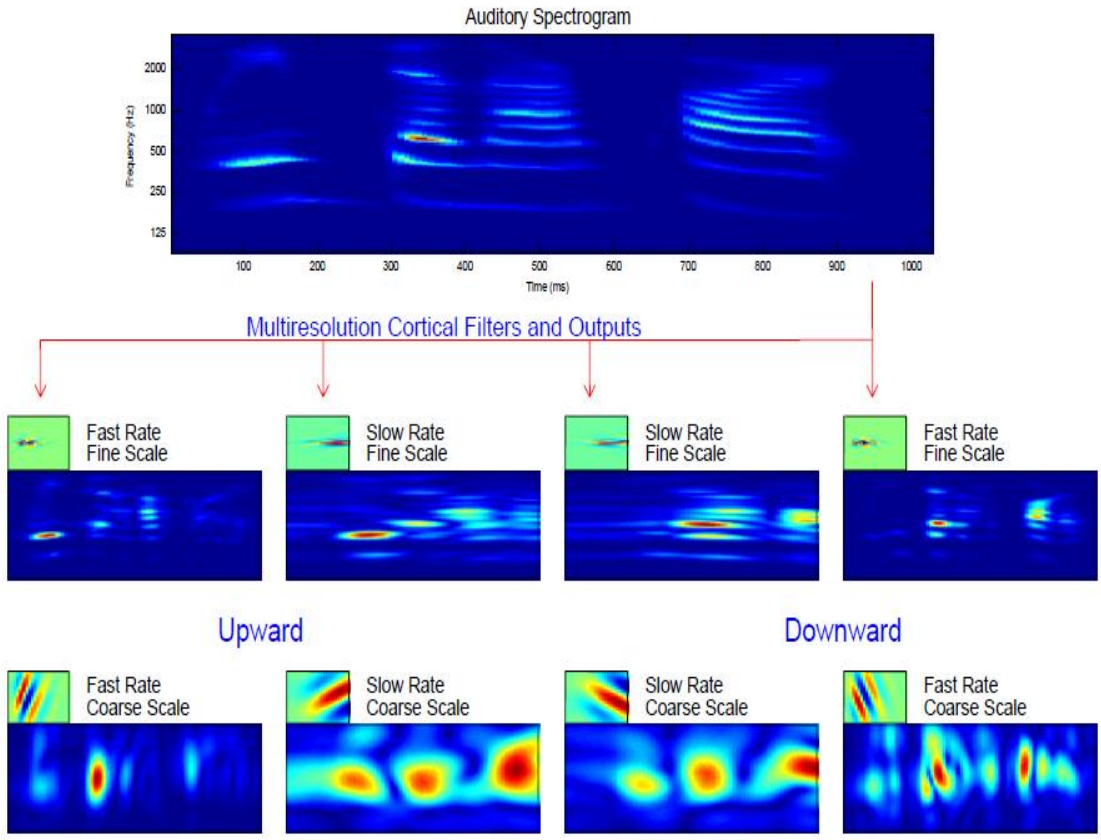


圖 13　語音封包通過大腦皮質層不同 rate-scale 二維調變濾波器之輸出結果
資料來源：[3]

圖 13 最上面的 Auditory Spectrogram 是輸入語音通過基底膜濾波器組的輸出結果，其頻率在對數上呈線性分佈，將這張 Auditory Spectrogram 通過大腦皮質層裡不同 rate-scale 的二維調變濾波器分析後的結果為下面八張小圖，其中 fast rate 為 high rate 濾波器所輸出的結果，此濾波器輸出的訊號在時間上的週期長度較短，slow rate 則為 low rate 濾波器輸出之結果，此濾波器輸出的訊號在時間上的週期長度則較長；而 fine scale 指的是 high scale 濾波器的輸出結果，通過此濾波器的訊號在頻率軸上的一個倍頻裡所包含的訊號週期長度較長，至於 coarse scale 則是 low rate 濾波器所輸出的結果，通過此濾波器的訊號在頻率軸上的一個倍頻裡所包含的訊號週期長度則較短。

2.2.1 語音和雜訊在高維度分析的結果

透過觀察各種訊號在 rate 以及 scale 上的分佈，可發現高斯白雜訊 (white noise)、嘈雜人聲 (babble noise) 跟語音有明顯的區別，如下圖 14~16：

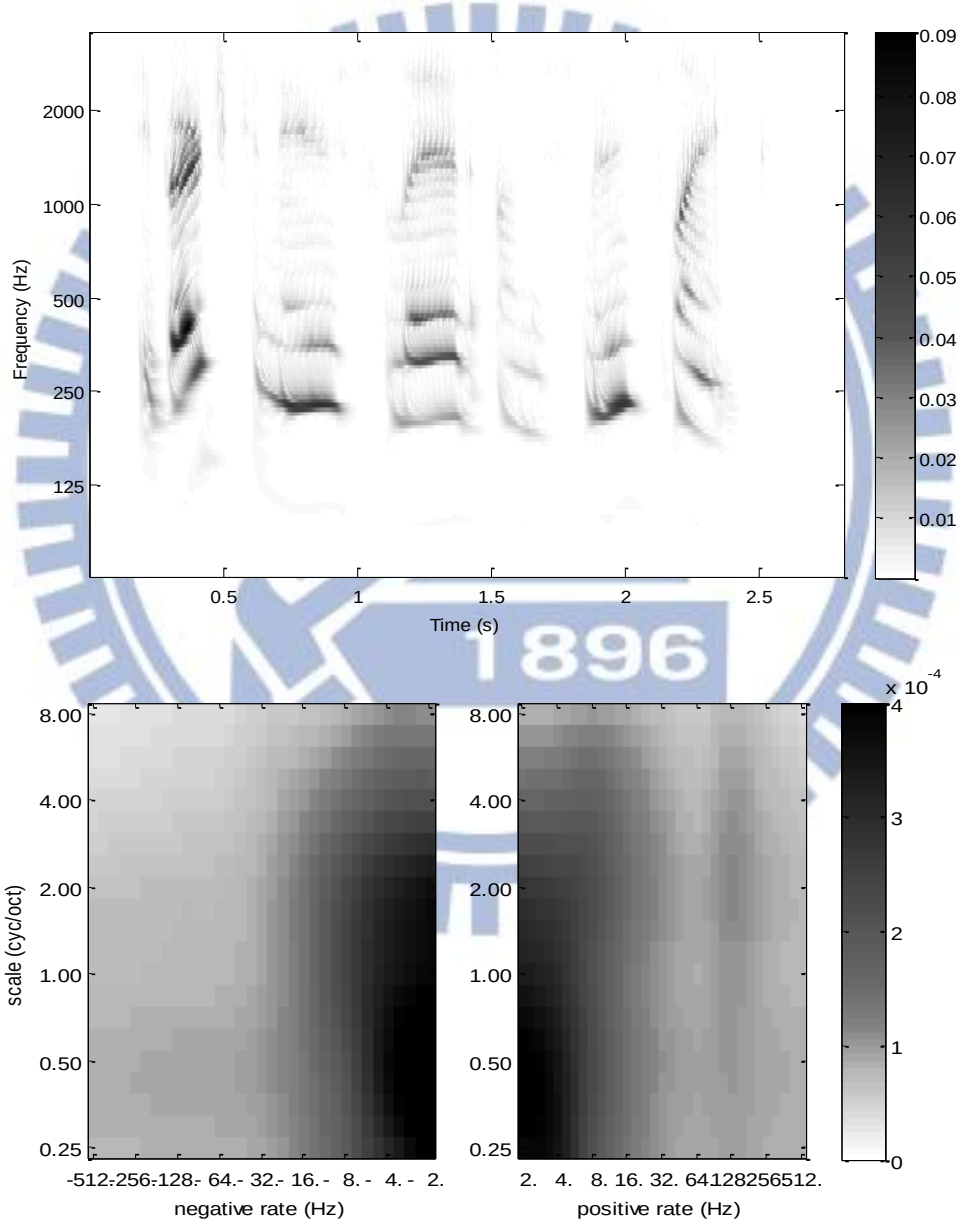


圖 14　乾淨語音 ”The birch canoe slid on the smooth planks” 以及在 rate-scale 上的分佈

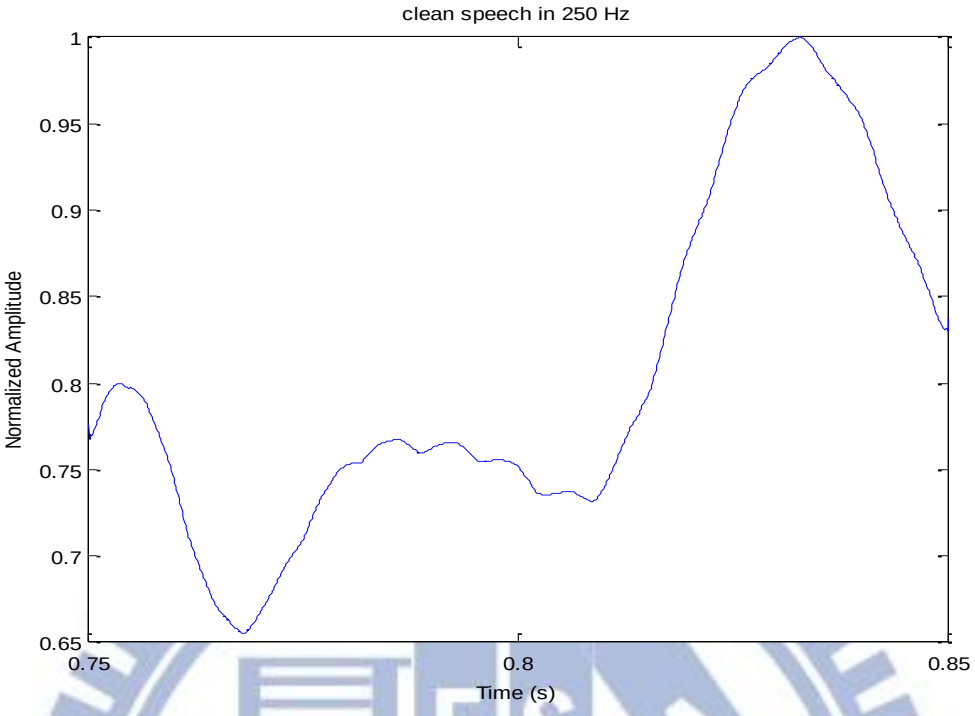


圖 14a　乾淨語音”The birch canoe slid on the smooth planks”在 250 Hz 上的訊號封包剖面圖

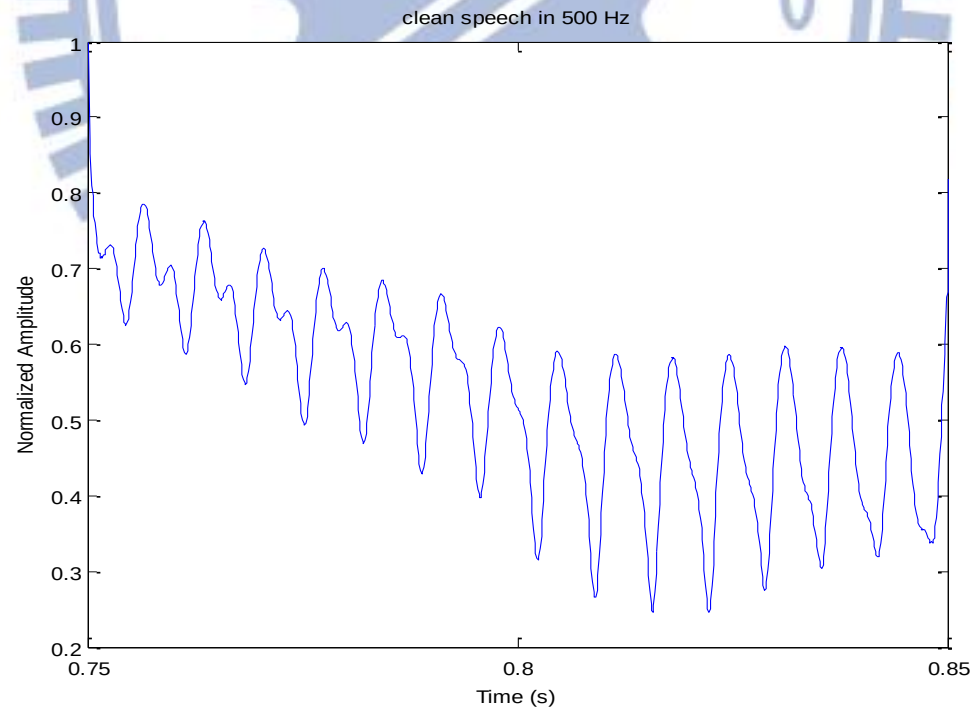


圖 14b　乾淨語音”The birch canoe slid on the smooth planks”在 500 Hz 上的訊號封包剖面圖

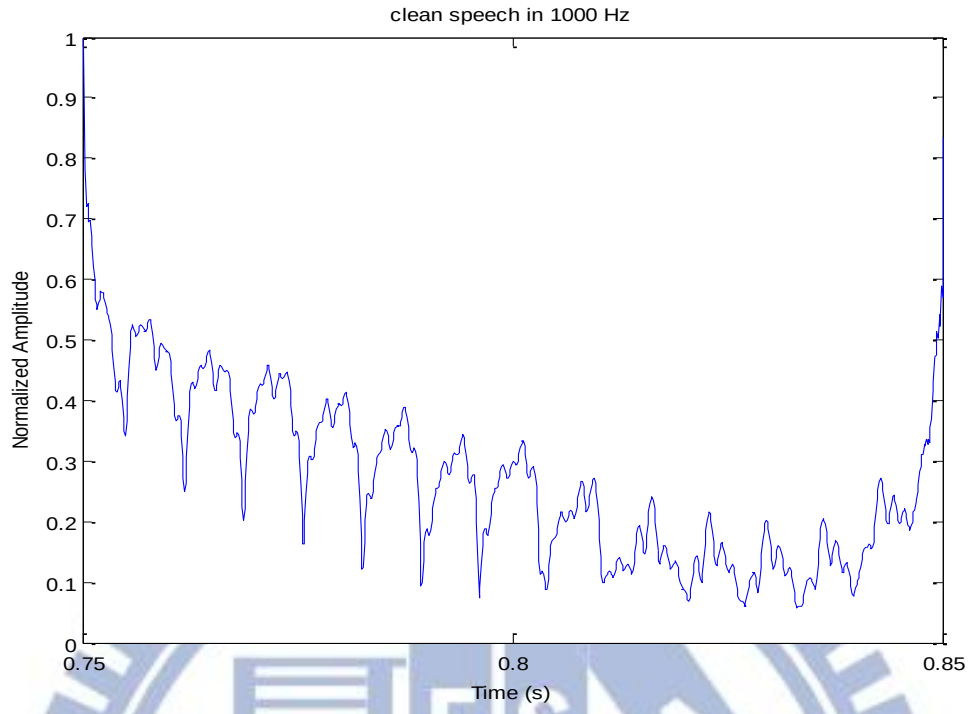
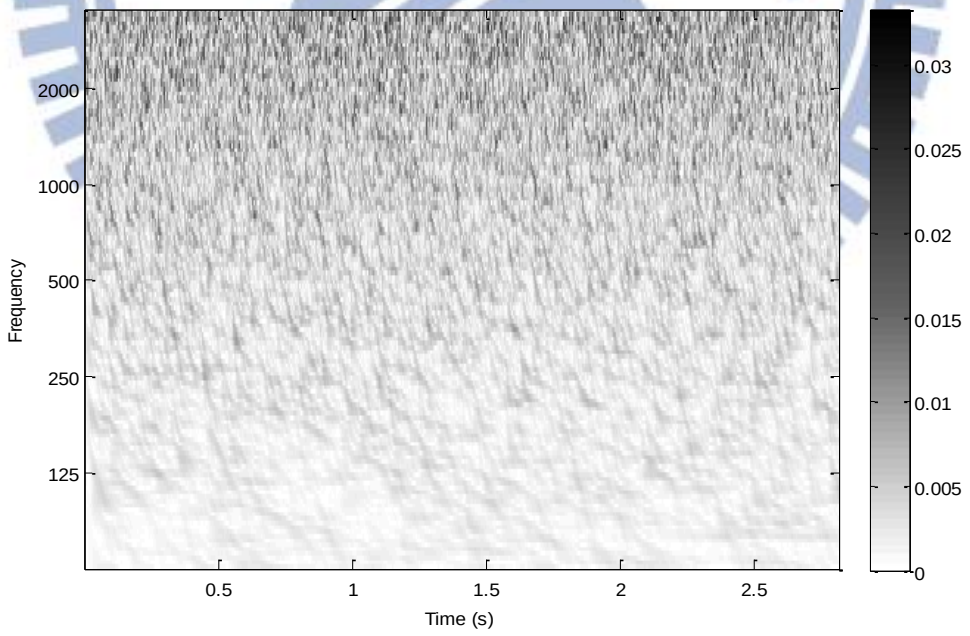


圖 14c　乾淨語音”The birch canoe slid on the smooth planks”在 1000 Hz 上的訊號封包剖面圖



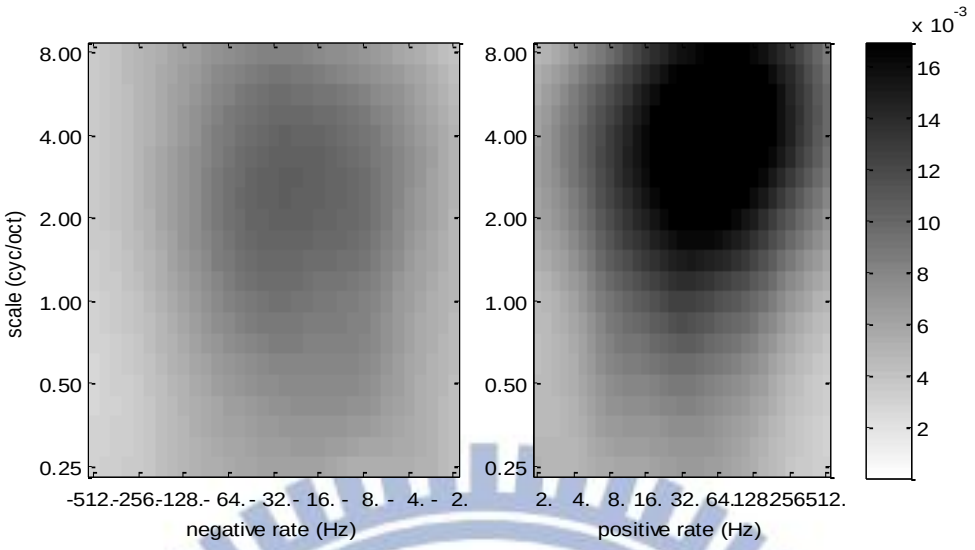


圖 15　高斯白雜訊，訊雜比 (SNR) 為 0 dB 以及在 rate-scale 上的分佈

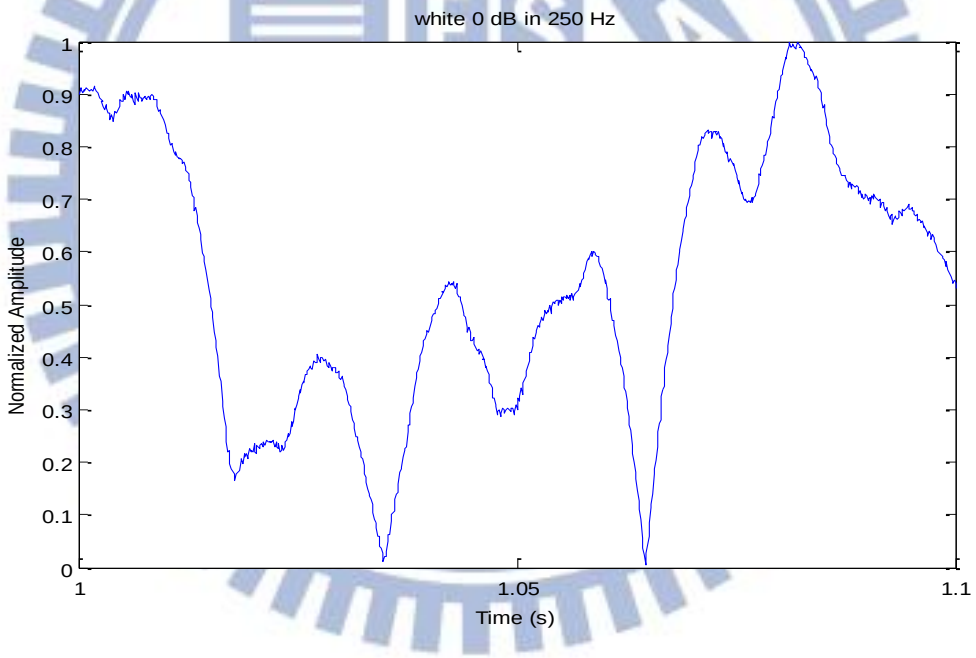


圖 15a　高斯白雜訊，訊雜比 (SNR) 為 0 dB 在 250 Hz 上的訊號封包剖面圖

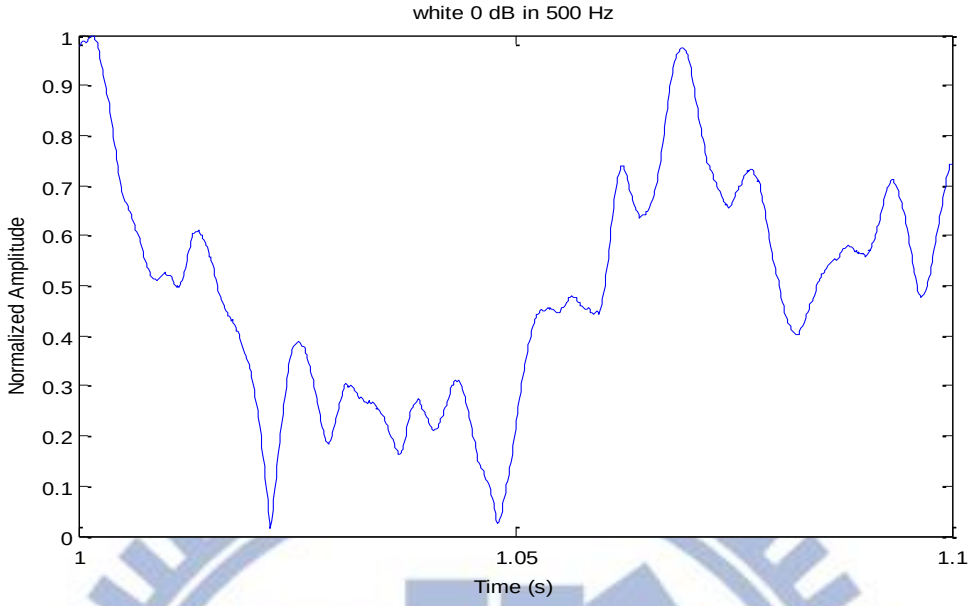


圖 15b　高斯白雜訊，訊雜比 (SNR) 為 0 dB 在 500 Hz 上的訊號封包剖面圖

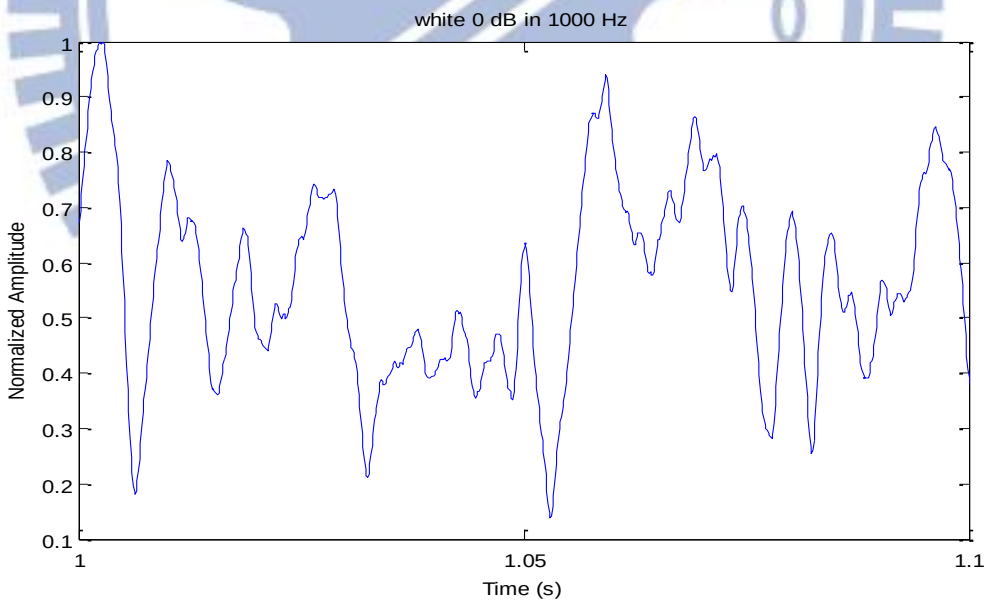


圖 15c　高斯白雜訊，訊雜比 (SNR) 為 0 dB 在 1000 Hz 上的訊號封包剖面圖

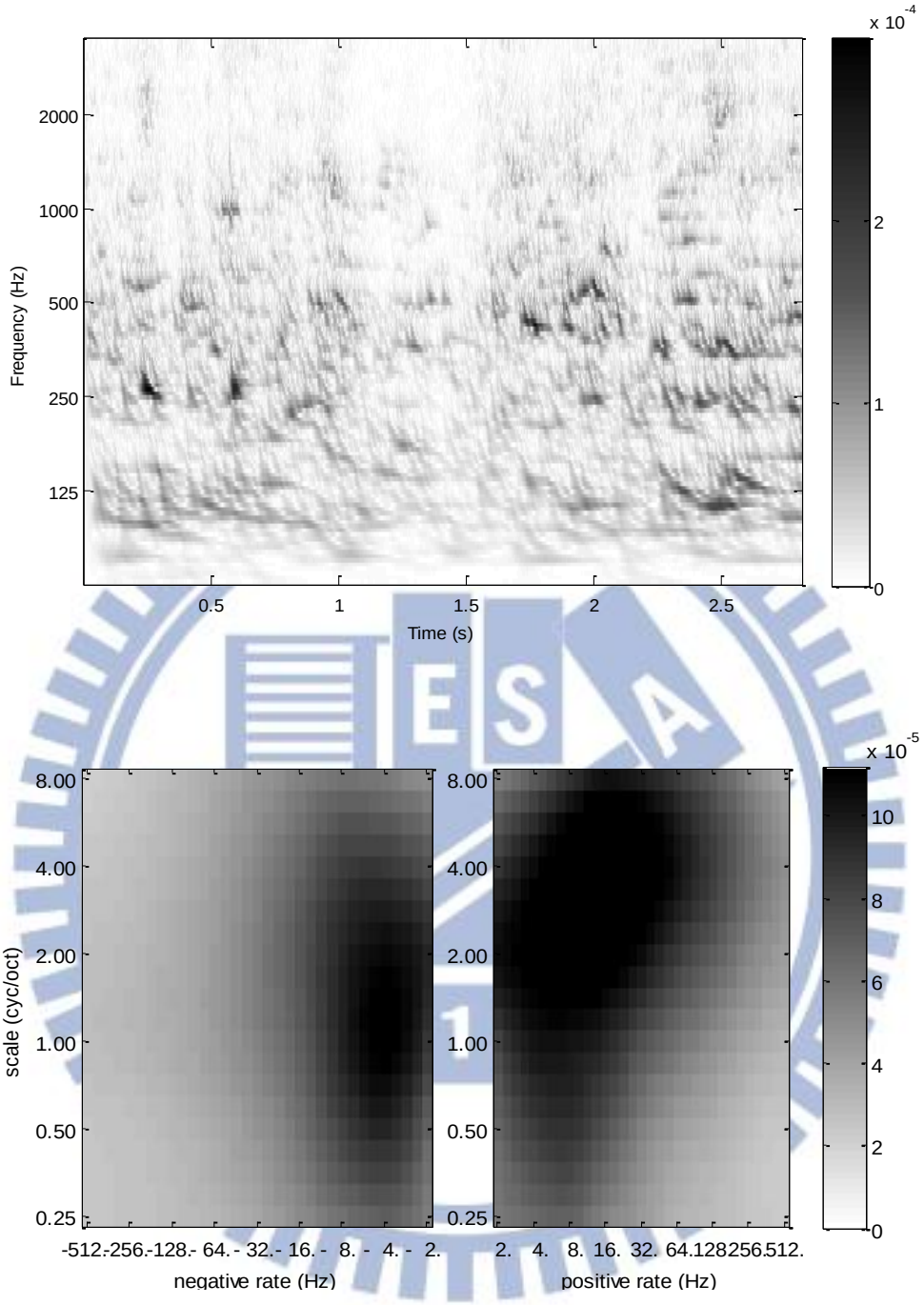


圖 16　嘈雜人聲，訊雜比 (SNR) 為 0 dB 以及在 rate-scale 上的分佈

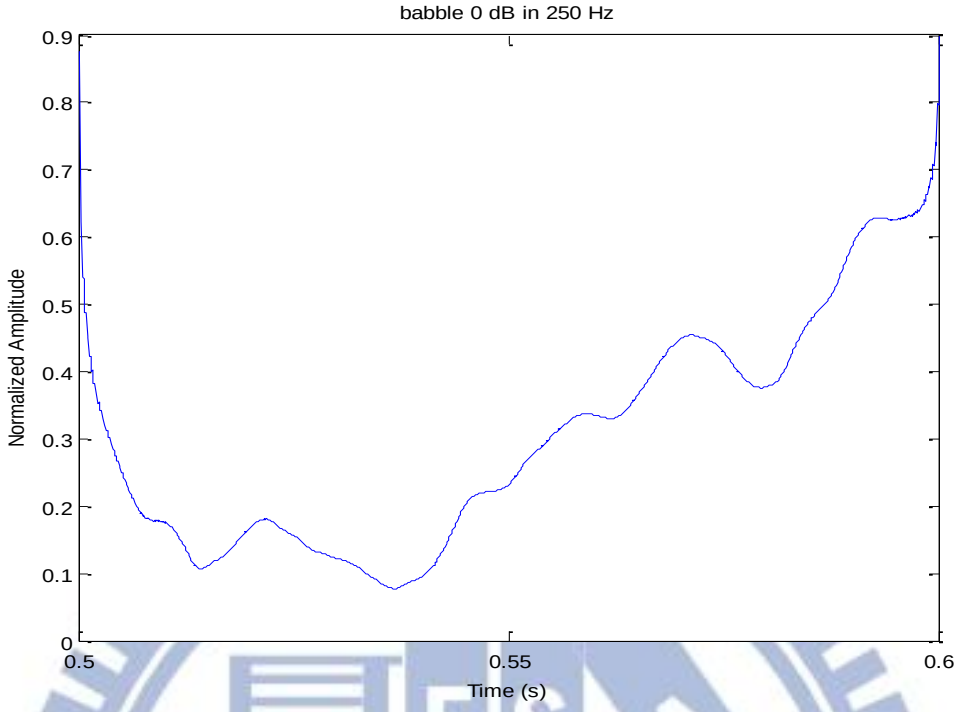


圖 16a　嘈雜人聲，訊雜比 (SNR) 為 0 dB 在 250 Hz 上的訊號封包剖面圖

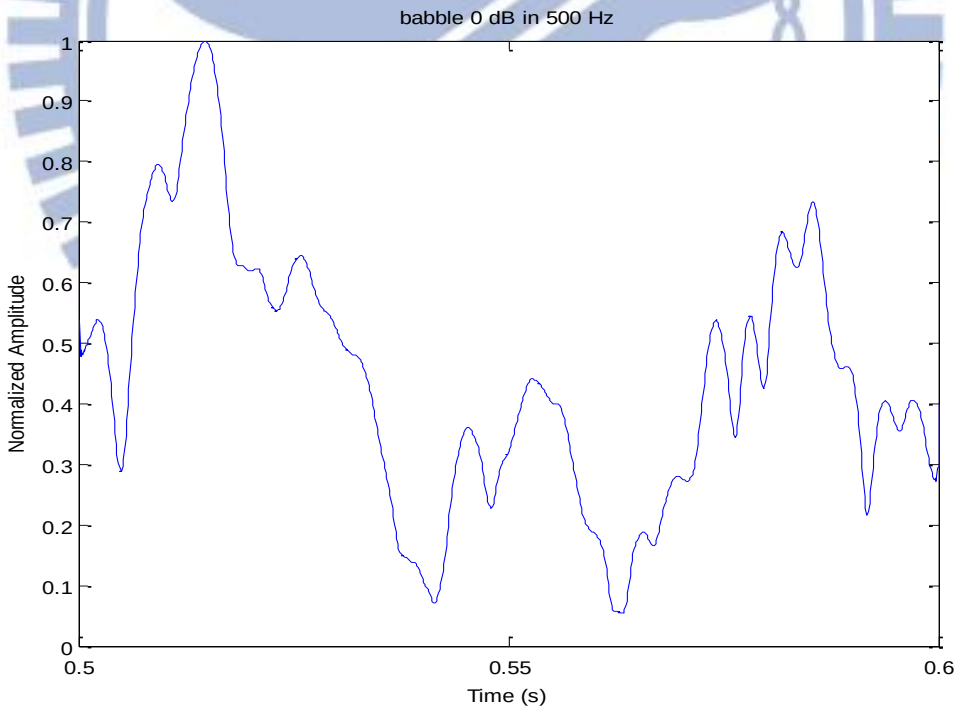


圖 16b　嘈雜人聲，訊雜比 (SNR) 為 0 dB 在 500 Hz 上的訊號封包剖面圖

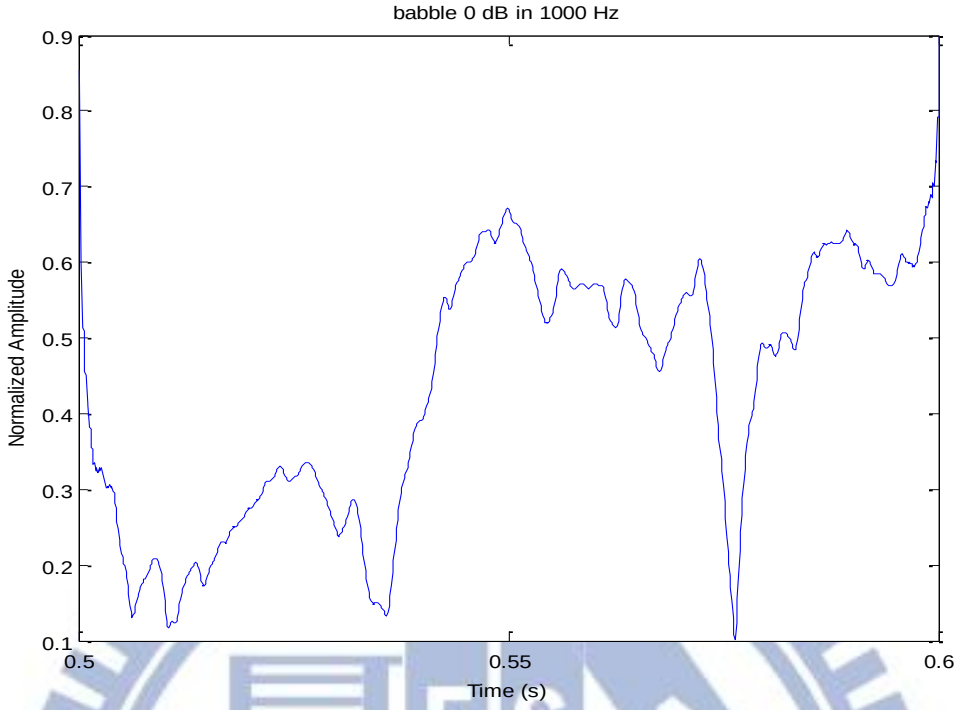


圖 16c 嘈雜人聲，訊雜比 (SNR) 為 0 dB 在 1000 Hz 上的訊號封包剖面圖

由圖 14~16 可觀察到，語音幾乎都分佈在 rate 16 Hz 之下，但在 rate 128 Hz 的地方也有能量分布，這可解釋 14b 與 14c 在封包變化上為何較快速且呈週期性，因為是由語音封包所含的音高 (pitch)資訊產生，而語音的音高在 rate-scale 分布上集中在 100~200 Hz 之間，此外在 rate 256 Hz 以上的部分則幾乎沒有能量分布，然而高斯白雜訊則是在 rate 256 Hz 之上都有能量，而在低 rate 部分幾乎沒有能量分布，因此可以 rate 256 Hz 以上的地方作為語音跟高斯白雜訊的分離依據，而嘈雜人聲雖然在高 rate 部分也有能量但在 rate 256 Hz 以上的能量分布卻相當稀少，且在低 rate 部分亦跟語音有重疊，因此在此環境下所作的語音雜訊分離效果較不顯著，而後兩者的雜訊皆無音高的資訊存在，因此在封包變化曲線上無此音高週期性的變化。

2.3 時域上訊號封包的頻率分析

從上圖 14~16 高維度分析訊號的結果來看，我們可得知高斯白雜訊幾乎分佈在高 rate 的部分，也就是說，在時域上來看，高斯白雜訊的調變能量隨著時間變化較快，如圖 17；至於嘈雜人聲的分佈在時域上來看跟語音的分佈是相當

近似的，可比較圖 18 與圖 19；語音則是分佈在低 rate 的部分，從時域上來看，語音的調變能量隨著時間變化較慢，如圖 19。

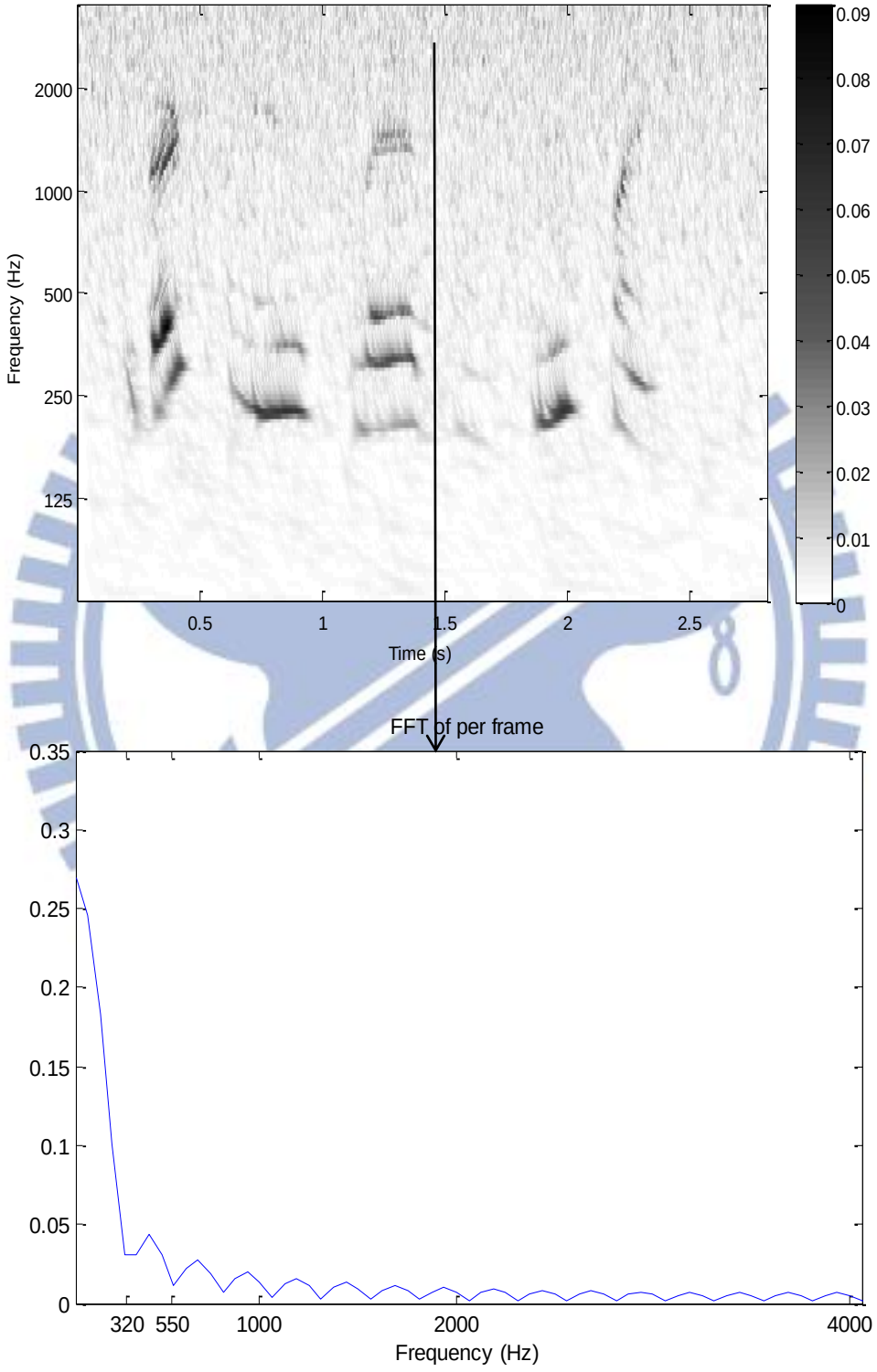


圖 17　高斯白雜訊在特定時-頻單點經傅立葉轉換後的顯示圖，訊雜比 (SNR) 0 dB

從圖 17 可看出，白雜訊的時域封包 320 Hz 之後的頻率成分其值偏高，特別是主葉與第一個副葉之間的波谷明顯偏高，而圖中主葉的最高點則是每個時-頻單點的直流值，此圖亦可看出高斯白雜訊集中分佈於頻譜圖中的高頻部分，屬於高頻雜訊。

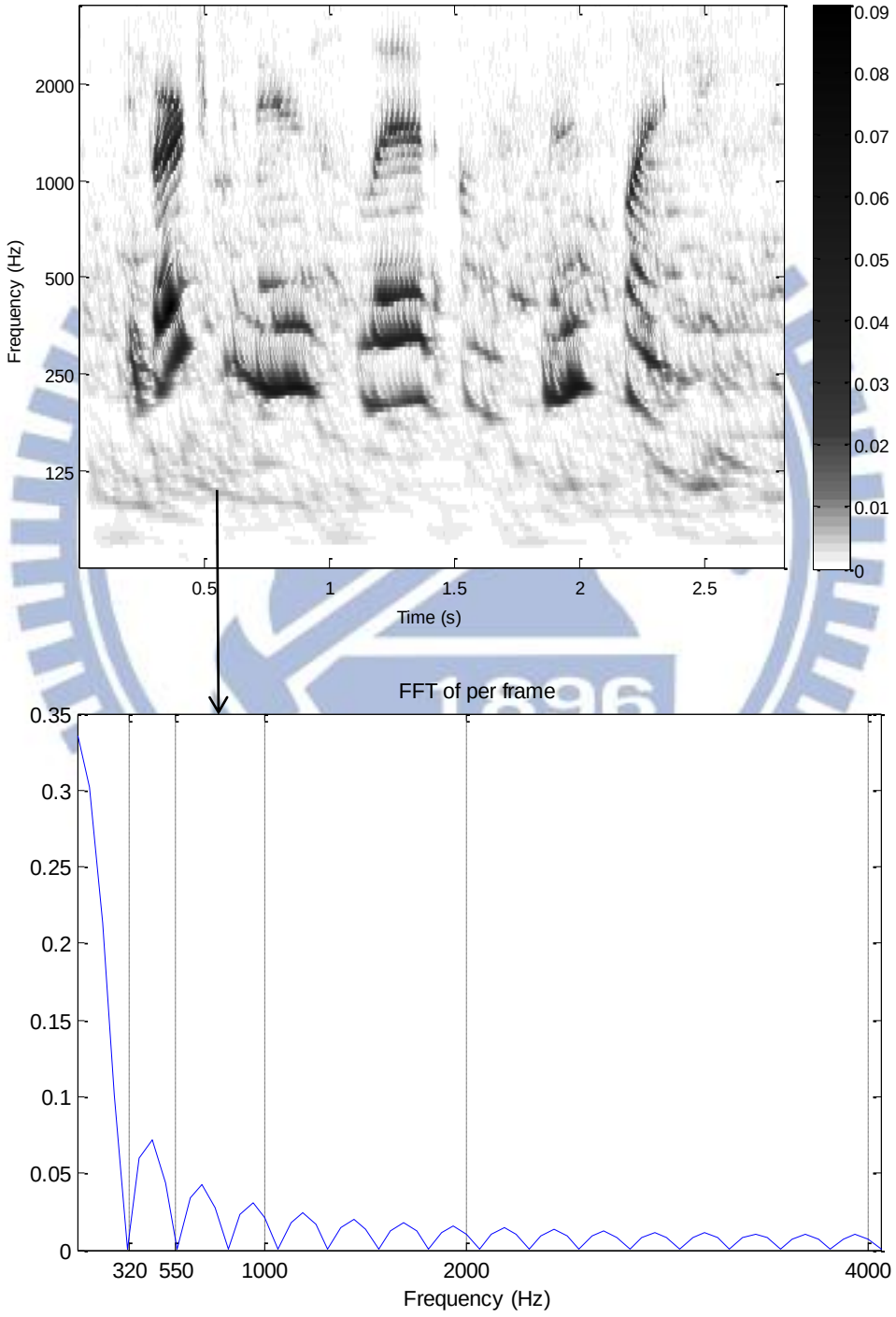


圖 18　嘈雜人聲在特定時-頻單點經傅立葉轉換後的顯示圖，訊雜比 (SNR) 0 dB

將圖 18 與圖 17、圖 19 相比，可看出在嘈雜人聲的時域上頻率分布跟語音的並無太顯著的差異，也因此在後面實驗中可看出嘈雜人聲的環境下其性能的改善比起高斯白雜訊並不顯著，但大部分時-頻單點仍可藉由主葉的最高點也就是每個時-頻單點的直流值來判定出語音與非語音部分。

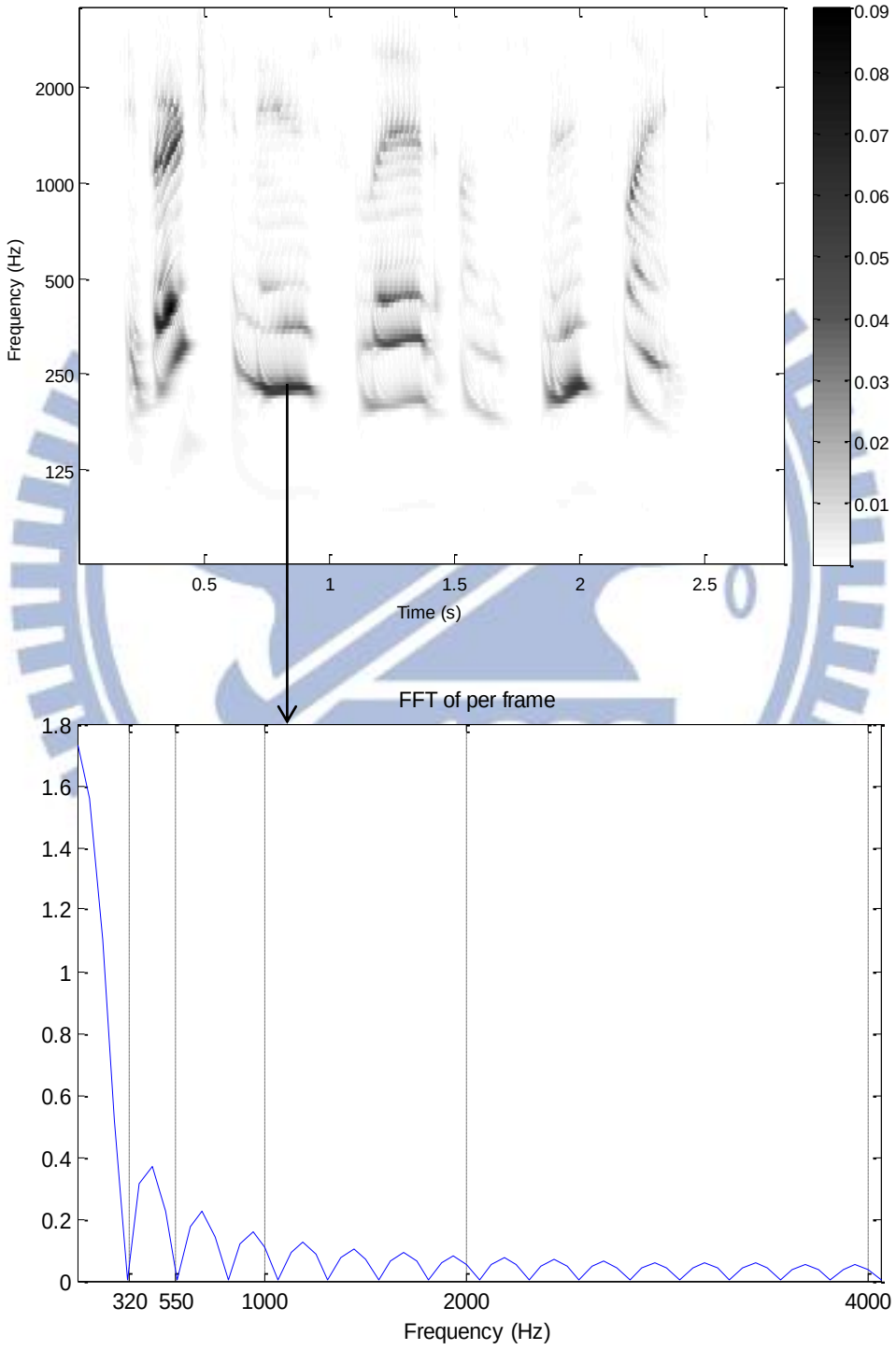


圖 19　乾淨語音 ”The birch canoe slid on the smooth planks” 在特定時-頻點經傅立葉轉換後的顯示圖

圖 19 可看出在 320 Hz 之後的頻率成分其值陡降，主葉 (main lobe) 與第一個副葉 (side lobe) 之間的波谷可與上圖 17 比較，發現語音的部分此波谷趨近於零，高斯白雜訊部分與前者相比則偏高。

三、應用在助聽器上的雜訊消除

圖 20 為本篇論文提出方法的系統架構圖：

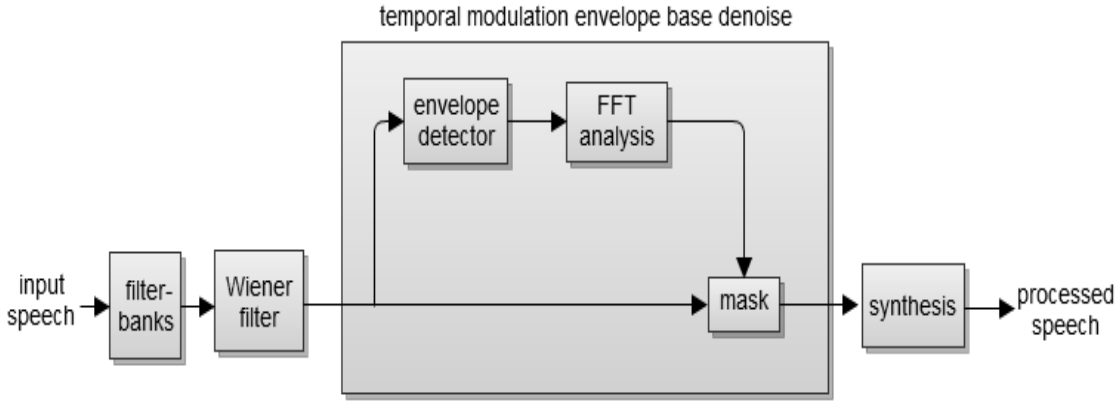


圖 20　系統流程方塊圖

本篇論文所提出的方法是基於分析-改善-合成 (analysis-modification-synthesis) 的架構去實行，如圖 1 所示。有很多語音增強系統也是基於此種架構。

3.1　分析階段 (analysis stage)

3.1.1　助聽器濾波器 (filter-banks)

有別於前一章所提到的基底膜濾波器組 (圖 10) 以及聽覺系統的模擬，助聽器系統以計算量為優先考量，不使用一百二十八個濾波器，而改採用六十四個濾波器，這六十四個濾波器使用的脈衝響應仍然模擬了第二章裡介紹的人耳聽覺中的各種特性，例：constant Q 的濾波器特性、中心頻率的分佈在對數頻率軸上

為等距以及鄰近頻率會產生遮蔽效應等等，其六十四個濾波器示意圖如圖 21：

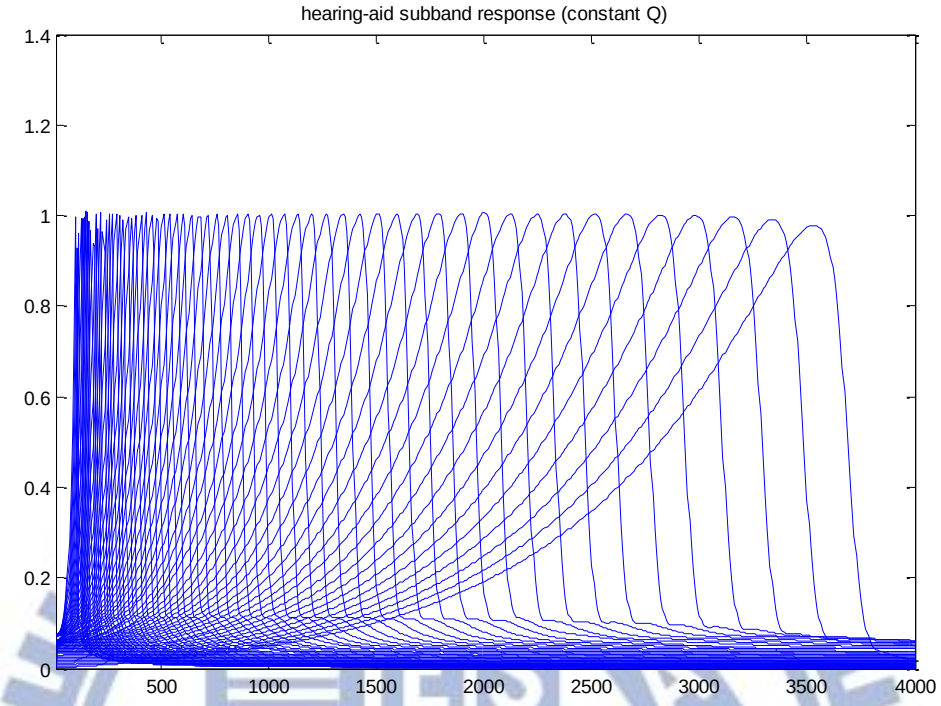


圖 21　應用於助聽器系統之濾波器

輸入的語音訊號經過此濾波器後將會被分頻為六十四個訊號。

3.1.2　維納 (Wiener) 濾波器

Wiener 濾波器是依據事前訊雜比 (priori SNR) 作為每個音框的遮蔽值，其值如下：

$$W(n,k)=SNR_{priori}(n,k)/(1+ SNR_{priori}(n,k)) \qquad (3\text{-}1)$$

n 是 1, 2, …, N 的音框數，k 則是 1, 2, …, 64 的子頻帶數，由 (3-1)式可推知，在語音越強的音框，其事前訊雜比值會越大，W(n)值也會越趨近於 1，對此音框所作的遮蔽也會越少以此來保留語音部分；相反地，如果特定音框之事前訊雜比越小，代表其語音成分越少，W(n)值也會越趨近於 0，對此音框所作的遮蔽也會越多，以達到雜訊消除的效果，至於事前訊雜比則主要是根據 (3-2)式推得：

$$SNR_{priori}(n,k)=(1-\alpha)\,SNR_{priori}(n-1,k)+\alpha\,SNR_{poster}(n,k) \qquad (3\text{-}2)$$

式子(3-2)中的 $SNR_{poster}(n)$為事後訊雜比 (posteriori SNR)，α 值通常為一極小值。

3.2 改善階段 (modification stage)

3.2.1 封包探測 (envelope detector)與快速傅立葉變換 (fast Fourier transform)

由於系統在進入下一個階段需要分析每個子頻帶上封包的頻率成分，以此作為對特定時-頻單點遮蔽的依據，因此利用 Hilbert 濾波器 [11]來探測每段訊號的封包。

在擷取出每個音框的封包後，在時域上對此封包訊號作快速傅立葉轉換 (FFT)，並進一步分析判斷，以此作為每個時-頻單點 (T-F unit) 遮蔽值的依據。

3.2.2 基於時域封包調變的雜訊消除技術 (modulation envelope based denoise)

從前面 2.3 節的圖 17、18 我們可知在時域上分析語音的封包頻率時於高頻處能量相較於雜訊較低，特別是在副葉 (side lobe) 的波谷更明顯，這也可從訊號的 rate-scale 看出 (圖 14~16)，語音主要分布在低 rate 的部分，至於雜訊則主要位於高 rate 的地方，因此我們根據此特點來判別出訊號裡有語音以及沒語音的部分，並且作進一步的遮蔽。

在這裡我們作了兩層的遮蔽，介紹如下：

- 第一層 energy 遮蔽 (the first mask)：依據頻譜圖上的每一個時-頻單點 (T-F unit)的直流 (DC) 值 (也就是在圖 14~16 裡主葉的最大值)設計一個臨界值 η_1，此臨界值是依據每一個子頻帶的前十個音框(通常為無語音的部分)的直流值的平均再乘上一定值 β 所得，只要大於 η_1 的音框就

判定為有語音的部分並乘上一趨近於 1 的值，小於 η_1 的音框則判定為無語音的部分 並乘上一趨近於 0 的值作壓抑，其式子為(3-3)：

if DC value of certain frame < η_1 * β of certain subband,
certain frame * (maximum of 320~550 Hz / sum of all positive frequency)
if DC value of certain frame > η_1 * β of certain subband,
certain frame* (1 - minimum of 0~320 Hz / maximum of 320~550 Hz)

(3-3)

由於仍有部分非語音音框的 DC 值因為受到雜訊汙染較多所以亦會高於 η_1，因此作了第二層的遮蔽將這些高於 η_1 的音框部分濾掉

- 第二層 amplitude modulation (AM) 遮蔽 (the second mask)：主要是依據第一層遮蔽的值，也就是式子(3-3)裡的 certain frame * (1 - minimum of 0~320 Hz / maximum of 320~550 Hz) 來進一步將有語音以及無語音的部分作分離，使用此式子主要是因為雜訊對於在時域上 300 Hz 之後的頻率成分影響最大，若是有語音的音框，0~320 Hz 的最小值與 320~550 Hz 的最大值相差甚遠，因此其比值會隨著語音成分越強而越小；相對地若為沒有語音的音框，若此音框受到雜訊汙染越嚴重，其比值則會越趨近於 1，因此採用 1 減去此兩者的比值最能準確判別出語音與非語音部分，但缺點在於受到雜訊汙染極小的頻帶，即使為非語音部分，其遮蔽值也會趨近於 1，這將導致我們不想要的過多剩餘雜訊，因此才需要根據直流值來判定語音及非語音的第一層遮蔽，而第二層的臨界值 η_2 的設計是依據每個子頻帶的前十個音框的第一層遮蔽值的平均所得到，其式子如下：

  if (1 - minimum of 0~320 Hz / maximum of 320~550 Hz) < η_2 of certain subband,
  certain frame * (maximum of 320~550Hz / sum of all positive frequency)
  if (1 - minimum of 0~320 Hz / maximum of 320~550 Hz) > η_2 of certain subband,
  certain frame * (1 - minimum of 0~320 Hz / maximum of 320~550 Hz)

  (3-4)

圖 22a-c 分別為語音、訊雜比 0dB 的高斯白雜訊、訊雜比 0dB 的嘈雜人聲在時域上封包經過傅立葉轉換後的頻譜圖：

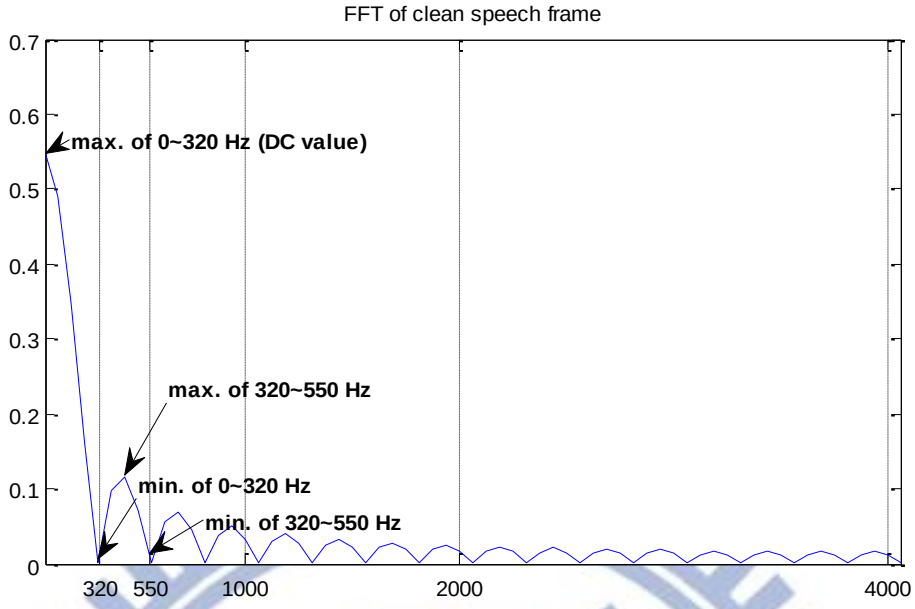


圖 22a　在時域上語音的封包經傅立葉轉換後的頻譜圖

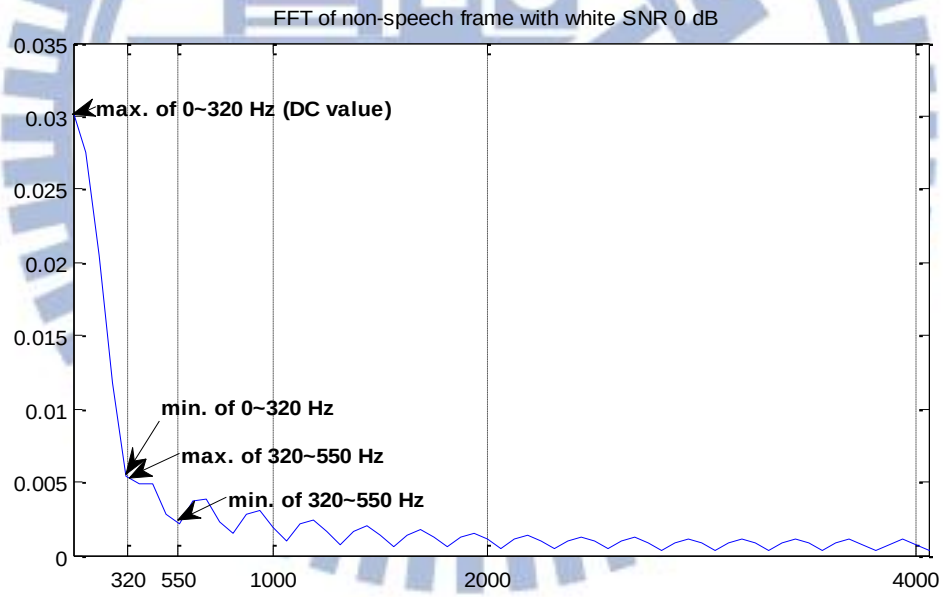


圖 22b　在時域上高斯白雜訊的封包經傅立葉轉換後的頻譜圖

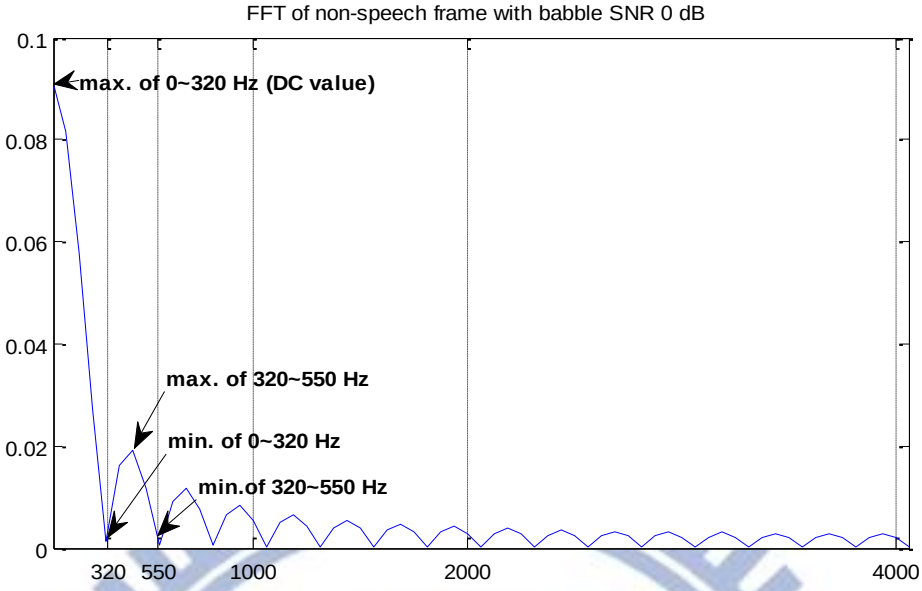


圖 22c　在時域上嘈雜人聲的封包經傅立葉轉換後的頻譜圖

以圖 22a 與圖 22b 來比較可觀察出，在純語音的部分，0~320 Hz 的最小值比上 320~550 Hz 的最大值其比值將會趨近於 0；而在沒有語音且受到高斯白雜訊的汙染時，這兩者的比值將會趨近於 1，以此作為判斷音框是否為語音部分並且作為遮蔽值的依據。

至於圖 22a 與圖 22c 的比較，可發現嘈雜人聲的調變封包在頻譜上的分布跟語音相當類似，因此對於受到嘈雜人聲污染的語音，我們主要是透過第一層遮蔽所考慮到的直流值作為依據，也就是圖中 0~320 Hz 裡的最大值，也因為第二層遮蔽對於嘈雜人聲的效能並不高，所以對於 Wiener 濾波器的改善也相當有限。

經過以上兩層遮蔽後，我們與理想的二分法遮蔽 (ideal binary mask)作比較計算了此種方法在語音的命中率 (1 hit rate) 以及語音和非語音的命中率 (1&0 hit rate)，並比較了其他不同的第二層參數 η_2 的設定值，其背景環境的設定為高斯白雜訊，訊雜比為 10 dB，採用的是 NOIZEUS 的三十段語句，這些語句由三位男性語者與三位女性語者所產生，其值如下表 1：

表 1　　不同參數的三十句語音平均命中率比較

White noise 10 dB	Average 1 hit rate	Average 1&0 hit rate
Parameter 1: 1- (min. of 0~320 Hz/max. of 320~550 Hz)	0.64	0.81
Parameter 2: 1- (max. of 320~550Hz/max. of 0~320 Hz)	0.53	0.8

Parameter 3: 1-(min. of 320~550Hz/max. of 0~320 Hz)	0.61	0.8

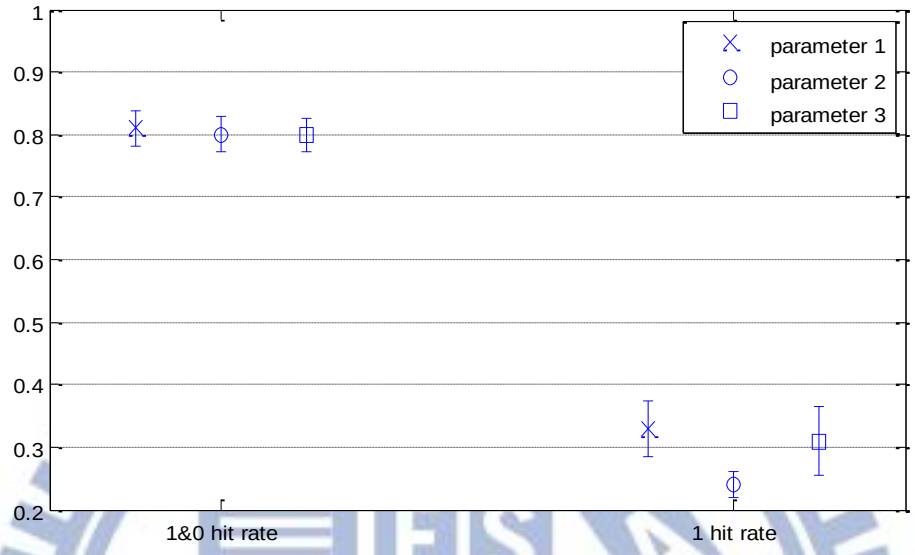


圖 23　各參數之三十句語音平均命中率與標準差

由表 1 可看出這三種參數設定在語音以及非語音的命中率是差不多的，但單就語音的命中率來看，參數一的性能比另外兩種要好，也因為參數二與參數三的語音命中率較低，所以在語音失真上也更為嚴重，因此本篇論文所提出的方法採用了參數一作為第二層的遮蔽值。

3.3　合成階段 (synthesis stage)

在經過處理後的訊號，我們接著利用 overlap and add (OLA)將這些長度四毫秒的音框合成回原本在各個子頻帶內的訊號長度。接著計算並取得助聽器濾波器裡六十四組不同脈衝響應的反函數，這裡反函數的計算僅是簡單將分析階段所使用的六十四組脈衝響應的極點 (poles) 與零點 (zeros) 作對調而取得。接著將六十四個子頻帶裡的訊號各自通過相對應的脈衝響應後再相加取得完整的語音。

四、實驗設計與結果分析

4.1 實驗背景

4.1.1 使用工具

實驗過程裡所使用的語料庫為 NOIZEUS [18]，總共有三十個語句，每一句時間長度都在 2.5~3 秒之間，這些語句由三位男性語者與三位女性語者所產生，本篇論文裡所表現的二維或更高維度的語音資料分析圖皆來自 NOIZEUS 的第一句”The birch canoe slid on the smooth planks”，此句來源為男性語者，而我們加入的背景雜訊則是從 NOISEX-92 取得，訊雜比在後面實驗會採用 0 dB、5 dB、10 dB 以及 15 dB，背景雜訊採用高斯白雜訊 (white noise)以及嘈雜人聲 (babble noise)。

本篇論文將比較傳統 Wiener filter [8]、Joint ST Wiener filter [7]、所提出的時域封包調變消噪技術以及所提方法與 Wiener 作結合的效能。以下將概略介紹 spectro-temporal subband Wiener filter：

- Joint spectro-temporal (ST) Wiener filter：
  將輸入語音經過 short time fourier transform (STFT) 得到二維的頻譜圖後，再進一步作四維的分析如圖 14a-c，並假設在每段訊號的第一秒為沒有語音的部分，於此部分估計雜訊的 rate-scale 分佈，再計算後面子頻帶內每個音框於 rate-scale 上的事前訊雜比，以此作為子頻帶內每個音框的遮蔽值，其式子如下：

$$W(f; t_n, \omega_i, \Omega_j) = P_S(f; t_n, \omega_i, \Omega_j) / (P_S(f; t_n, \omega_i, \Omega_j) + \alpha P_N(f; \omega_i, \Omega_j)) \quad (4\text{-}1)$$

將此遮蔽值乘上對應到的時-頻單點上的 rate-scale 圖，再還原回二維的頻譜圖後，透過 overlap and add (OLA)還原回語音訊號，此做法將 Wiener 同時作在 spectro-temporal 上，而本篇所提及的方法是 Wiener 先在 spectrum 作過加強後，再利用 temporal 上的 energy 與 AM 兩層遮蔽作進一步的雜訊消除，等於是先在 spectro 作完再於 temporal 作，此兩種方法的比較也將列於後面，而只對 spectro 作的 Wiener 以及本篇提出的只對 temporal 作未與 Wiener 結合的系統也都會在後面一併比較。

本篇論文所使用的評分方式為客觀評分 (objective evaluation)，分別採用 perceptual evaluation speech quality (PESQ) [9]以及 Itakura-Saito distance (IS dist.) [10]來評估四種不同方法之間的效能，PESQ 的原理就是將原本未被汙染的語音跟處理過後的語音來比較它們頻譜圖的差異，再以估計 mean opinion score (MOS) 的評分法來呈現此差異。PESQ 跟 MOS 的相關值可大於 0.9，因此在一定程度上可代表主觀的受試者測驗結果。 (其中 MOS 是主觀評分 (subjective evaluation)，主要是請受試者對於所聽到的語音品質作 1~5 的評分，再把每個受試者的分數平均起來，分數越高代表語音品質越好。) 而 IS dist. 的計算式為：

$$ISD_{x\hat{x}} = \frac{E_{\hat{x}}}{E_x} - \ln\frac{E_{\hat{x}}}{E_x} - 1 \tag{4-2}$$

式子 (4-2)裡的 E_x 指的是輸入語音訊號的預測誤差能量 (prediction error power)，式子的計算意義主要是在計算兩個語音訊號之間的線性估測係數 (linear prediction coefficients) 的差距，以此來算出處理後的語音相較原始的乾淨語音的失真程度。

4.1.2 參數設定

下表為實驗中所採用的各種參數：

表 2 實驗中各參數設定

取樣頻率 (sampling rate)	8000 Hz
音框長度 (frame size)	4 ms

音框位移量 (frame shift)	2 ms
快速傅立葉轉換點數 (FFT points)	128
濾波器個數 (filter banks)	64

下表 3 為比較在不同音框長度情況下，輸入語音長度為 2.5 秒，PESQ 分數為 NOIZEUS 三十個語句的平均分數，環境為高斯白雜訊，訊雜比則為 10 dB，時域封包消噪與 Wiener 濾波器結合之系統性能與計算速度：

表 3　不同音框長度下的系統性能與計算時間評比

frame size	PESQ	cost time
2 ms	2.55	14.82 s
4 ms	2.65	24.97 s
8 ms	2.64	46.41 s

從表 3 裡可看出音框長度選取 4 毫秒會使得系統的性能達到最佳，在 8 毫秒時其性能與 4 毫秒差不多，但消耗時間明顯較少，雖然 2 毫秒的音框長度計算時間較短，但分數也略差於 4 毫秒的音框長度，因此在後面的實驗裡，皆已 4 毫秒的音框長度去作處理。

表 4　不同傅立葉轉換點數下的系統性能與計算時間評比

FFT points	PESQ	cost time
64	2.45	24.04 s
128	2.65	24.97 s
256	2.69	27.14 s

從表 4 可看出，若傅立葉點數越多，其性能越好，但所需要的計算時間也會越多，因此在性能與計算時間折衷下我們選擇了 128 點作為接下來實驗所用到的參數。

表 5　不同濾波器個數下的系統性能與計算時間評比

filterbanks	PESQ	cost time
32	2.04	11.86 s
64	2.65	24.97 s
128	2.81	55.58 s

從表 5 可看出，若濾波器個數越多，PESQ 分數越高，但計算時間也會大幅增加，主要是因為將各個濾波器輸出之訊號還原回語音時會產生無法避免的失真，當濾波器個數越多，失真就會越少，但相對地計算量也會增加，因此在以下實驗裡我們會選用 64 個濾波器組對輸入語音作分頻。

4.2　實驗結果與分析

表 6　高斯白雜訊汙染之語音在不同 β 值所對應的 PESQ 分數

β	0 dB	5 dB	10 dB	15 dB
2.1	1.96	2.33	2.63	2.84
2.5	1.98	2.34	2.66	2.86
2.9	1.99	2.36	2.65	2.86
3.3	1.99	2.37	2.68	2.87
3.7	1.97	2.37	2.67	2.86

表 7　嘈雜人聲汙染之語音在不同 β 值所對應的 PESQ 分數

β	0 dB	5 dB	10 dB	15 dB
2.1	1.91	2.3	2.58	2.82
2.5	1.94	2.29	2.59	2.83

2.9	1.95	2.29	2.6	2.84
3.3	1.94	2.3	2.62	2.87
3.7	1.93	2.3	2.63	2.87

表 6 與表 7 列出本篇論文所提出的時域封包消噪法在不同背景雜訊下不同 β 值所對應的語音品質。

由以上兩表可看出，在少部分的參數調動上，其性能變動幅度不大，β 值越大代表遮蔽越多，這在訊雜比較差的情況下會造成更多的語音失真，但 β 值若越小，所能壓抑的雜訊也會越少，因此在參數選取上須找到最佳折衷，若未來應用於硬體上，則可視情況需求調整參數大小，以下實驗所使用的 β 值為 2.9。

下圖 24 是以感知聽覺為基礎所畫的受到高斯白雜訊污染的結果頻譜圖：

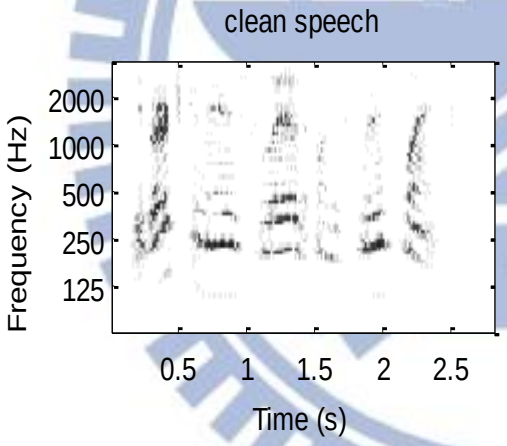


圖 24a

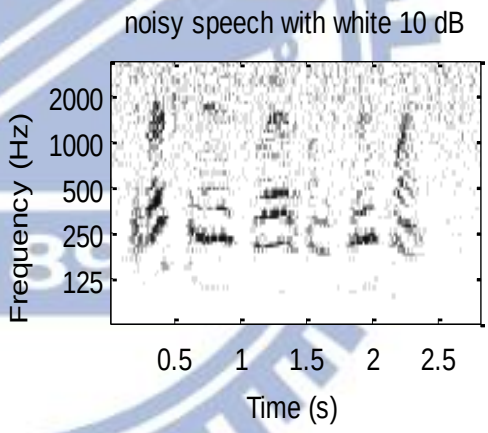
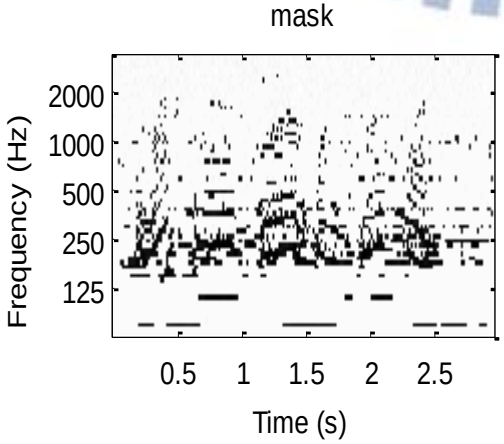


圖 24b



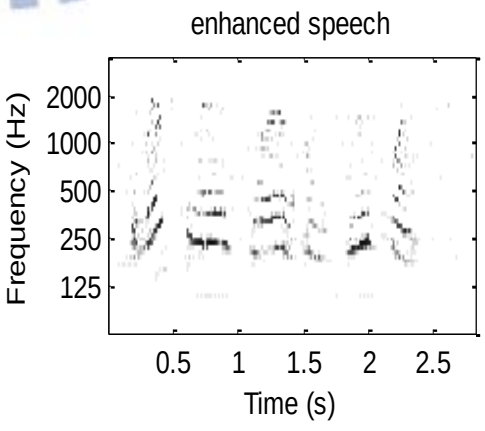


圖 24c　　圖 24d

圖 24a 為乾淨語音”The birch canoe slid on the smooth planks”，圖 24b 為高斯白雜訊汙染訊雜比 10 dB 的語音頻譜圖，圖 24c 為本篇論文所提方法所計算出的二維遮蔽圖，圖 24d 則為圖 24b 乘上圖 24c 所得之結果

下圖 25 則是以感知聽覺為基礎所畫的受到嘈雜人聲汙染的結果頻譜圖：

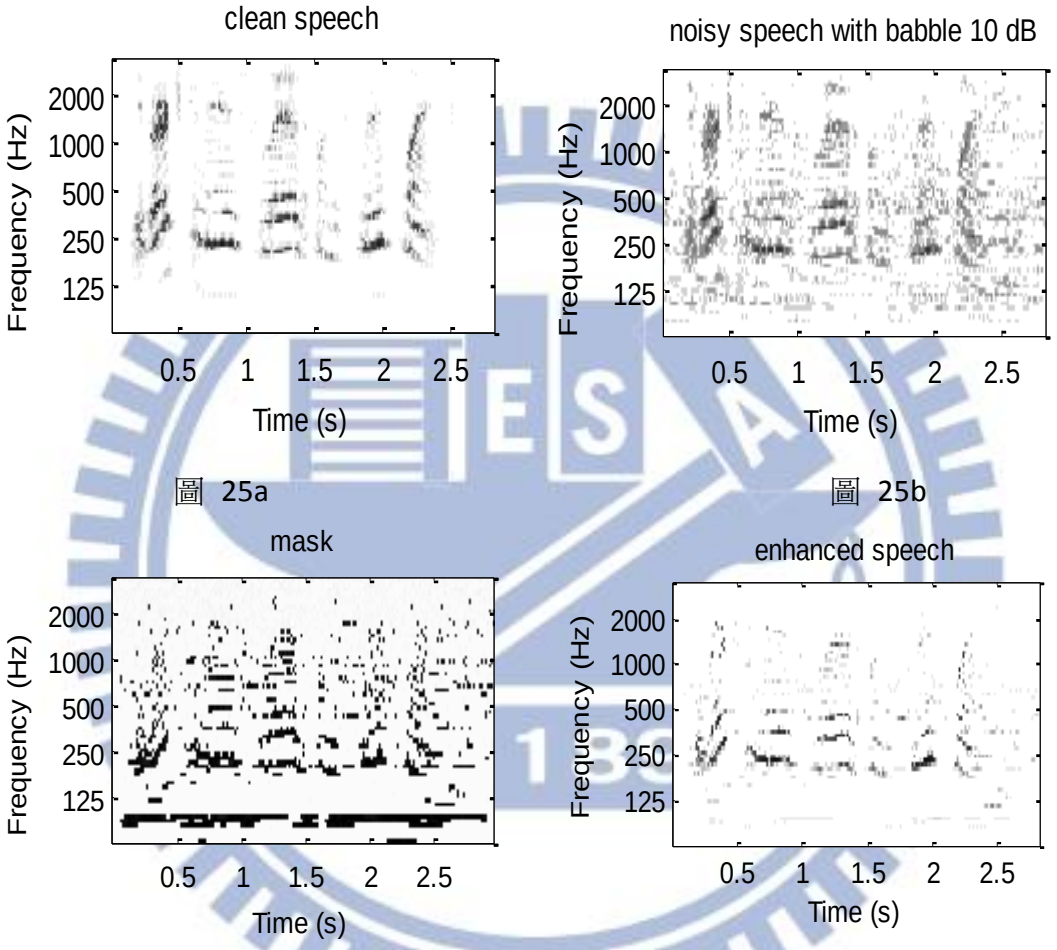


圖 25a　　圖 25b

圖 25c　　圖 25d

圖 25a 為乾淨語音”The birch canoe slid on the smooth planks”，圖 25b 為嘈雜人聲汙染訊雜比 10 dB 的語音頻譜圖，圖 25c 為本篇論文所提方法所計算出的二維遮蔽圖，圖 25d 則為圖 25b 乘上圖 25c 所得之結果

從圖 24b 與圖 25b 我們可知道高斯白雜訊主要分佈在高頻處，而嘈雜人聲則主要在中低頻，我們亦可看出在高斯白雜訊汙染的情況下所得到的二維遮蔽圖比起受到嘈雜人聲污染的二維遮蔽圖較為乾淨，這是由於嘈雜人聲在低 rate 也有分佈，因此其封包頻譜的形狀跟語音相似，導致本篇論文所提及的方法效果有限，我們可從語音和非語音以及語音兩種命中率來看：

表 8　　高斯白雜訊與嘈雜人聲背景下的命中率比較

background noise	white noise with 10 dB	babble noise with 10 dB
1&0 average hit rate	0.81	0.78
1 average hit rate	0.64	0.58

接下來我們利用 PESQ 來評估四種要比較的系統性能，其結果如下：

表 9　　在高斯白雜訊的背景下各系統的 PESQ 平均分數

SNR	0 dB	5 dB	10 dB	15 dB
noisy speech	1.58	1.85	2.16	2.5
Wiener filter	1.91	2.27	2.56	2.76
Joint spectro-temporal Wiener filter	2.26	2.57	2.84	3.08
proposed method	1.98	2.28	2.52	2.69
Proposed method combined with Wiener filter	1.99	2.36	2.65	2.86

由表 9 我們可看出 joint spectro-temporal Wiener filter 效能最好，本篇論文所提出的方法結合 Wiener 濾波器其次，而 Wiener 濾波器與本篇論文提出的時域封包調變消噪方法的性能相比之下，在訊雜比較優的情況下，Wiener 濾波器的性能略優，然後當訊雜比較差時，本篇論文的方法則比較好。

詳細的平均 PESQ 分數與不同語句間的 PESQ 標準差繪於下圖 26：

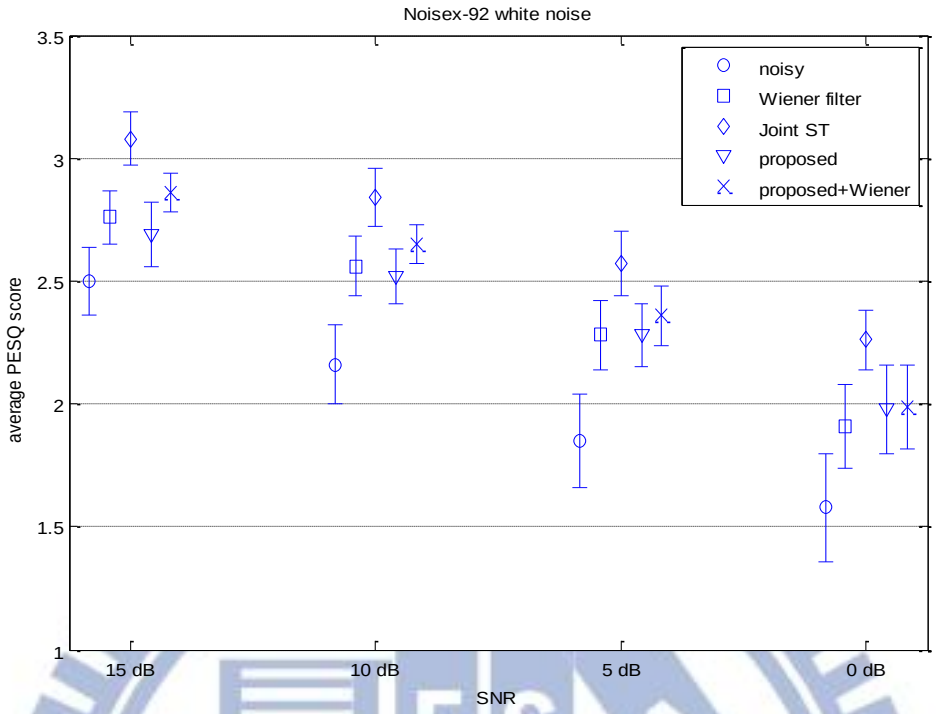


圖 26　高斯白雜訊在不同 SNR 之輸入語音於不同系統處理後的 PESQ 平均與標準差

接著考慮在嘈雜人聲為背景雜訊的情況下，各系統的性能：

表 10　在嘈雜人聲的背景下各系統的 PESQ 平均分數

SNR	0 dB	5 dB	10 dB	15 dB
noisy speech	1.85	2.18	2.5	2.82
Wiener filter	1.9	2.23	2.52	2.75
Joint spectro-temporal Wiener filter	1.95	2.3	2.65	2.98
proposed method	1.77	2.13	2.43	2.63
Proposed method combined with Wiener filter	1.95	2.29	2.6	2.84

從表 10 我們可發現 joint spectro-temporal Wiener filter 所增加的效能幅度不比在高斯白雜訊的情況，這是由於此方法壓抑了高 rate 與高 scale 的部分，但嘈雜人聲事實上也有分佈於低 rate 與低 scale 跟語音重疊的地方，因為此雜訊特性跟語音相近，所以導致此方法的效能有限，也讓本篇論文提出的方法結合 Wiener 濾波器的性能與 joint spectro-temporal Wiener filter 的性能更為相近。

詳細的平均 PESQ 分數與不同語句間的 PESQ 標準差繪於下圖 27：

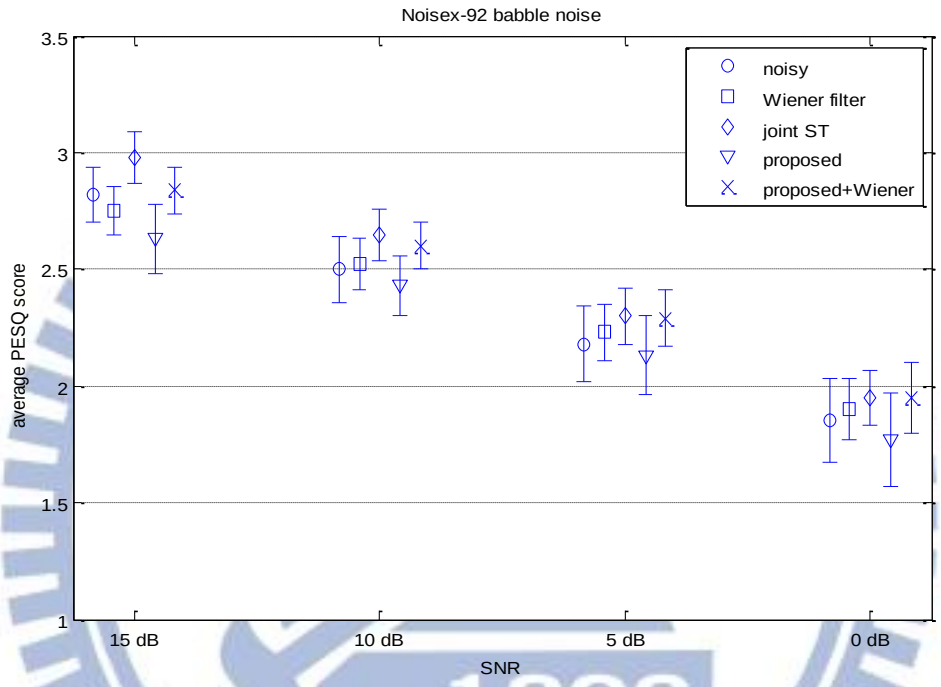


圖 27　嘈雜人聲在不同 SNR 之輸入語音於不同系統處理後的 PESQ 平均與標準差

從圖 27 我們可發現本篇論文所提出之時域封包消噪法在不同語句間的標準差隨著訊雜比越差其標準差值也會越大，主要原因在於我們所作的兩層遮蔽是隨著不同的子頻帶而改變其臨界值 η_1 與 η_2，假若汙染語音的背景雜訊大多分佈在中低頻 (例如：嘈雜人聲)，會導致中低頻帶的臨界值 η_1 與 η_2 比較高，壓抑此部分頻帶的訊號也會比較多，而主要語音部分如果也大多分佈在中低頻帶，將造成不少的語音失真；相對地高斯白雜訊多分佈在高頻帶，若輸入之語音也以分佈在高頻為主，則對此語音造成的失真也會很多，因此結合 Wiener 濾波器一起作語音增強，可降低被雜訊污染最多的子頻帶的臨界值 η_1 與 η_2，而使得這些子頻帶內的壓抑較少，也可以避免過多的語音失真。

接著我們用 Itakura-Saito distance (IS dist.) 來比較本篇所提的方法、本篇所提的方法與 Wiener 濾波器結合以及傳統的 Wiener 濾波器方法這三種系統之間的性

能，此評估系統主要是在計算語音失真的程度，其平均 IS dist.結果如下表 11 所示：

表 11　　在高斯白雜訊的背景下各系統的平均 IS dist.

SNR	0 dB	5 dB	10 dB	15 dB
noisy speech	6.04	5.25	4.41	3.63
Wiener filter	3.96	3.15	2.48	1.92
proposed method	3.42	3.01	3.17	3.38
proposed method combined with Wiener filter	3.39	2.61	2.08	1.6

由表 11 我們可看出，我們所提出之系統結合 Wiener 濾波器的效能最優，而 energy 與 AM 兩層遮蔽消噪則是在訊雜比越差的情況下效能反而越好，正是因為前面有提到的因為語音的主要分佈頻帶與雜訊主要分佈的頻帶一樣，會使得臨界值 η_1 與 η_2 過高，壓抑的值也會過大，導致不少的語音失真，但隨著訊雜比越差，臨界值 η_1 與 η_2 所增加的幅度反而不多，遮蔽的部分相較於訊雜比高的環境還要少，因此失真也較少，但剩餘雜訊也因此較多，IS dist.是主要量測語音失真程度的系統，因此在訊雜比較差的環境時 IS dist.結果較好，然後 PESQ 除了量測語音失真亦有考慮剩餘雜訊的問題，因此像這樣因為遮蔽少失真少但剩餘雜訊卻較多的情況也會導致 PESQ 分數不高。

詳細的平均 IS dist.與不同語句間的 IS dist.標準差繪於下圖 28：

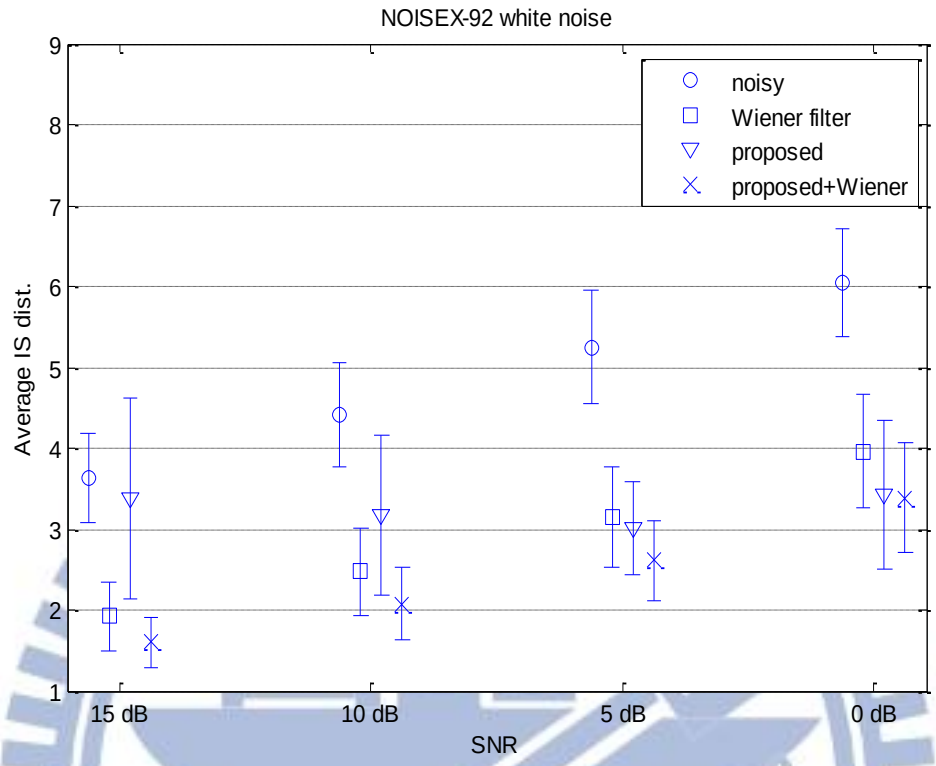


圖 28　高斯白雜訊在不同 SNR 之輸入語音於不同系統處理後的 IS dist.平均與標準差

從圖 28 發現，本篇論文所提出之 energy 與 AM 兩層遮蔽 (proposed)去掉少數幾句 IS dist.特別大的輸入語音之後其標準差值仍略大，主要原因是由於 IS dist.是計算時域上訊號之間的線性估測係數的距離，因此對於幾乎沒有語音分佈的高頻帶上因為遮蔽所造成的失真也考慮在內，使得部分輸入語音經系統處理後，其 IS dist.計算所得的結果相差較多。

下面是背景雜訊為嘈雜人聲時各個系統輸出的平均 IS dist.結果表示：

表 12　　在嘈雜人聲的背景下各系統的平均 IS dist.

SNR	0 dB	5 dB	10 dB	15 dB
noisy speech	4.26	3.56	2.91	2.36
Wiener filter	3.46	2.71	2.15	1.73
proposed method	4.04	4.02	4.05	4.3
Proposed method combined with Wiener filter	2.94	2.31	1.75	1.35

表 12 裡，本篇論文所提出的 energy 與 AM 遮蔽結合 Wiener 濾波器的方法為最佳，而只有 energy 與 AM 遮蔽的系統在四種訊雜比的情況下皆為最差，這是由於第二層的 AM 遮蔽無法準確分辨出語音以及非語音部分，因為嘈雜人聲在 AM 的形狀跟語音相當類似，所以只能由 energy 遮蔽作分辨，而 energy 遮蔽在語音失真與剩餘雜訊的選擇上難以達到一個較好的折衷，導致在嘈雜人聲的環境下，若要去除背景雜訊，勢必會造成可察覺的語音失真，因此本篇論文所提出之時域封包消噪法，要消除跟語音有類似特性的背景雜訊上其效能並不理想。

詳細的平均 IS dist.與不同語句間的 IS dist.標準差繪於下圖 29：

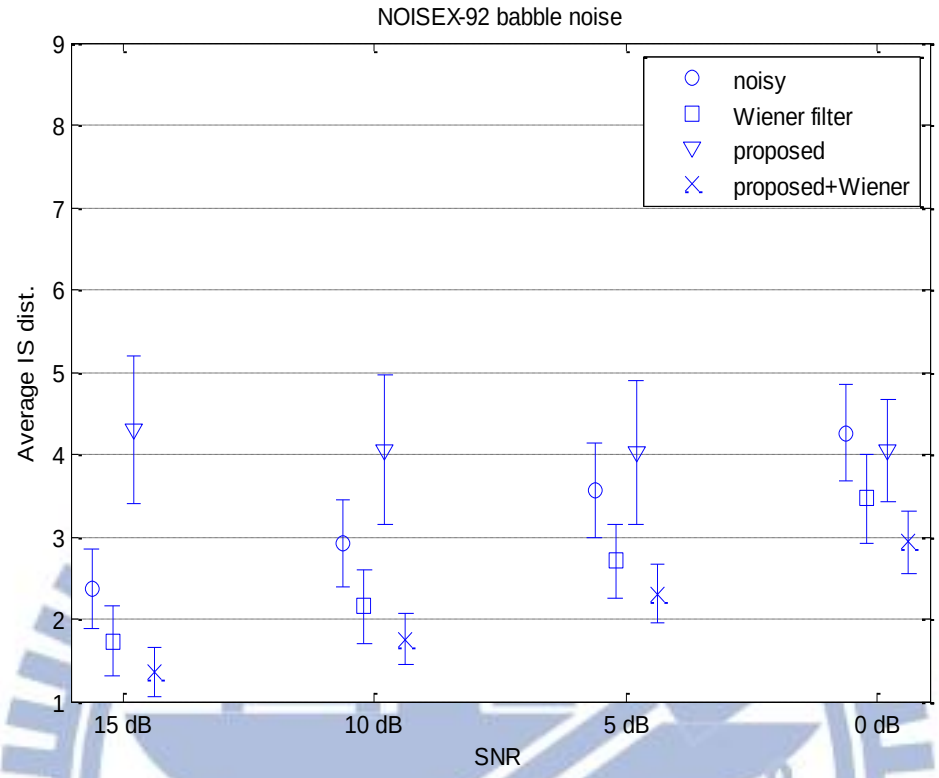


圖 29　嘈雜人聲在不同 SNR 之輸入語音於不同系統處理後的 IS dist.平均與標準差

從圖 29 我們可發現在去掉少數幾句 IS dist.特別大之輸入語音後，proposed 的方法在嘈雜人聲的環境下跟在高斯白雜訊的環境下一樣都是標準差值偏大，主要原因正是 IS dist.對於較無語音分佈的高頻帶所作的遮蔽造成的失真程度也計算在內，而作在嘈雜人聲環境下的 proposed 方法因為只能根據第一層的 energy 遮蔽來消除雜訊，因此其性能比起作在高斯白雜訊的情況還要來的更差，但若與 Wiener 濾波器結合後，可顯著改善傳統 Wiener 濾波器的性能。

下表 13 為實驗中比較過的四種系統計算時所耗的總時間：

表 13　各系統計算所需時間，輸入之語句長度為 2.5 秒左右

System	cost time (s)
Wiener filter	4.26
joint spectro-temporal Wiener filter	28.06
proposed method	22.28
Proposed method combined with Wiener filter	24.07

雖然跟 Wiener 濾波器相比，energy 與 AM 兩層遮蔽的方法仍比較慢，但比起 joint spectro-temporal Wiener，其計算速度略快，這是由於後者有進入四維分析處理，計算複雜度較高，且需要完整的語音資訊才能開始計算，這將會導致語音處理上的延遲，因此無法達成即時性的計算，而 energy 與 AM 遮蔽則只有在二維部分作分析處理，且計算過程亦能符合即時性的需求，用來運算的硬體規格為：Intel (R) Core (TM) i7-2600 CPU @ 3.40 GHz，至於用來計算 joint spectro-temporal Wiener 的硬體規格則為：Intel (R) Core (TM)2 Quad CPU Q 9400 @ 2.66 GHz。

下表 14 是 Wiener 濾波器、本篇論文所提出的方法以及此方法與 Wiener 結合的系統單一音框內所使用的乘法運算子次數之比較：

表 14　各系統之乘法運算子數目比較

System	Number of multiplies
Wiener filter	3547
proposed method	29376
Proposed method combined with Wiener filter	32923

五、結論與未來展望

5.1 結論

本篇論文所提出的方法是基於聽覺感知模型所分析出來的語音在四維的分佈上大都集中在低 rate，而雜訊則主要在高 rate 的結果所作的遮蔽，但與其他相同原理所作的方法不同處在於本篇論文並未作到四維的分析，只在時域上對封包作快速傅立葉轉換後分析其調變頻譜，並應用 energy 與 amplitude modulation (AM)的原理作了兩層遮蔽將語音以及非語音部分作分離並加以遮蔽，而參數 β 值亦可依現實情況的需求作調整，如要求語音失真要壓至最低，則可調低 β 值，相反地若要求雜訊盡可能壓抑就將 β 值調高。

由於本篇論文的方法是先在 spectro 利用 Wiener 濾波器作一次語音增強後再於 temporal 作 energy 與 AM 的兩層遮蔽，屬於 spectro 與 temporal 分開作的方法，與單作在 spectro 的 Wiener 濾波器、單作在 temporal 的 AM 與 FM 兩層遮蔽以及利用 Wiener 濾波器同時作在 spectro-temporal 上的方法比較，從實驗結果可得知，本篇論文所提出之方法雖然僅次於 Wiener 濾波器同時作在 spectro-temporal 上，但卻能滿足即時性運算的需求。

5.2 未來展望

從上一章的實驗結果中不難發現，本篇論文所提出的時域封包調變方法仍有相當多有待改善的地方：

- 會隨著不同語句輸入及不同雜訊，在性能上有大幅度變動的特性，將來必須設計出適應不同訊雜比以及不同背景雜訊的參數。

- 對於 rate 分佈特性跟語音相近的背景雜訊 (例如：嘈雜人聲)，需要更能有效分辨出兩者差異的參數設定，並考慮到時間消耗的需求不使用到高維的分析處理。
- 必須讓音樂雜訊 (musical noise) 殘留更少且不會因此造成語音失真，現階段的實驗中曾用通過低通濾波器對判定為非語音的音框作 smooth，但這會讓有些誤判為非語音的音框產生語音失真，導致評估時 PESQ 分數下降，IS dist.的距離則增加。
- 不須與 Wiener 濾波器結合就可達到穩定且良好的性能，或者與其他方法結合，試著在性能上有更好的突破，例如本篇論文所提之方法並未對相位 (phase)作處理，也許可與頻譜相位補償法 (phase spectrum compensate) [28]作結合。
- 計算時間上仍需要盡可能地縮短，減少運算子數目以及音框數量，未來打算對時-頻單點作判定，判別為非語音部分則統一乘上一極小值遮蔽，不須每個時-頻單點都計算其遮蔽值，如此一來將可省下不少計算時間。

參考文獻

[1] Tai-Shih Chi, Powen Ru and Shihab A. Shamma, “Multiresolution spectrotemporal analysis of complexsounds” J. Acoust. Soc. Am., vol. 118, no. 2, pp. 887-906, 2005.

[2] H. Sheikhzadeh, R. L. Brennan and H. Sameti, “Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications” Proc. IEEE ICASSP, pp. 808-811, 1995.

[3] Tai-Shih Chi, class notes of Auditory and Acoustical Information Processing, Department of Communication Engineering, National Chiao-Tung University, Taiwan, 2011.

[4] Nima Mesgarani and Shihab Shamma, “Denoising in the domain of spectrotemporal modulations” EURASIP Journal on Audio, Speech, and Music Processing Volume 2007.

[5] Tai-Shih Chi, Ting-Han Lin and Chung-Chien Hsu, “Spectro-temporal modulation energy based mask for robust speaker identification” J. Acoust. Soc. Am. 131 (5), pp. 368-374, 2012.

[6] Chung-Chien Hsu, Ting-Han Lin and Tai-Shih Chi, “FFT-based spectro-temporal analysis and synthesis of sounds” Proc. IEEE ICASSP, pp. 5388-5391, 2011.

[7] Chung-Chien Hsu, Tse-En Lin, Jian-Hueng Chen and Tai-Shih Chi, “Spectro-temporal subband wiener filter for speech enhancement” Proc. IEEE ICASSP, pp. 4001-4004, 2012.

[8] P. C. Loizou, Speech Enhancement: Theory and Practice (CRC, New York, 2007).

[9] Antony W. Rix, John G. Beerends, Michael P. Hollier and Andries P. Hekstra,

"Perceptual evaluation of speech quality (PESQ) –a new method for speech quality assessment of telephone networks and codes" Proc. IEEE Acoustics, Speech,and signal processing, pp. 749-752, 2001.

[10] Juang, B.-H., "On the Itakura-Saito measures for speech coder performance evaluation" AT&T Bell Laboratories Technical Journal, 63,8, pp. 1477-1499, 1984.

[11] S. M. Schimmel and L. E. Atlas, "Coherent envelope detection for modulation filtering of speech"
Proc. IEEE Acoustics, Speech,and signal processing, vol 1, pp. 221-224, 2005.

[12] Sofia Ben Jebara, "A perceptual approach to reduce musical noise phenomenon with wiener denoising technique" Proc. IEEE ICASSP, pp. 49-52, 2006.

[13] Md. Jahangir Alam, Sid-Ahmed Selouani and Douglas O'Shaughnessy, "An improved perceptual speech enhancement technique employing a psychoacoustically motivated weighting factor" IEEE ASRU, pp. 266-270, 2009.

[14] A. Amehraye, D.pastor and A. Tamtaoui, "Perceptual improvement of wiener filtering" Proc. IEEE ICASSP, pp. 2081-2084, 2008.

[15] Chang Huai YOU, Soo Ngee KOH and Susanto RAHARDJA, "An MMSE speech enhancement approach incorporating masking properties" Proc. IEEE ICASSP, pp. 725-728, 2004.

[16] Firas Jabloun and Benoit Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement" IEEE Trans. Acoustics, Speech, Audio Processing, Vol. 11, No. 6, pp. 700-708, 2003.

[17] Chang Huai YOU, Susanto RAHARDJA and Soo Ngee KOH, "Perceptual kalman filtering enhancement" Proc. IEEE ICASSP, pp. 461-464, 2006.

[18] Hu, Y. and Loizou, P, "Subjective evaluation and comparison of speech enhancement algorithms" speech communication. 49, pp. 588-601, 2007.

[19] H. Hirsch, and D. Pearce, "The Aurora Experimental Framework for the

Performance Evaluation of Speech Recognition Systems under Noisy Conditions." *ISCA ITRW ASR2000*, Paris, France, pp. 18-20, 2000.

[20] Yi Hu and Philipos C. Loizou, "A perceptually Motivated Approach for speech enhancement" IEEE Trans. Acoustics, Speech, Audio Processing, Vol. 11, No. 5, pp. 457-465, 2003.

[21] Te-Won Lee and Kaisheng Yao, "Speech enhancement by perceptual filter with sequential noise parameter estimation" Proc. IEEE ICASSP, pp. 693-696, 2004.

[22] Hong You and Abeer Alwan, "Temporal modulation processing of speech signals for noise robust ASR" Proc. Interspeech, pp. 36–39, 2009.

[23] Kuen-Shian Tsai, Li-Hui Tseng, Cheng-Jung Wu and Shuenn-Tsong Young, "Development of a mandarin monosyllable recognition test" Ear & Hearing, vol. 30, No. 1, pp. 90-99, 2009.

[24] Rainer Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics" IEEE Tran. Acoustics, Speech, Audio Processing, Vol. 9, No. 5, pp. 504-512, 2001.

[25] Thomas Esch and Peter Vary, "Exploiting temporal correlation of speech and noise magnitude using a modified kalman filter for speech enhancement" ITG-Fachtagung Sprachkommunikation, pp. 8-10, 2008.

[26] Esfandiar Zavarehei and Saeed Vaseghi, "Speech Enhancement in temporal DFT trajectories using kalman filters," Proc. of INTERSPEECH, Lisbon, Portugal, pp. 2077-2080, 2005.

[27] Thomas Esch and Peter Vary, "Speech enhancement using a modified kalman filter based on complex linear prediction and supergaussian priors, " Proc. of ICASSP, Las Vegas, USA, pp. 4877-4880, 2008.

[28] Kamil Wojcicki, Mitar Milacic, Anthony Stark, James Lyons and Kuldip Paliwal, "Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement" IEEE Signal
Process. Lett., vol. 15, pp. 461–464, 2008.

[29] Stephen So, Kamil K. Wojcicki, James G. Lyons, Anthony P. Stark, Kuldip K. Paliwal, “Kalman filter with phase spectrum compensation algorithm for speech enhancement.” Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 4405–4408. ,2009.

[30] Esfandiar Zavarehei, Saeed Vaseghi, and Qin Yan, “SPeech enhancement with kalman filtering the short-time DFT trajectories of noise and speech” EURASIP, 2006.

[31] Afshin Rezayee and Saeed Gazor, “An adaptive KLT approach for speech enhancement” IEEE Tran. Acoustics, Speech, Audio Processing, Vol. 9, No. 2, pp. 87-95, 2001.

[32] Stephen So, Kuldip Paliwal, “Suppressing the influence of additive noise on the kalman filter gain for low residual noise speech enhancement.” Speech Commun. 53 (3), pp. 355–378, 2010.

[33] S.D. Apte and Shridhar, “An efficient speech enhancement algorithm using conjugate symmetry of DFT” Electrical Engineering and Control, LNEE 98, pp. 695-701, 2011.

[34] Ching-Ta Lu, Kun-Fu Tseng and Chih-Tsung Chen, “Reduction of residual noise using directional median filter” IEEE CSAE., pp. 475-479, 2011.

[35] Sriram Ganapathy, Samuel Thomas and Hynek Hermansky, “Temporal envelope subtraction for robust speech recognition using modulation spectrum” IEEE ASRU., pp. 164-169, 2009.

[36] James G. Lyons and Kuldip K. Paliwal, “Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement.” Proc. ISCA Conf. Internat. Speech Comm. Assoc. (INTERSPEECH), pp. 387–390, 2008.

[37] Tiago H. Falk, Svante Stadler, W. Bastiaan Kleijn and Wai-Yip Chan, “Noise suppression based on extending a speech-dominated modulation band,” Interspeech, pp. 970-973, 2007.

[38] Wen-Rong Wu and Po-Cheng Chen, “Subband kalman filtering for speech

enhancement” IEEE Tran., analog and digital signal processing, Vol. 45, No. 8, 1998.

[39] Kuldip Paliwal, Kamil Wo´jcicki, Belinda Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain.” Speech Comm. 52 (5), pp. 450–475.

[40] Stephen So and Kuldip K. Paliwal, “Modulation-domain kalman filtering for singlechannel speech enhancement.” Speech Comm. 53 (6), pp. 818–829, 2011.

[41] Mark Marzinzik and Birger Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics” IEEE Tran. Acoustics, Speech, Audio Processing, Vol. 10, No. 2, pp. 109-118, 2002.

[42] Jing-Dong Chen, Jacob Benesty, Yiteng (Arden) Huang and Simon Doclo, “New insights into the noise reduction wiener filter” IEEE Tran. Audio, Speech, Language Processing, Vol. 14, No. 4, pp. 1218-1234, 2006.