國立交通大學

電信工程學系

碩 士 論 文

藉由感知特徵對語音品質做客觀的評量

Objective Assessment of Speech Quality by Perceptual Features

研 究 生：顏廷宇

指導教授：冀泰石　教授

中 華 民 國 九 十 七 年 七 月

藉由感知特徵對語音品質做客觀的評量

Objective Assessment of Speech Quality by Perceptual Features

研 究 生：顏廷宇　　Student：Ting-Yu Yen

指導教授：冀泰石　　Advisor：Tai-Shih Chi

國 立 交 通 大 學
電 信 工 程 學 系
碩 士 論 文

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

Communication Engineering

July 2008

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 七 年 七 月

藉由感知特徵對語音品質做客觀的評量

研究生：顏廷宇　　指導教授：冀泰石 博士

國立交通大學電信工程學系碩士班

中文摘要

在本論文中，我們使用一個同時考慮時間和頻率上變化的人耳聽覺模型來對語音品質做客觀的評量。我們研究目的是希望可以準確的預測聽者對於語音品質主觀的平均意見分數。客觀的評量主要分爲兩種方法：一種是侵入式，另一種是非侵入式。首先，我們會在兩個聽覺感知階段觀察和分析乾淨的語音、加上背景雜訊的語音、以及經過各種不同語音壓縮標準的語音，第一個階段是人耳到中腦的頻譜估計，第二個階段是中腦到大腦皮質聽覺區對時域和頻域同時做分析。其次，我們將從這兩個階段，擷取出在人耳感知上可能影響聽者判斷語音品質好壞的特徵當作參數來對語音品質做客觀評量，這三個特徵分別是－理解性、自然性、基頻失真。最後，我們使用複迴歸分析的方法，將三個特徵參數對語音品質影響的關係做結合，希望藉由這三個基本的特徵參數讓我們能對語音品質的好壞做快速並可靠的評量。

關鍵詞：語音品質、侵入式、非侵入式、特徵參數、理解性、自然性、基頻失真

Objective Assessment of Speech Quality by Perceptual Features

Student: Ting-Yu Yen Advisor: Dr. Tai-Shih Chi

Department of Communication Engineering

National Chiao Tung University

Abstract

In this study, a joint spectro-temporal auditory model was utilized to assess speech quality objectively. In this model, the first stage is to mimic early cochlear functions of the spectrum estimation and the second stage is to mimic cortical functions of the multi-dimensional spectrum analysis. The goal of this study is to predict subjective mean opinion score (MOS).

Objective speech quality assessment can be done by two methods：intrusive and non-intrusive. In this study, firstly, we observe and analyze patterns of the clean speech, the noisy speech with different background noise, and the degraded speech through different codecs at two auditory stages. Secondly, we will derive an objective estimate of the MOS from data-driven perceptual parameters which are believed to reflect people's judgment on speech quality. Four perceptual parameters considered are intelligibility, naturalness, and pitch distortion. Finally, we use multiple regression analysis to combine the relationship between speech quality and these perceptual parameters, and then obtain our predicted MOS. We then demonstrate the MOS can be characterized quickly and reliably by these three perceptual features.

Keywords：joint spectro-temporal, speech quality, MOS, intrusive, non-intrusive, intelligibility, naturalness, pitch distortion

誌　　謝

在交大電信研究所這兩年，真的誠摯感謝我的指導教授冀泰石老師不厭其煩幫我在研究上耐心的指導和解惑，研究所念碩士和大學部求學時思考和學習模式有相當大的差異，因爲冀老師不時的經驗傳承和熱心協助，我才能很快的適應研究所生活。老師帶研究生的方式非常自由和輕鬆，不強迫我們一定要待在實驗室做研究，但希望我們能養成自動自發學習並自我要求；在剛升上碩一的暑假，老師除了叮嚀要先看一些書和論文，也鼓勵我們應該多出去走走增廣見聞。每次的meeting 都在輕鬆愉快的氣氛下進行，老師會跟我們一起討論思考並解決問題。除了課業和研究上的問題，老師說生活上問題也可以找他諮詢；老師總是保持微笑，讓我們學習到凡事與人相處要樂觀開朗。

另外，我還要感謝實驗室其它同學以及學弟們，每個人都相當熱心而且好相處，實驗室的氣氛非常融洽和愉快，在實驗室這兩年的點點滴滴更是我人生中最值得珍惜的時光；然後我還要感謝我的爸媽，謝謝你們多年來的辛苦養育之恩，有你們的協助我才能支付在外地求學龐大的生活開銷；最後我還必須感謝女朋友的支持、體諒和包容，我才能沒有後顧之憂的在新竹完成繁重的學業和研究。

我會帶著大家的鼓勵、期許和祝福在未來的人生裡繼續努力和學習，真的非常感激所有曾經幫助過我的朋友們，謝謝你們。

目　　錄

表目錄

圖目錄

符號說明

ITU-T：International Telecommunication Union---
telecommunication standardization sector

MOS ：mean opinion score

ACR ：Absolute Category Rating

PCM ：Pulse Code Modulation

VAD ：Voice Activity Detection

PESQ：Perceptual Evaluation of Speech Quality

MNRU：Modulated Noise Reference Unit

STMI ：Spectro-temporal modulation index

第一章　導論

1.1　研究動機

語音傳輸在有線與無線通信網路是相當普及的應用，隨著電腦、通訊、多媒體等資訊科技成熟的發展，現代的電話網路變的越來越複雜，除了固有傳統的公眾交換電話網路 (PSTN)，各種不同型態的電信網路在我們生活中被廣泛的使用，例如行動電話網路的 GSM、UMTS、CDMA 以及透過網際網路傳輸語音的 VoIP 等。因此，可能降低通信品質的因子就變的相當多而且無法預測。尤其當各種不同的電信網路相連接時，我們想要再去了解並估算各個通信系統的元件或連接方式對於通信品質的衝擊就變的更加困難。另外，爲了增加傳輸系統的容量，各種低位元傳輸率的語音壓縮技術不斷的推陳出新，但語音壓縮的程度越高，語音需要編碼解碼的時間就增加，語音訊號傳送過程的延遲也會增加，爲了維持語音訊號經過壓縮後的通話品質，系統的運算能力必須提高，相對的，系統的複雜度和成本也隨壓縮程度的增加而提高。如何在有限的系統資源時，仍能維持高的通話品質，在通話品質和系統成本間取得平衡更是通信系統設計者所關心的問題。因此，如何有效率評估傳輸系統接收端通話品質的好壞顯然相當重要。

最可靠的語音品質評量方式，是找一大群受試者直接來聽各種不同要測試的語音，並把語音聽起來的品質好壞分成五個等級做評分，得到一個主觀的平均意見分數；但顯然的，這種主觀評估語音品質的方式相當耗費人力、金錢與時間，尤其當複雜的通信網路架構有一小部份地方改變了，整個主觀評量語音品質好壞的實驗又必須重做一次，所以主觀評量方式在實際上並不可行。因此，我們希望能發展出一套客觀評量語音品質好壞的系統以取代耗時及高成本的主觀評量方式，藉由語音訊號在聽覺感知上的分析與研究，得到準確的語音品質評量分數。

1. 2 研究方向

在過去好幾年來，各種不同客觀評量語音品質的方法持續發展和研究 [1]，從 1990 年開始，客觀評量語音品質的發展，開始朝向以心理聲學爲基礎所發展出來的感知模型得到的參數距離，取代原本以波形或發聲模型爲基礎的參數距離。一個成功的客觀語音品質評量方法不僅可以快速提供TTS (text-to-speech)合成器或行動電話網路通道上語音傳輸品質的監控，更希望能將此方法擴展到對多媒體音樂品質評量。

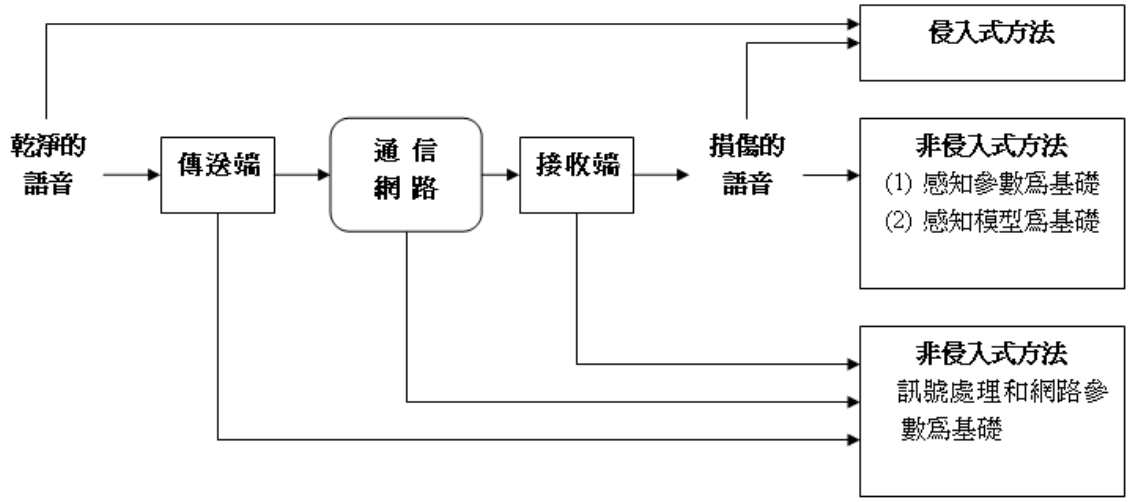


圖 1-1：客觀評量語音品質的三種方法

客觀評量語音品質好壞分爲兩大方式：第一種是侵入式 (intrusive) 方法，另一種是非侵入式 (Non-intrusive) 方法。侵入式方法必須有傳送端乾淨的語音信號當作參考，再和接收端經過損傷的語音信號做比較，比較的方式是從乾淨和損傷的信號抽取出人類發聲系統或聽覺系統的參數，將兩者參數間距離的差異作爲訊號失真程度的評量，目前 ITU-T 已提出一套國際標準 PESQ [2]，其它已經發表的研究可以參考 [3] [4]。

非侵入式的方法主要可以分成兩種，一種是以模型爲基礎，另一種則是以參數爲基礎；這兩種方法都沒有傳送端乾淨的語音訊號作爲參考，只能針對接收端受到損傷的訊號作爲評分的依據。

模型爲基礎的方法主要是先從多個語音資料庫建立乾淨語音的模型，再和損傷的語音比對兩者間模型上的不一致性程度，作爲語音品質好壞的評量標準；近幾年來相關的研究相當多，例如：以感知線性預估 (PLP) 係數和 PLP 倒頻譜建立的向量量化技術 [5]、以發聲腔道建立的模型 [6]、以 PLP 建立的高斯混合模型 [7]，目前 ITU-T 也在前幾年提出一個以模型爲基礎的國際標準 P. 563 [8]。

參數爲基礎的方法分爲兩種，第一種是針對語音訊號從傳送端、經過通信網路、再到接收端，經過了哪些訊號處理和網路損傷。例如：不同的語音壓縮標準、語音壓縮前經過的 VAD (voice activity detection) 處理、接收端耳機和喇叭的好壞、背景雜訊的訊雜比高低、網路傳輸過程發生時間延遲、封包遺失、回音等。相關的研究有 [9] [10]，ITU-T 也在前幾年提出一個國際標準 G. 107 [11]。

另一種以參數爲基礎的方法並不是對訊號或是網路本身加以分析，而是從人的感知反應來分析，例如：同時考慮人耳發聲和聽覺的重要特性，對語音訊號經過人耳感知處理後的頻譜加以分析[12] [13] [14]。由前人的研究知道，頻譜上隨著時間變動的封包特性和聽者判斷語音品質好壞有關係，於是便可藉由分析時間軸上的封包特性來評量語音品質。我們的研究便是從低階的人耳感知行爲出發，進而延伸到高階的大腦認知行爲，希望藉由低階的人耳感知模型伴隨著高階的大腦認知模型，可以有效並正確的客觀評量語音品質好壞。

1. 3 研究方法

我們客觀評量語音品質的方法，會先從低階的人耳感知聽覺開始再擴展到高階的人類大腦認知與解析。我們使用一個結合頻、時域 (joint spectro-temporal) 的

人耳聽覺運算模型，用此聽覺模型觀察並分析語音訊號。

此模型是根據已知生物物理現象的聽覺系統及大腦皮質聽覺區單一神經元的反應而建立。這個多重解析的聽覺模型 [15] 可用來說明由低階的聽覺感知行爲，再上達直至大腦皮質聽覺區的完整路徑。在我們的研究方法裡，我們著重在高階的大腦認知模型的發展，並將模型運用在語音品質評量上。這個認知模型應該有多重的維度，其中我們會先探討的兩個維度是語音的理解性(Intelligibility)[16][17]及自然性(Naturalness)[18][19]，再試著探索其它維度，希望能夠確認更多和語音品質相關的維度，並將這些維度納入我們的認知模型(Cognitive model)[20]。

此外，本研究除了探討大腦對語音品質認知的兩個維度，我們尚在聽覺模型中觀察到，經過壓縮處理後的語音以及加上背景雜訊的語音，在基頻能量分佈區域有相當大的失真，而這部份的失真和語音品質的好壞有高度的相關；最後，考慮到語音在經過網路傳送過程中可能發生位元、音框和封包遺失的問題，我們對於接收到的語音訊號在時間軸上的不連續現象也做了處理和分析，並將此時間軸上的不連續現象造成語音品質下降的關係用一個數學模型來表示 [21] [22]。

1.4 章節概要

第一章 導論：本篇論文的研究動機、研究方向、研究方法以及章節概要。

第二章 感知聽覺系統與模型

第三章 語音資料庫介紹、主觀評量與客觀評量語音品質方法的國際標準

第四章 客觀的侵入式語音品質評量方法

第五章 客觀的非侵入式語音品質評量的方法

第六章 結果討論與未來展望

第二章　感知聽覺系統與模型

2. 1 人耳生理解剖學上的構造

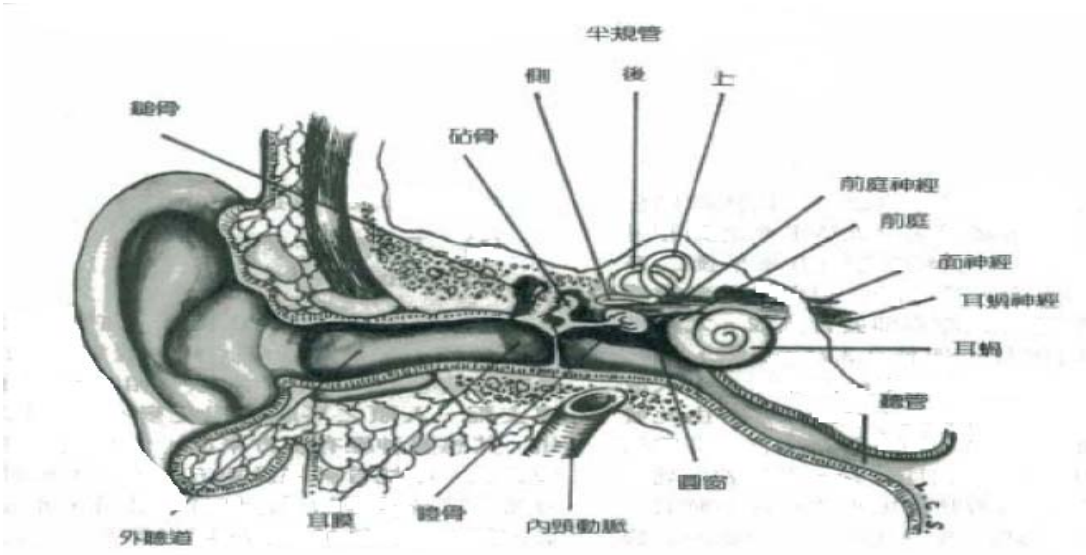


圖 2-1：人耳生理解剖學上的構造
資料來源：張斌耳鼻喉科學

2.1.1　外耳

耳廓到鼓(耳)膜這段屬於外耳，耳廓用來判斷聲音來源的方向；耳廓到鼓膜間的通道叫做耳道，耳道是個共振腔，第一個共振頻率大約 3KHz，共振的作用造成此頻率附近的聲波被放大，因此人耳聽覺對 3kHz 附近的頻率聲音特別敏感。

2.1.2　中耳

從耳道傳進來的聲波會碰擊鼓膜，鼓膜會以和聲波相同的頻率內外震動，鼓膜的震動會啓動三小聽骨(錘骨、砧骨和鐙骨)的運動，三小聽骨扮演著槓桿和壓力波交換器的角色，將低壓的鼓膜振動轉換成在表面積很小的卵圓窗上的高壓聲

音振動，這種壓力波的轉換是必要的，因爲聲波在外耳時是在空氣中傳送，過了卵圓窗後來到內耳，是在組織液(液體)中傳送。在中耳時，聲音資訊仍以壓力波的形式存在，傳到內耳的耳蝸，聲波會被轉換成神經衝動再向上傳至中腦。

2.1.3 內耳

內耳最重要的聽覺構造就是耳蝸(Cochlea)，耳蝸內的基底膜(basilar membrane)可對傳送進來的聲波做頻率成份分析，基底膜上的柯蒂氏器(Organ of corti)有內毛細胞和外毛細胞，可將聲波資訊轉換成神經衝動傳至大腦。

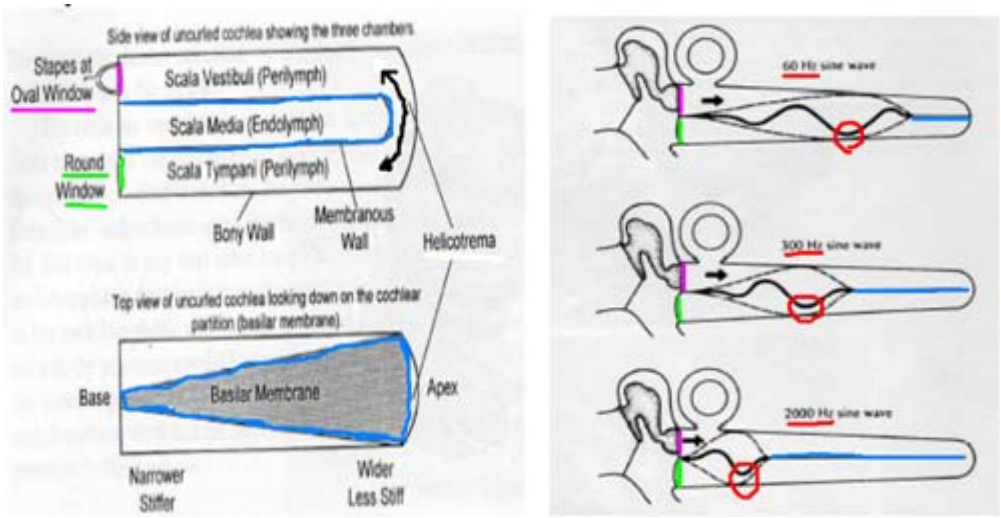


圖 2-2：基底膜的位置及其隨著不同頻率的聲波以行進波方式震動
資料來源：Hearing Physiology Handouts

基底膜是一個貫穿整個耳蝸的膜狀結構。由圖 2-2 可以看到，壓力波從卵圓窗(Oval window)進來，內耳的液體壓力產生變化，基底膜因爲液體壓力產生行進波(travelling wave)的震動，基底膜的震動會在和聲波頻率共振的位置上形成最大的波動幅度。最後，壓力波傳到圓窗導致其突起，得到壓力的釋放。基底膜總長度大約 35mm；在靠近中耳進來的地方，比較狹宰和僵硬，相反的，在內耳深處，會比較寬闊有彈性。比較狹宰和僵硬的地方，可以感受較高頻率的聲波；比較寬

闊有彈性的地方，感受到較低頻率的聲波。基底膜的共振頻率範圍大約20-20kHz，即人類的正常聽覺的頻率範圍。

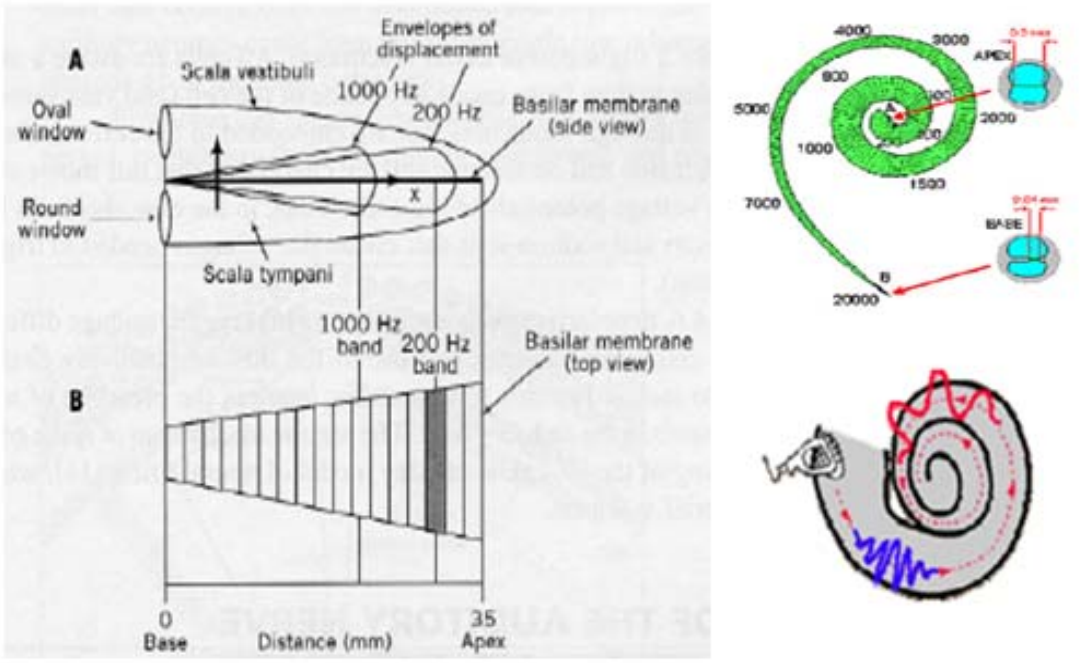


圖 2-3：基底膜外窄內寬的形狀及其感受不同頻率聲波的位置
資料來源：Hearing Physiology Handouts

毛細胞(Hair cells)規則分佈於基底膜上面，靠近耳蝸中心的稱爲內毛細胞(Inner hair cells)，遠離耳蝸中心的稱爲外毛細胞(Outer hair cells)。內毛細胞是感受器細胞，和聽神經纖維形成突觸相連，其主要的功能在將內耳液體機械震動的壓力波轉換成神經訊號的動作電位，再由聽神經傳至大腦。

基底膜行進波的運動會在和聲波頻率共振的位置上產生受激態，但在此位置前面(較高頻率位置)也會形成壓抑態；人耳接收到的聲波通常是許多頻率所組成，因此複雜頻率的聲波造成基底膜上連續的行進波，導致相鄰近位置的內毛細胞其功能有彼此互相壓抑的現象，這種現象正可以說明人耳有頻率遮蔽的效應。例如：一個頻率 800Hz、音強(intensity)80dB 的聲音會在其頻率附近形成遮蔽曲線，這條曲線左右並不對稱，往高頻下降慢但往低頻下降快，此頻率聲音會拉高附近頻率聲音的聽覺閥形成遮蔽閥，導致其它聲音必須有更高音強才聽的到。

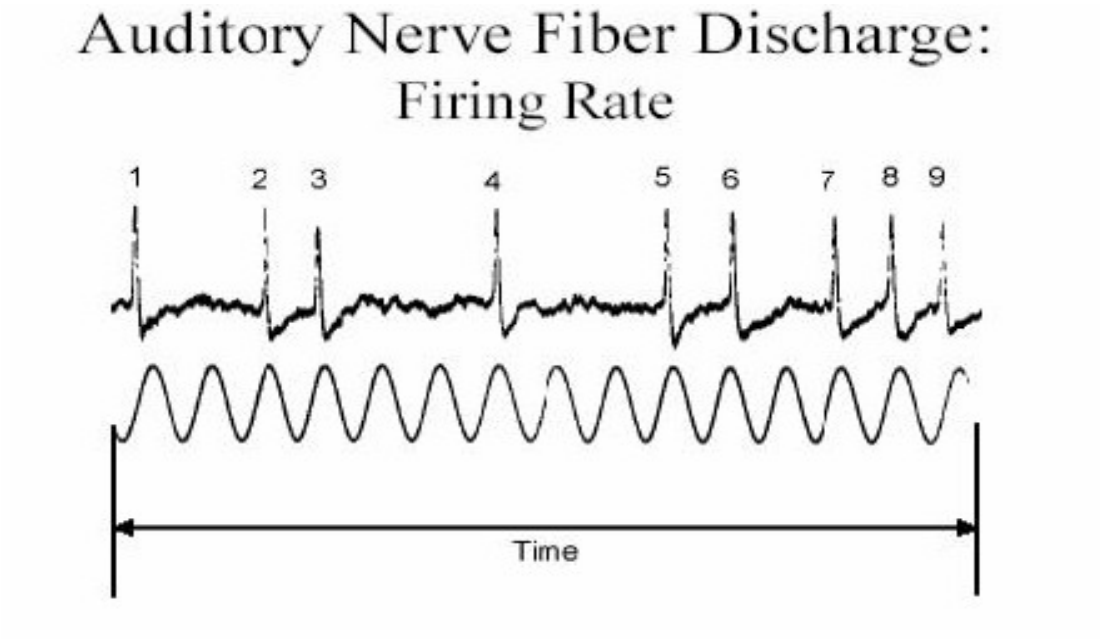


圖 2-4：單頻率的聲音和聽神經細胞發射神經衝動關係圖

資料來源：Hearing Physiology Handouts

當內毛細胞將聲波資訊由壓力波轉成動作電位後，聲音訊息變成電位訊號沿著聽神經繼續傳到大腦；但當某神經元在維持一段時間的動作電位後，神經元需要休息才能繼續發射神經衝動。從圖 2-4 我們可以看到，某個單頻率的聲波經由基底膜的振動讓內毛細胞接收到訊息後，在正常情況下，內毛細胞會在聲波的同一相位發射神經衝動，但若聲波的頻率太高，內毛細胞發射神經衝動的速度便無法跟的上聲波頻率，因此在某些週期，便會遺失一部份神經衝動。內毛細胞是否會正常發射神經衝動不僅和聲波的頻率有關，也和聲波的音強有關，若聲波的振幅不夠大，神經細胞感受到的刺激不夠強，也不會發射神經衝動。

我們從生物學上已經發現，人的聽覺對於聲波在時間上變化的解析度或敏感度越來越低。在耳朵的階段，內毛細胞神經衝動發射速率最高約 5 kHz，這也正可解釋我們人耳大多對~5 kHz 以下的聲音有較強的反應。到了中腦的聽神經細胞，大概只能處理頻率最高 1kHz 的訊號；最後到了大腦，只能處理頻率 100 多 Hz 的訊號了。而所有這些人耳感知特性在我們的感知聽覺模型中都有考慮進去。在下一章節裡，將詳細介紹我們的感知聽覺模型。

2. 2 感知聽覺模型

我們採用的感知聽覺模型其基礎是建立在：哺乳動物對於聲音的處理從耳朵到大腦的聽覺路徑。此路徑主要包含兩種轉換：第一個階段是頻譜的估計(Spectrum Estimation)，第二個階段是頻譜的分析(Spectrum Analysis)。第一個階段是初期耳蝸階段(Early cochlear stage)，模擬人耳到中腦的信號轉換結果，對人耳在低階的感知聽覺上有完整的描述。第二個階段模擬中腦到大腦對訊號在時域和頻域一起分析(joint spectro-temporal modulation)。

2.2.1 初期耳蝸感知階段

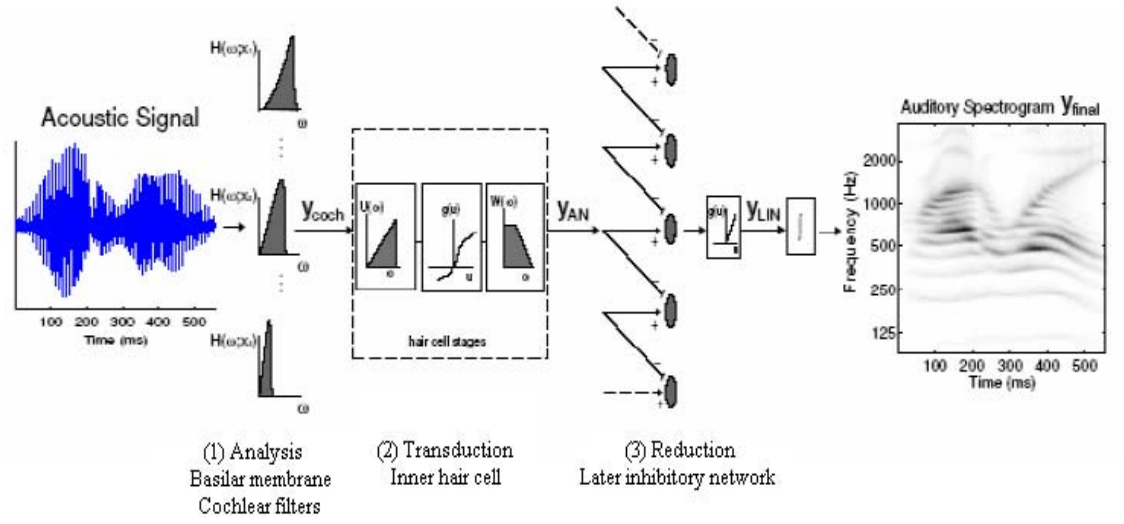


圖 2-5：聲波信號經過耳朵到中腦的感知特性得到聽覺頻譜圖

$$y_{coch}(t,x) = s(t) *_t h(t;x) \quad (1)$$

$$y_{AN}(t,x) = g(\partial_t y_{coch}(t,x)) *_t \omega(t) \quad (2)$$

$$y_{LIN}(t,x) = \max(\partial_x y_{AN}(t,x), 0) \quad (3)$$

$$y_{final}(t,x) = y_{LIN}(t,x) *_t \mu(t;\tau) \quad (4)$$

初期耳蝸階段主要分爲三個程序：第一個程序是模擬耳蝸的基底膜經由行進波的振動分析聲波訊號的頻率，第二個程序是模擬耳蝸的內毛細胞將基底膜震動的壓力波轉換成神經衝動的電位訊號，第三個程式是模擬內毛細胞彼此互相壓抑的效果。

第一個程序是做頻率分析，由於人耳對聲音高低的感覺是和基本頻率的對數成正比，因此我們在對數的頻率軸上，等分成一組 128 個有互相重疊(overlapping)的帶通濾波器(cochlear filter bank)，這些濾波器的中心頻率除以頻寬必須等於常數 Q，我們設定Q值爲 4。由 式(1) ，s(t)是時域的聲波信號，h(t; x)是帶通濾波器的頻率響應，x的不同代表基底膜上不同位置，對應不同的共振頻率，$*_t$ 表示時域上的褶積(convolution),這組濾波器可以分析 5.3 個倍頻(octave)，每個倍頻裡包含 24 個濾波器。隨著想要分析的聲波訊號其取樣頻率的不同，我們可以調整這些濾波器的中心頻率，進而在人耳可聽到的頻率範圍 20Hz~20kHz移動這組濾波器來分析。圖 2-6 以取樣速率 8kHz的濾波器組爲例作說明。

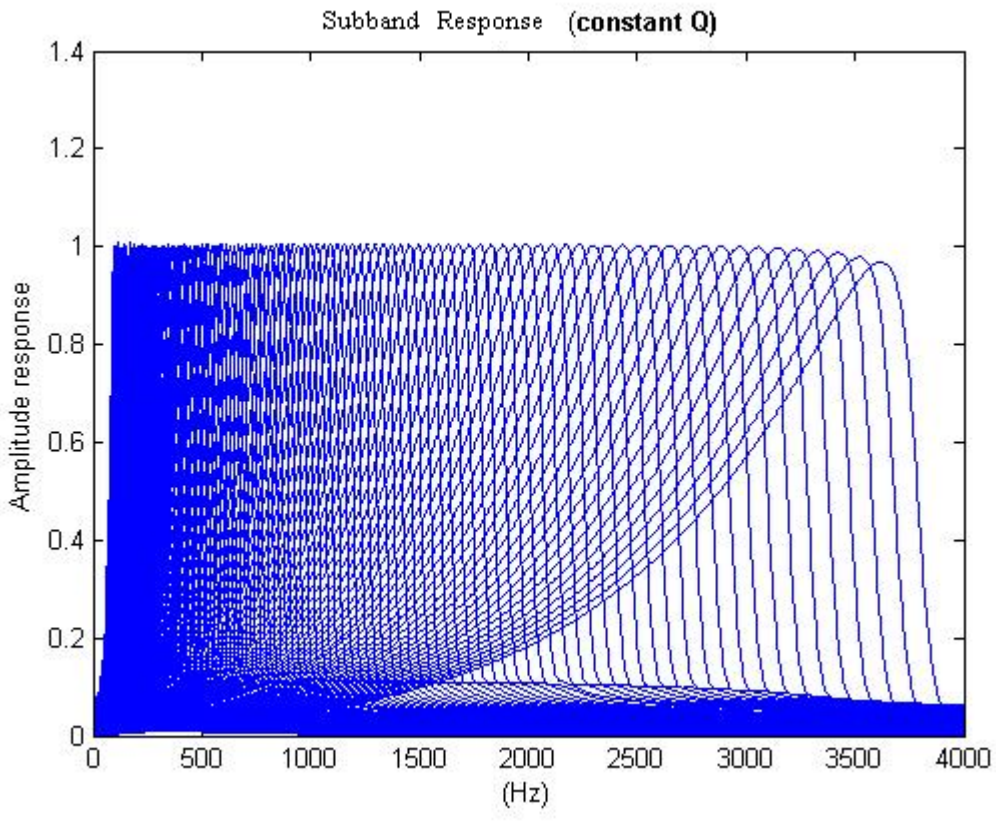


圖 2-6：在 0Hz 至 4kHz 有 128 個濾波器，最低的濾波器中心頻率是 90Hz。

第二個程序是內毛細胞做轉導，可以分爲三個步驟：第一個步驟我們用一個高通濾波器做微分模擬聲波的壓力波轉換成速度；第二個步驟我們用一個 sigmoid 函數 $g(u) = 1/(1+e^{-u})$ 模擬內毛細胞的保護作用；第三個步驟我們用一個低通濾波器模擬內毛細胞的神經衝動發射速率。

我們由這三個函數微觀的功能也可以來解釋人耳聽覺巨觀的現象：因爲人耳對音強的感知並非呈線性，而是較接近對數曲線，因此第二個步驟中的 sigmoid 函數功能可以解釋音強(intensity)到響度(loudness)間壓縮過程。一般人耳比較敏銳的頻率範圍大約 200Hz 至 4kHz，因此第一個步驟和第三個步驟一起看成一個帶通濾波器，這個濾波器可以解釋人耳聽覺的臨界曲線(hearing threshold)。由式(2)我們可以清楚看到這些函數功能，其中的 $\omega(t)$ 代表積分視窗。

第三個程序是內毛細胞會有彼此壓抑的現象，我們使用一階的差分濾波器沿著 x 軸做差分後再接著半波整流器模擬此現象得到 y_{LIN}，由式(3)可看到函數的功能。這現象可以解釋人耳聽覺的頻率遮蔽效應(frequency masking)。

從耳朵到中腦，神經細胞對於訊號在時間軸上的變化越來越遲鈍(temporal dynamics reduction)，中腦聽覺神經細胞的神經衝動發射速率大約只剩 1kHz，因此我們用積分視窗 $\mu(t;\tau) = e^{-t/\tau} \cdot u(t)$ 模擬此現象。由式(4)，τ 是時間常數(time constant)，$u(t)$ 是單位步階函數(unit step function)，時間常數 τ 可以隨不同的分析目的做變化，其數值的倒數表示低通濾波器的 3dB 頻寬。例如：我們想要在中腦的頻譜圖上至少看到 1kHz 的聲音變化，時間常數 τ 通常是取 2ms。

由圖 2-7 可以看到，左邊是耳朵上呈現的頻譜圖 y_{LIN}，右邊是中腦上呈現的頻譜圖 y_{final}。在 y_{LIN} 我們可以從頻率軸上看到有三個不同頻率的聲音和其波紋的變化，也可以從時間軸上由聲音的波紋計算其頻率；但從 y_{final} 我們只能從頻率軸上看到 250Hz、1kHz 和 4kHz 有聲音，但 4kHz 聲音的波紋變化(fine structure)就看不到了，表示中腦的聽覺神經細胞對 4kHz 聲音已經分辨不出來。如果想要在中腦保留更高頻率的聲音變化，時間常數 τ 就要設定比較小。

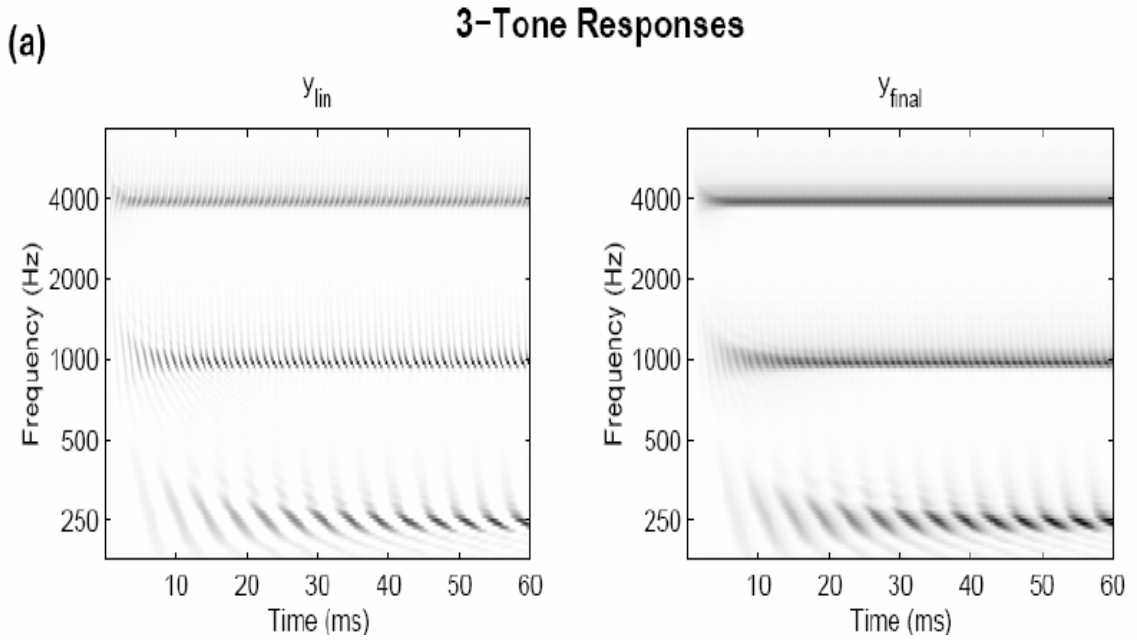


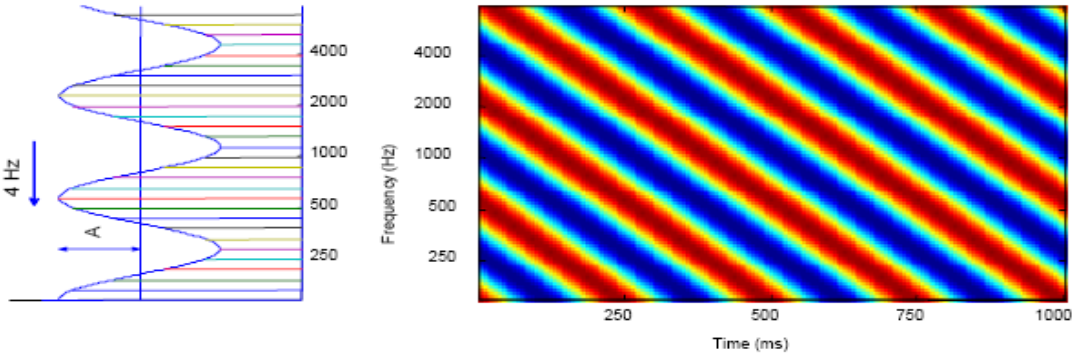
圖 2-7： y_{LIN} 是耳朵的頻譜圖， y_{final} 是中腦的頻譜圖

2.2.2 大腦皮質聽覺區解析階段

對於經過耳朵和中腦感知後得到的頻譜圖，大腦皮質聽覺區階段會對頻譜圖進一步的分析。從工程學的角度來看，大腦基本上是把中腦輸出的頻譜圖拿來直接當作兩個維度的圖案(pattern)做處理。因此我們將大腦每個神經細胞對於這個二維圖案輸入訊號的輸出訊號當作二維的脈衝響應(Spectro-Temporal Receptive Field)，不同的神經細胞有不同的二維脈衝響應。這個脈衝響應不但對二維圖案對頻率上的變化有選擇性，對於時間上的變化也有選擇性；我們可以假設：大腦對於來自中腦的二維圖案進一步做了萃取的動作。

因此，我們用一組結合頻域和時域的調變濾波器(joint spectro-temporal modulation filters)來模擬大腦皮質聽覺區對頻譜圖的解析，藉由這種多重解析，聲波的所有資訊都可清楚在分析結果中看到。從數學的觀點，結合頻域和時域的多重解析結果可以視爲頻譜圖經過一群二維的帶通濾波器的輸出。因此，頻譜圖所包含的聲音訊息會被大腦皮質不同的濾波器以不同的解析度來做編碼處理。

從訊號處理觀點，若將大腦皮質聽覺區當作系統，而來自中腦的頻譜圖當做輸入訊號，吾人可以藉由系統的輸出訊號了解大腦皮質聽覺區是如何解析頻譜圖。因爲系統的輸入，也就是頻譜圖，是二維的輸入訊號，因此模型提出者設計了同時包含特定頻域和時域變化的信號，稱之爲移動波紋刺激源(moving ripple stimulus)來模擬頻譜圖。當然，人聲的頻譜圖是由許多特定頻域和時域變化的信號所組成；將許多不同的結合頻域和時域變化的信號經過我們的二維帶通濾波器多重解析的輸出結果發現，和生物實驗上哺乳動物大腦皮質聽覺區對於聲音訊號的反應圖案相當接近，因此假設我們所使用的感知聽覺模型可以模擬真實大腦皮質聽覺區最外層的函數功能，希望此模型輸出結果更加貼近人耳對聲音真實的感受。

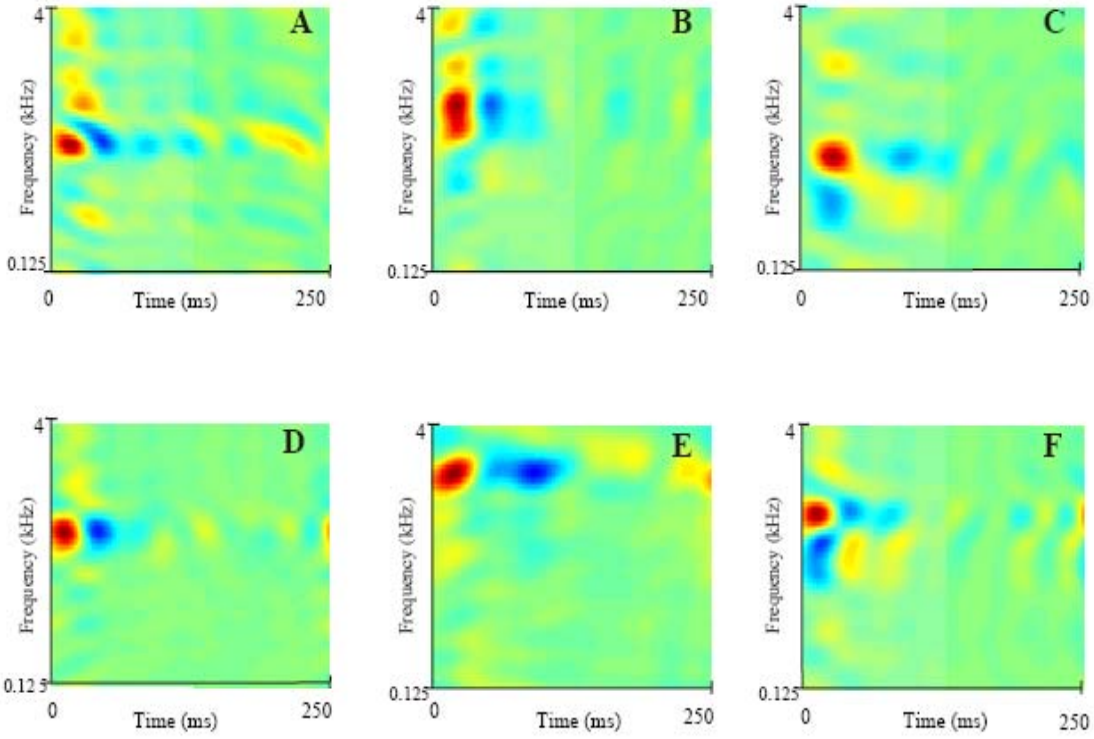


Ripple velocity (rate) : 4 Hz

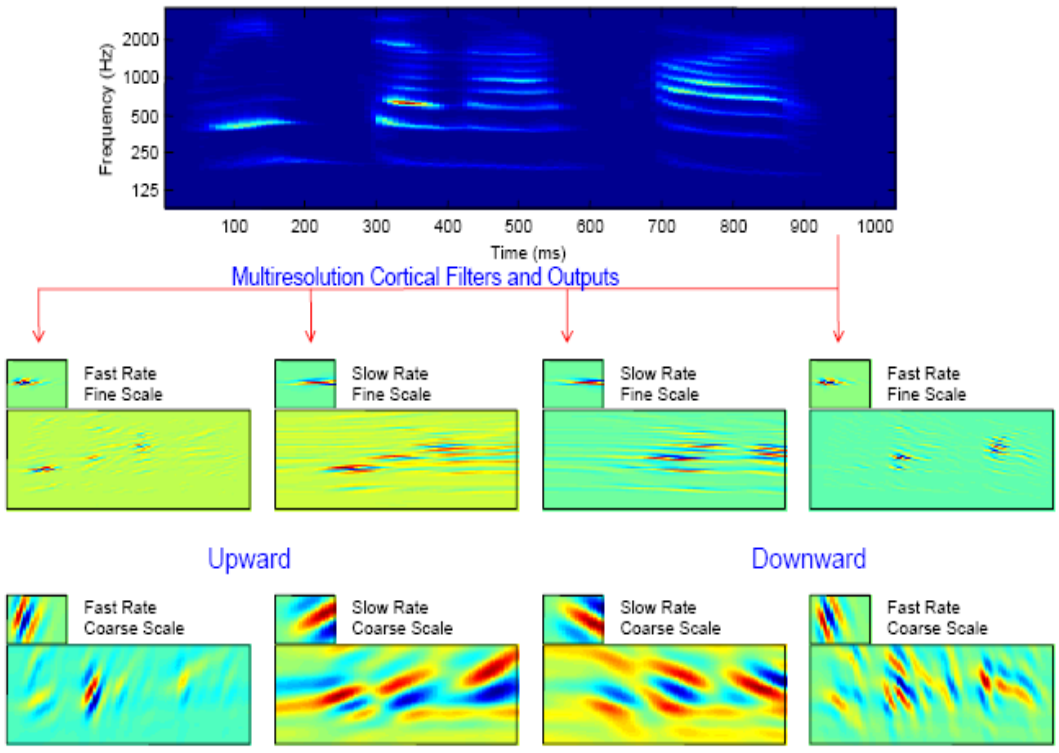
Ripple density (scale): 0.5 cyc/oct

圖 2-8： 移動波紋刺激源

從圖 2-8 可以看到時域和頻域上不同變化的移動波紋訊號。從橫軸時間軸上來看，每 250ms 發生一個週期的變化，因此時域上的變化是 4Hz，我們定義的名詞是 rate，單位是：Hz；從縱軸對數頻率軸上來看，每 2 個倍頻發生一個週期的變化，因此每個倍頻只包含 0.5 個週期變化，我們定義的名詞是 scale，單位是：cycle/octave，octave 表示一個倍頻。



圖：2-9：生物實驗上哺乳動物「貂」的大腦皮質聽覺區對於聲音的反應圖案



圖：2-10：我們的模型對於頻譜圖上不同 rate 和 scale 的變化做解析

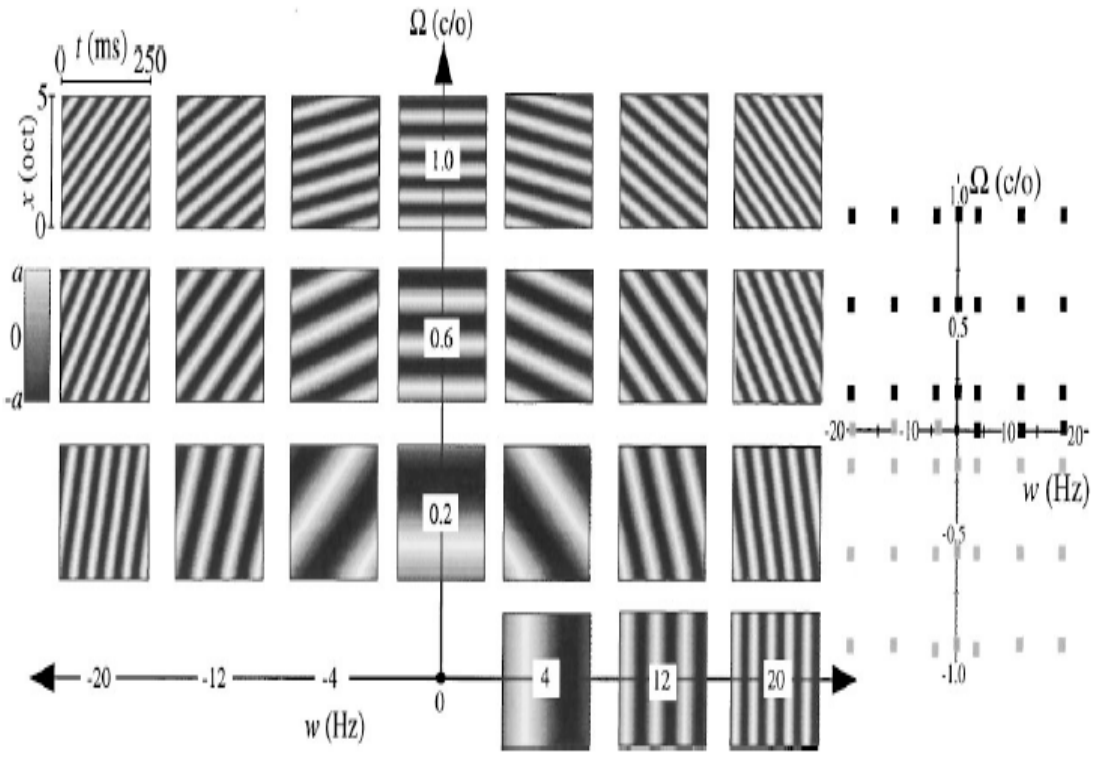


圖 2-11： 不同時域和頻域變化的水波紋，對應不同 rate 和 scale 的二維位置

資料來源：D.J. Klein, 2000 [23].

我們所使用的模型是假設不同的神經細胞對於不同 rate 和 scale 變化的頻譜圖成份有不同的二維脈衝響應，因此將頻譜圖經過一群二維的帶通調變濾波器(見圖 2-12)得到結合頻域和時域的分析結果。可以由 圖 2-11 清楚看到，不同頻域和時域變化的移動波紋對應至不同 rate 和 scale 的二維位置，每個移動波紋的時間總長度是 250ms，頻率變數以 x 表示，總共包含 5 個倍頻。

圖 2-11 是個六維的圖像，其中的五個維度分別是頻率(octave 或 frequency)、時間(time)、頻率變化(cycle/octave)、時域變化(Hz)以及能量大小(magnitude)。其中，Ω 表示頻率變化，我們稱之爲 scale；ω 表示時間變化，我們稱之爲 rate。另外，因爲大腦中的神經細胞對於聲音頻率的上升或下降也有選擇性，因此我們使用的感知聽覺模型對此定義了第六個維度，也就是正的 rate 和負的 rate；正的 rate 表示對頻率的下降有反應，負的 rate 表示對頻率的上升有反應，一般來說，大腦皮質聽覺區對正的 rate 反應較爲強烈。

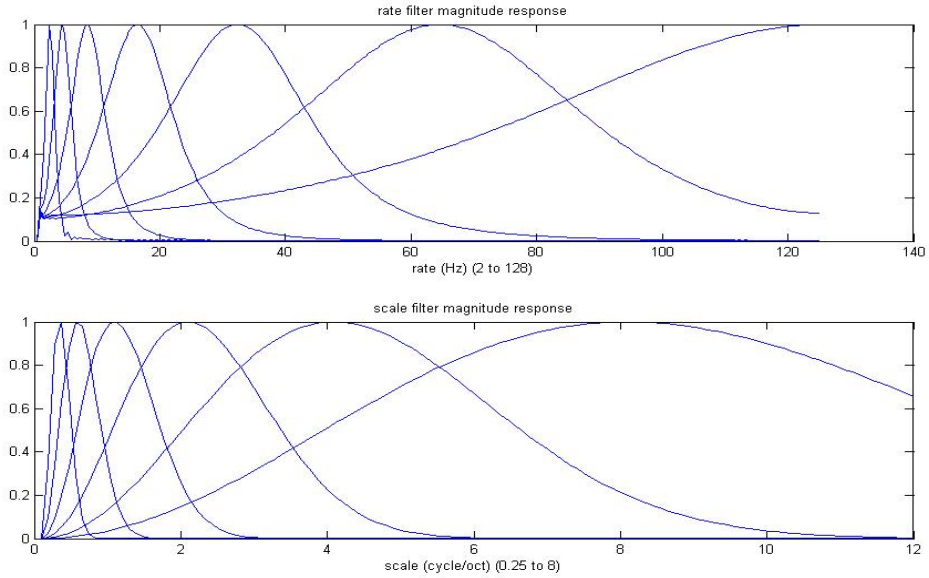


圖 2-12： 對不同 rate 和 scale 有選擇性的濾波器的頻率響應圖

由圖 2-13 清楚看到，頻譜圖(A)，經過大腦皮質聽覺區不同神經元的解析，得到六維的輸出圖像；對頻率和時間取平均後，得到二維的 Rate-Scale 圖像(B)。從(B)看到在 Rate=4、Scale=1~4 的反應特別強烈，我們可以說(A)包含比較強的這種頻域-時域變化的二維信號。

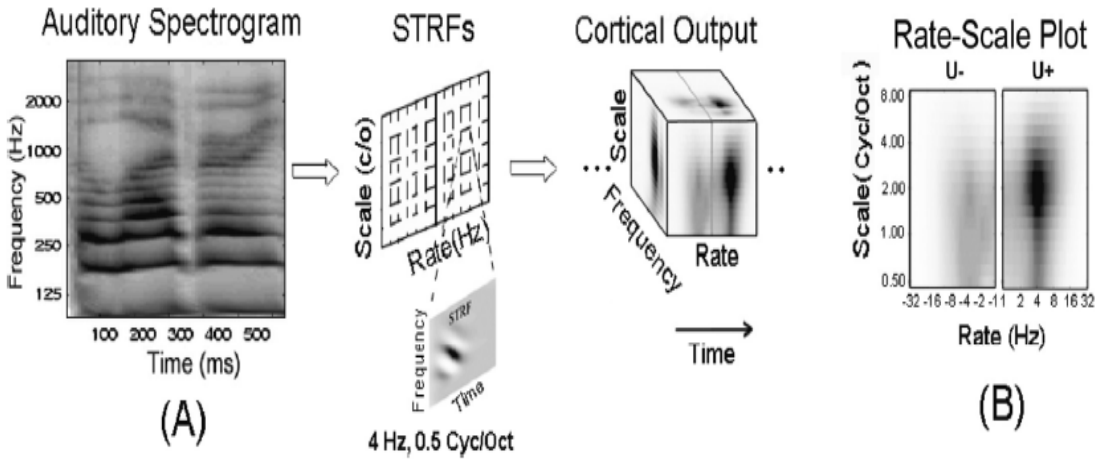


圖 2-13： 頻譜圖經過頻域-時域的解析後在 Rate 和 Scale 得到的二維圖像

資料來源：N. Mesgarani, 2006 [24]

第三章　語音資料庫介紹、主觀評量與客觀評量語音品質方法的國際標準

3. 1 語音資料庫介紹

3.1.1　TIMIT

TIMIT(TIMIT Acoustic-Phonetic Continuous Speech Corpus)語音資料庫內含630位以美式英語爲母語人士的語音資料，此語音資料由德州儀器(TI)和麻省理工學院(MIT)共同錄製，音訊的格式爲PCM，取樣速率爲16kHz，取樣值的量化程度爲16bits，男性人數佔了70%，女性人數佔了30%。

資料庫中包含訓練語料和測試語料。訓練語料是326位男性和136位女性共462個人一起錄製，每人各錄製10句，故共有4620句語音，總時間長度是3小時49分10秒；測試語料是112位男性和56位女性共168個人，每人各錄製10句，故共有1680句語音，總時間長度是1小時23分51秒。TIMIT資料庫內包含了八個方言區(Dialect Region)，每個方言區的男女生可參考表 3-1。

表 3-1：TIMIT語料庫中，不同方言區的男女生人數分佈狀況

TIMIT				
dr	Dialect Region	Male	Female	Total
dr1	New England	31	18	49 (8%)
dr2	Northern	71	31	102 (16%)
dr3	North Midland	79	23	102 (16%)
dr4	South Midland	69	31	100 (16%)
dr5	Southern	62	36	98 (16%)
dr6	New York City	30	16	46 (7%)
dr7	Western	74	26	100 (16%)
dr8	Army Brat	22	11	33 (5%)
	Total	438 (70%)	192 (30%)	630 (100%)

3.1.2 ITU-T Supp.23

這個編碼語音資料庫 [25] 是在 1998 年由 ITU-T 公佈並開始使用。語料庫包含三個實驗，第一個實驗是將乾淨語音經過各種不同無線傳輸標準(Codec)的處理；第二個實驗是考慮了環境雜訊(Background noise)因素，將乾淨語音加上各種不同雜訊，例如：辦公室吵雜聲、街道喧鬧聲、汽車噪音、高斯白雜訊等；第三個實驗考慮了傳輸通道造成訊號失真的效應，包含了語音的音框(frame)或位元(bit)可能隨機(random)或連續(burst)遺失的情形。

音訊的格式爲 PCM，取樣速率爲 16kHz，取樣值的量化程度爲 16bits，音框大小爲 10ms，每句語音的時間長度是 8 秒，位元傳輸率是 8kbps。每個實驗各有 4 位語者，實驗一有 44 種條件狀況，實驗二有 7 種狀況，實驗三有 50 種狀況，因此實驗一共有 176 句語音，實驗二共有 28 句語音，實驗三共有 200 句語音。每個實驗包含各種不同的語言，例如：法語、德語、美語、日語、挪威語、義大利語等。我們這次的研究使用的語料來自實驗一和實驗三的美語，評量其經過各種不同條件狀況後的語音品質。實驗一和實驗三包含的條件狀況請參考 表 3-2、3-3。

表 3-2：實驗一

Condition	1st codec	2nd Codec.	3rd Codec	dB Q
C1	G.729			
C2	G.729	G.729		
C3	G.729	G.729	G.729	
C4	G.726			
C5	G.726	x 4		
C6	G.728			
C7	G.711			
C8	GSM-FR			
C9	IS-54			
C10	JDC-HR			
C11	G.729	G.726		
C12	G.729	G.728		
C13	G.729	GSM-FR		
C14	G.729	IS-54		
C15	G.729	JDC-HR		

表 3-2：實驗一(續)

C16	G.726	G.729		
C17	G.728	G.729		
C18	GSM-FR	G.729		
C19	IS-54	G.729		
C20	JDC-HR	G.729		
C21	G.729	G.729	GSM-FR	
C22	G.729	G.729	IS-54	
C23	G.729	G.729	JDC-HR	
C24	G.729	G.726	GSM-FR	
C25	G.729	G.728	GSM-FR	
C26	GSM-FR	G.729	G.729	
C27	IS-54	G.729	G.729	
C28	JDC-HR	G.729	G.729	
C29	GSM-FR	G.726	G.729	
C30	GSM-FR	G.728	G.729	
C31	GSM-FR	IS-54		
C32	IS-54	JDC-HR		
C33	JDC-HR	GSM-FR		
C34	GSM-FR	G.729	IS-54	
C35	IS-54	G.729	JDC-HR	
C36	JDC-HR	G.729	GSM-FR	
C37	MNRU			5
C38	MNRU			10
C39	MNRU			15
C40	MNRU			20
C41	MNRU			25
C42	MNRU			30
C43	MNRU			35
C44	MNRU			50

表 3-3：實驗三

Cond No.	Codec	Trans-codings	Noise Type	Error Type	Error Rate (%)
1	G.729	1	Clean	–	–
2	G.729	1	Clean	Random Frame	3
3	G.729	1	Clean	Random Frame	5
4	G.729	1	Clean	Bursty Frame	3
5	G.729	1	Clean	Bursty Frame	5
6	G.729	1	Vehicle	–	–
7	G.729	1	Vehicle	Random Frame	3
8	G.729	1	Vehicle	Random Frame	5
9	G.729	1	Vehicle	Bursty Frame	3
10	G.729	1	Vehicle	Bursty Frame	5

表 3-3：實驗三(續)

11	G.729	1	Street	–	–
12	G.729	1	Street	Random Frame	3
13	G.729	1	Street	Random Frame	5
14	G.729	1	Street	Bursty Frame	3
15	G.729	1	Street	Bursty Frame	5
16	G.729	1	Hoth	–	–
17	G.729	1	Hoth	Random Frame	3
18	G.729	1	Hoth	Random Frame	5
19	G.729	1	Hoth	Bursty Frame	3
20	G.729	1	Hoth	Bursty Frame	5
21	G.729	2	Clean	–	–
22	G.729	3	Clean	–	–
23	G.729	2	Clean	Random Frame	3, 3
24	G.729	3	Clean	Random Frame	3, 0, 3
25	G.729	2	Clean	Bursty Frame	3, 3
26	G.729	3	Clean	Bursty Frame	3, 0, 3
27	G.729	2	Vehicle	Random Frame	3, 3
28	G.729	2	Vehicle	Bursty Frame	3, 3
29	G.729	1	Clean	Random Bit	1
30	G.729	1	Clean	Random Bit	3
31	G.729	1	Clean	Random Bit	5
32	G.729	1	Clean	Random Bit	10
33	G.729	1	Clean	Burst Frame/Random Bit	3, 1
34	G.729	1	Clean	Burst Frame/Random Bit	3, 3
35	G.729	1	Clean	Burst Frame/Random Bit	3, 5
36	G.729	1	Clean	Burst Frame/Random Bit	3, 10
37	G.726	1	Clean	–	–
38	G.726	1	Vehicle	–	–
39	G.726	1	Street	–	–
40	G.726	1	Hoth	–	–
41	MNRU	1	Clean	Q = 10dB	–
42	MNRU	1	Clean	Q = 15dB	–
43	MNRU	1	Clean	Q = 20 dB	–
44	MNRU	1	Clean	Q = 25 dB	–
45	MNRU	1	Clean	Q = 30 dB	–
46	MNRU	1	Clean	Q = 50 dB	–
47	Direct	–	Clean	–	–
48	Direct	–	Vehicle	–	–
49	Direct	–	Street	–	–
50	Direct	–	Hoth	–	–

資料來源：ITU-T, Supp. 23 to P series. [25]

3. 2 主觀評量語音品質的方法

主觀評量語音品質的方法有四種，分別是 A-B test、MOS、DMOS 和 Conversational test，關於此四種方法的優缺點可參考 表 3-4。我們這次的研究所要比較的主觀評量方法是 MOS(Mean Opinion Score)，由 ITU-T 在 1996 年所提出，是一種絕對分類標準(Absolute Category Rating)[26]，分類的標準可參考 表 3-5。

表 3-4：主觀評量語音品質的方法

Test	Advantages	Disadvantages
A-B test	• Simple • Low cost	• Only relative rating
MOS (ACR)	• Accurate • Well defined procedure	• Costly • Does not test effect of delay etc.
DMOS (DCR)	• Accurate • Well defined procedure	• Costly • Only relative scoring • Does not test effect of delay etc.
Conversational test	• Close to real situation • Tests effect of delay etc.	• Time consuming and costly • Needs full implementation

資料來源："Measuring Voice Quality, "GLOBAL IP SOUND, 2006

表 3-5：ITU-T P.800, Mean Opinion Score(MOS)

語音品質	分數
非常好	5
好	4
普通	3
尚可	2
很糟	1

3. 3 客觀評量方法的國際標準

ITU-T 在 2001 年 2 月和 2004 年 5 月分別提出兩套客觀式評量語音品質的方法，前者是侵入式的評量方法 P.862 [2]，後者是非侵入式的評量方法 P.563 [8]，兩者均適用於窄頻帶(Narrow-band)3.1kHz 以下的電話網路系統。

3.3.1 ITU-T P.862 (PESQ)

P.862 是侵入式的客觀評量語音品質方法，亦是大家所熟知的 PESQ(Perceptual evaluation of speech quality)，其適用範圍可參考：表 3-6。PESQ 的計算模型考慮相當多因素，例如：語音壓縮或經通道傳輸造成的失真、傳輸過程的時間延遲、通道的雜訊以及通道發生音框遺失或位元錯誤情形。

表 3-6：PESQ 評量語音品質所考慮的因素

Test factors	Coding/network technologies	Measurement applications
Coding distortions Transmission/packet loss errors Multiple transcodings Environmental noise * Time warping (variable delay)	Waveform codecs (e.g. G.711, G.726, G.727) CELP/hybrid codecs at 4kbit/s and above (e.g. G.728, G.729, G.723.1) Mobile codecs and systems (e.g. GSM FR, EFR, HR, AMR; CDMA EVRC, TDMA ACELP, VSELP; TETRA)	Live network testing Network planning Codec evaluation/selection Equipment selection Codec/equipment optimisation

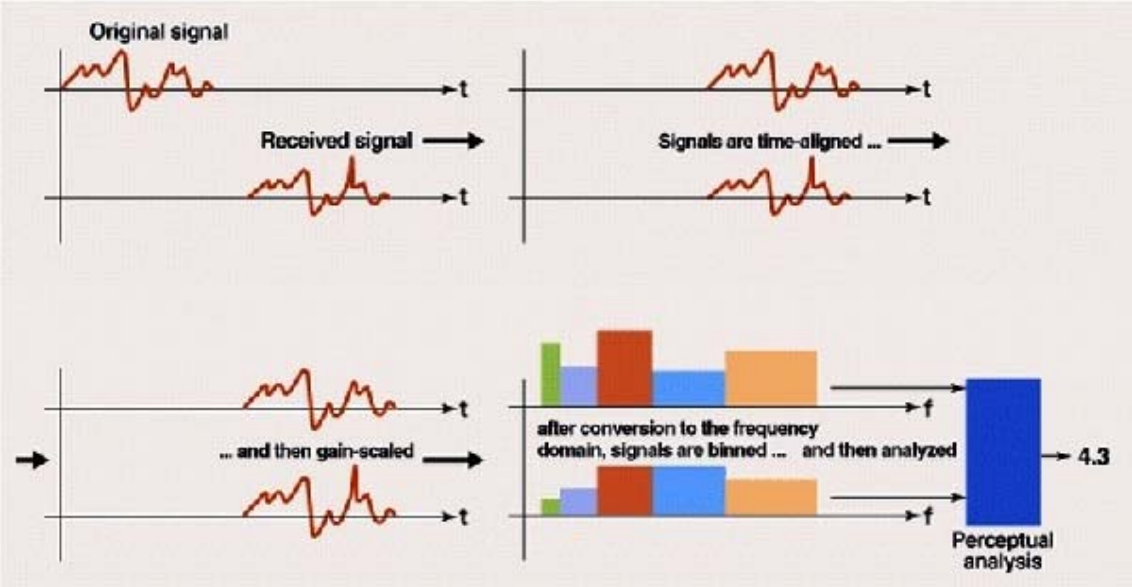
資料來源：J. G. Beerends, "Psychoacoustic model", 1998

PESQ 尚有許多和語音品質好壞有關但尚沒有考慮進去的因素，例如：響度的損失(loudness loss)、回音(echo)、側音(sidetone)、串音(crosstalk)、VAD 的影響等。

而且 PESQ 只能處理窄頻帶的電話網路通訊以及單方向(listening only)語音品質，對於寬頻(wide-band)以及雙方向(conversational)的語音品質或甚至是音樂(audio)品質仍無法得到準確可靠的評量結果。

PESQ 是侵入式評量方法，需要來源端的乾淨語音和接收端經過損傷的語音做比較以得到人耳感知聽覺上的語音品質分數；PESQ 模型主要的參作程序有：時間對位、振幅大小校準、頻率的壓縮(Bark scale)以及響度的壓縮，最後在頻譜上做分析比對得到預測的語音品質分數。PESQ 的基本概念可參考：圖 3-1。

圖 3-1：PESQ 的基本概念與方法



資料來源：http://www.eedesign.com.tw,作者 P. Denisowski

由圖 3-1 可看到，右上角圖是將兩個訊號做時間上的對位，左下角圖是振幅大小的調整，最後將時域上信號轉換至頻域上信號，再依據人耳聽覺與頻率之間的非線性相關，使用 Bark scale 在頻帶上做不同的解析，由右下角圖可看到低頻端的頻帶頻寬較窄而高頻帶的頻寬較寬，最後在頻譜上做感知分析得預測的 MOS。

PESQ 可能遇到的問題在於：電話網路傳輸的延遲時間相當不固定，不同封包的時間延遲長度不同導致時間扭曲(time-warping)，造成接收訊號和原來傳送訊

號長度不同，因此在做時間對位時，必須藉由調整兩個訊號間的交叉相關性(cross-correlation)達到最大，來校正兩個訊號間的時間差；另外，在振幅校準時也不易處理，因爲接收端的訊號經過的哪種損傷和衰減實際上並不清楚，如果只是藉由放大訊號來調整訊號的振幅或整句語音的功率，可能存在許多問題。

3.3.2 ITU-T P.563

P.563 是非侵入式的客觀評量語音品質方法 [8]，只使用接收端經過損傷的語音訊號來預測語音品質分數而不需要傳送端的原始乾淨訊號；P.563 演算法的模型是建構在人的口腔發聲系統和人耳聽覺的感知系統的特性。此演算法的目的在於希望能預測窄帶(3.4kHz 以下)語音訊號的主觀品質，這些語音訊號經過電話網路傳輸後可能伴隨很多損傷，例如：背景雜訊、網路元件的頻率響應(filtering)、不同的時間延遲以及因爲通道傳輸發生錯誤或語音壓縮編解碼所造成的失真。

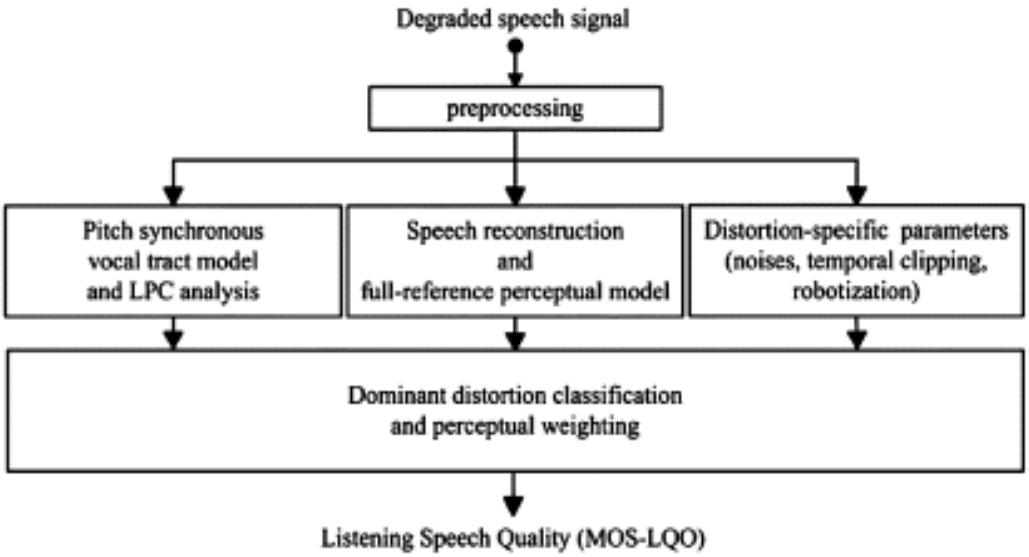


圖 3-2：P.563 演算法架構完整描述

資料來源：ITU-T Rec., P.563 [8]

因爲非侵入式方法只有接收端的訊號可作爲評分依據，因此對於接收端訊號必須先做一些假設。P.563 演算法分成三個階段循序進行，分別是預先處理階段(preprocessing)、失真評估階段(distortion estimation)以及感知映照階段(perceptual mapping)。因爲沒有單一的方法能夠將所有失真問題一起處理，P.563 綜合三個基本原則來評估所有失真造成的效應，這三個基本原則分別是：從口腔管(vocal tract)建立發聲系統模型、從接收到的損傷語音重建一個乾淨的參考語音以及確認並評估特定失真(例如：temporal clipping、robotizationa、noise)的影響；因爲當很多損傷或失真情況同時發生時，人類聽者會專注在最主要的(dominant)失真情況上，因此 P.563 最後使用失真相依的權重(distortion-dependent weighting)將所有失真參數一起考慮得到預測的語音品質分數。整個 P.563 的演算架構可參考 圖 3-2。

P.862.1 [27]是 ITU-T 在 2003 年提出，將原始 PESQ 的分數透過一個函數轉換到更接近 MOS 的方法。表 3-7 列出 P.563 和 P.862.1 對於語音資料庫 ITU-T Supp.23 語音品質評量分數的效能優劣比較，由 表 3-7 看到，P.563 的評量結果和 MOS 的平均相關性約 0.89，而 P.862.1 的評量結果和 MOS 的平均相關性約 0.95；P.563 的結果比較不準確是可以預期的，因爲它並沒有來源訊號當作評分參考。

表 3-7：P.563 和 P.862.1 使用 Supp.23 database 的評量結果

8 kbps ITU & ETSI standard codecs interworking	**P.563**	**P.862.1**
French	0.885	0.947
Japanese	0.842	0.957
American English	0.902	0.968
Channel Errors and background noise		
French	0.886	0.904
Italian	0.854	0.964
Japanese	0.929	0.943
American English	0.916	0.934
Average	**0.888**	**0.945**

資料來源：ITU-T Rec., P.563 [8]

P.563 適用範圍相當廣泛，可參考 表 3-8，但其仍有不足的地方。首先，因爲 P.563 在評量時，會將語音訊號分爲低基頻(low-pitched)和高基頻(high-pitched)兩類，再使用不同的參數和權重對兩種不同類的語音做評分；因此，倘若要評分的語音音高變化非常大，則 P.563 的評分結果就不可避免地相當糟糕。其次，因爲 P.563 的評分機制裡沒有語義分析(semantic analysis)，倘若語音中有一個字(word)全部遺失，P.563 的分數不會受到任何影響，但實際主觀的 MOS 卻會相當低。最後，P.563 是設計來評量窄頻帶(3.4kHZ)、取樣速率 8kHz 的語音，對於現在或將來更寬頻的語音傳輸更普遍的應用，原有 P.563 就顯得不足。

表 3-8：P.563 的相關應用範圍

Test factors
Characteristics of the acoustical environment (reflections, different reverberation times)
Environmental noise at the sending side
Characteristics of the acoustical interface of the sending terminal
Remaining electrical and encoding characteristics of the sending terminal
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment with CELP codecs
Bit rates if a codec has more than one bit-rate mode
Transcodings
Effect of varying delay on listening quality in ACR tests
Short and long term time warping of speech signal
Transmission systems including echo cancellers and noise reduction systems under single talk conditions and as they will be scored on an ACR scale
Coding technologies
Waveform codecs, e.g., G.711; G.726; G.727
CELP and hybrid codecs ≥4 kbit/s, e.g., G.728, G.729, G.723.1
Other codecs: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP, TDMA-VSELP, TETRA
Recommended application scenarios
Live network monitoring using digital or analogue connection to network
Live network end-to-end testing using digital or analogue connection to the network
Live network end-to-end testing with unknown speech sources at the far end side

資料來源：ITU-T Rec., P.563 [8]

第四章　客觀的侵入式語音品質評量方法

4.1 背景知識

我們客觀的侵入式語音評量方法，是從人的發聲系統(articulatory system)和聽覺系統(auditory system)特性，找出並量化其和語音品質好壞的關聯。在前人的研究實驗裡 [12]，我們可以得到一些線索和知識，語音信號的時間變動包跡(temporal envelope)和人耳聽覺上的感知特性有關。從時間變動包跡的頻譜圖上，在低頻部分，我們可以看到語音信號的口腔變動頻率成分大小，在高頻部份，也可以看到因聲帶震動造成的基頻和泛頻的成分大小，見圖 4-1。

時間變動包跡經複立葉分析後的頻率稱之爲調變頻率(modulation frequency)，從前人的研究結果 [28] 知道，人耳對時間變動包跡的頻率響應可看作一個低通濾波器，截止頻率大約是 50Hz，這表示人耳的聽覺靈敏度對於語音信號在較高調變頻率成分的變化已是相當遲鈍。另外，由於人發聲系統的機械是震動速度大約限制在 2Hz 到 30Hz，因此在 [12] 中假設，倘若時間變動包跡的調變頻率成分大小在 30Hz 到 50Hz 間比一般正常的乾淨語音增加許多，則這些額外增加的頻率成分實際上並非人聲，而且可能影響語音品質好壞，所以，我們可以藉由這個調變頻率範圍的能量變化當作一個特徵線索，來評量語音品質好壞。

圖 4-1 的(a)是一段語音，取出其中的 250ms 語音做時間變動包跡分析得到(b)，在(b)中可看到在這段時間裡有個母音，基頻大約 100Hz。(c)是(b)的頻譜圖，圖(c)中在比較低頻的部份，口腔振動頻率成份在 2~8Hz 相當高。而在比較高頻的部份，我們可以看到基頻和泛頻週期性出現的成份，100Hz、200Hz、...等，圖(c)橫軸的調變頻率用對數方式呈現。圖 4-2 是兩句乾淨的女聲信號以及另外加上不同 SNR 的 MNRU(modulated noise reference unit)得到時間變動包跡的頻譜圖，由圖中清楚看

到，乾淨人聲在調變頻率超過 30Hz 以後成份大大降低，但經過 MNRU 處理的人聲，在超過 30Hz 以後的頻率成分隨著 SNR 的降低而增加，因此藉著計算調變頻率在 30Hz 到 50Hz 間的能量變化當作評量語音品質的指標。

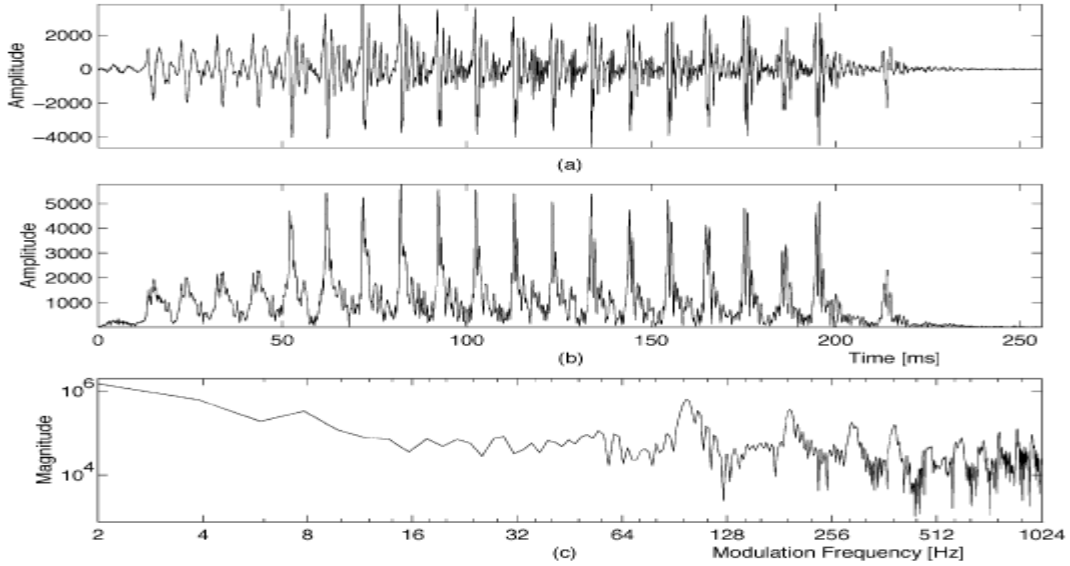


圖 4-1：(a)語音信號，(b)時間變動包跡，(c)時間變動包跡的頻譜圖

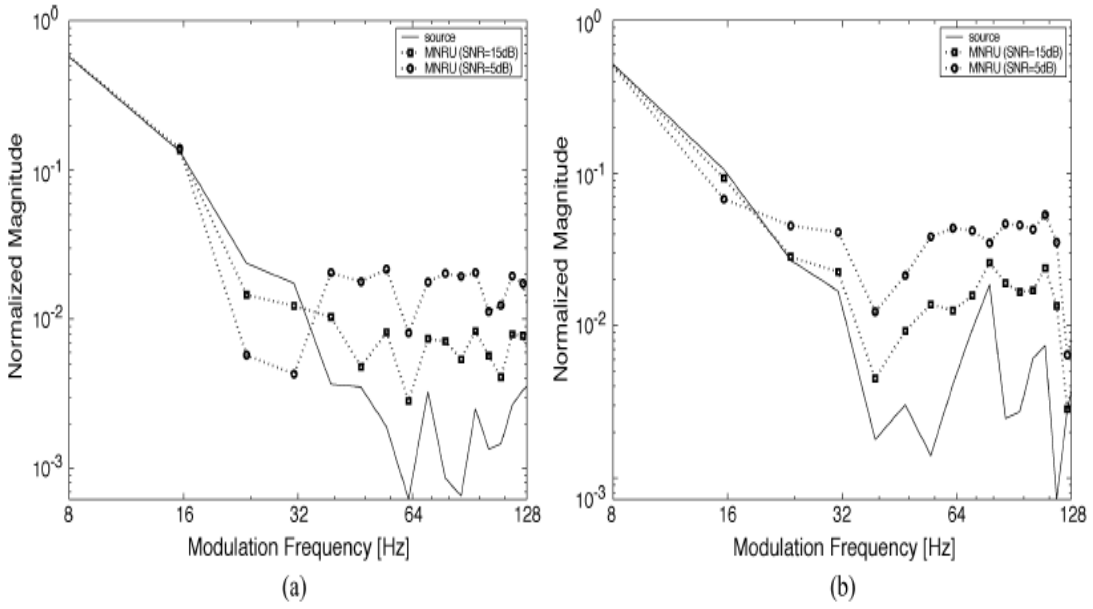


圖 4-2：乾淨語音經過不同 SNR 的 MNRU 處理後的時間變動包跡頻譜圖

資料來源：D.-S. Kim, 2004 [12]

4.2 研究方法

我們的客觀侵入式語音評量方法，是先觀察並了解乾淨的人聲在我們使用的人耳聽覺模型上的圖像(pattern)，再從不同維度去分析乾淨的人聲，最後界定出人聲在各個維度上出現的範圍。藉由觀察語音信號在大腦皮質聽覺區(cortical domain)的圖像，除了能看出男聲和女聲在不同維度有各自不同的響應大小，也能區分出雜訊可能出現的範圍以及乾淨語音經過編碼解碼(codec)後的圖像差異。

從圖 4-3 到圖 4-6 可看到，男聲和女聲在 Rate 維度大概在 2 到 16Hz 能量最強烈，反應了口腔振動頻率範圍，男女聲相差不多；但是在 Scale 維度可看出男女聲相當不同，反應了男生和女生不同的聲帶振動頻率造成泛頻間隔的密度大小。從圖 4-3 到圖 4-4 可看到，女聲在 scale 維度的能量在 Scale 是 2 到 4(cyc/oct)地方最爲強烈，而男聲在 Scale 維度的能量強烈部分在 Scale 是 0.5 到 4(cyc/oct)的地方都有，從圖 4-5 可看到女聲的圖像在 Scale 是 2 到 4(cyc/oct)顏色最深，從圖 4-6 可看到男聲的圖像在 Scale 是 0.5 到 4(cyc/oct)顏色比較深但均勻；從圖 4-7 到圖 4-8 更可清楚分辨男聲和女聲在 Scale 維度上的差異，男生因爲基頻較低造成泛頻間隔較密，因此在 Scale 維度比較高的地方(8 cyc/oct)也有能量分佈。

一般關心雜訊和干擾時會先從高斯白雜訊(white noise)討論起，因此我們想先了解白雜訊在我們所使用的模型在不同的維度上有怎樣的圖像。圖 4-9 到圖 4-10 我們可看到，在 Rate 維度上，白雜訊在低 Rate(16Hz 以下)能量很弱，在 Rate 超過 16Hz 以上，能量顏色慢慢加深，在愈高頻時出現能量分佈愈強。在 Scale 維度上，白雜訊在低 Scale(1 cyc/oct 以下)能量顏色很淡，在 Scale 是 2、4、8(cyc/oct)時才出現很強能量。從 Rate 和 Scale 兩個維度的能量分佈我們可以了解，白雜訊的能量集中在 Rate 和 Scale 比較高的區域，而這個區域對語音品質好壞有相當的影響，因爲這區域影響了語音信號的子音和基頻。

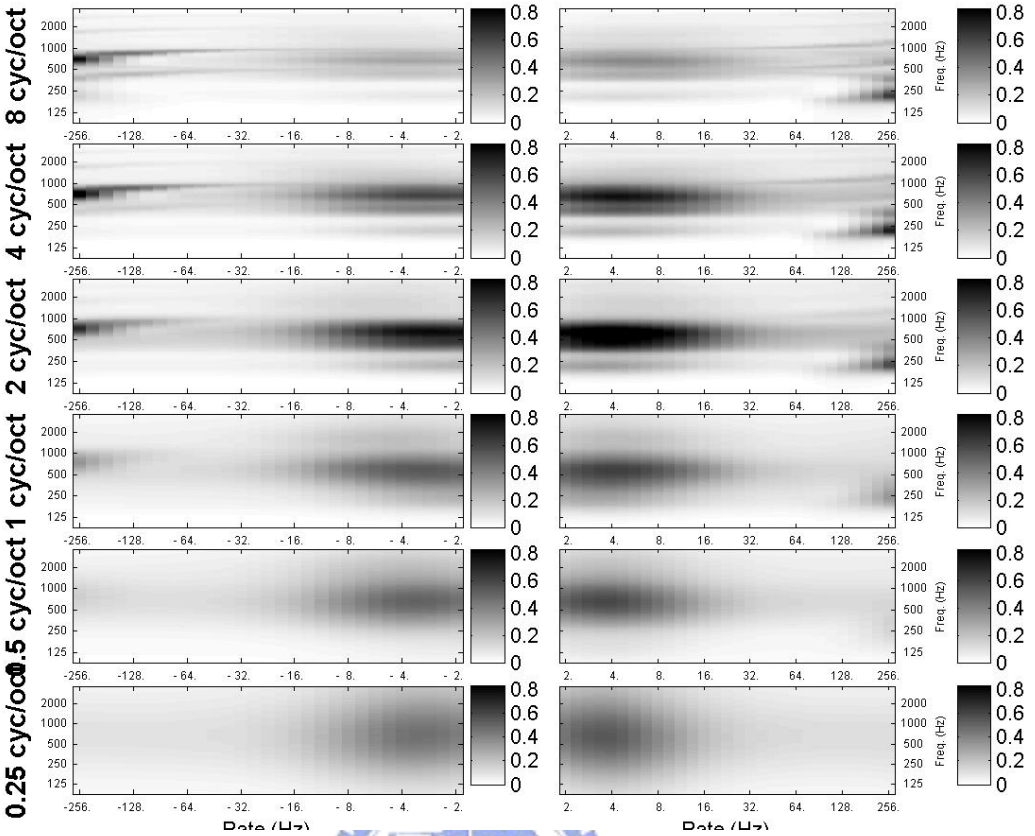


圖 4-3：長時段平均女聲在 Rate-Scale-Frequency domain 上的圖像

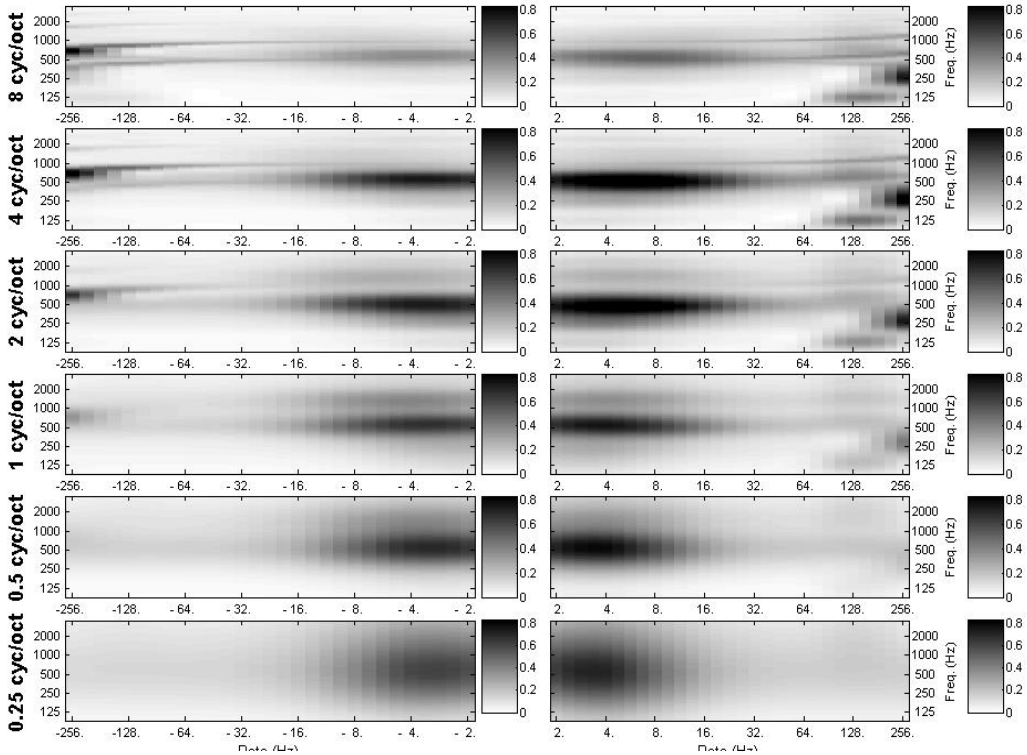


圖 4-4：長時段平均男聲在 Rate-Scale-Frequency domain 上的圖像

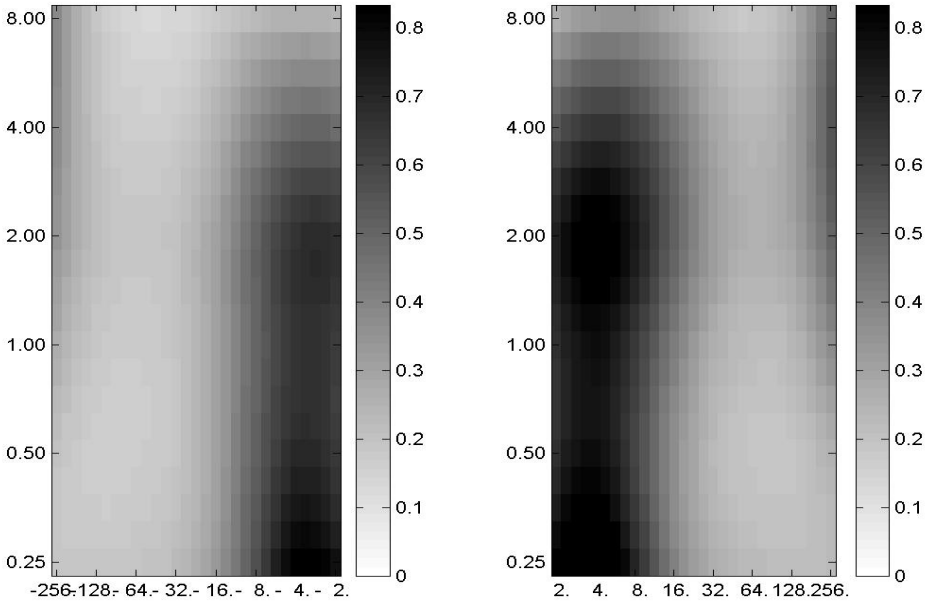


圖 4-5：長時段平均女聲在 Rate-Scale domain 上的圖像

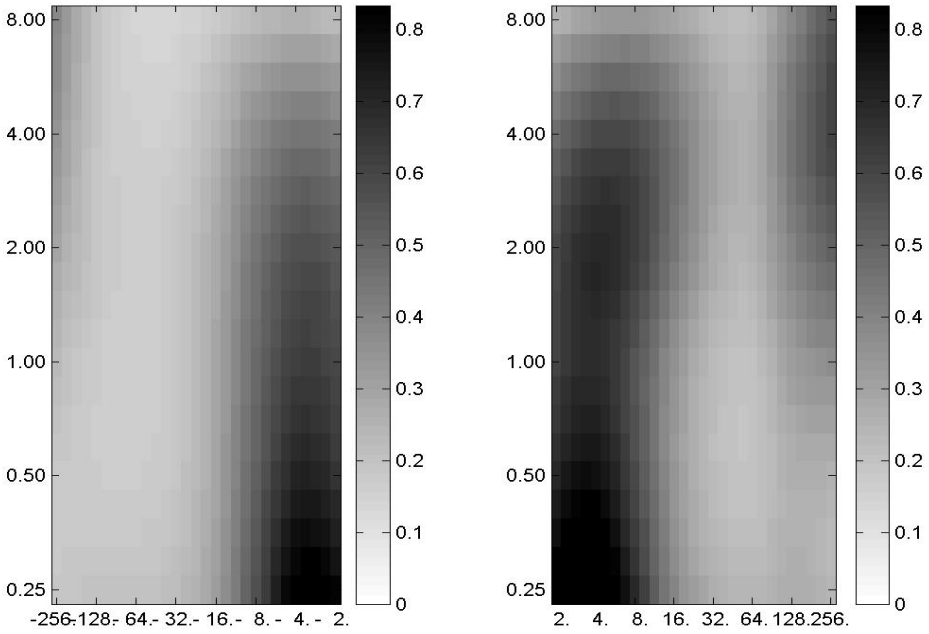


圖 4-6：長時段平均男聲在 Rate-Scale domain 上的圖像

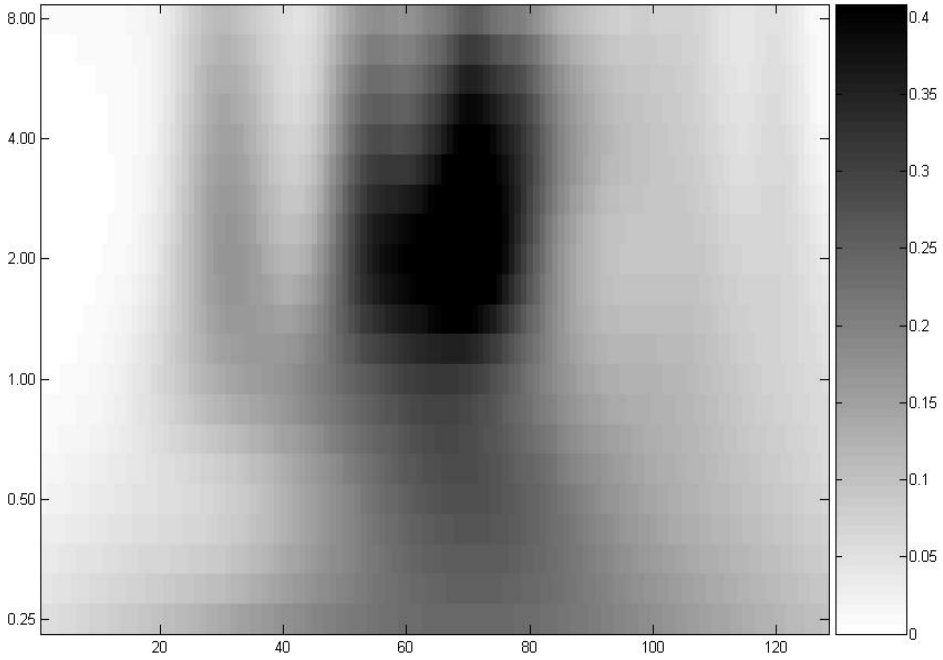


圖 4-7：長時段平均女聲在 Freq-Scale domain 上的圖像

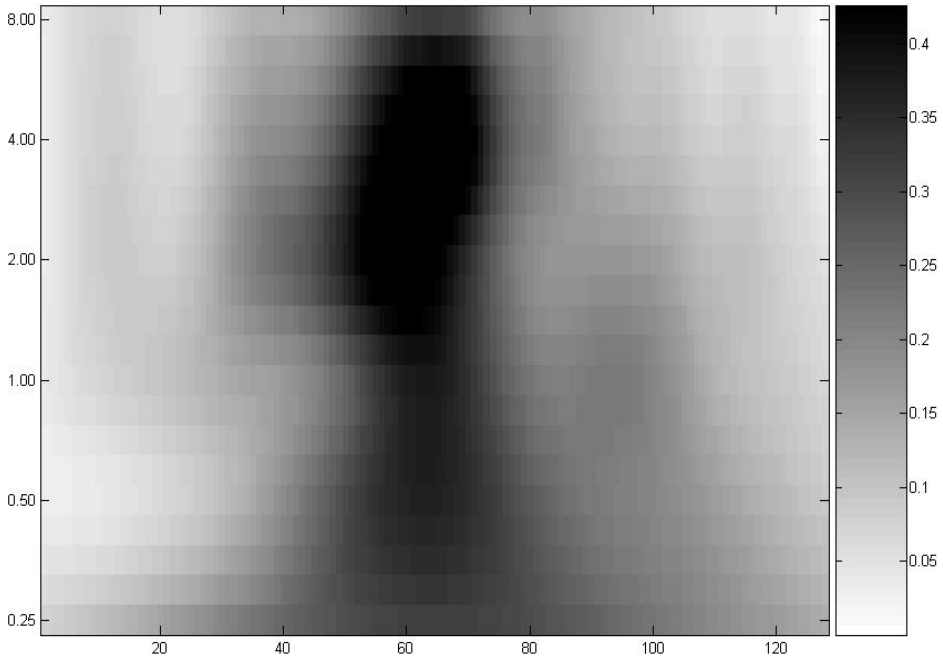


圖 4-8：長時段平均男聲在 Freq-Scale domain 上的圖像

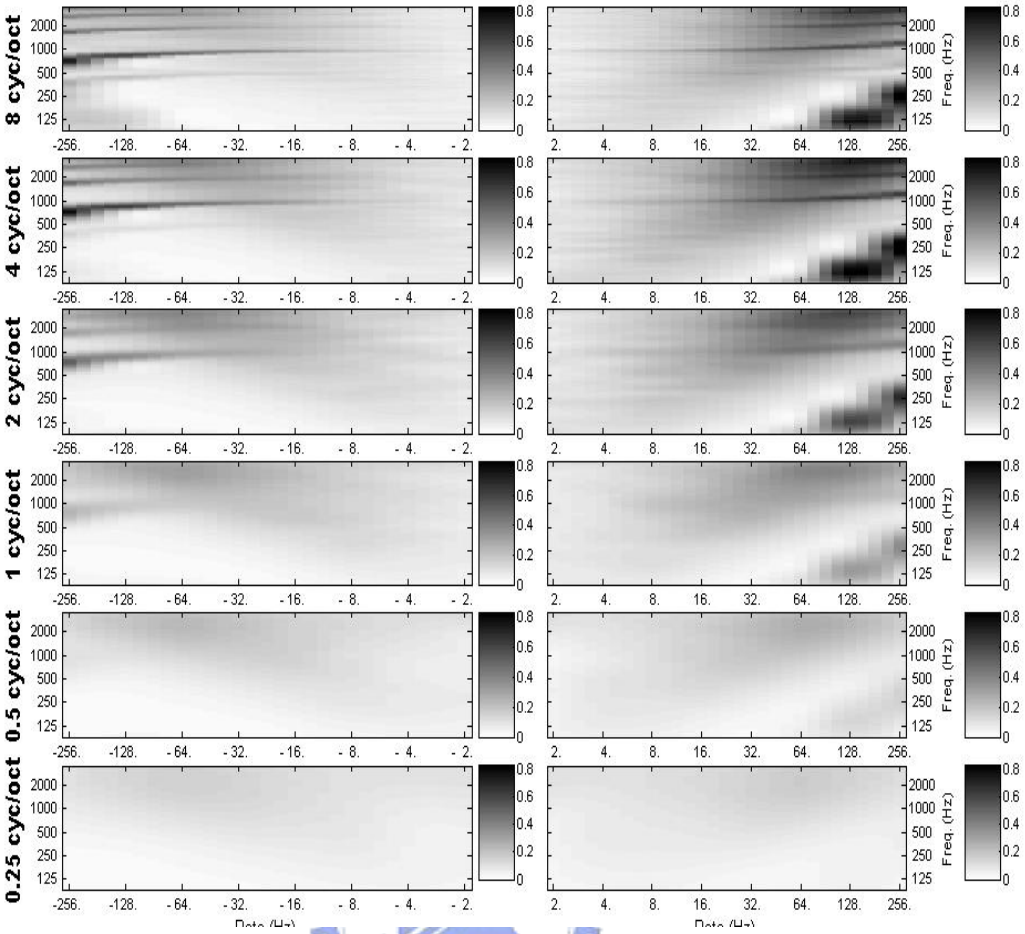


圖 4-9：white noise(SNR=5dB)在 Rate-Scale-Frequency domain 上的圖像

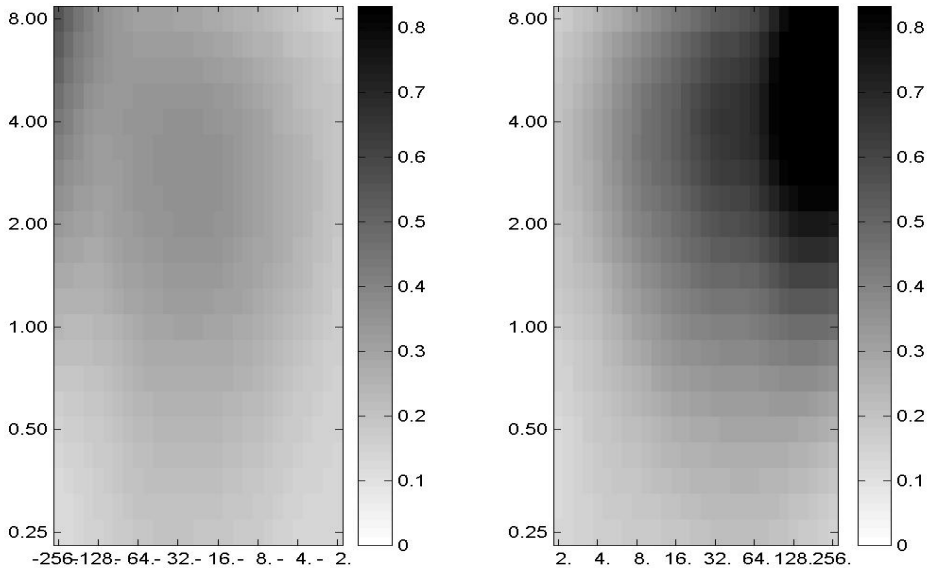


圖 4-10：white noise(SNR=5dB)在 Rate-Scale domain 上的圖像

當我們了解乾淨的人聲以及高斯白雜訊在感知聽覺模型上的圖像後，對我們評量語音品質好壞有很大的幫助，因爲我們可以很清楚的在模型的圖像上，看出哪些是語音而哪些是干擾或雜訊。更進一步的，我們想要探討真實的電話網路通信系統上，乾淨的語音經過不同的語音壓縮編解碼機制後，在感知聽覺模型上的圖像會產生何種變化，而這些變化是否會造成語音品質的下降。

因此我們先觀察一句乾淨的女生語音，從圖 4-11 看到，該女聲在 Rate 小於 16Hz 以下以及 Scale 小於 4cyc/oct 有很強能量，而在比較高 Rate 地方，可看出該女聲基頻大小在 250Hz 附近相當強烈，能量的顏色很深。

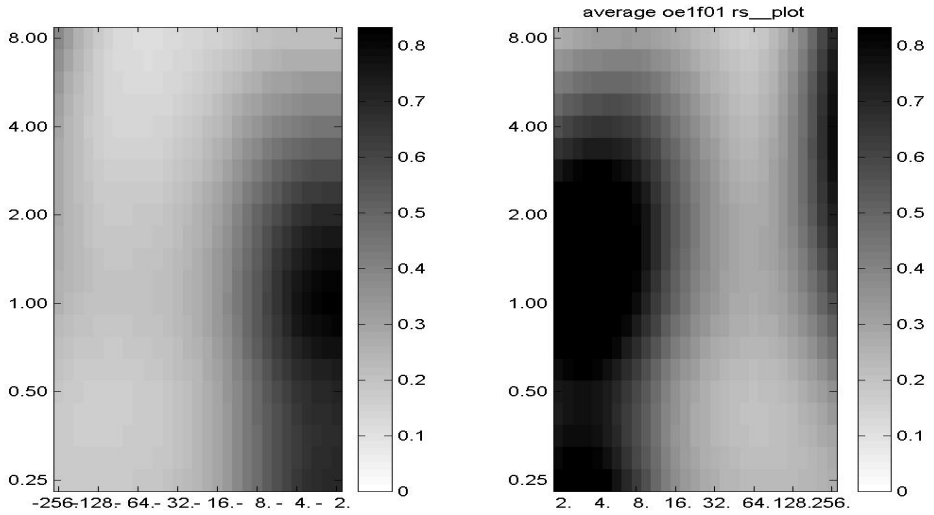


圖 4-11：乾淨女生語音在 Rate-Scale domain 上的圖像

接著我們觀察這句乾淨的女生語音分別經過語音壓縮處理以及加入 MNRU(Q=5dB)後在聽覺模型上的圖像呈現。從圖 4-12 看到，經過 codec 後的語音，在原本人聲的能量範圍(Rate 小於 32 以及 Scale 小於 4)的地方顏色變淡且分散了，而反應人聲基頻能量範圍(Rate 在 128 到 256 之間)的顏色變的相當淡，和原本乾淨語音在基頻很強能量相差很懸殊，我們可以說原本語音的基頻被破壞的相當嚴重，也因而導致語音品質下降。

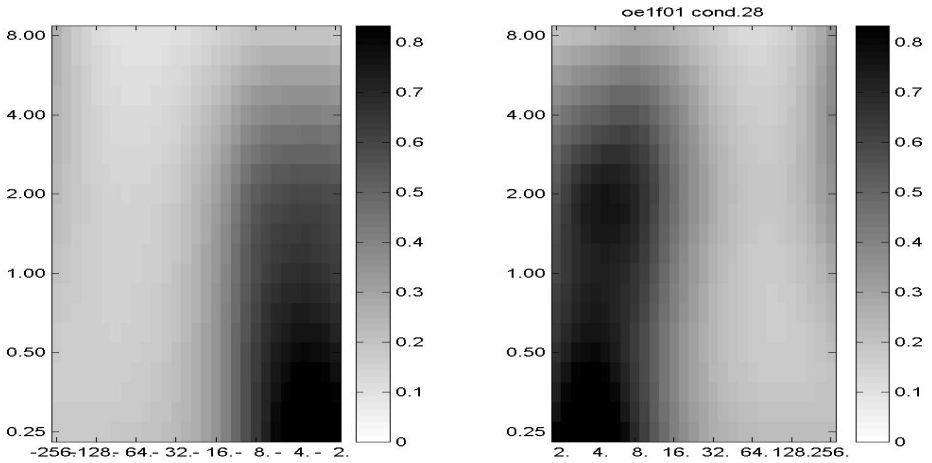


圖 4-12：乾淨女生語音經過 codec 後在 Rate-Scale domain 上的圖像

從圖 4-13 看到，加上 MNRU 的語音和原本乾淨的女聲相比，在原本人聲口腔頻率範圍(Rate 小於 32 和 Scale 小於 4)以及人聲基頻能量範圍(Rate 在 128 到 256 之間)的能量似乎都被保留下來，但我們發現，在有一大塊區域，也就是 Rate 在 16 到 128 之間以及 Scale 在 2 到 8 之間，出現相當多非人聲的能量，這塊區域的能量會隨著 Q 值得下降而增加，這個觀察和前人的研究結果 [12] 圖 4-2 相當一致，因此我們可以將這個區域的能量大小當作評量語音品質好壞的指標。

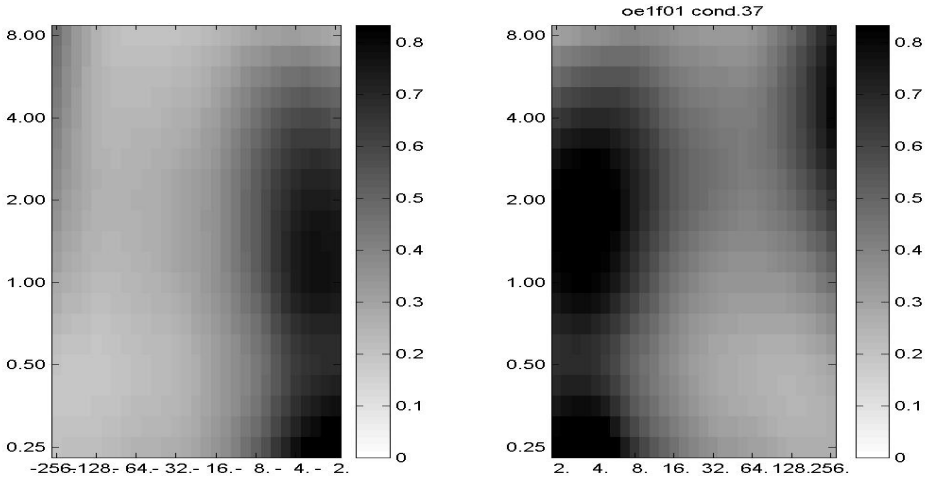


圖 4-13：乾淨女生語音經過 MNRU(Q=5dB)後在 Rate-Scale domain 上的圖像

4. 3 研究結果

從前面兩個章節的背景知識和觀察結果我們知道，經過各種損傷的語音，我們可以藉由觀察它們在感知聽覺模型上的圖像來判斷和評量語音品質好壞。而且，藉由感知聽覺模型，我們可以在 Rate 和 Scale 兩個維度上找到一個能量分佈區域，即 Rate 在 32Hz 到 256Hz 以及 Scale 在 2cyc/oct 到 8cyc/oct 之間。因此，我們藉由計算不同時間點，乾淨語音和損傷語音在此特殊區域的能量變化，來評量語音品質好壞，然後和 MOS 對照(mapping)並將結果和 PESQ 做比較。

圖 4-14 中的每一資料點是 ITU-T Supp.23 實驗一同一句女生語音經過各種不同 codec 處理後，我們計算其在特殊區域的能量變化數值，並將此能量數值和其 MOS 的關係畫出來，嘗試不同階次的迴歸分析得到預測的 MOS。這句女生語音經 42 種不同 codec 處理後，我們預測的 MOS 和真實 MOS 有 88%以上的準確度。

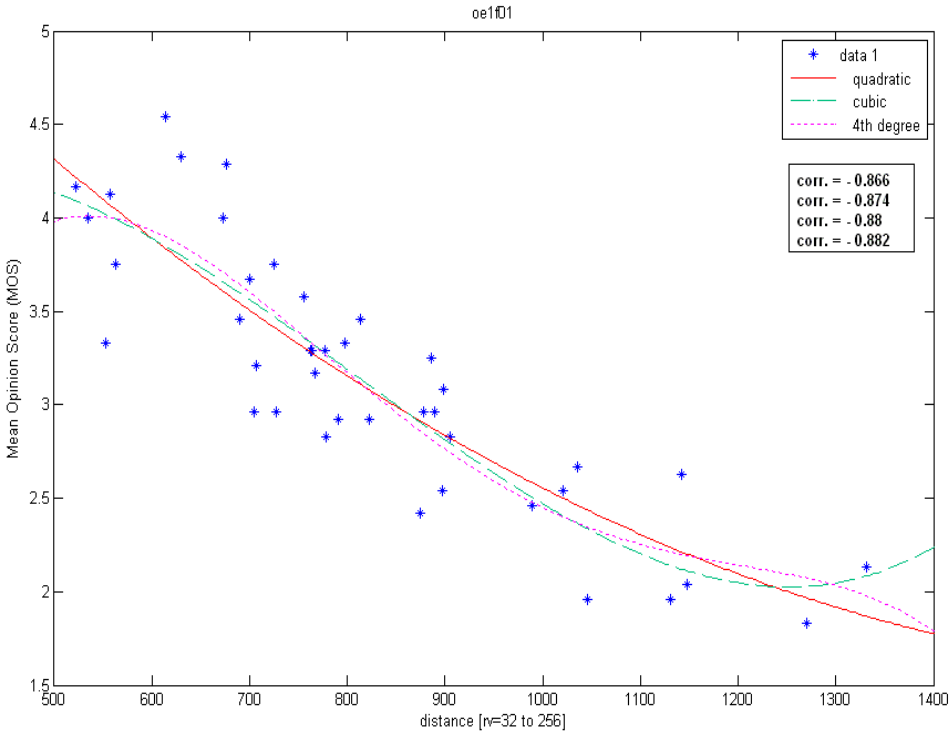


圖 4-14：特殊區域的能量變化經不同階次迴歸後和 MOS 的相關程度

圖 4-15 中的每一資料點是 ITU-T Supp.23 實驗一同一句男生語音經過各種不同 codec 處理後，我們計算其在特殊區域的能量變化數值，並將此能量數值和其 MOS 的關係畫出來，嘗試不同階次的迴歸分析得到預測的 MOS。這句男生語音經 42 種不同 codec 處理後，我們預測的 MOS 和真實 MOS 有 92%以上的準確度。

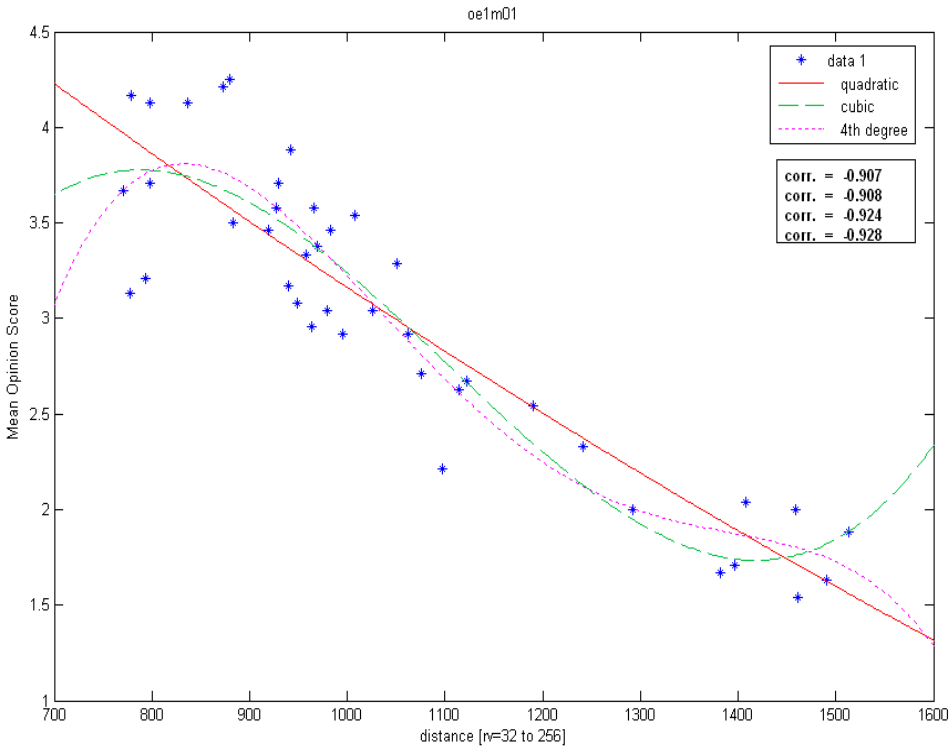


圖 4-15：特殊區域的能量變化經不同階次迴歸後和 MOS 的相關程度

最後，我們將 ITU-T Supp.23 實驗一的四位語者(兩男兩女)經過 42 種 codec 處理的語音都做了相同分析，並將分析結果和 ITU-T P.862(PESQ)的結果做比較。從表 4-1 看到，對於 ITU-T Supp.23 實驗一的四位語者，我們預測的 MOS(pre_MOS)和真實 MOS 最高的相關係數可達 0.928，最差是 0.795，平均是 0.868；而 PESQ 最高是 0.923，最差是 0.866，平均是 0.902。若將我們的預測結果和 PESQ 的預測結果相除做比較，平均可高達 96%的相關程度，由此可看出我們客觀的侵入式預測 MOS 方法的效能和 PESQ 相當。

Supp.23	pre_MOS	P.862 (PESQ)	pre_MOS/PESQ (%)
male 1	0.928	0.916	101
male 2	0.795	0.866	92
female 1	0.882	0.904	98
female 2	0.868	0.923	94
Average	0.868	0.902	96

表 4-1：我們預測的 MOS 以及 PESQ 的分數，兩者和 MOS 的相關程度

第五章　客觀的非侵入式語音品質評量方法

我們的客觀非侵入式語音品質評量方法是在兩個聽覺感知階段做觀察和分析，希望從人耳低階感知反應上擷取出可能影響聽者在高階認知判斷語音品質的特徵參數來對語音品質做客觀評量，這三個特徵分別是－理解性、自然性、基頻失真。最後，我們使用最小平方法將三個特徵參數和語音品質的關係做結合，希望藉由這三個基本的特徵參數能讓我們對語音品質的好壞做快速並可靠的評量，並將評量結果和 ITU-T P.563 做比較。

5. 1 理解性

5.1.1　背景知識

聽者對於語音品質好壞與否是由很多抽象維度組成的認知(cognition)來決定，理解性(intelligibility)是語音品質(quality)的其中一樣重要特性(attribute)，通常對於一句語音，我們會先聽看看，看能不能聽的懂這個語者在說什麼(what)，是否能了解這句語音所要傳達的意思或資訊，而能聽懂多少的，我們稱之爲理解性。當我們聽懂語者的話之後，我們才會再去聽看看這句話是否好聽(how)，而好不好聽必須再由其它抽象維度來確定(identify)並量化(quantify)，比如這句話聽起來自不自然(naturalness)、清不清晰(clarity)等。因此，評量語音品質好壞的首要程序，就是要先了解並量化語音的理解性和語音品質間的關係。

計算理解性高低的研究從 1969 年就開始，但仍有相當多的問題尚待克服，例如：語音信號經過傳輸通道後發生相位抖動(phase jitter)或相位平移(phse shift)等問

題，由 T.-S. Chi 等人 在 1999 年 [16] 提出的一個結合頻率和時間變化的調變轉移函數(spectro-temporal modulation transfer function)，使用感知聽覺模型來分析語音信號，並將其應用在語音理解性高低的評量。在 [16] 的研究裡，使用 TIMIT 語料庫中的 380 句話，其中包含 240 句男生和 140 句女聲，先觀察男聲和女聲在 Rate-Scale-Time 上一段長時間的平均(long-term average)，看人聲的語音信號在 Rate-Scale 上的圖像呈現如何；研究結果發現，圖像上有某塊區域是語音信號最重要的可感知調變區(critical perceptible modulations)，這塊區域的範圍是：Rate 在 4Hz 到 8Hz 之間而 Scale 在 4 cyc/oct 以下。並使用結合頻域和時域的調變指數(Spectro-temporal modulation index,STMI)評量理解性高低。

5.1.2 研究方法

我們評量理解性高低的方法是使用 [16] 提出的結合頻域和時域的調變指數(STMI)，這個方法的前提是假設乾淨語音和損傷語音在 Rate-Scale 上的圖像都有，再使用式(5)一種有點像計算相似性(similarity)或相關性的方法來評量損傷語音和乾淨語音的差異，相關性是沿著時間軸的變化來計算；客觀的非侵入式評量語音品質時，我們並沒有乾淨語音當作參考訊號，因此我們是由 TIMIT 語料庫中的 4620 句訓練語料，經過感知聽覺模型的分析，得到所有語料在 Rate-Scale 上長時間平均(long-term average)的圖像呈現，我們將此人聲的長時間平均圖像當作乾淨的語音訊號，和損傷語音沿著時間軸上的 Rate-Scale 能量變化計算相似性，最後除以所有 Rate(ω)和 Scale(Ω)維度，得到式(6)的 STMI(ρ)。

$$\rho(\Omega,\omega)=\frac{\left\langle \mathrm{SR_C}(t;\Omega,\omega)-\mu_{\mathrm{SR_C}}(t;\Omega,\omega),\mathrm{SR_n}(t;\Omega,\omega)-\mu_{\mathrm{SR_C}}(t;\Omega,\omega)\right\rangle}{\left\|\mathrm{SR_C}(t;\Omega,\omega)-\mu_{\mathrm{SR_C}}(t;\Omega,\omega)\right\|\bullet\left\|\mathrm{SR_n}(t;\Omega,\omega)-\mu_{\mathrm{SR_C}}(t;\Omega,\omega)\right\|} \tag{5}$$

$$\rho=\frac{1}{\mathrm{N}_\Omega\bullet\mathrm{N}_\omega}\sum_{\Omega}\sum_{\omega}\rho(\Omega,\omega) \tag{6}$$

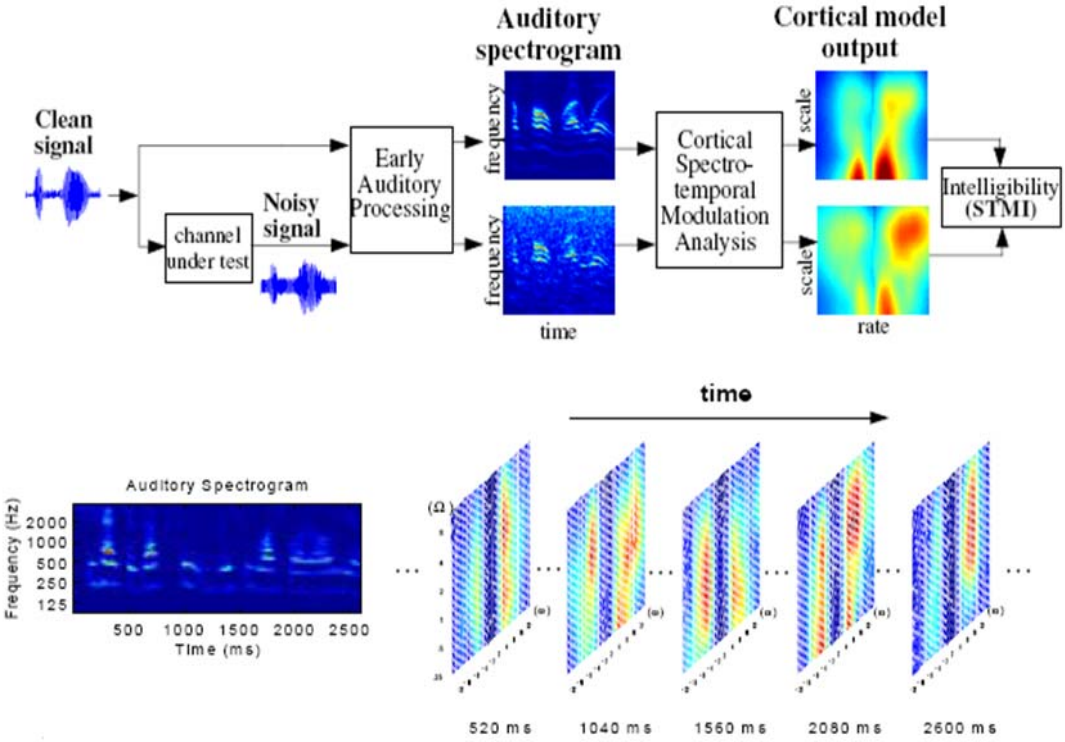


圖 5-1：結合頻域和時域的調變指數(STMI)評量理解性的方法

資料來源：T.-S. Chi, 1999 [16] ; M. Elhilali, 2003 [17]

從圖 5-1 的上圖可看到，乾淨語音和雜訊語音先經過低階的耳朵感知處理(Early Auditory Processing)得到聽覺頻譜圖，再到大腦皮質聽覺區做結合頻域和時域的調變分析，最後使用 STMI 計算得到雜訊語音的理解性。在此當作計算參考的乾淨語音我們是用人聲在 Rate-Scale-Time 的長時間平均來取代，而雜訊是用不同 SNR 的高斯白雜訊。從圖 5-1 的下圖可看到，STMI 的相似性計算方式，是將雜訊語音在不同時間點的 Rate-Scale plot 一個一個的和人聲長時間平均的 Rate-Scale plot 比較相似性，最後得到的 STMI 數值當作理解性高低的指標。

5.1.3 研究結果

在這個實驗裡，我們假設不同 SNR 的雜訊語音經過 PESQ 計算得到的預測數

值是原本真實的 MOS。圖 5-2 中共有 121 個資料點，每個資料點代表不同 SNR 的雜訊語音，SNR 從-15dB 到 45dB，每隔 0.5 產生一個雜訊語音。我們將 STMI 當作理解性高低的量化數值，從圖 5-2 中的橫軸我們可以看到，隨著 SNR 增加，PESQ 數值也跟著增加，表示語音品質跟著提高；另外，從縱軸我們可以看到，隨著 SNR 增加，STMI 數值也跟著增加，表示語音的理解性也跟著提高。

值得注意到的是，當 SNR 增加到超過 10dB 左右，STMI 數值已漸趨平緩不再變動，此時相對應到的 PESQ 數值約在 2 附近。從圖 5-2 我們可以看到，當語音品質(PESQ)上升時，語音的理解性(STMI)也跟著上升；從另外一個角度來看，語音理解性增加，語音品質也會增加，但當理解性增加到極限時，語音品質仍會隨著 SNR 的增加繼續提高。這表示理解性確實是決定語音品質好壞的其中一個要素，但理解性高低只能用來預測語音品質在 M0S 小於 2 的分數，而 MOS 超過 2 以上時，需要其它要素來說明，這個實驗結果和 [29] 研究結果一致，理解性只是影響聽者評量語音的許多要素之一。

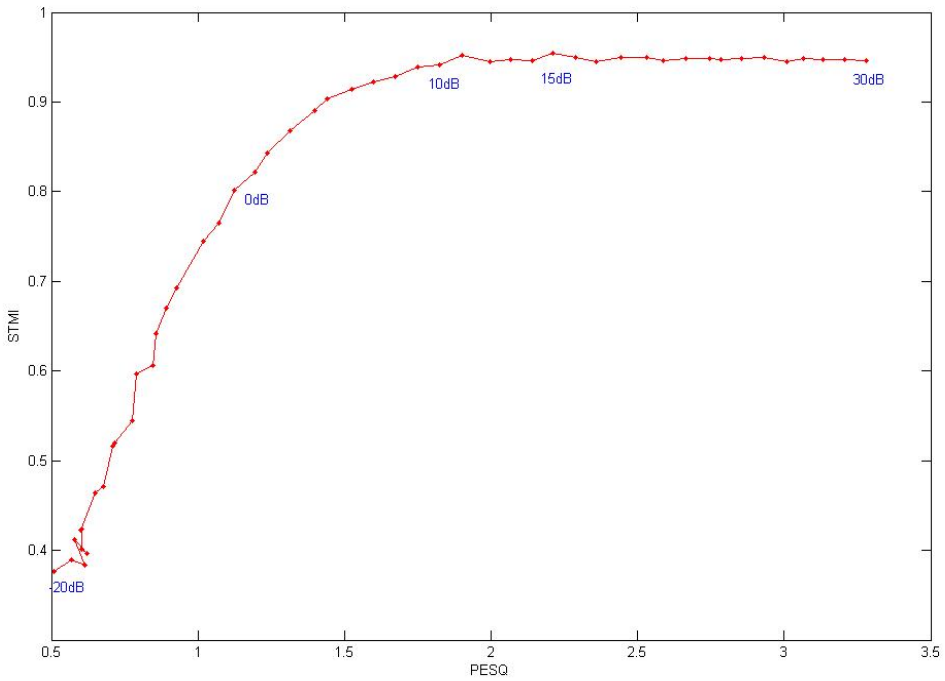


圖 5-2：加上不同 SNR 白雜訊的雜訊語音，其 STMI 數值和 PESQ 的關係

5. 2 自然性

5.2.1 背景知識

語音的自然性(naturalness)概念從 1951 年就被 Parrish 提出，當時對自然性的定義是：聽者對於這句語音的感覺是否正常(normal)或自然(natural)。從 Sanders 等人在 1981 年的研究 [19] 知道，語音品質的評量包含三個要素：理解性、清晰度(clarity)以及自然性；從前人的研究結果 [18] 知道：較快的說話速率(word/min)而且沒有停頓或暫停(pause)的語音和較慢的說話速率而且加上停頓的語音相比較，前者會聽起來會比較自然；當說話速率等於或超過 180(word/min)時，這句語音聽起來會比較自然。

在 [18] 研究結果裡提到重要的一點就是，前人的研究專注於語音的理解性，雖然比較慢的說話速率可以增加理解性，但卻也會降低自然性，因此想要同時評量這兩個特性應該從不同角度做探討。在上一章節裡，我們是在比較低 Rate(2Hz 到 16Hz)和低 Scale(0.25 cyc/oct 到 4 cyc/oct)的區域，也就是跟語音的母音成分有關，來評量語音的理解性；在這個章節裡，我們先假設英語語音平均每個音節(syllable)有 3 個音素(phoneme)，倘若一分鐘的語音超過 180 個字，則平均每秒鐘會有 3 個字；因爲每個字至少有一個音節，因此平均每秒鐘就有至少 9 個音素，爲了不跟母音的區域重疊，我們取比較大的 Rate，也就是 32Hz 到 64Hz。 此外，因爲語音的子音成分在 Scale-Rate 的圖像和高斯白雜訊很像，在 Scale 比較大的區域有較強的反應，所以我們 Scale 的範圍是取 2cyc/oct 到 8cyc/oct。

5.2.2 研究方法

在這個章節裡，我們沿用上一章節評量語音理解性的 STMI 方法，只是在 Rate-Scale plot 上比較相似性時，Rate 的範圍是取 32Hz 到 64Hz，Scale 的範圍是取 2 cyc/oct 到 8 cyc/oct，其它計算方法同 5.1.2 小節。此外，我們從 TIMIT 語料庫中選取一句語音，觀察這句語音在加上不同 SNR 的高斯白雜訊時，語音的理解性和自然性有何變化，以及這兩個特性和語音品質有何相關性。

5.2.3 研究結果

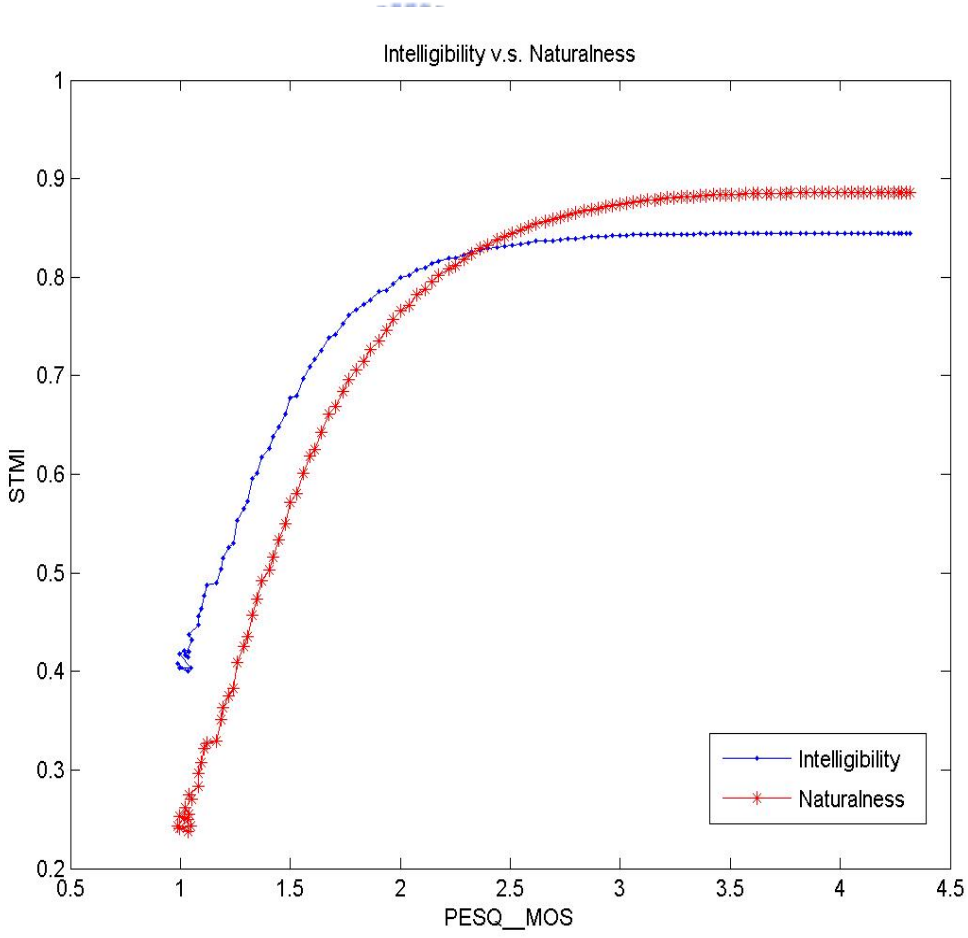


圖 5-3：加上不同 SNR 高斯白雜訊的語音，其理解性和自然性隨 MOS 的變化

從圖 5-3 我們可以看到，自然性和理解性隨著 MOS 的不同有不同的變化趨勢。從 5.1.3 節的結果我們知道，當 MOS 等於 2 附近，理解性的上升趨勢已漸趨平緩，而在圖 5-3 中可以看到，代表自然性的曲線在 MOS 超過 2 之後仍然維持線性的上升趨勢，一直上升到 MOS 超過 3 附近，這條曲線才漸趨平緩。

由此結果我們可以推論說，語音的品質確實和理解性以及自然性有關係，也映證了 Sanders 等人的研究 [19]：這兩種特性對於語音品質的影響並不相同。在語音品質比較差的時候(MOS 較低時)，理解性顯得比較重要；在語音品質比較好的時候(MOS 較高時)，自然性顯得比較重要。此外，理解性的高低只能用來預測語音品質在 M0S 在 1 到 2 附近的分數，而自然性的高低卻能用來預測語音品質在 M0S 在 1 到 3 附近的分數；因此，在 MOS 等於 2 到 3 時，自然性對於決定語音品質就是相當重要的要素。

5. 3 基頻失真

5.3.1 背景知識

從第四章客觀的侵入式語音評量方法的研究結果知道，在 Rate-Scale domain 上有塊區域和聽者在評量語音品質有相當大的關連性，這塊區域是 Rate 取在 32Hz 到 256Hz，而 Scale 是取在 0.25 cyc/oct 到 8 cyc/oct；而這塊區域裡的其中一部分，和人聲的基頻高低息息相關，也就是 Rate 取在 128Hz 到 256Hz、Scale 維持 0.25 cyc/oct 到 8 cyc/oct 這個區域。當乾淨的語音經由各種不同的 codec 處理、加上不同的背景雜訊以及在語音在電話網路傳送過程可能發生封包、音框、位元遺失的損傷，這些損傷都會導致語音在這塊區域有相當大的失真，我們在 5-3 這節就是要探討基頻失真和語音品質好壞有何關聯。

5.3.2 研究方法

首先，我們從 Supp.23 語料庫中挑選出四種經過不同損傷的語音，分別是：

(1) experiment 1：oe1m1723.out (G.729 x G.729 x JDC-HR)

(2) experiment 3：oe3m4216.out (G.729 + Hoth noise)

(3) experiment 3：oe3m4519.out (G.729 + Hoth noise + Bursty Frames 3%)

(4) experiment 3：oe3m5232.out (G.729 + clean + random bit 10%)

然後，我們觀察乾淨語音和損傷語音經由感知聽覺模型得到的頻譜圖以及在 Rate-Scale domain 上呈現的圖像，研究不同損傷會造成圖像有何異狀。

最後，我們仿照第四章客觀的侵入式語音評量方法，在 Rate 是 128Hz 到 256Hz 和 Scale 是 0.25 cyc/oct 到 8 cyc/oct 這個區域計算基頻能量失真；因爲是非侵入式方法，所以我們並沒有各種不同損傷語音的原始乾淨語音，因此我們是先建立一個乾淨人聲在『Rate 是 128Hz 到 256Hz 以及 Scale 是 0.25 cyc/oct 到 8 cyc/oct』這區域能量分布的簡單模型，將損傷語音和此模型做比較來估算基頻能量失真。

從圖 5-4 到圖 5-11 我們清楚看到，無論是經過何種損傷的語音，基頻的能量分佈都遭到相當程度的破壞。因爲我們選取的是一個男生語音，基頻應該在 100 到 200Hz 附近；從頻譜圖上可看到，乾淨語音的頻譜圖，基頻能量明顯且完整，但經過損傷的語音，基頻能量幾乎都看不到了，只有泛頻部份有保留下來，但也變的相當模糊，我們認爲在頻譜圖上的失真現象可能和語音品質好壞有關。另外，從 Rate-Scale 圖像上也可清楚看到，原本乾淨的語音，在 Rate 是 128Hz 到 256Hz 之間以及 Scale 大於 1.0 cyc/oct 的區域，能量的顏色都相當深黑，但經過各種損傷的語音，這塊區域的顏色都變淡了，我們可以說基頻能量受到相當程度的破壞。

因此我們針對這塊區域做研究，探討這塊區域的能量變化和語音品質有何相關性；對 Supp.23 實驗一的 4 位語者和實驗二的 4 位語者總共 376 句經不同損傷的語音信號做基頻失真分析。

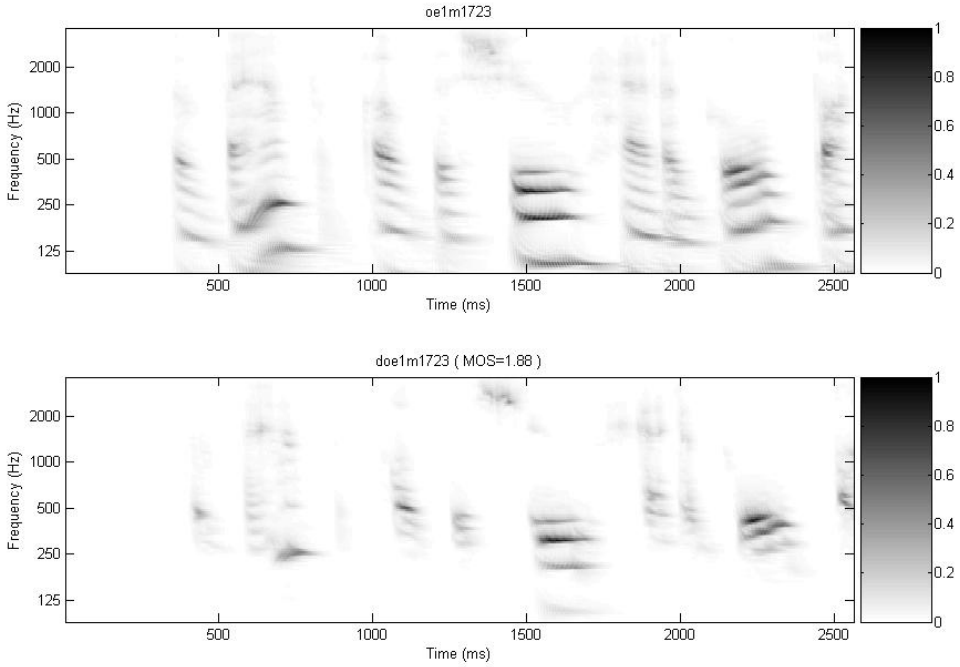


圖 5-4：上圖是乾淨語音，下圖是經過三種 codec 處理過後的損傷語音

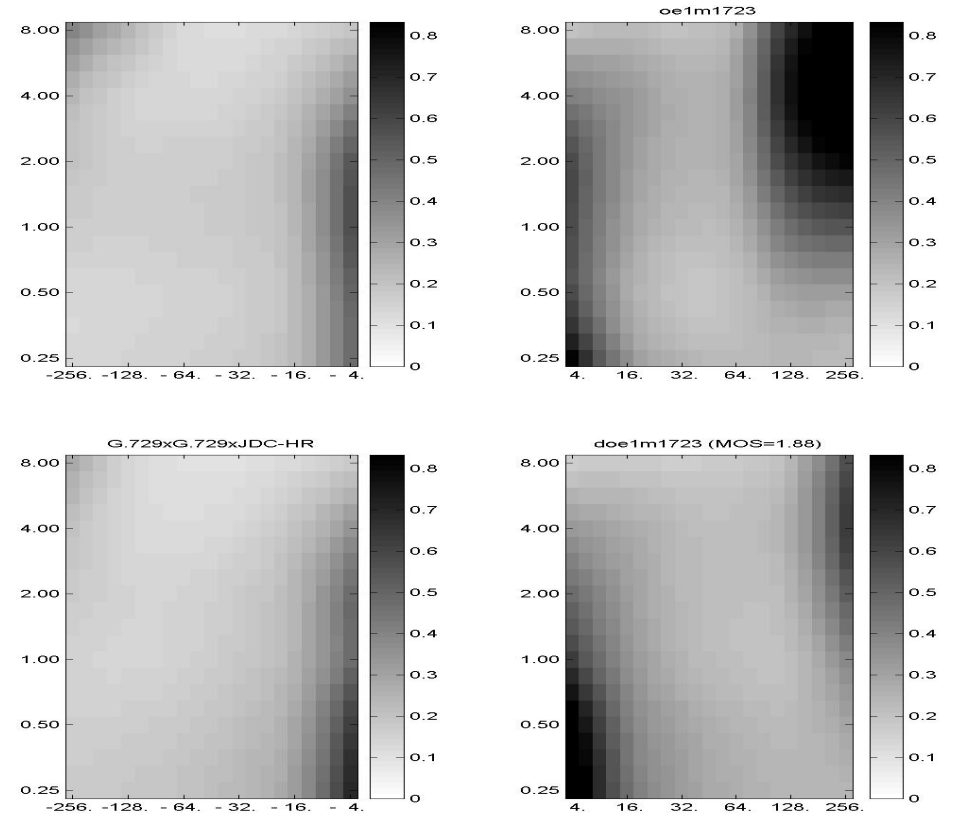


圖 5-5：上兩圖是乾淨語音，下兩圖是經過三種 codec 處理過後的損傷語音

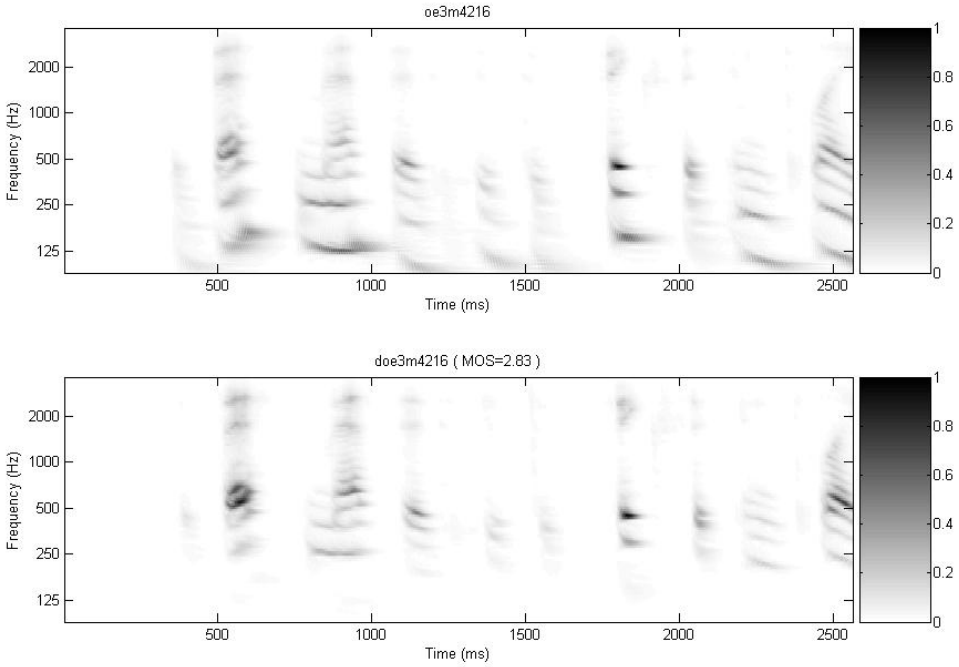


圖 5-6：上圖是乾淨語音，下圖是損傷語音(G.729 + Hoth noise)

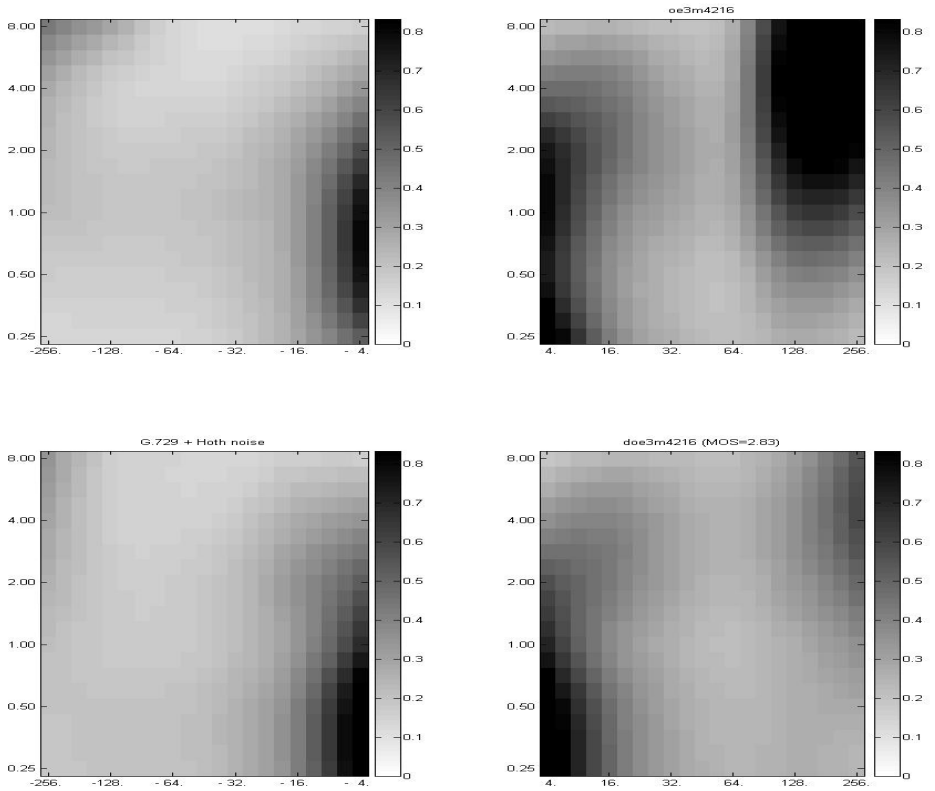


圖 5-7：上兩圖是乾淨語音，下兩圖是損傷語音(G.729 + Hoth noise)

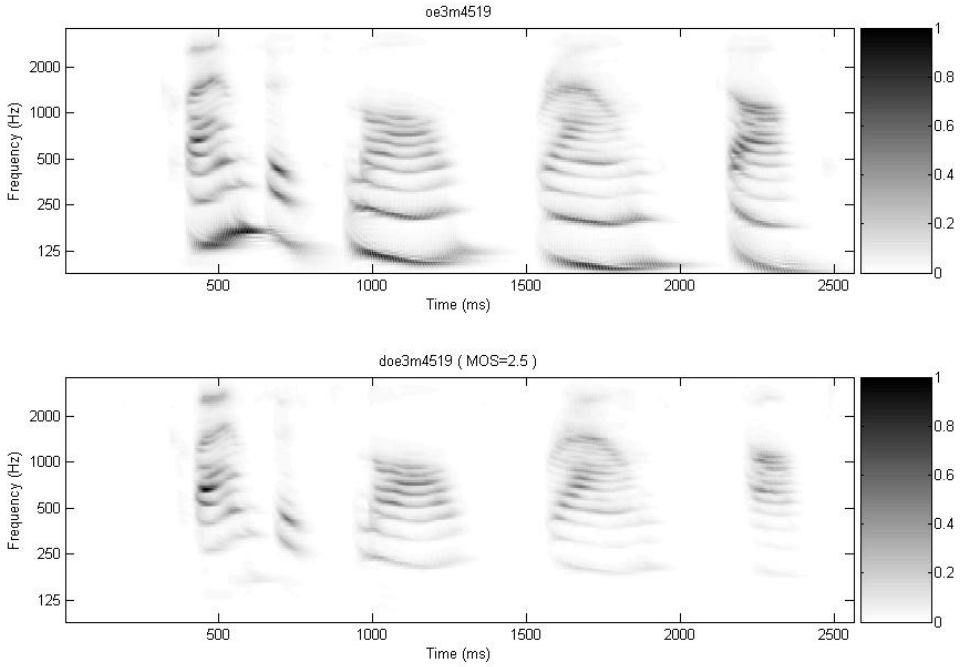


圖 5-8：上圖是乾淨語音，下圖是損傷語音(G.729+Hoth noise+Burst Frame 3%)

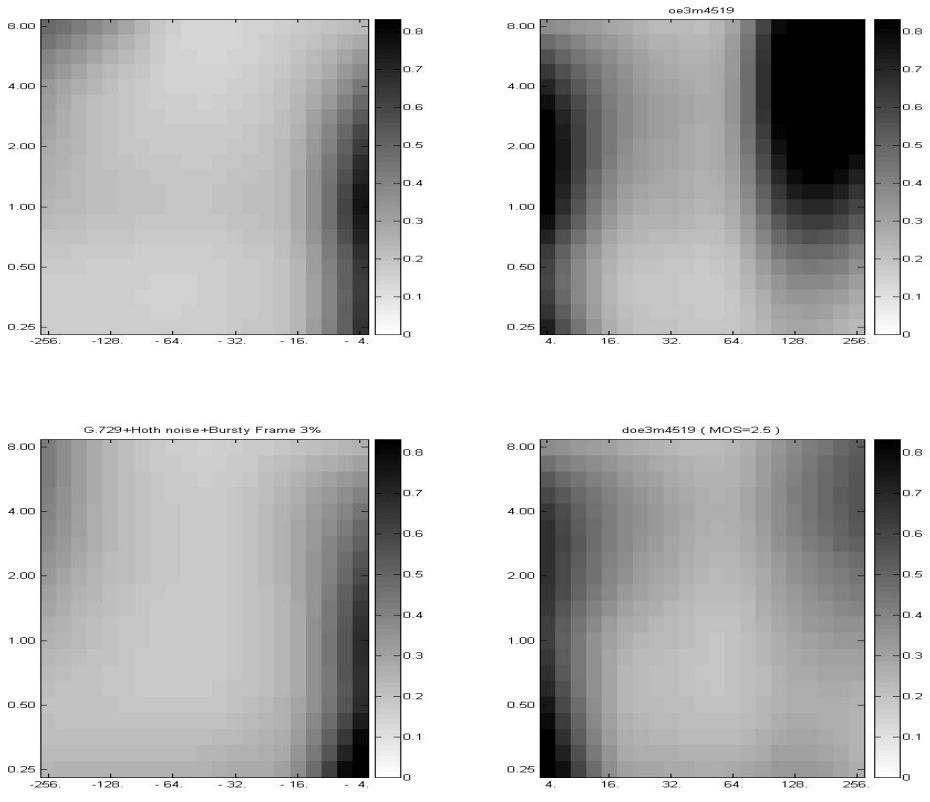


圖 5-9：上兩圖是乾淨語音，下兩圖是損傷語音(G.729+Hoth noise+Burst Frame 3%)

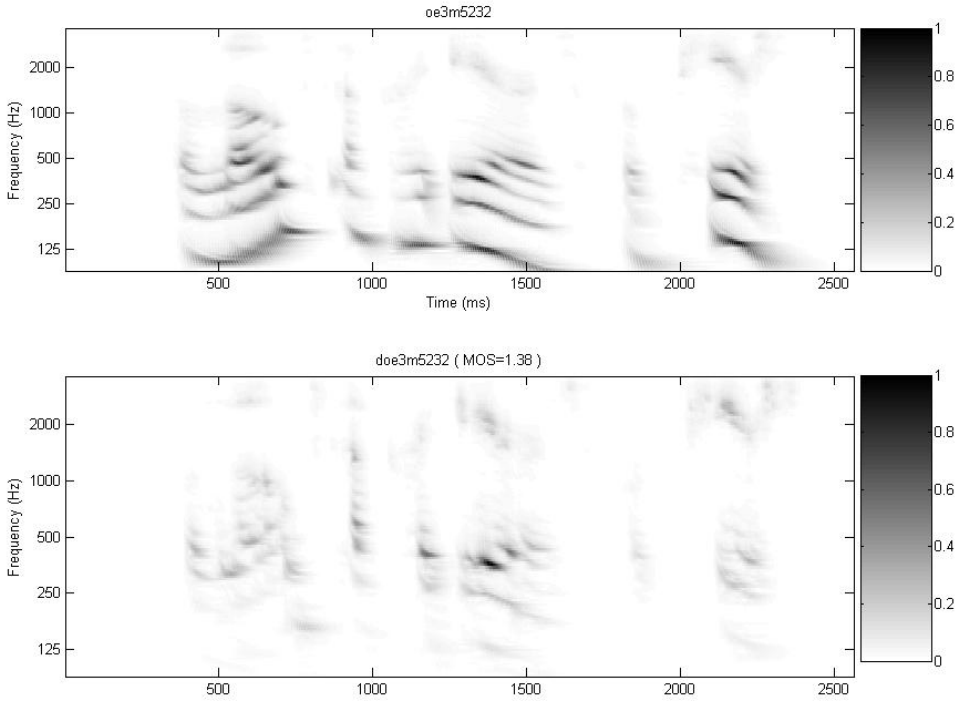


圖 5-10：上圖是乾淨語音，下圖是損傷語音(G.729+clean+random bit 10%)

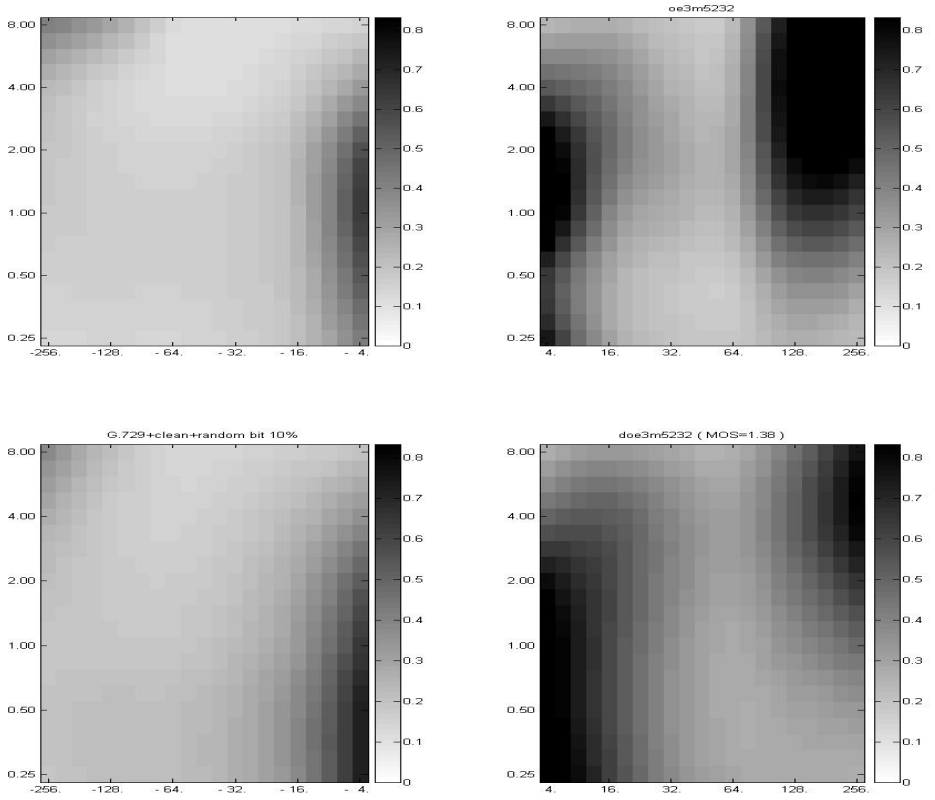


圖 5-11：上兩圖是乾淨語音，下兩圖是損傷語音(G.729+clean+random bit 10%)

5.3.3 研究結果

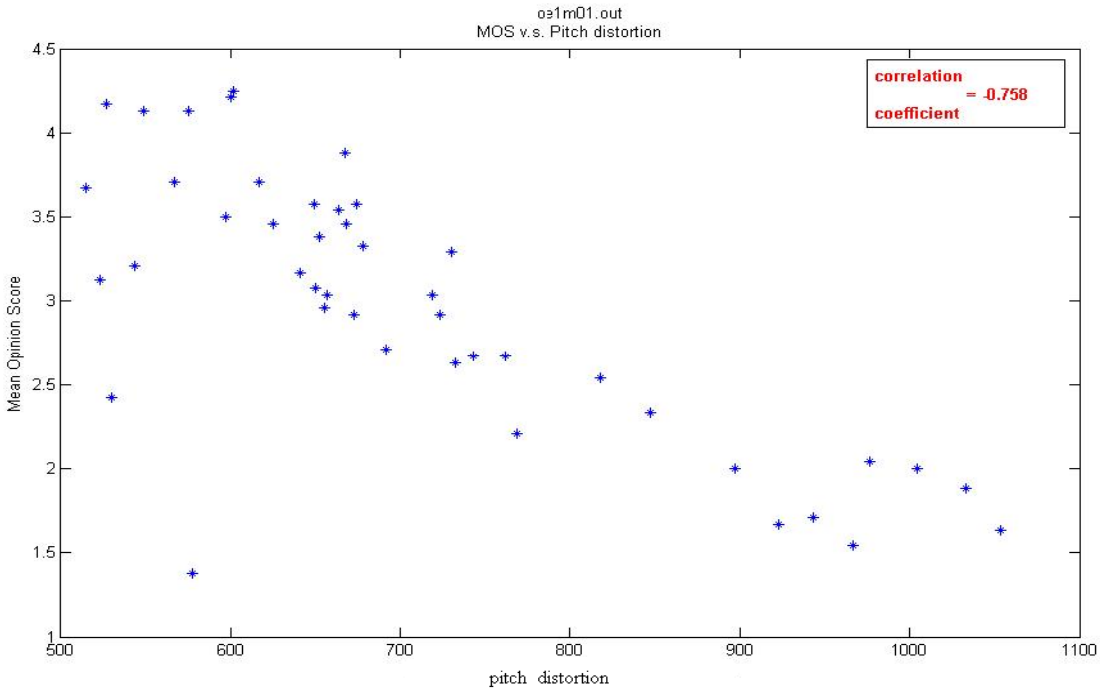


圖 5-12：實驗一的 44 種損傷語音，其基頻失真程度和 MOS 的關係

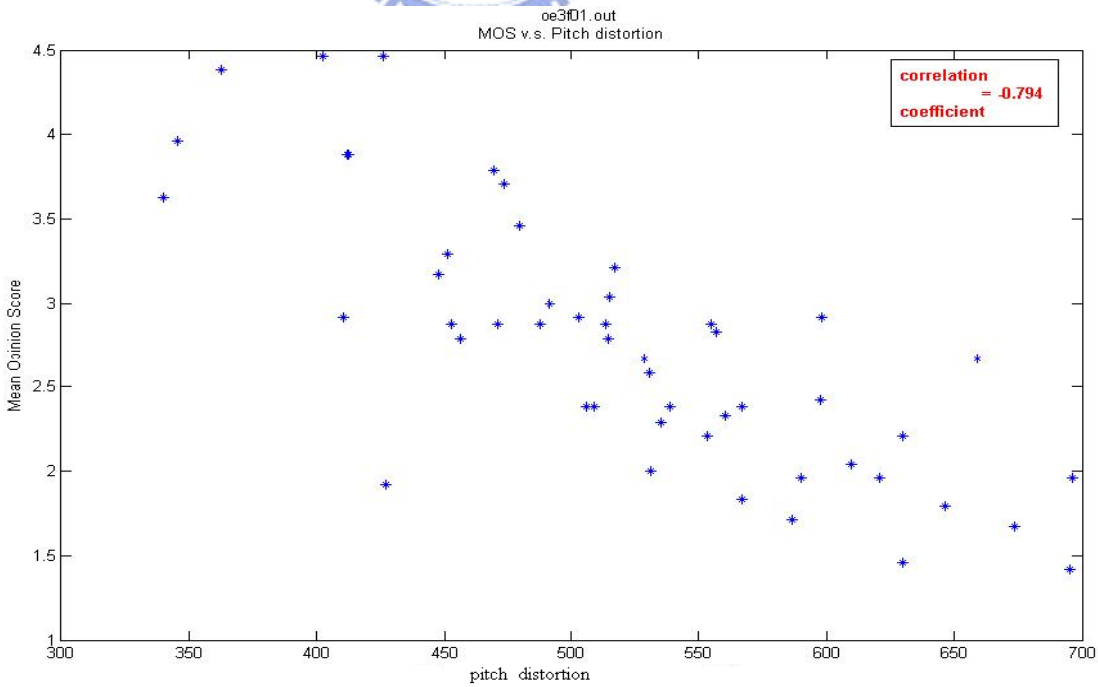


圖 5-13：實驗三的 50 種損傷語音，其基頻失真程度和 MOS 的關係

基頻能量失真與MOS的相關程度關係			
Experiment 1	相關係數	**Experiment 2**	相關係數
female 1	-0.699	female 1	-0.794
female 2	-0.714	female 2	-0.671
male 1	-0.758	male 1	-0.747
male 2	-0.577	male 2	-0.718
Average	-0.687	**Average**	0.708

表 5-1：基頻能量失真與語音品質好壞的相關係數

5. 4 綜合三種感知特徵評量語音品質

5.4.1 研究方法

由前三節得到的語音信號三種感知特徵：理解性、自然性、基頻失真來做客觀非侵入式語音品質評量；由前三節的方法，我們可分別得到三種感知特徵經由計算後得到的量化數值，再用最小平方法將三者對語音品質好壞的影響做結合，得到最後我們預估的語音品質分數(pre_MOS)。

我們使用 Supp.23 語料庫中實驗一和實驗三中的語料，對其中四位男性與四位女性經過各種損傷的語音做分析，分別得到各別預估的語音品質分數，最後再和 ITU-T 客觀評量方法的國際標準 P.563 做比較，預估的效能用相關係數表示。比較結果可參考表 5-2 到表 5-5。

5.4.2 研究結果

實驗一：兩位女性語者

exp 1	female 1			female 2		
condition	MOS	P.563	pre_MOS	MOS	P.563	pre_MOS
c1	4	2.82	3.62	3.46	3.55	3.73
c2	3.29	3.39	3.26	2.83	3.31	3.13
c3	2.96	2.97	3.00	2.04	3.02	2.24
c4	4	3.54	4.07	3.96	3.56	3.88
c5	2.92	2.95	3.23	3	3.27	2.91
c6	3.75	2.88	4.01	3.46	3.14	3.85
c7	4.29	4.38	3.43	4.21	4.28	3.62
c8	3.46	2.68	3.53	3.58	2.30	3.85
c9	3.46	2.72	3.05	3.58	3.30	3.02
c10	3.25	3.17	2.95	2.67	3.09	2.28
c11	3.58	3.26	3.37	3.71	3.10	3.64
c12	3.75	2.61	3.41	3.5	3.14	3.03
c13	3.29	2.71	3.35	3.33	3.12	3.33
c14	3.17	2.99	3.30	3.04	3.09	2.65
c15	2.63	3.00	2.23	1.92	2.45	1.73
c16	3.33	3.10	3.18	3	2.86	3.43
c17	3.67	3.23	3.51	3.25	2.65	3.09
c18	3.21	2.98	3.42	2.92	3.33	3.10
c19	2.96	2.84	2.87	3.08	3.16	2.51
c20	2.54	3.51	2.64	1.92	2.74	1.91
c21	2.54	2.20	2.95	2.71	2.47	2.32
c22	2.83	2.77	2.83	1.63	2.63	2.07
c23	1.96	2.53	2.19	2.08	2.36	1.91
c24	3.29	2.57	3.43	2.88	1.54	3.10
c25	2.92	2.47	3.23	2.83	2.57	3.07
c26	3.08	2.64	2.86	2.67	2.89	2.85
c27	2.46	3.16	2.59	2.04	1.48	2.15
c28	1.83	2.06	1.85	1.58	2.06	1.46
c29	2.83	2.81	3.29	3.25	2.32	3.28
c30	2.96	2.74	3.47	3.08	3.00	2.98

c31	2.96	2.46	3.43	2.92	2.79	3.67
c32	2.67	2.94	2.54	1.71	1.66	1.71
c33	1.96	2.07	2.53	1.38	1.83	2.17
c34	2.42	2.65	2.99	2.38	2.84	2.87
c35	2.13	2.21	1.69	1.38	2.18	1.33
c36	2.04	2.40	2.11	1.5	2.22	2.33
c40	3.33	3.66	3.99	3.29	3.22	3.67
c41	4.13	3.90	3.91	4.08	3.59	3.46
c42	4.17	3.90	4.06	4.21	4.03	3.93
c43	4.33	3.83	3.70	4	4.18	3.62
c44	4.54	4.11	3.81	4.08	4.04	3.27
相關係數		**0.742**	**0.855**		**0.759**	**0.878**

表 5-2：實驗一兩個女聲的 MOS、P.563 預估分數、我們的預估分數

實驗一：兩位男性語者

exp 1	male 1			male 2		
condition	MOS	P.563	pre_MOS	MOS	P.563	pre_MOS
c1	3.88	3.47	3.30	4.29	3.43	3.87
c2	3.54	3.23	3.22	3.88	2.99	3.46
c3	2.63	2.55	2.77	2.96	2.92	3.20
c4	3.71	3.72	3.43	4.04	3.57	3.75
c5	3.5	2.83	3.65	3.33	3.16	4.01
c6	3.67	3.34	3.95	3.96	3.36	4.15
c7	4.21	3.63	3.61	4.54	3.79	3.89
c8	3.13	3.42	4.00	3.92	3.00	4.00
c9	3.46	2.85	3.44	4.17	3.39	3.61
c10	2.33	3.20	2.40	3.33	3.42	3.10
c11	3.58	3.01	3.42	4.08	3.68	3.80
c12	2.96	2.71	3.39	3.92	3.46	3.84
c13	3.08	2.79	3.37	3.21	2.76	3.68
c14	2.71	2.73	2.99	3.92	2.86	3.45
c15	2	2.11	1.71	2.83	2.76	2.46
c16	3.58	2.91	3.28	4	3.38	3.44
c17	3.38	2.60	3.35	4.17	3.14	3.63
c18	3.33	2.87	3.27	4.21	3.11	3.63

c19	3.29	2.77	3.09	3.79	3.14	3.22
c20	2.04	2.56	1.74	2.67	2.91	2.61
c21	2.67	2.20	2.84	2.79	2.61	3.68
c22	2.21	3.40	2.85	3.13	2.86	2.96
c23	1.88	2.26	1.53	2.17	2.58	2.33
c24	3.04	2.34	3.31	3.46	2.97	3.81
c25	2.92	2.82	3.24	2.83	2.58	4.00
c26	2.92	2.46	3.10	2.96	2.74	3.04
c27	2.54	1.74	2.55	2.96	3.07	2.59
c28	1.63	2.35	1.51	1.79	2.14	2.11
c29	3.46	2.92	3.23	3.46	2.95	3.63
c30	3.17	2.74	3.30	3.33	2.79	3.57
c31	3.04	2.77	3.05	3.54	2.83	3.52
c32	2	2.75	2.21	2.46	2.79	2.60
c33	1.67	2.32	2.00	2.63	2.64	3.38
c34	2.67	2.92	2.80	2.83	2.41	2.95
c35	1.71	2.12	1.95	2.08	2.60	2.11
c36	1.54	1.26	1.67	2.04	2.71	2.84
c40	3.71	3.19	3.83	3.5	3.99	4.05
c41	4.13	3.73	3.89	4.33	3.84	4.14
c42	4.25	4.07	3.58	4.38	3.82	4.33
c43	4.13	3.75	3.71	4.5	3.78	4.25
c44	4.17	3.62	3.93	4.5	4.15	4.19
相關係數		**0.769**	**0.908**		**0.810**	**0.801**

表 5-3：實驗一兩個男聲的 MOS、P.563 預估分數、我們的預估分數

實驗三：兩位女性語者

exp 3	female 1			female 2		
	MOS	P.563	pre_MOS	MOS	P.563	pre_MOS
c1	3.88	3.18	3.60	3.75	3.19	2.72
c2	3.46	3.17	3.14	2.75	3.11	3.21
c3	3.71	3.52	3.21	2.71	2.76	2.72
c4	3.79	3.43	3.23	3.79	3.36	3.28
c5	2.88	3.39	3.36	3.42	3.14	3.23
c6	3.04	2.54	2.77	2.75	2.65	2.27

c7	2.67	2.46	2.64	2.58	2.20	2.12
c8	2.38	2.33	2.84	2.08	2.36	2.14
c9	1.96	2.55	2.26	2.79	2.06	1.70
c10	1.46	2.13	1.96	1.92	2.15	2.00
c11	2.58	2.65	2.71	2.46	2.45	1.66
c12	2.38	2.68	2.89	1.88	2.85	1.90
c13	1.71	2.37	2.30	2	2.32	1.85
c14	2.83	2.39	2.52	1.92	2.84	2.21
c15	2.38	2.75	2.44	1.63	2.72	2.11
c16	2.88	2.74	2.83	2.25	2.51	2.32
c17	2.21	2.69	1.96	2.38	2.56	2.59
c18	2.21	2.81	2.54	1.71	2.18	2.08
c19	2.33	2.19	2.57	2.08	2.26	2.29
c20	1.83	2.33	2.45	2.25	2.20	2.26
c21	3	3.37	3.04	3.38	3.48	2.59
c22	2.67	3.02	1.75	2.33	3.05	2.33
c23	2.92	2.69	2.19	2.25	2.79	2.75
c24	1.96	2.62	1.46	1.88	2.82	2.01
c25	1.96	2.50	2.05	2.67	2.52	3.09
c26	2.42	3.02	2.21	1.63	2.42	1.71
c27	2.04	2.63	2.12	1.63	2.44	2.22
c28	1.42	2.35	1.48	1.67	1.94	1.57
c29	3.29	3.09	3.35	2.75	2.69	3.34
c30	2.29	2.73	2.71	1.79	2.43	2.42
c31	1.79	2.11	1.86	1.54	2.60	2.14
c32	1.67	2.06	1.65	1.17	2.28	1.48
c33	3.21	2.81	2.83	3.33	2.86	3.03
c34	2.88	2.94	2.62	2.63	2.64	3.40
c35	2.38	2.76	2.77	2.04	2.63	3.11
c36	2	2.88	2.80	2.21	2.45	2.91
c37	3.88	3.15	3.62	3.83	3.73	3.29
c38	2.88	2.75	3.07	3.17	2.91	2.92
c39	2.79	2.43	2.78	2.88	2.52	2.69
c40	2.92	2.70	2.81	2.79	2.54	3.10
c43	3.63	3.56	4.20	3.54	3.04	3.69
c44	3.96	3.82	4.19	4.29	3.82	3.49
c45	4.46	4.15	3.55	4.17	3.84	3.75

c46	4.38	3.65	4.03	4.54	4.43	3.48
c47	4.46	3.82	3.73	4	3.55	3.85
c48	3.17	3.00	3.28	3.25	3.11	2.87
c49	2.79	2.86	3.16	2.75	3.32	2.95
c50	2.88	3.54	3.14	3.13	3.19	3.50
相關係數		**0.839**	**0.850**		**0.811**	**0.782**

表 5-4：實驗三兩個女聲的 MOS、P.563 預估分數、我們的預估分數

實驗三：兩位男性語者

exp3	male 1			male 2		
	MOS	P.563	pre_MOS	MOS	P.563	pre_MOS
c1	3.88	3.16	3.61	3.83	3.18	3.29
c2	3.04	3.15	3.44	3.33	3.16	3.07
c3	3.17	2.88	2.95	3.79	3.28	3.63
c4	3.5	2.53	3.25	3.33	3.31	3.43
c5	3.58	3.04	3.30	2.88	3.04	3.60
c6	3.08	2.74	3.15	2.96	2.66	3.23
c7	2.88	2.41	2.79	2.71	2.86	2.41
c8	2.58	2.69	2.61	2.46	2.38	2.31
c9	2.96	2.41	2.98	1.92	2.45	1.85
c10	2.29	2.57	2.25	2.79	2.84	2.95
c11	2.83	2.69	2.85	3.08	2.87	3.03
c12	2.63	2.60	2.37	2.63	2.44	2.85
c13	2.29	2.65	2.09	2.42	2.71	2.49
c14	2.25	2.76	2.47	2.13	2.64	2.52
c15	1.92	2.71	2.28	1.71	2.71	2.15
c16	2.67	2.54	2.48	2.83	2.57	2.78
c17	2.83	2.75	2.73	2.63	2.87	2.80
c18	2.71	2.46	2.31	2.46	2.92	2.87
c19	2.42	2.95	2.83	2.5	2.88	2.49
c20	2.58	2.39	2.48	2.67	2.84	2.88
c21	2.63	2.50	3.30	3.75	3.29	2.99
c22	2.33	1.95	1.97	3.29	2.85	2.53
c23	2.25	2.59	2.74	3.08	2.17	2.58
c24	2.67	2.55	2.37	2.46	2.62	2.01

c25	1.96	2.38	2.33	1.58	2.82	1.81
c26	2.63	1.91	2.42	1.88	2.86	1.87
c27	1.63	1.97	1.76	2.08	2.41	1.96
c28	1.5	2.23	1.57	1.38	2.32	1.80
c29	3.46	2.98	3.10	3.42	3.42	2.46
c30	2.88	2.56	3.49	2.79	2.81	2.72
c31	2.04	3.29	2.68	2.58	2.79	2.96
c32	1.5	2.17	2.17	1.38	2.75	2.51
c33	3.54	2.88	3.48	3.67	3.23	3.49
c34	2.29	3.05	2.58	3.54	3.18	3.11
c35	2.29	3.01	2.86	2.79	2.98	3.37
c36	2.25	2.62	2.16	2.13	2.93	3.08
c37	4.04	4.00	3.72	4.13	3.67	3.88
c38	2.96	2.68	2.76	3.13	2.95	3.05
c39	2.75	2.87	2.82	3.13	2.74	3.57
c40	2.67	2.90	3.06	2.96	3.16	3.16
c43	3.46	3.31	4.23	3.58	3.32	4.08
c44	4.17	3.85	3.82	4.33	3.98	3.63
c45	4.25	3.72	3.63	4.38	3.72	4.04
c46	4.46	3.53	3.66	4.5	3.85	4.32
c47	4.38	3.18	4.16	4.29	3.85	4.19
c48	3.17	3.01	2.27	3.21	3.32	2.86
c49	3.04	2.83	2.79	3.17	2.83	2.79
c50	3.33	3.14	3.53	3	3.29	3.22
相關係數		**0.707**	**0.845**		**0.789**	**0.826**

表 5-5：實驗三兩個男聲的 MOS、P.563 預估分數、我們的預估分數

第六章 結果討論與未來展望

6. 1 結果討論

6.1.1 客觀的侵入式語音品質評量方法

我們客觀的侵入式語音評量方法，是使用感知聽覺模型以及前人的研究結果發現【12】，在 Rate 和 Scale 兩個維度上找到一個能量分佈區域，這個特殊區域是 Rate 在 32Hz 到 256Hz 以及 Scale 在 2cyc/oct 到 8cyc/oct 之間；藉由計算並累加不同時間點的乾淨語音和損傷語音在此特殊區域的能量變化來評量語音品質好壞，然後和 MOS 對照(mapping)並將結果和 ITU-T P.862(PESQ)做比較。

從表 4-1 可看到，對 ITU-T Supp.23 實驗一的四位語者，我們預測的 MOS 和真實 MOS 最高的相關係數可達 0.928，最差是 0.795，平均是 0.868；而 PESQ 最高是 0.923，最差是 0.866，平均是 0.902。若將我們的預測結果和 PESQ 的預測結果相除做比較，平均可高達 96%的相關程度，由此可看出我們客觀的侵入式預測 MOS 方法的效能和 PESQ 相當。

6.1.2 客觀的非侵入式語音品質評量方法

我們客觀的非侵入式語音品質評量方法是是使用感知聽覺模型在兩個聽覺感知階段做觀察和分析，希望從人耳低階感知反應上擷取出可能影響聽者在高階認知判斷語音品質的特徵參數來對語音品質做客觀評量，這三個特徵分別是－理解性、自然性、基頻失真。

我們評量理解性高低的方法是使用結合頻域和時域的調變指數(STMI)，在語

音信號最重要的可感知調變區(critical perceptible modulations)，這塊區域的範圍是 Rate 在 4Hz 到 8Hz 之間而 Scale 在 4 cyc/oct 以下作計算。評量自然性高低的方法亦是使用 STMI，Rate 的範圍是取 32Hz 到 64Hz，Scale 的範圍是取 2 cyc/oct 到 8 cyc/oct。評量基頻失真程度的方法是在 Rate 是 128Hz 到 256Hz 以及 Scale 是 0.25 cyc/oct 到 8 cyc/oct 這區域，先建立一個乾淨人聲能量分布的簡單模型，將損傷語音和此模型做比較來估算基頻能量失真。

從 5-1 節的研究結果得到，理解性確實是決定語音品質好壞的其中一個要素，但理解性高低只能用來預測語音品質在 M0S 小於 2 的分數，而 MOS 超過 2 以上時，需要其它要素來說明，這個實驗結果和【29】研究結果一致。

從 5-2 節的研究結果得到，代表自然性的曲線在 MOS 超過 2 之後仍然維持線性的上升趨勢，一直上升到 MOS 超過 3 附近，這條曲線才漸趨平緩。由此結果可以推論，語音的品質確實和自然性有關係，也映證了 Sanders 等人的研究【19】。

從 5-3 節的研究結果知道，無論是經過何種損傷的語音，基頻的能量分佈都遭到相當程度的破壞；乾淨語音的頻譜圖，基頻能量明顯且完整，但經過損傷的語音，基頻能量幾乎都看不到了，只有泛頻部份有保留下來，但也變的相當模糊，研究結果得到：基頻能量失真和 MOS 有接近負 0.7 的高度負相關性。

由第五章得到的語音信號三種感知特徵：理解性、自然性、基頻失真經計算後得到的量化數值，再用最小平方法將三者對語音品質好壞的影響做結合，得到最後我們預估的語音品質分數(pre_MOS)，並將結果和 ITU-T 客觀評量方法的國際標準 P.563 做比較，可參考表 6-1。

我們這次使用 ITU-T Supp.23 的實驗一和實驗三的語料庫，並沒有把 MNRU(Q=5dB 和 10dB)的情況考慮進來，因爲在我們計算基頻失真的程度時，若加入 MNRU(Q=5dB 和 10dB)兩個情況，會大大降低我們預估 MOS 的效能，但在研究完成期限內，我們尚未發現問題所在，可能在計算過程有忽略了哪些重要資訊，這是我們後續還必須克服和想辦法解決的問題。

相關係數 / 實驗一	P.563	pre_MOS
Female 1	0.742	0.855
Female 2	0.759	0.878
Male 1	0.769	0.908
Male 2	0.810	0.801
Average	0.77	0.861
相關係數 / 實驗三	P.563	pre_MOS
Female 1	0.839	0.850
Female 2	0.871	0.782
Male 1	0.707	0.845
Male 2	0.789	0.826
Average	0.802	0.826

表 6-1：P.563 預估 MOS 和我們預估 MOS 的效能評比，以相關係數表示

6.2 未來展望

我們的研究是使用感知聽覺模型，以一種感知特徵參數爲基礎的方法做語音品質評量，這些參數並不是從訊號或是網路本身抽取出來，而是從人的感知反應來分析，同時考慮人耳發聲和聽覺的重要特性，對語音訊號經過人耳感知處理後的頻譜加以分析。我們的研究便是從低階的人耳感知行爲出發，進而延伸到高階的大腦認知行爲，希望藉由低階的人耳感知模型伴隨著高階的大腦認知模型，可以有效並正確的客觀評量語音品質好壞。我們這次的研究對於感知特徵只抽取出三個參數，分別是理解性、自然性以及基頻失真，但人腦對語音品質好壞的高階認知仍有相當多未知參數尚待開發，將來的目標是希望能確認更多和語音品質有關的感知特徵參數，並推導出數學模型，找出它和語音品質的關係，並將其量化以預測 MOS。

參考文獻

[1]. A. W. Rix, J.G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, “Objective Assessment of Speech and Audio Quality-Technology and Applications,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 6, Nov.2006.

[2]. “Perceptual evaluation of Speech quality, an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” ITU-T Rec. P.862, 2001.

[3]. T. P. Barnwell, “Improved objective quality measures for low bit speech compression,” National Science Foundation, Final Technical Report, 1985.

[4]. S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, “Objective Measures of Speech Quality,” Englewood Cliffs, NJ : Prentice-Hall, 1998.

[5]. C. Jin and R. Kubichek, “Vector quantization techniques for output-based objective speech quality,” IEEE Int. Conf. Acoust., Speech, Signal Process. Atlanta, GA, pp.491-494, 1996

[6]. P. Gray, M. P. Hollier, and R. E. Massara, “Non-intrusive speech quality assessment using vocal tract models,” Inst. Elect. Eng. Proc. Vis. Image Sig. Process.,vol. 147, no. 6, pp. 493-501, 2000.

[7]. T. H. Falk and W.-Y. Chan, “Nonintrusive speech quality estimation using Gaussian mixture models,” IEEE Sig. Process. Letters, vol. 13, no. 2, pp.108-111, 2006.

[8]. “Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications,” ITU-T Rec. P.563, 2004.

[9]. A. Raja, R. M. A. Azad, C. Flanagan, and C. Ryan, “Real-Time, Non-intrusive Evaluation of VoIP,” EuroGP LNCS 4445, pp. 217-228, 2007.

[10]. L. Ding, Z. Lin, A. Radwan, M. S. El-Hennawey, and R. A. Goubran,

"Non-intrusive single-ended speech quality assessment in VoIP," Speech Communication 49, pp. 477-489, 2007.

[11]. "The E-model, a computational model for use in transmission planning," ITU-T Rec. G.107, 2002.

[12]. D.-S. Kim, "A cue for objective speech quality estimation in temporal envelope representations," IEEE Signal Processing Lett., vol. 11, no. 10, pp.849-852, Oct. 2004.

[13]. A. Raake, "Does the content of speech influence its perceived sound quality？" in Proc. 3rd Int. Conf. on Language Resources and Evaluation, vol. 4, pp. 1170-1176, 2002.

[14]. D.-S. Kim, "ANIQUE：An auditory model for single-ended speech quality estimation," IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp.821-831, Sep. 2005.

[15]. T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am., vol. 118, no. 2, pp. 887-906, 2005.

[16]. T. Chi, Y. Gao, C. G. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc.Am.,vol. 106, no. 5, pp. 2719-2732, 1999.

[17]. M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index(stmi)for assessment of speech intelligibility," Speech Communication, vol. 41,

no. 2-3, pp.331-348, 2003.

[18]. A. Ratcliff, S. Coughlin, and M. Lehman, "Factors influencing ratings of speech naturalness in augmentative and alternative communication," ISAAC, vol. 18, Mar. 2002.

[19]. W. Sanders, C. Gramlich, and A. Levine, "Naturalness of synthesized speech,"

University-level computer-assisted instructions at Stanford：1968-80(pp. 487-501)

[20]. J. G. Beerends, “Modelling cognitive effects that play a role in the perception of speech quality,” in Workshop “Speech Quality Assessment”, Bochum, Germany, pp. 1-9, 1994.

[21]. S. D. Voran, “Perception of temporal discontinuity impairments in coded speeh-A proposal for objective estimators and some subjective test results,” Int. for Telecommunication Sciences, 2003.

[22]. L. Ding, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, “Measurement of the effects of temporal clipping on speech quality,” IEEE Trans. on instrumentation and measurement, vol. 55, no. 4, Aug. 2006.

[23]. D.J. Klein, D.A. Depireux, J.Z. Simon, and S.A. Shamma, “Robust spectrotemporal reverse correlation for the auditory system：Optimizing stimulus design,” Journal of Computational Neuroscience 9, 85-111, 2000.

[24]. N. Mesgarani, M. Slaney, and S. A. Shamma, “Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations,” IEEE Trans. on audio, speech, and language processing, vol. 14, no. 3, May 2006.

[25]. “ITU-T coded-speech database,” 1998, Supp.23 to P series Rec., ITU-T.

[26]. “Methods for subjective determination of transmission quality,” 1996, ITU-T.

[27]. “Mapping function for transforming P.862 raw result scores to MOS-LQO,” ITU-T Rec. P.862.1, 2003.

[28]. N. F. Viemeister, “Temporal modulation transfer functions based on modulation thresholds,” J. Acoust. Soc. Amer., vol. 66, pp. 1364-1380, 1997.

[29]. K. Hustad, “Intelligibility differences for three listener groups,” Journal of Speech and Hearing Research, 41, 744-752, 1998.