國立交通大學

電信工程學系

碩士論文

感知訊號非侵入式客觀語音品質測量

Model-based Non-Intrusive
Objective Speech Quality Measurement
Using Perceptual Parameters

研究生:余尚儒

指導教授:冀泰石 教授

中 華 民 國 九 十 七 年 七 月

感知訊號非侵入式客觀語音品質測量

Model-based Non-Intrusive Objective Speech Quality Measurement Using Perceptual Parameters

研 究 生:余尚儒　　student: Shang-Ju Yu

指導教授:冀泰石　　Advisor: Tai-Shih Chi

國立交通大學

電信工程學系碩士班

碩士論文

A Thesis
Submitted to Institution of Communication Engineering
College of Electrical and Computer Engineering
National Chiao Tung University
In Partial Fulfillment of the Requirement
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

感知訊號非侵入式客觀語音品質測量

研究生:余尚儒　　指導教授:冀泰石 博士

國立交通大學

電信工程學系碩士班

摘要

語音品質的評測一直為通訊系統的重要議題。由於早期的主觀語音品質測量需要耗費較多的人力與金錢，而有了客觀性(Objective Base)語音品質測量方法的需求。又實際的語音品質測量中，常缺乏原始語音訊號，因此，無需原始訊號即可判斷語音品質的非侵入式(Non-intrusive)語音品質測量正符合此需求。其優點除了不需耗費太多人力外,更可做即時且有效率的品質評斷。本文主要是嘗試用人耳模型模擬人類聽覺系統，從接收到的訊號中抽取聽覺參數，而做一客觀的非侵入式語音品質測量。

我們將利用聽覺參數特性提出一聲音變化偵測器(voice activity detector，VAD)演算法，以此演算法將語音分類成:母音(voice)，子音(unvoice)，及無聲(inactive)部分。接著，在經過人耳聽覺模型(Auditory Model)後的頻譜中求取倒頻譜參數(Cepstral coefficients)，在此我們稱為聽覺倒頻譜參數(Auditory Cepstral coefficients，ACC)。為了能夠無需參考訊號即可做語音品質判斷，我們以高斯混合模型(Gaussian Mixture Model)將無雜訊語音利用聽覺倒頻譜參數訓練出一乾淨語音之模型。

在語音品質評測的部分，將經過不同通道及不同編碼技術的語音以聲音變化偵測器做分類，並求其聽覺倒頻譜參數。接著將此參數與乾淨語音之高斯混合模型做比對，由比對高斯分佈後的對數機率分佈函數(log-pdf)做為與理想乾淨語音之差距，並做適當的回歸函數(regression function)將此種差距量化為語音評分，最後將求出的語音評分與實際由實驗者測出的評分做相關性的比對，以驗證此方法。

Model-based Non-Intrusive Objective Speech Quality Measurement Using Perceptual Parameters

Student: Shang-Ju Yu Advisor: Dr. Tai-Shih Chi

Department of Communication Engineering

National Chiao Tung University

Abstract

Assessing speech quality is an important issue in modern communication systems. The *subjective* speech quality measurements in early days involve much human resource and money such that the need of an *objective* speech quality measurement emerges. In addition, original speech signals are not always available when measuring speech quality in practical world. Many non-intrusive methods, which do not require original signals in judging the speech quality, are newly developed to meet this criterion. Such *non-intrusive* methods do not cost much human resource while being used for the real-time quality test with great efficiency.

The main theme of this work is to extract perceptual parameters from an auditory model, which mimics the signal processing principles in the human auditory pathway, and build an objective speech quality measurement without reference signal.

First, we propose a voice activity detector (VAD) algorithm by using the perceptual parameters from the auditory model. This VAD algorithm detects three basic categories in speech signals: voice, unvoice and inactive. Next, we acquire the auditory cepstral coefficients (ACC) to be the non-intrusive quality judging parameter. A Gaussian Mixture Model (GMM) is used to build the statistical template of the clean signal to represent the absent reference signal.

When measuring the quality of speech from different channels and codecs, the VAD is first utilized to distinguish distorted speech into three categories. Then, ACC parameters are extracted and compared to the statistical templates of the clean speech. The log-probability density function (log-pdf) is used to represent the distance between clean and degraded speech signals. Finally, a regression function is used to map the overall distances from those three categories to the subjective quality scores. The correlation between our objective measures and the subjective measures are examined to validate our approach.

致謝

首先，我要感謝我的指導教授冀泰石教授，讓我了解做研究的態度與方法，也因為教授的態度讓整個實驗室的風氣充滿和諧，並能在學習中發覺興趣和成就，從教授身上看到對研究的執著與熱情以及對學生生活上與課業上的關愛。與教授相處兩年，無論是在研究知識上與人生態度皆獲益良多。

其次，我要謝謝指導過我的學長，包括語音實驗室的江振宇學長及楊智合學長，感謝他們讓我對語音處理有所了解,並感謝指導程式上的困難與解決方法,另外也感謝 922 室的許學長，一起和你討論數學及統計上的概念是很有啟發的過程。

也要謝謝實驗室的同學及學弟們，一起相處兩年，從你們身上看到自己很多沒有的優點，也體會到自己應改的缺點和習慣，有了你們讓我研究生活多采多姿。

最後要謝謝我的家人，提供我生活上的需求，能讓我專心於自己研究上面，學習上不曾給我壓力並鼓勵我吸取多方面的知識。

目錄

表目錄

圖例目錄

第一章 緒論

1.1 研究動機

語音品質測量(Speech Quality Measurement)在現今有線(Wire)或無線(Wireless)通話中均扮演著重要的腳色。為了能夠達到即時且有效率的測量，發展客觀式(Objective)的評估成為研究人員們長期以來的目標。客觀式評估的優點在於不需實際的受測者去給予語音品質評分，而是藉由一套演算法機制來取代人的判斷，早期的做法是將污染的語音訊號(degraded signal)與原始的參照訊號(reference signal)做比對，然而實際即時通話中，受話端通常只能得到污染的語音訊號而難以得到原始訊號，為了克服此問題，近年來研究人員紛紛著手於非侵入式的評斷系統之研究(Non-Intrusive base, Output base or No Reference base)(或稱之為無對照式的評斷)[1]。

影響語音品質測量的因素極多，包含參數選擇、不同傳輸通道、不同編碼技術、以及發話端的背景音等。在接收端的非侵入式客觀品質評分則需要考慮以上各因素，才能使評分結果與人的判斷結果相近。

本論文概念是緣於一最原始及最準確的語音品質評測是來自於人耳聽覺，因此在模擬人耳聽覺系統的模型[2][25]中抽取的參數，將可自然地做為評斷語音訊號破壞程度的最基本元素。並測試及驗證聽覺參數是否能符合判斷語音品質的功能，並與其他參數做比較，探討相對的缺失及優點，進而使以後欲使用聽覺模型的使用者了解本論文參數的優點及所該補足之處。

1.2 研究背景

1.2.1 評分機制

真實的語音品質測量，是將錄製的語音訊號撥放給一群人聆聽，並由這群聽者給予此聲音的品質好壞評分[3][4]。最常用的評分機制為「平均意見分數」(Mean Opinion Score，MOS)，藉由每位評分者之「絕對分類計分」(Absolute Category Ratings，ACR)的平均值來得到，在絕對分類計分(ACR)中，聽者直接聽取受汙染之語音訊號(degraded signal)並直接給予評分，分數為一分到五分無小數點之分數;另有「衰減平均意見分數測試」(Degraded Mean Opinion Score，DMOS)，在此測試中，聽者在聽取受汙染之語音訊號前先聽取無雜訊的原音，再經由將每位評分者之「衰減分類計分」(Degradation Category Ratings)取平均值求得，此測試方式通常用於音響(Audio)之品質評測。另有「比較分類計分」(Comparison Category Ratings，CCR)的測試方式，相較於前者「衰減分類計分」(DCR)，此評分同樣是將後者語音訊號比較前者語音訊號，不同處在於「比較分類計分」(CCR)中，聽者第一次聽到的可能是有雜訊語音，亦即前後無噪音或有噪音之語音，撥放次序並無固定。[5][6]

Grade	ACR(MOS)	DCR(DMOS)	Grade	CCR
			3	Much Better
5	Excellent	Inaudible	2	Better
4	Good	Audible, but not annoying	1	Slightly Better
3	Fair	Slightly annoying	0	About the Same
2	Poor	Annoying	-1	Slightly Worse
1	Bad	Very annoying	-2	Worse
			-3	Much Worse

表 1-1 評分用語及涵義

語音品質測試實際上包含：收聽品質測試(Listening Quality Test)[3][4][5][6]、音響品質測試(Audio Quality Test)[4][8][9]、及對話品質測試(Conversation Quality Test)[4][7]。本論文主要探討最為一般狀況的收聽品質測試。

1.2.2 評分模型分類

依照不同的分類方式，語音品質測量模型可分為以下幾種

一、評測品質者為「人」或「電腦演算法」:

(1) 主觀式評測(subjective test):其優點在於得到分數為人真實的平均意見，因此在受測者數量多的情況下最準確。其缺點在於耗時(time-consuming)且需要大量金錢確定錄音設備及錄音環境的穩定。

(2) 客觀式評測(objective test):其優點在於有效率，且不需實際聽者去評分，也避免了聽者所處的接收環境之影響，更切合實際上的運用及需求。其缺點在於，客觀式演算法尚無法完全模擬人的判斷行為。此外，客觀式評測所能判斷的狀況不一，有時需視應用的要求做適時的修正。

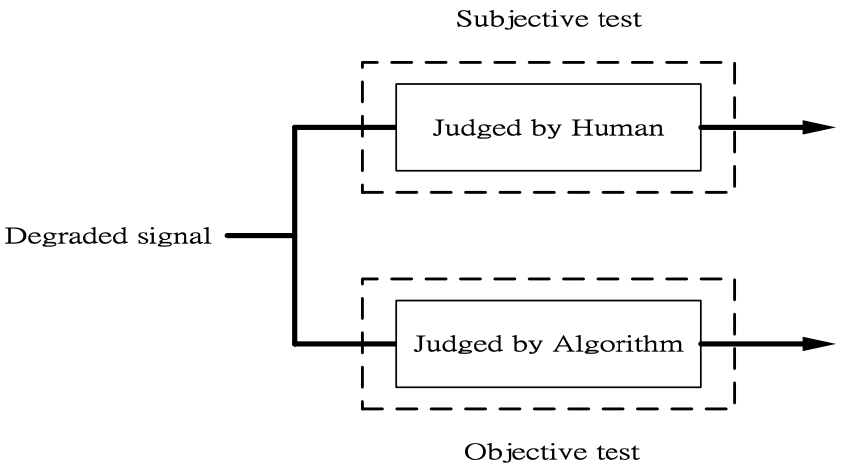


圖 1.1

二、評測品質時「是否有原始語音訊號(無汙染訊號)」:

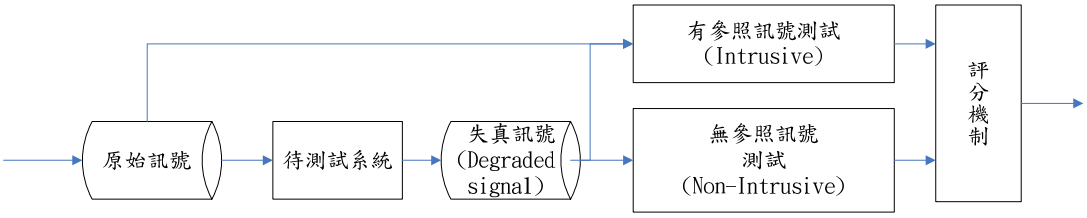


圖 1.2 Intrusive & Non-Intrusive Methods.

(1) 有參照(原始)訊號狀況下之測量(Intrusive Method, Reference base, Input-Output Base):

此為最基本的客觀式評測方法，由於有對照訊號，因此專家們著重於語音參數的選取與評分方式，然而此方法較不符合實際需求，例如評測語音品質的環境不一與遠距離通訊時接收端不易取得原始訊號，使得此方法對於新的通訊技術(ex.VoIP[12])不見得適用。

(2) 無參照訊號狀況下之測量(Non-Intrusive Method, Non-Reference base, Output Base):

無參照訊號狀況下之測量，顧名思義測量語音品質時沒有原始訊號，優點在於運用上較實際，在演算法設計良好之情況下可做及時的品質監測，缺點在於設計上較困難，且難以完全應付不同噪音狀況，又可分為以下兩種:

a. 模型式(Model base, black box approach):

主要的概念為選取適當的參數後，建立統計或機率式的乾淨語音模型(ex. 隱藏式馬可夫模型(HMM)[10]、高斯混合模型(GMM)[11])。當接受汙染的語音(degraded speech signal, signal after channel)後，擷取其中參數並與資料庫中乾淨語音比對出差異，由此差異性來判別品質好壞

b. 參數式(parametric base, glass box approach):

與上者最大的差別在於，參數式不做資料庫模型的訓練，直接由受汙染語音中，擷取認為造成影響語音品質因素的參數，直接由此參數作評分。優點在於本方法較直覺，但缺點是影響人對語音評判好壞的因素眾多(ex. Suddenly mute, harmonicity, naturalness, intelligibility…)，因此，目前仍有許多研究著重於發掘影響語音品質的因素。

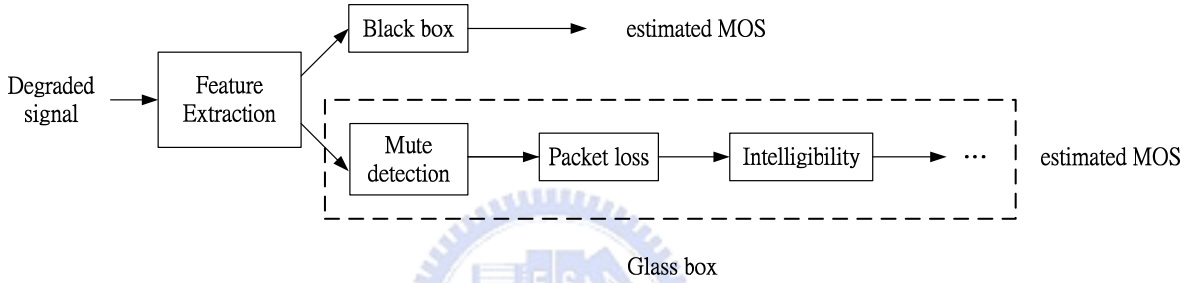


圖 1.3 Black box & Glass box

· 1.3 前人相關研究

有參照(原始)訊號狀況下之測量(Intrusive Method):

最早期為利用波形比對(waveform-comparison algorithm)來決定語音品質好壞,例如:訊雜比(SNR)、分割訊雜比(SSNR: segmental signal to ratio)，這些方式運算複雜度較低，且演算法簡單，但缺點是人的判斷並單純非是由訊雜比的關係來給予高低分，此外，此種方式面對不同編碼或通道失真(distortion)無法做有效率的辨別。頻域上(Frequency-domain)的判斷[13]，例如 IS 評測(Itakura-Saito measure)及頻域失真判斷(SD-Spectral Distortion)接著被廣泛運用。然而，僅考慮頻域上的判斷方式對訊號在時間上的平移(shift)非常敏感，但在人的感知上卻未必能被察覺，雖然後來有結合時間軸上的波形比對與頻域上的判斷，但越來越多人利用感知上的模型來做為模擬人的判斷機制中最基本元件。

目前已有大量的感知方法運用於有參照訊號下之測量，這些方法結合了人類的感知系統(Perceptual System),包含基本的聽覺生理與心理現象。除了九零年初期的 BSD(bark spectral distortion)[14]，以及感知語音品質測量(PSQM)[15]，九零年末期更有人提出了 MNB(Measuring normalizing block)[16][17]的方法，直到西元 2001 年國際電信聯盟標準化部門(ITU-T)提出了標準 PESQ[18](Perceptual Evaluation of Speech Quality)，亦是建構於人類的基本聽覺感知系統之上。

無參照訊號狀況下之測量(Non-Intrusive Method)

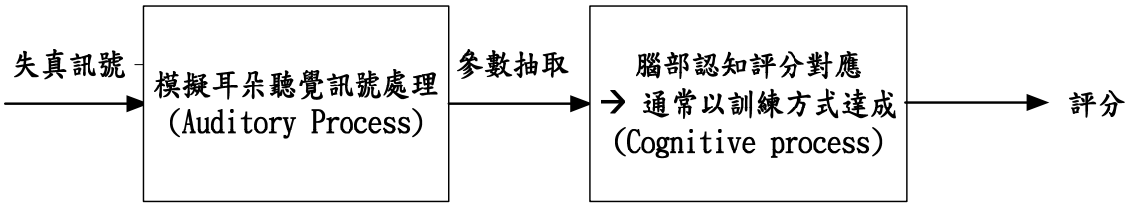


圖 1.4 無參照式概念圖

無參照訊號狀況下的測量方法，從九零年代末期就有許多人對此方法研究躍躍欲試，尤其是網路通訊的運用，無論無線(wireless)傳輸或有線(wire)傳輸[19]，聲音在網路傳輸的品質逐漸引起關注。在 P.Gray 與 M.Hollier 所提出的方法中[20]，為了使失真訊號有比對的依據，他們利用了口腔模型(Vocal-Track Model)來分析語音;此外亦有研究人員利用感知線性預估參數(PLP Coefficients)來做語音訊號品質測量時的基本參數[21]。如圖 1.4 所示，現今此種測量方法大多在第一部分以訊號處理的方式模擬聽覺效應，而第二部分則是將處理過的參數做比對，比對的資料通常由大量的乾淨語音作訓練而得。

近來，已有人使用高斯混和模型(Gaussian Mixture Model)[22]及隱藏式馬可夫模型(Hidden Markov Model)做為評分時的比對模型，亦有許多較新穎的參數式測量(parametric base)[23]方式誕生。

1.4 本論文研究方向及概要

本論文研究方向主要為「無參照訊號下模型式語音品質測量」，而參數的擷取主要利用到 NSLtool[24]，利用 NSLtool 將訊號作處理後模擬耳朵接受到的訊號，再將接收到的訊號轉成頻譜圖(Spectrogram)，由其中求取倒頻譜係數(Cepstral Coefficients)，除了驗證此係數與梅爾倒頻譜係數效能差別(MFCC)，並與含有正確 MOS 值的語音資料庫作簡單的關係比對，如圖 1.5 所示。

本論文主要將整段語音分為三大部分，母音(voice)、子音(unvoice)、及無聲(inactive)，在此以 NSLtool 求出之頻譜圖(Spectrogram)，依照不同的音框在不同密度的頻域濾波庫(frequency filter bank)解析之能量，來判別是母音、子音、或無聲。語音品質判斷流程圖如下(詳細說明及參數設定會在後面章節提及):

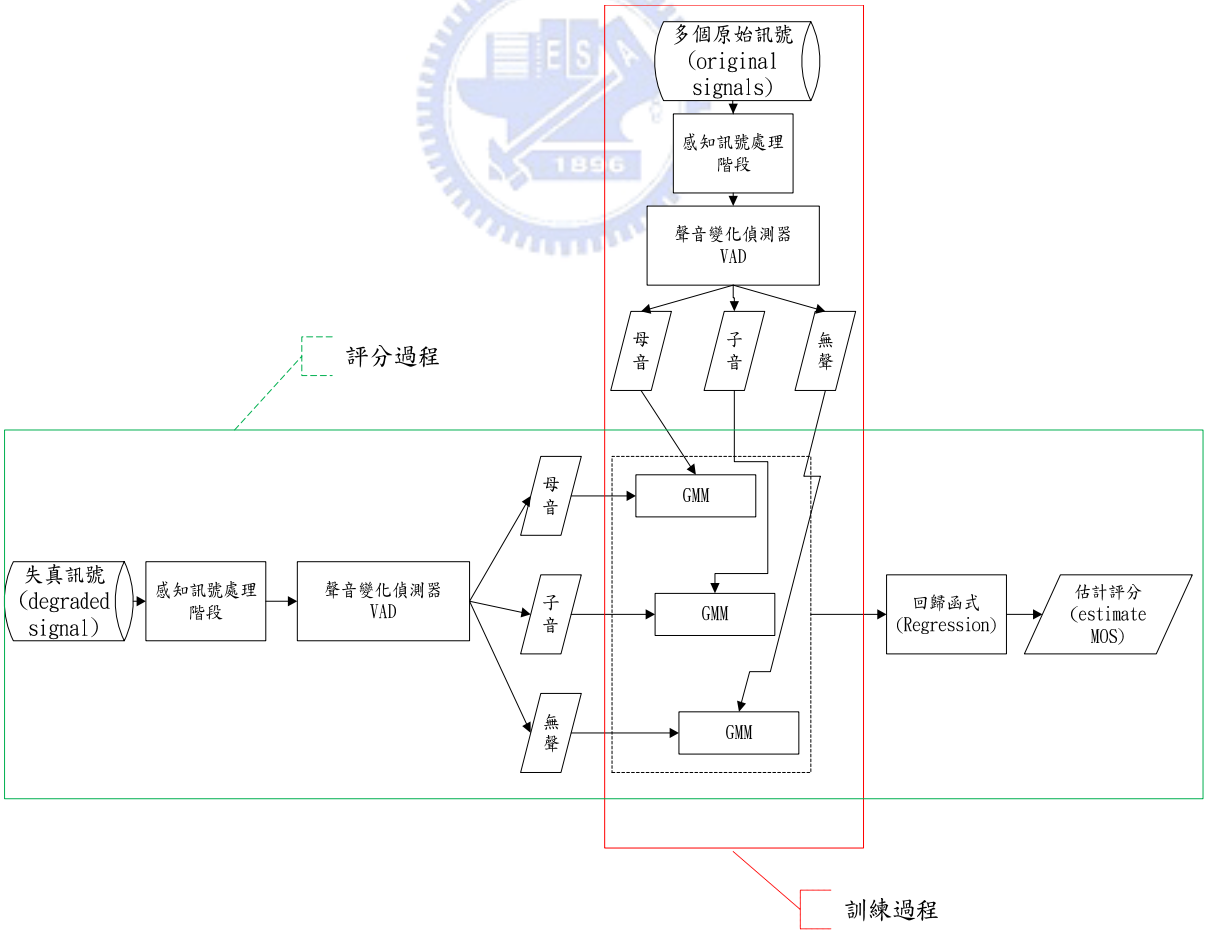


圖 1.5 論文流程圖

1.5 章節概要說明

本論文第二章將闡述感知訊號的背景，包括 NSLtool 的說明及運用，在第二章後半部，將說明擷取參數的過程及觀察，由觀察中了解參數之特性與差異。第三章中，將利用到部分在第二章中的觀察，做為聲音變化偵測器的辨認參數，並比較不同方式下，聲音變化偵測器的效能。第四章，則利用高斯混和模型(GMM)將乾淨的語音由第二章討論的參數作訓練，並描述訓練流程。第五章，前半部是做一侵入式測量，後半部則做非侵入式測量，以前面 2~4 章為基礎，利用基本的回歸函式與實際的平均意見分數(MOS)作比較。最後，在第六章，探討參數的可用性比較結果以及未來建議的改進方法。

第二章 感知訊號背景與參數選擇

2.1 聽覺頻譜圖

聽覺模型[2][25]

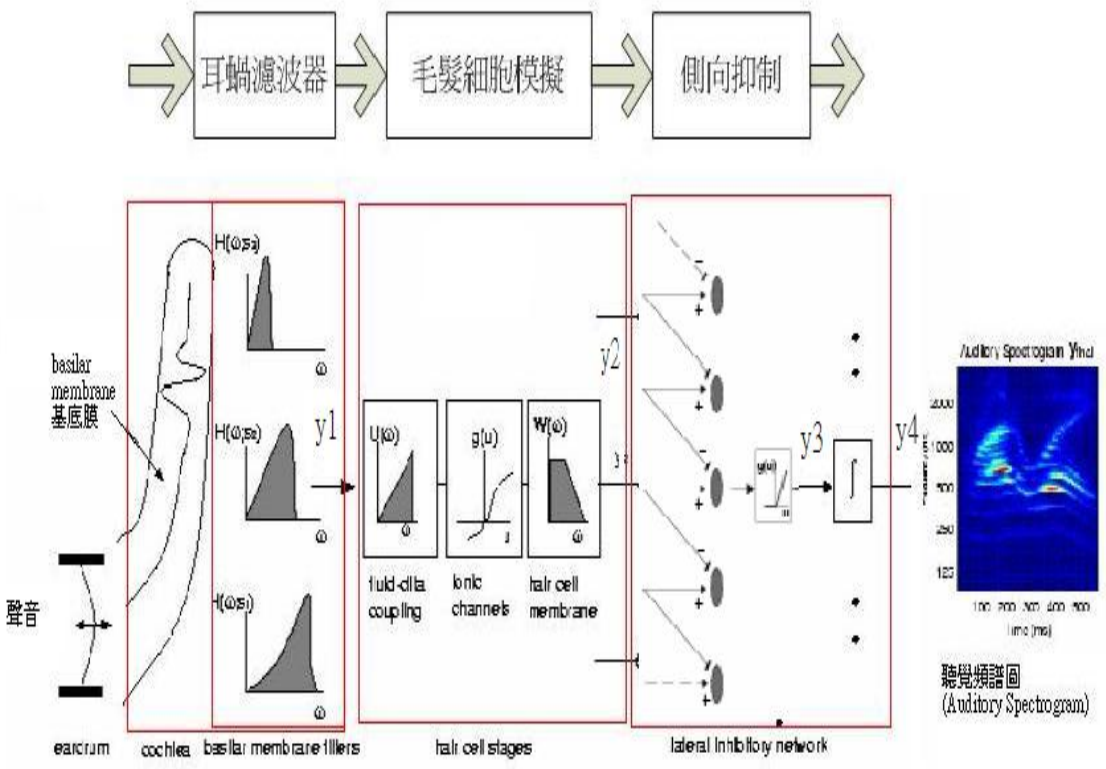


圖 2.1 聽覺感知模型之模擬與流程

上圖為 NSLtool[24]的第一階段聽覺模型流程，使用函式為 wav2aud，主要功能為模擬人耳接收到的訊號所顯示的*聽覺頻譜圖*(Auditory Spectrogram)。以下將圖 2.1 分三個階段介紹:

2.1.1 耳蝸濾波器組階段(cochlea filter bank stage)

2.1.2 毛髮細胞模擬階段(hair cell stage)

2.1.3 側向抑制網路階段(lateral inhibitory network stage)

· 2.1.1 耳蝸濾波器組(cochlea filter bank)

生理現象

聲音以空氣為媒介傳達至外耳殼，經過耳道至中耳之鼓膜，利用空氣壓力原理震動鼓膜，動能經鼓膜轉至三小聽骨做一能量緩衝的作用，再由三小聽骨震動卵圓窗，將能量轉換成液壓，而內耳中的耳蝸即是由充滿淋巴液的空腔所形成。

耳蝸外型為一個蝸牛式的捲曲狀，若我們將其拉直為一直線來觀察，可發現耳蝸內含一基底膜(Basilar membrane)，淋巴液受液態壓力的震動在基底膜上形成一行進波(Traveling wave)，行進波依液壓震動的不同，在基底膜不同的位置上形成振幅大小不同的共振，形成此現象的原因為，基底膜的厚度及寬窄不同，基底膜的頂部厚度較厚且寬，而基底膜的底部較扁且薄，因此當外耳接收到的聲音，將空氣動能轉換成機械動能(三小聽骨受鼓膜推擠而敲動)並轉換為液壓動能(淋巴液的流動)，此能量的轉換可視為對耳朵的保護，並將能量逐層削減避免脆弱的神經受刺激過大而受傷，而基底膜的構造，造成不同的頻率在基底膜不同的位置上共振，基底膜內部之細胞會將此訊息傳至大腦，此階段做了細胞分工，亦作了分頻解析的動作，期共振頻率之範圍由 20hz 至 20000hz，而此即人類聽覺頻率範圍。

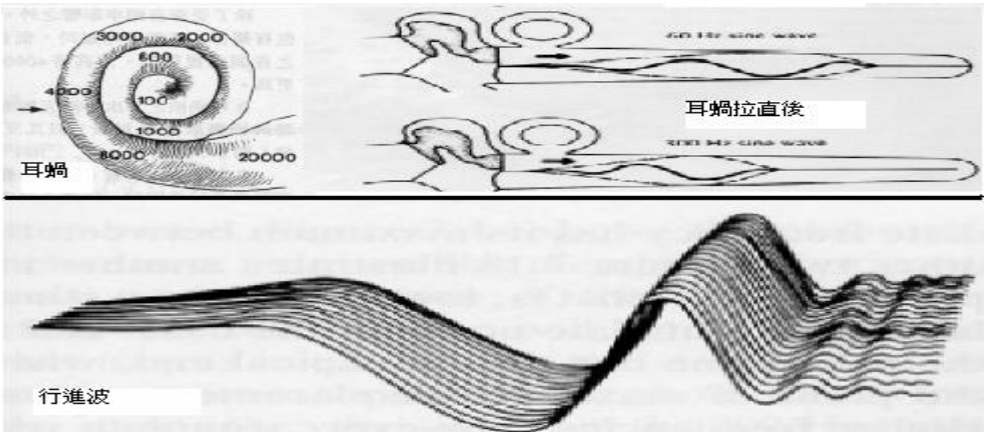


圖 2.2 耳蝸與行進波(traveling wave)的示意圖

程式模擬

$$y_{coch}(t,x) = s(x) * h(t,x) \qquad (2\text{-}1)$$

式(2-1)為模擬耳蝸濾波器之函式，其中 s(t)為收到的語音訊號，並與 h(x,t)作 convolution 之動作，其中 t 為時間，x 為特定的頻帶(可視為基底膜上的頻率共振位置)，耳蝸濾波器 h(x,t)是以一濾波組(filter bank, 或稱濾波庫)來模擬，濾波器的中心頻率在對數頻率(log-probability)上呈線性分佈，共有 128 個不同解析度的濾波器在此濾波組中，並設計為一倍頻中有 24 組濾波器(24 filters/octave)。為了使不同取樣頻率的訊號可用，濾波組之中心頻率可變，濾波組中的每個濾波器符合 Q-常數定律(Constant Q)，亦即中心頻率與頻寬比例成一定值，如式(2-2)所示:

$$central\_frequency / bandwidth = Q(cons\tan t) \qquad (2\text{-}2)$$

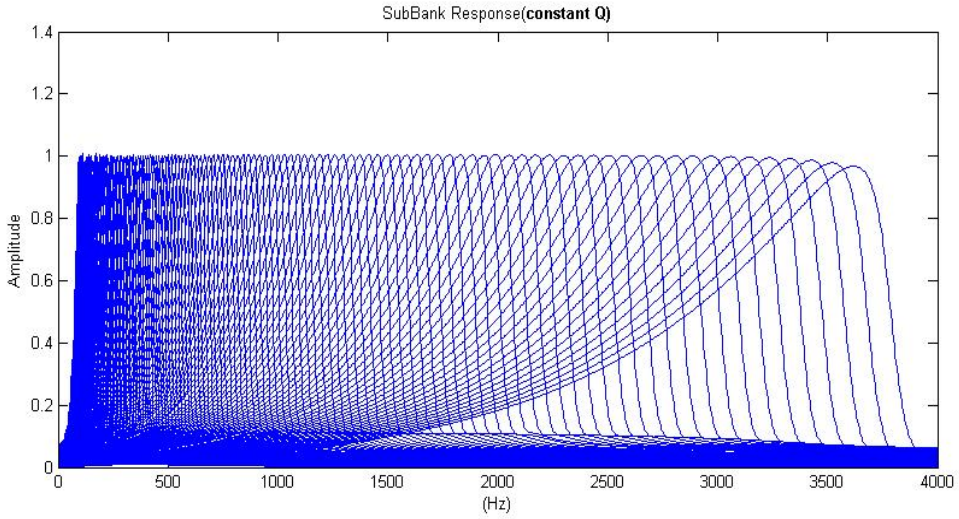


圖 2.3 模擬耳蝸的濾波組:利用濾波組(filter bank)模擬耳蝸的濾波器組，每一濾波器皆符合 Q-常數定律

· 2.1.2 毛髮細胞模擬階段(hair cell stage)

生理現象

基底膜上含有許多的毛髮細胞(Hair cells)，毛髮細胞在未受刺激時即隨時發射電位訊號，當基底膜受到共振時，基底膜拉扯毛髮細胞，造成毛髮細胞發射離子電位訊號之速率(firing rate)提高，以此方式將該基底膜位置的頻率訊息傳至腦部。在此一階段，可視為液壓動能轉換為電位能(神經訊號)之過程，毛髮細胞又分為內毛細胞(Inner hair cells)和外毛細胞(Outer hair cells)，內毛細胞與多個神經纖維連結，當液壓動能拉扯其毛髮時，刺激了內毛細胞所接觸的神經纖維而形成電位；外毛細胞較長，具有避免基底膜共振過大的保護作用，對耳蝸的調節與保護機制有關。

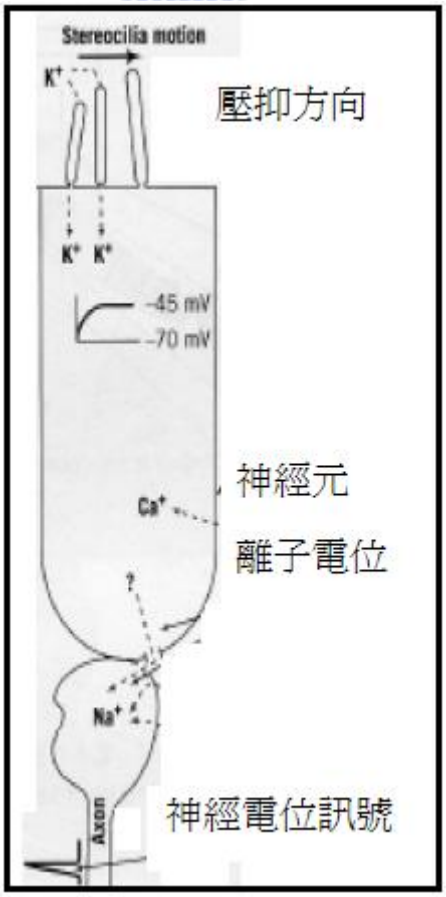


圖 2.4 毛髮細胞示意圖:毛髮細胞受液壓的移動導致離子電位差的產生，由液壓的機械移動轉換成神經細胞訊號之過程

程式模擬

$$y_2 = g(\partial y_1(t,x)) * w(t) \qquad (2\text{-}3)$$

式(2-3)模擬將液壓轉換成神經電位的過程，先經過一高頻濾波器來模擬液壓的變化，而由於能量的轉換非線性，因此，之後再用一 sigmoid 的函式($g(u)=1/(1+e^{-u})$)來表達此非線性之能量轉換，同時也針對毛髮細胞的保護作用(神經電位的發射率，不會線性上昇，而會有所飽和)做了模擬，最後再以一低通濾波器做一平滑化效果，來模擬神經的最高發射速率。

· 2.1.3 側向抑制網路階段(lateral inhibitory network stage)

生理現象

毛髮細胞(Hair cells)的最高神經發射速率為 4000hz，而中腦的神經接收速率大約為 1000hz 左右，毛髮細胞對應鄰近的毛髮細胞(基底膜相鄰的頻帶)會有壓抑的現象，在實驗中所測得的中腦反應速率較慢，因此在聽覺神經傳達訊息至中腦階段，可視為一平滑化(smoothing)的過程，讓聲音訊息簡化並使腦部接收的訊息更健全(robust)。

程式模擬

$$y_3(t,x) = \max(\partial y_2(t,x), 0) \qquad (2\text{-}4)$$

$$y_4(t,x) = y_3(t,x) * \mu(t,\tau) \qquad (2\text{-}5)$$

式(2-4)及(2-5)模擬毛髮細胞將訊息傳達至中腦時的簡化現象，(2-4)中說明基底膜某區域受激共振時，鄰近區域的毛髮細胞會受壓抑，此現象與聲音心理學中人耳頻率的遮蔽效應相符，(2-4)中以一階的對頻率微分來模擬此現象；(2-4)後半段取出信號之包跡(envelope)再結合(2-5)之時域上的積分，來模擬神經訊號傳至中腦平滑化的過

程，其中 $\mu(t;\tau) = e^{-t/\tau} \cdot u(t)$，u(t)為步階函數(unit-step function)，τ 為時間常數(time-constant)，一般以 8ms 來對應 1000hz 之神經發射速率，在 NSLtool 中可視情況做 1ms~128ms 的調整。

經過『2.1.1』、『2.1.2』、『2.1.3』所提之三步驟後，我們可得到一聽覺頻譜圖(Auditory Spectrogram)，頻譜圖中橫軸為時間，縱軸為頻率，由於人對頻率的感覺為對數線性(log-linear)關係，因此縱軸的頻率以倍頻的方式顯示。

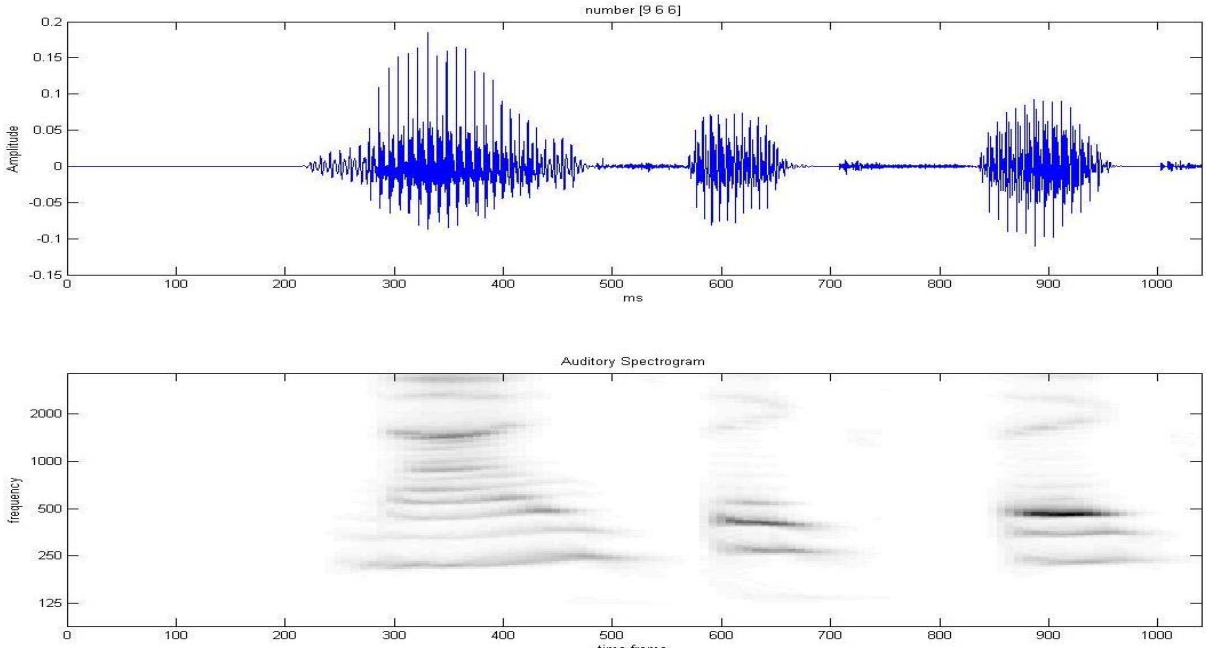


圖 2.5 聽覺頻譜圖

圖 2.5 聽覺頻譜圖表示時域軸上的聲音訊號，及所對應的聽覺頻譜圖(Auditory Spectrogram)，圖中所說的字為 nine-six-six，x-軸為時間，單位為 ms，y-軸為頻率軸，等間隔為倍頻的頻率區間(人耳對頻率辨析之感知為對數線性關係)，由時間軸及聽覺頻譜圖皆可看出，「nine」發音時間較長，且能量較大(時間軸上的振幅較大且聽覺頻譜圖中顏色較鮮豔)。

2.2 腦部感知階段

聽覺頻譜圖屬於聽覺系統中，較前級之階段，主要模擬耳朵的聽覺生理及心理現象，而腦部感知階段，則屬於後級之階段，作用為將耳朵到中腦的資訊做解析。

生物之實驗中，以探針植入哺乳鼠類，並以不同的單頻聲音給予生物聽覺之刺激，觀察中發現不同的大腦皮質位置有不同的反應(response)，某些大腦區域對聲音時間的變化反應快，某些區域對聲音訊號中頻域的變化反應快，相對地，亦有某些區域對聲音訊號的時間變化或頻率變化有較慢的反應，有鑑於此，生物的基本反應可以用兩維度的基底表示式代表，分別為訊號的時間變化反應速率，及訊號的頻率變化反應速率。

經由以上的實驗，發現大腦的皮質細胞反應，可以用時間變化(rate)和頻域變化(scale)的不同來模擬，如下圖所示:

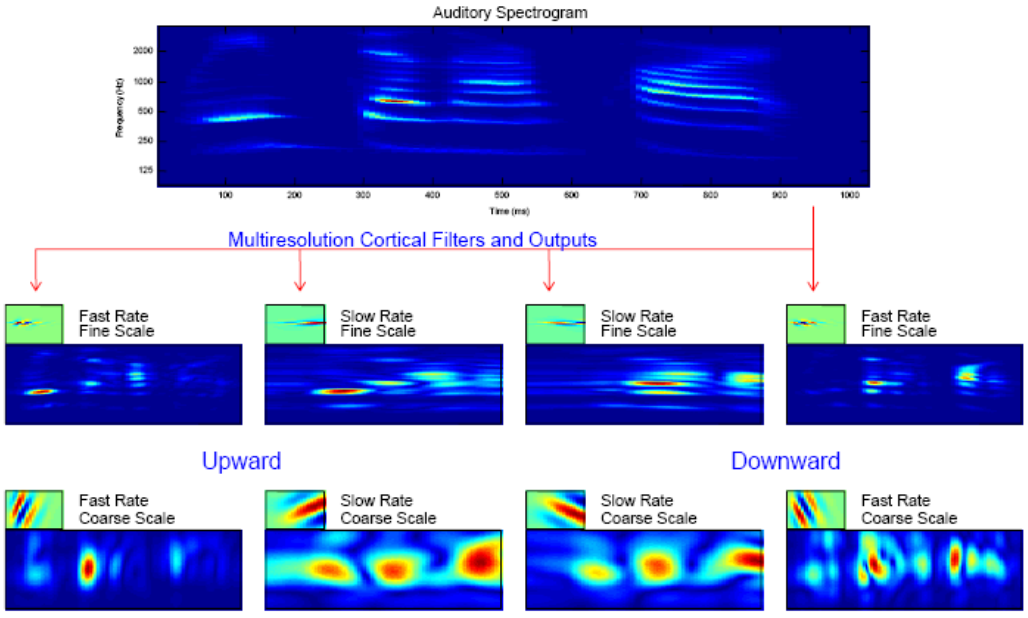


圖 2.6 不同時域變化(rate)及頻域變化(scale)解析圖

由圖 2.6 中可發現，原聽覺頻譜圖經過不同 rate-scale 的二維解析後，各有不同的反應(意同聲音訊號經過不同大腦皮質區得到不同的反應)，稱之為 STRF (Spectro-Temporal response field，頻域-時域變化反應區)，在此我們可視為不同的 rate 及不同的 scale 為大腦區域索引值，並對至不同的大腦皮質區，在 NSLtool 中，是使用 aud2cors 函式，函式中以聽覺頻譜圖為輸入，得到四維度的輸出(時域(time)-頻域(frequency)-時間變化(rate)-頻域變化(scale))，我們可設定函式中的 rate 及 scale 值大小，rate 值越大，越可由 STRF 中看見時間軸上能量變化，同樣地，scale 值越大，越可由反應 STRF 中看見頻率軸上能量變化。

· 2.3 參數抽取

上節介紹完聽覺頻譜圖的(Auditory Spectrogram)的原理後，本節將介紹，如何從聽覺頻譜圖抽取出我們想要的參數，並加以利用，由 1-4 之圖 1.6 可知，流程中說明，在作參數抽取後，將由參數作基本的聲音變化偵測器(Voice Activity Detector)，此外在訓練高斯混和模型(Gaussian Mixture Model, GMM)時亦須要一組參數來做比對。因此，在參數抽取的部分，可分為兩大類:

(1) 語音模型訓練及判別語音品質之參數。

(2) 聲音變化偵測器的參數。

· 2.3.1 正規化(Normalization)

由於不同語料庫的音檔長短及能量不同，因此在任何語音進入參數抽取步驟前必須正規化(Normalization)，以避免因為不同音檔中不同能量大小導致抽取之參數無法以同一標準做比較，本論文以高斯正規化使程式讀取語音檔後形成平均值為 0 變異數為 1 的高斯分佈，式子如下(使用程式為 unitseq):

$$X' = \frac{X - mean(X)}{std(X)} \qquad (2\text{-}6)$$

X 原始資料， X’ 正規化後之資料。

· 2.3.2 判別語音品質之參數

在判別語音品質參數的選擇上，自以往有很多種不同的選擇，例如，梅爾倒頻譜參數(Mel Frequency Cepstral Coefficients)、PLP 參數(Perceptaul Linear Prediction Coefficients[26])…，而這些參數的共同點為，皆對人耳前端的反應做了基本的模擬，並用離散的逆傅立業轉換(IDFT)或餘旋轉換(DCT)，將頻譜端的波形與人耳特性結合並取出，以下作 MFCC, PLP, ACC 的流程比較:

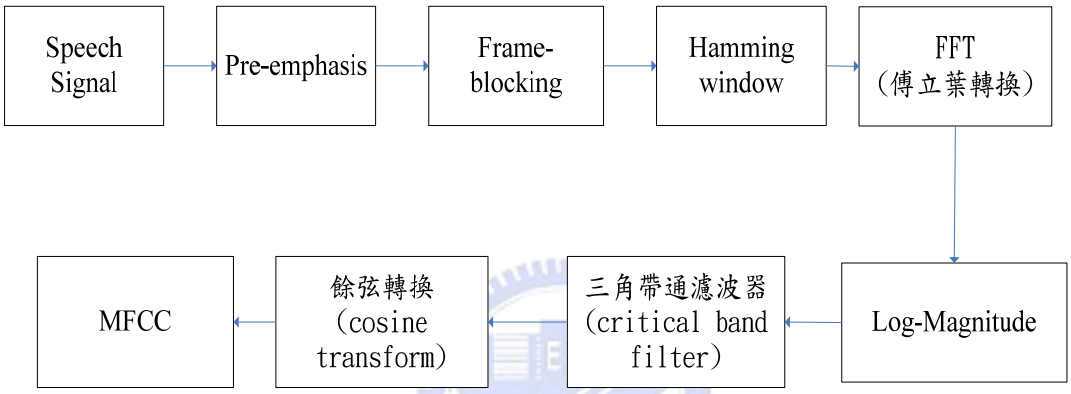


圖 2.7 MFCC 流程

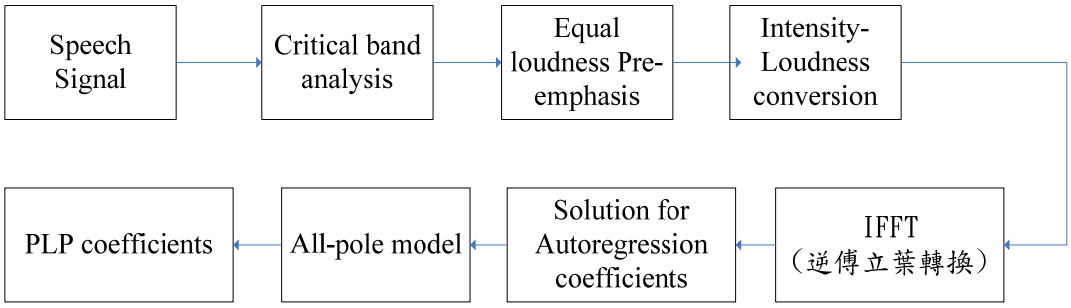


圖 2.8 PLP 流程

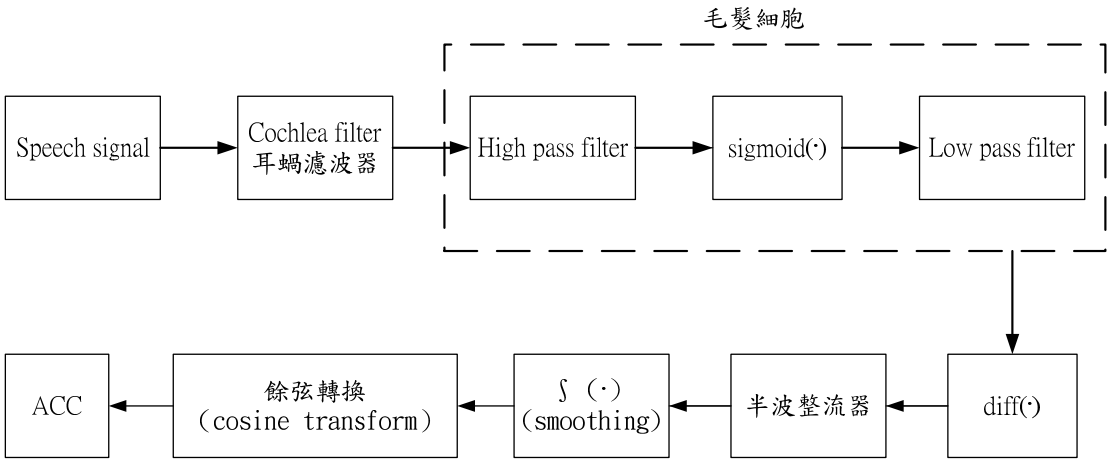


圖 2.9 ACC 流程

由 2.1 節我們已知聽覺頻譜圖所包含的聽覺資訊，而由上頁可知，聽覺倒頻譜參數(ACC, Auditory Cepstral Coefficients)的擷取，其後面之步驟與 MFCC 的取法類似，即是用餘弦轉換(Cosine-Transform)來模擬語音能量的曲線變化，並且可以捉取語音諧波成分(harmonic parts)之變化。

· 2.3.3 聲音變化偵測器(VAD，Voice Activity Detector)之參數

在本論文中，聲音變化偵測器扮演重要的角色，除了一開始訓練乾淨語料時，須從乾淨語音中分出 ”母音(voice)” 、”子音(unvoice)” 、”無聲(inactive)” ，再加以訓練出乾淨之模型外，在接收到失真語料時，也需要經由聲音變化偵測器，分出”母音” ，”子音” ，” 無聲” 三類，在進入各自的乾淨模型測試。

聲音變化偵測器的參數，是由聽覺頻譜圖(Auditory Spectrogram)中抽出，事實上，我們若以肉眼辨識頻譜圖中的特徵，可由頻譜圖中頻率高低、諧波成分、發聲時間長短、與能量大小等方式來看出「發聲與否」和「子音或母音」。

實作方式上，為了能夠抽取語音的能量特徵，同時必須兼具能夠提出頻率的能量分佈特性，先將聽覺頻譜圖切割成多個時間音框(time-frame)，以每 16ms 取一個音框，

接著對每個音框作頻率的解析，將每個音框以不同「頻寬密度(scale)」的濾波器組(濾波庫，filter banks)做濾波的動作，不同濾波庫所得到的輸出會因「音框的語音特性(諧波能量的疏密、能量大小)」和「濾波庫的頻寬疏密」而有所不同，再將每個音框在不同濾波庫所解析出之能量作為一項特徵，如此，不同頻寬密度的濾波庫，就有不同的能量特徵，而每個音框就會有多個不同的能量特徵，形成一多維向量，此多維向量即為聲音變化偵測器之參數。

更多詳細的過程及驗證，在第三章中討論。

第三章 聲音變化偵測器(VAD)

3.1 基本語音分類

基本的語音分類，可分「發聲」、「無聲」，更可將發聲部分細分「子音」、「母音」，進階更可探究不同的母音及子音，本論文依最基本的分類，目標為利用聲音變化偵測器，分出「子音」、「母音」、「無聲」，使用的語料為 ITU-T. Supplement p.23[32]，並且以英文語料作為討論。

3.1.1 聲音變化偵測器(Voice Activity Detector)

聲音變化偵測器之功能:

聲音變化偵測器(Voice Activity Detector)運用於現今無線及有線傳輸中[27]，主要是利用無聲音的壓縮(silence compression)，使得傳輸中無須傳輸無聲音部分，並且將平均位元速率(average bit rate)降低，在特殊協定下的手機傳輸(e.g. cellular radio system under DTX mode, GSM or UMTS)，可以使得範圍內手機使用者增加並降低手機功率消耗(power consumption)，除此之外，在語音辨認系統(robust speech recognition)及助聽設備(hearing aids devices)中在抑制噪音(noise reduction)的需求下亦常常需要使用到聲音變化偵測器[28]。

以往聲音變化偵測器之偵測方法:

最常見的作法為將語音在時間軸上作切割，使語音形成多個音框(time-frames)，並根據「決策規則」(decision rules)將每個音框作判斷[29]，然而由於在訊雜比(SNR)極低的情況下，單一的決策機制往往使得噪音被判成語音，因此有許多研究著重於多維的判斷機制(Mix-decision base)[30]或是適應性(adaptive)的判斷規則[31]。

3.2 本論文聲音變化偵測器之概念

一般聲音變化偵測器(Voice Activity Detector，VAD)的功能，目的在於能夠清楚切割語音部分及非語音部分，而由於本論文研究將聲音分為「母音」、「子音」、「無聲」三部分，因此不只要切割出語音部分和非語音部分，更需要從語音部分分出「子音」及「母音」，而早期的聲音變化偵測器主要是利用能量門檻值(energy threshold)來判別語音及無聲，而子音與母音的分別必須從頻域上加以考慮，所以在參數上必須至少能代表:

(1)單位音框能量狀況。

(2)音框本身在頻域上的資訊。

為了滿足上述兩點，並且需要包涵人耳聽覺資訊，我們由聽覺頻譜圖(Auditory Spectrogram)作為判斷的第一階段，前一章已介紹聽覺頻譜圖較能代表人耳的聽覺狀態，在此前提下，聲音變化偵測器的功效應該更能代表人耳判別有聲及無聲的能力。詳細步驟如下:

<1>將聽覺頻譜圖做時間軸的切割，以 16ms 為單位不重疊的切割時間軸上的音框，一般聲音變化偵測器音框切割通常以 4ms~16ms 為單位，在驗證時切割時間以 8ms 為單位，為了配合 ACC(Auditory Cepstral Coeficients)(見章節 2.2)，實際使用以 16ms 為單位，圖示如下:

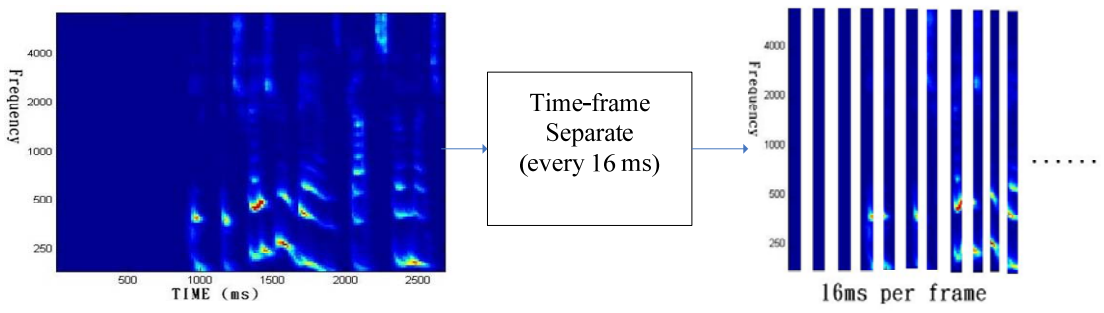


圖 3.1 時間軸切割示意圖

<2>將每個音框做頻域上的能量分析，以不同的 scale，解析音框之能量，在此可視為以不同頻寬密度的濾波器組(filter banks)將音框做濾波動作，亦可視為用不同頻域解析密度去解析單一音框在頻域上的分佈，在此使用到 NSLtool 中的 aud2cors 函式，函式中可輸入不同的 scale，來解析單一音框的原有頻率分佈。Scale 單位為(cycle/octave, 一倍頻中能解析出多少諧波能量, ex. Scale=2 指一倍頻中能解析出兩個諧波能量)，見圖 3-2。由圖中可發現，當 scale 越高(頻域解析密度能力越高)，則越能精細地描繪該音框在頻域軸上的變動。Scale 的選擇上，選擇了 23 維，matlab 表示式為 scale=2. ^(-1.4:0.2:3)，scale 超過 8 時會過度解析導致失真，因此 scale 最多只取到 8。

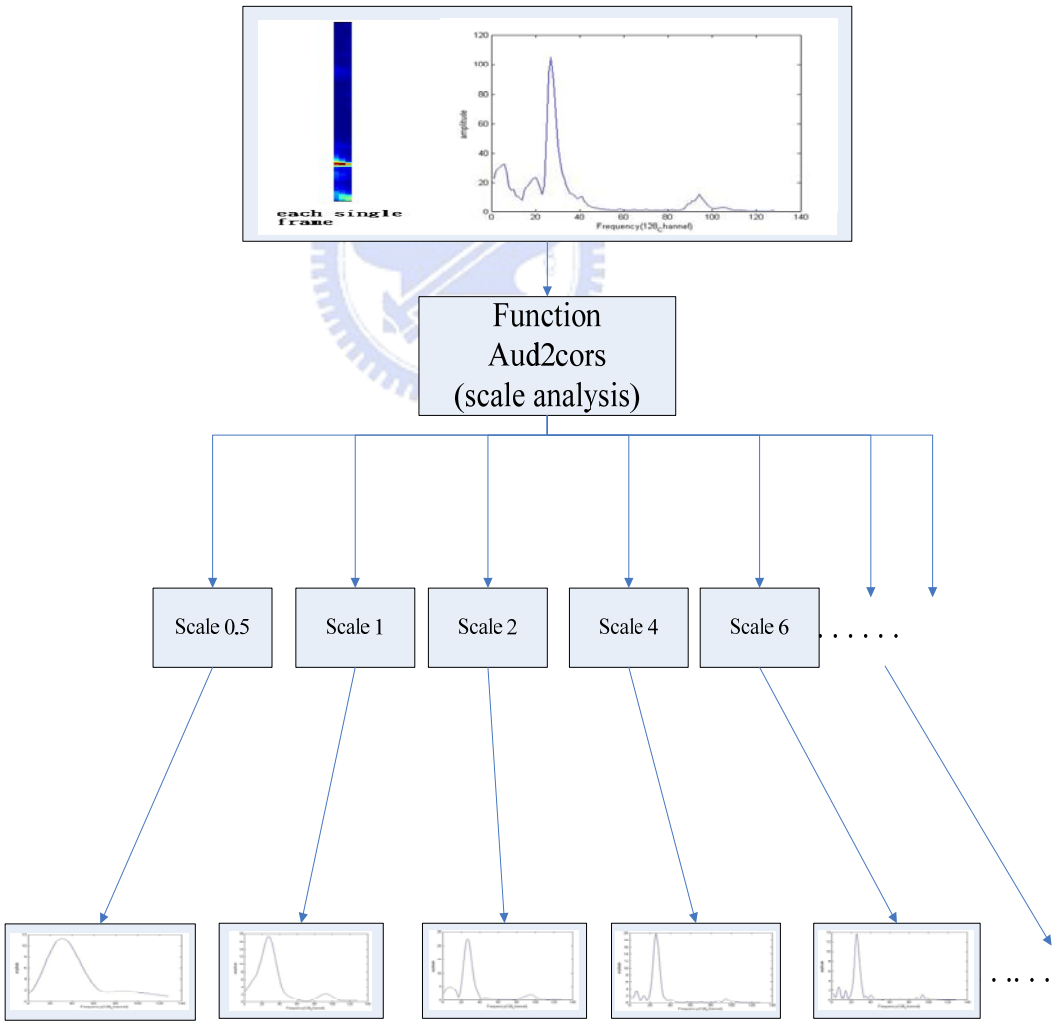


圖 3.2 頻率解析示意圖

<3>將不同 scale(頻域解析密度能力)所求出之頻域分佈圖，取絕對值，以解決虛數部分(imaginary parts)的問題，接著再取平均值，以求取該 scale 所解析出的平均振幅(可視為平均能量)，23 個 scale 所解析的頻域分佈圖此時變成 23 點，因此每個音框形成 23 維的向量，圖示如下:

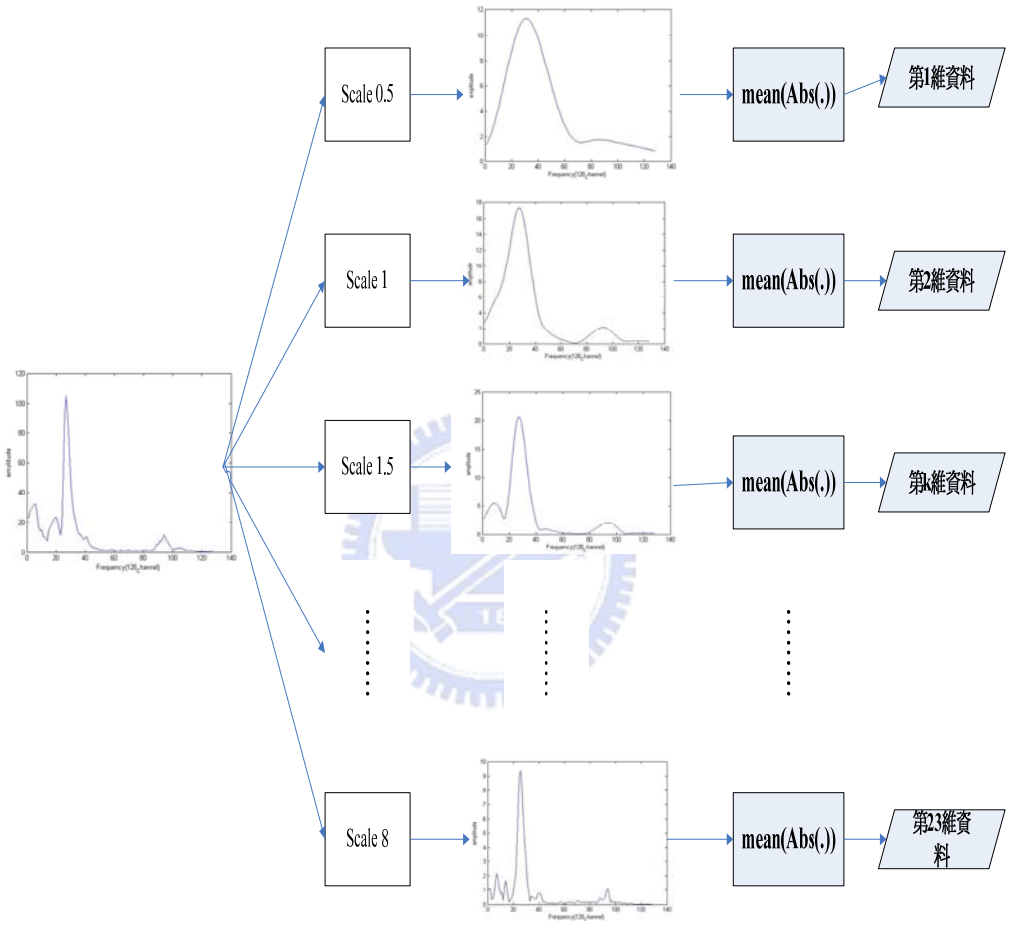


圖 3.3 VAD 參數抽取圖

<4>聽覺頻譜圖上的每一 16ms 之音框，經過上述三步驟的轉換，每一音框會形成一 23 維之向量，如下圖所示:

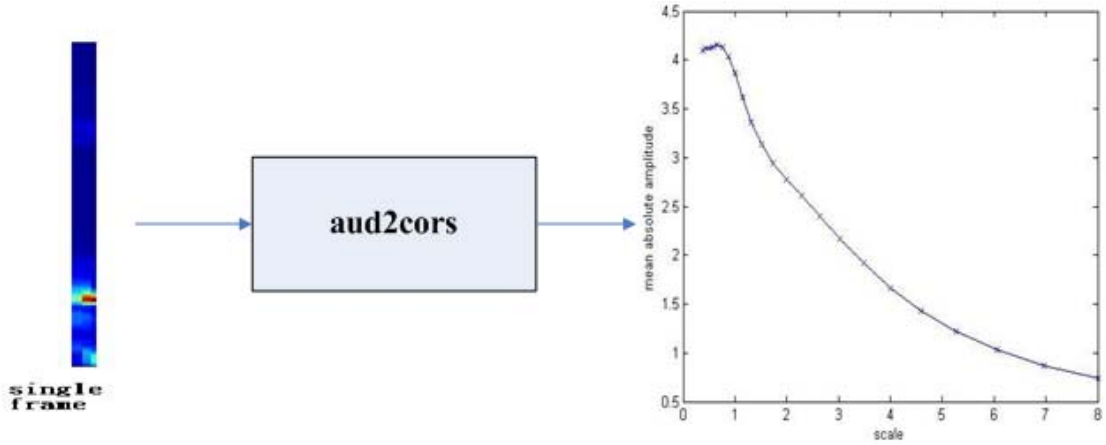


圖 3.4 音框與參數示意圖

此 23 維向量，除了包含了單一音框的能量情況，同時因為不同的頻率解析密度下，也包含了該音框的頻率分佈資訊。

▪ 3.3 參數觀察與設定

本章前兩節已說明了聲音變化偵測器的基本功能與用途，並探討了 23 維參數的運用原因及概念，本節將對 23 維的參數做基本的觀察與闡述，在語音品質評測中本論文是以三個類別來評分，本節即利用分出的三個類別來觀察參數的變化，以單一句乾淨的語音(時間長度為 8 秒)，先經過 2.2.1 之正規化(Normalization)的過程，再用手動的方式標示母音、子音、及無聲，觀察 23 維聲音變化偵測器的參數變化，接著，以不同訊雜比(SNR)的白噪音(white noise)作簡單測試，觀察三類別參數的變化程度。

參數設定:每一個音框(time frame)時間長度為 16ms，且無重疊部分(non-overlap)。加入白噪音(white noise)測試時，SNR=[-6 0 6 12 18 24]。

3.3.1 是母音音框的觀察，所畫圖形為所有母音音框*平均後*的圖形；3.3.2 是子音音框的觀察，所畫圖形是所有子音音框*平均後*之圖形；3.3.3 是無聲音框的觀察，所畫圖形是所有無聲音框*平均後*之圖形。3.3.4 任意單一音框之觀察。

- 3.3.1 母音音框觀察

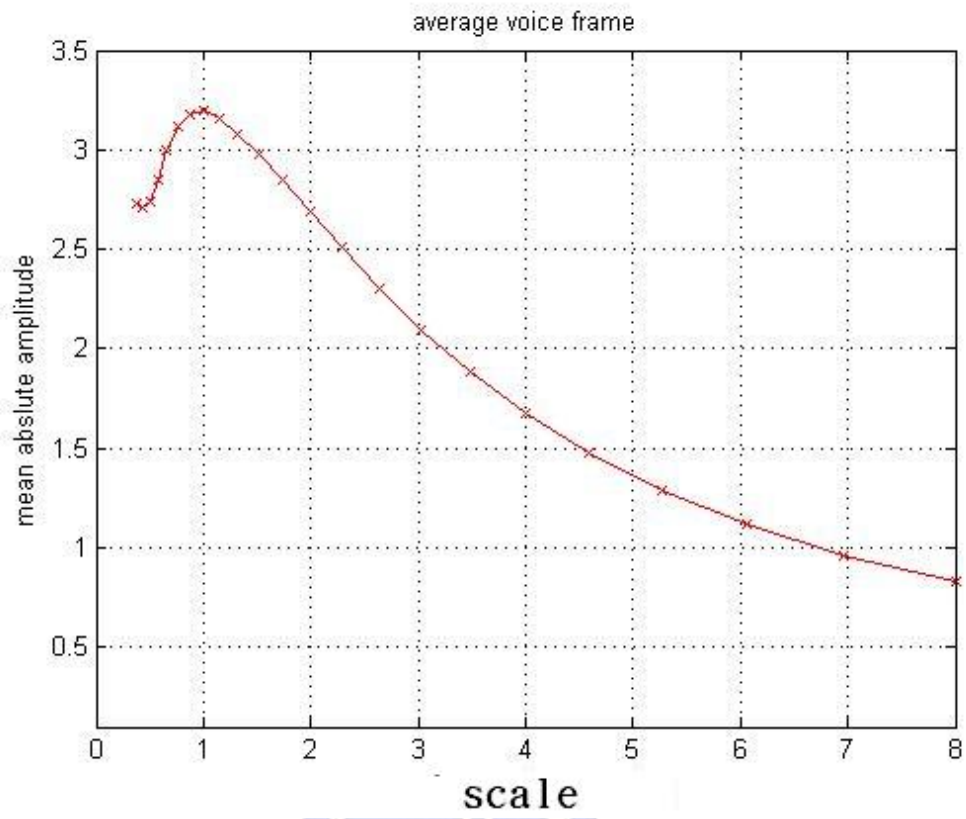


圖 3.5 平均乾淨的母音音框

由圖 3.5 可看出，在母音部分的 23 維 scale 向量中，其特徵是會有一個峰值(peak value)存在，第二個特徵為當 scale 在 1~4 之間時，其 y 軸之值(mean absolute amplitude, 在此我們可將其視為能量特徵)都超過 1，造成此現象的原因，是由於母音能量或諧波(harmonic)發生範圍多數在低頻，往往集中在頻率 1000hz 以下，且音框中諧波(harmonic)疏密在低頻時相較於子音音框之諧波較寬，因此當我們以低頻域解析密度去分析時就可捉取其大部分能量。

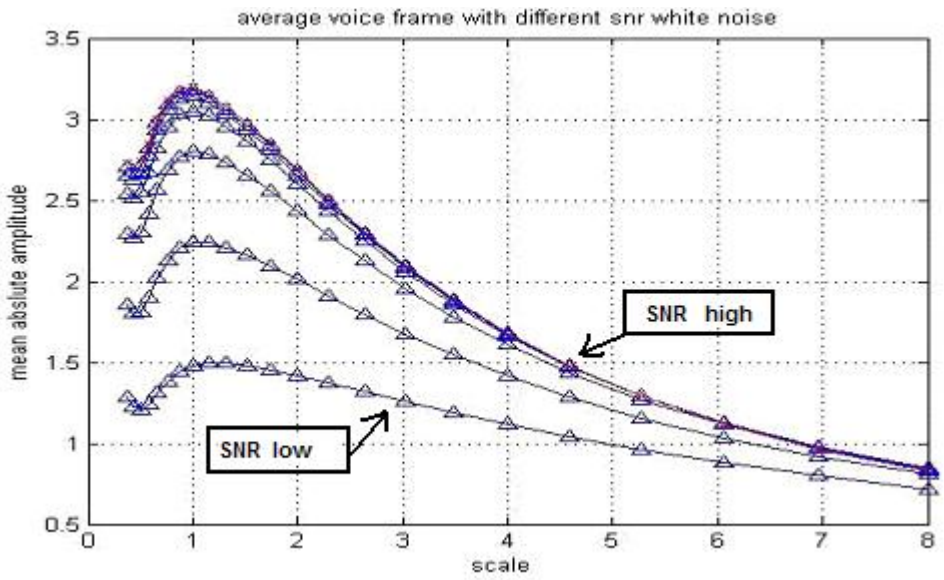


圖 3.6 不同訊雜比下之白噪音母音音框，紅色線(最上面一條粗體線)為乾淨母音所畫出之曲線

圖 3.6，顯示母音音框加入不同訊雜比(SNR)後的平均分佈，可看出隨著白噪音(white noise)的能量增加，曲線往下遞移，而曲線中最高值與最低值的坡度隨著訊雜比(SNR)降低而平緩，造成此現象的主因為，當雜訊比重在母音音框增加時，低頻解析出的諧波結構能量相對的降低，因此造成坡度平緩，此外，由於我們使用 2.2.1 節所提之高斯正規化，母音音框能量會稍有減少，所以圖形會有隨著訊雜比(SNR)降低而下降平緩的趨勢。

· 3.3.2 子音音框觀察

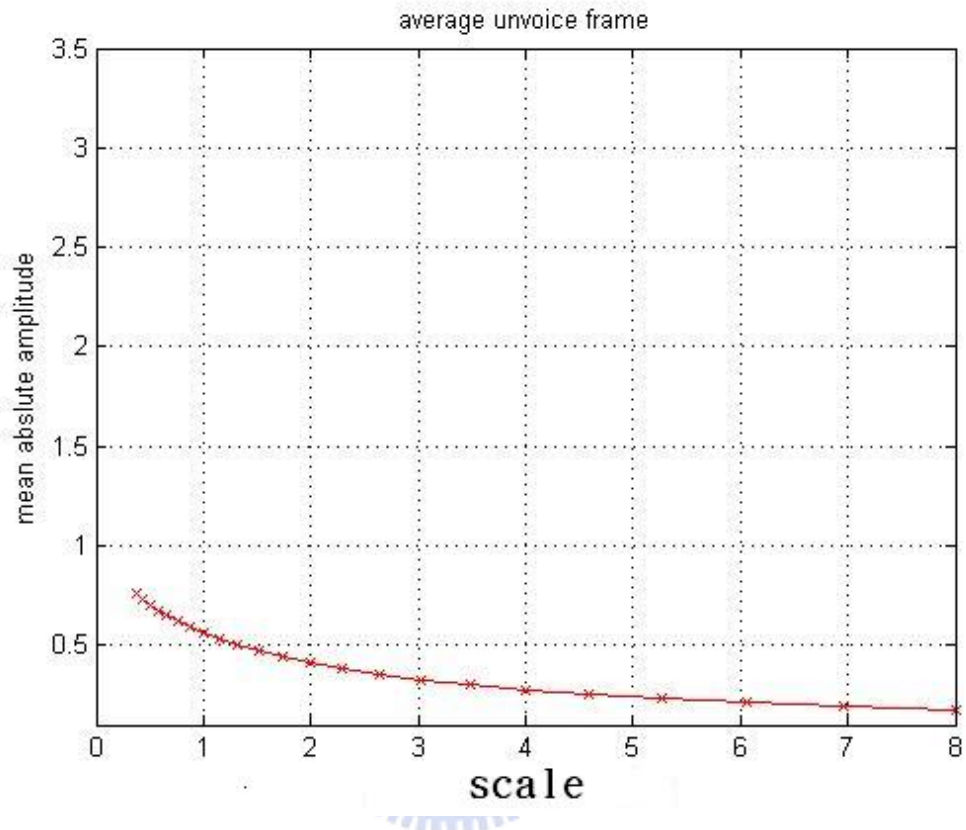


圖 3.7 平均乾淨的子音音框

圖 3.7，顯示無雜訊子音音框，在 23 維的 scale 向量中其特徵是 y 軸之值坡度較平緩，其能量也沒有母音音框高(參照圖 3.5)，由於子音有許多摩擦音之成份，造成其高頻解析出之能量較多，諧波(harmonic)密度較高，所以圖形上低 scale 與高 scale 所解析出之能量沒有太大差異。

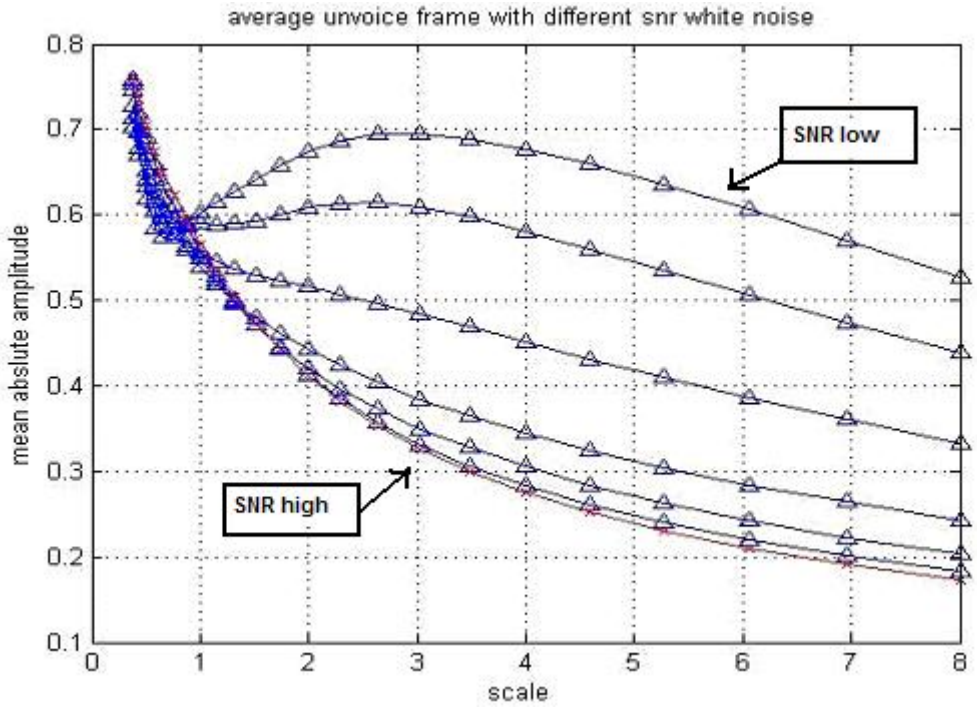


圖 3.8　不同訊雜比下之白噪音子音音框,紅色線(最下面一條線)為乾淨子音所畫出之曲線

圖 3.8，顯示子音音框加入不同訊雜比(SNR)後的平均分佈，可看出隨著白噪音(white noise)的能量增加,曲線往上遞增,基於原子音音框高頻成份增加,且使用 2.2.1 節的高斯正規化，其能量會多數集中在高頻，且由於高頻的諧波密度很高，因此在高 scale 時解析能量會增多，所以圖形坡度會有隨訊雜比(SNR)降低而往右上翹高的趨勢。

- 3.3.3 無聲音框觀察

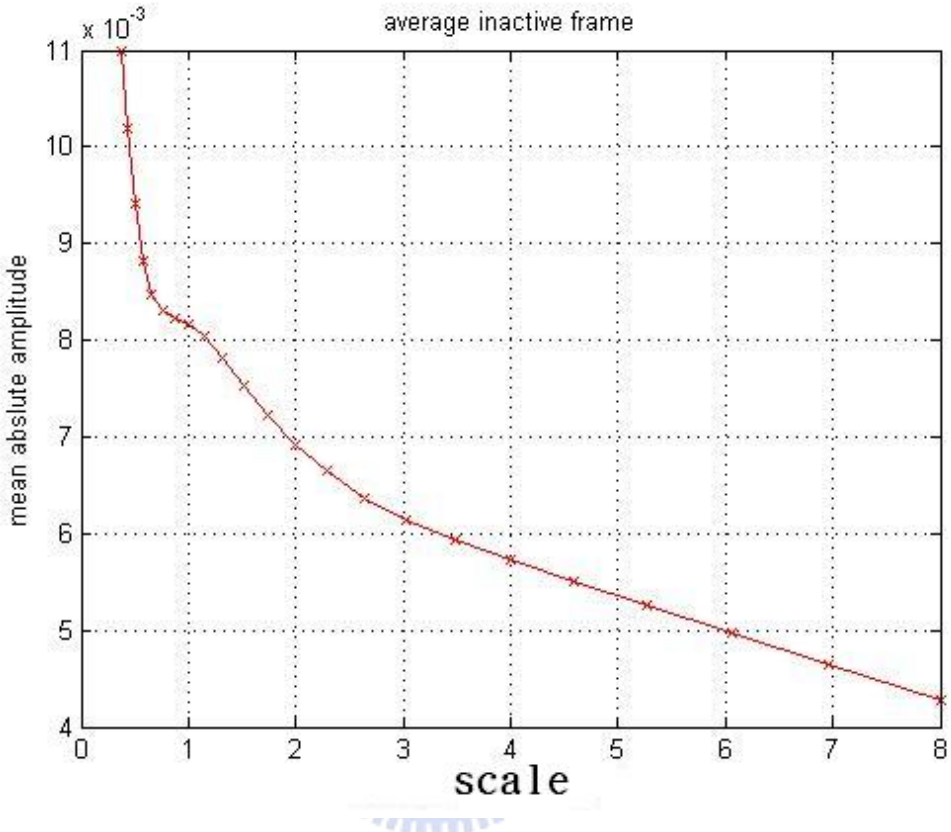


圖 3.9 平均乾淨的無聲音框

圖 3.9，顯示無雜訊之無聲音框，在 23 維的 scale 向量中，最大的特徵即是能量極小，左方的 y 軸是 10 的-3 次方為單位。

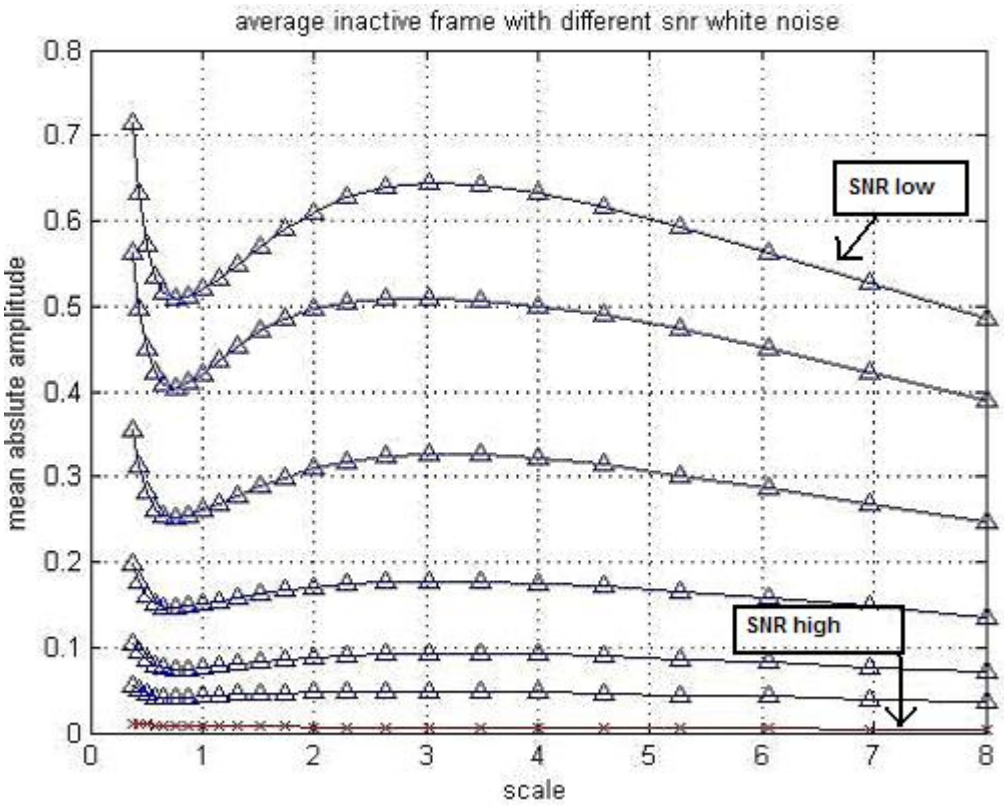
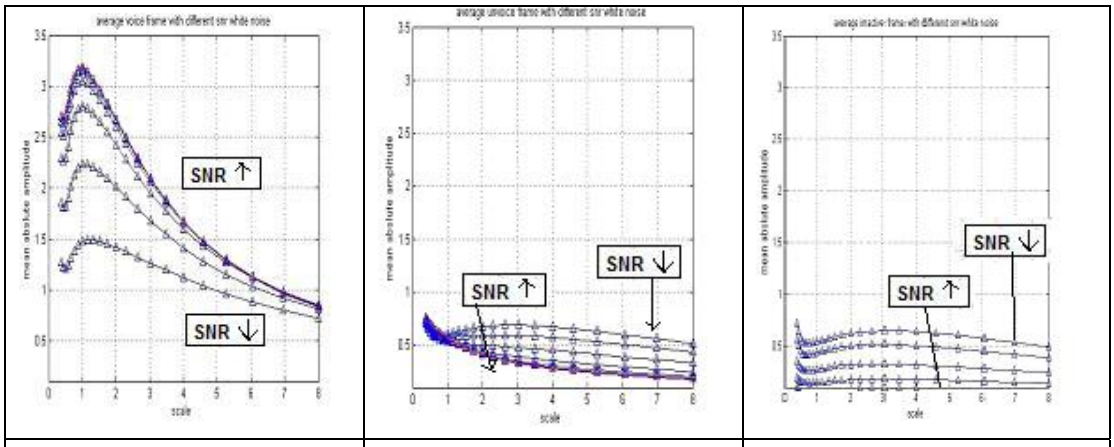


圖 3.10 不同訊雜比下之白噪音無聲音框，紅色線(最下面一條線)為無聲所畫出之曲線

圖 3.10,顯示無聲音框加入不同訊雜比(SNR)後的平均分佈,圖形中顯示訊雜比(SNR)越低時，無聲音框音加入了雜訊能量有所提升，23 維 scale 向量有隨訊雜比(SNR)降低而向上遞增之狀況。

以下將圖 3.8，圖 3.9，圖 3.10，Y 軸座標統一規格後(Y=0~3.5)並排，以便比較母音、子音、無聲加入不同訊雜比(SNR)後 23 維 scale 分佈。



母音	子音	無聲

圖 3.11 (3.6、3.8、3.10)之圖形比較

圖 3.11 正規化後，不同訊雜比下的平均[母音]、[子音]、[無聲]音框

· 3.3.4 任意單一音框之觀察

前三小節顯示了母音、子音、無聲在無噪音的狀況下「平均後」圖形分佈，並提供了加不同程度的雜訊(white noise)後，23 維線條趨勢與變化，事實上，母音及子音之總類繁多，能量所發生在的頻域範圍也不盡相同，以上三小節是由三類別(voice、unvoice、inactive)總體的巨觀觀察，實際上若從 16ms 的音框來看也會有巨觀現象的特性，以下三頁附錄以 16ms 音框為單位的母音、子音、及無聲在 23 維 scale 下分佈情況。

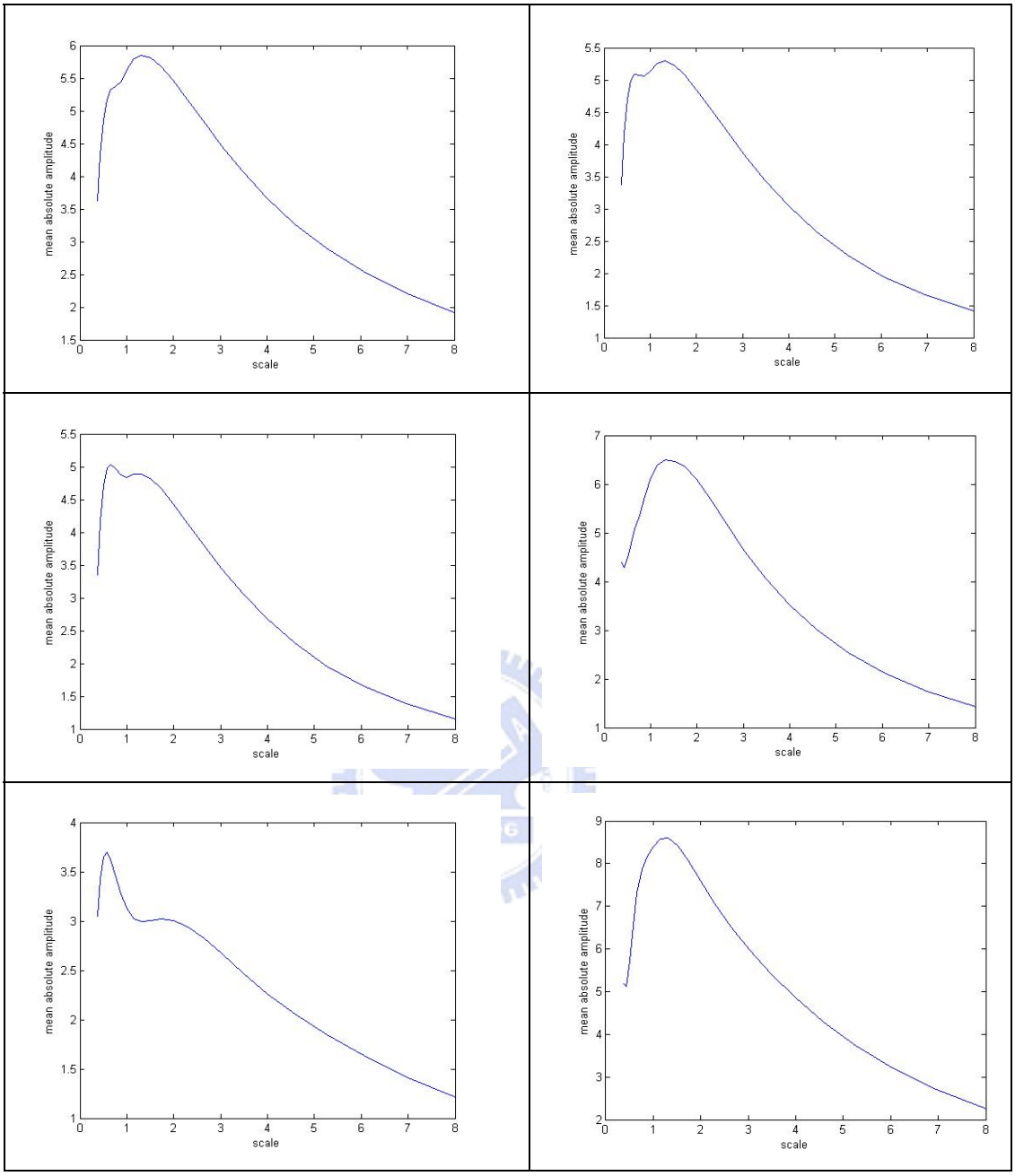


圖 3.12 單獨母音音框觀察

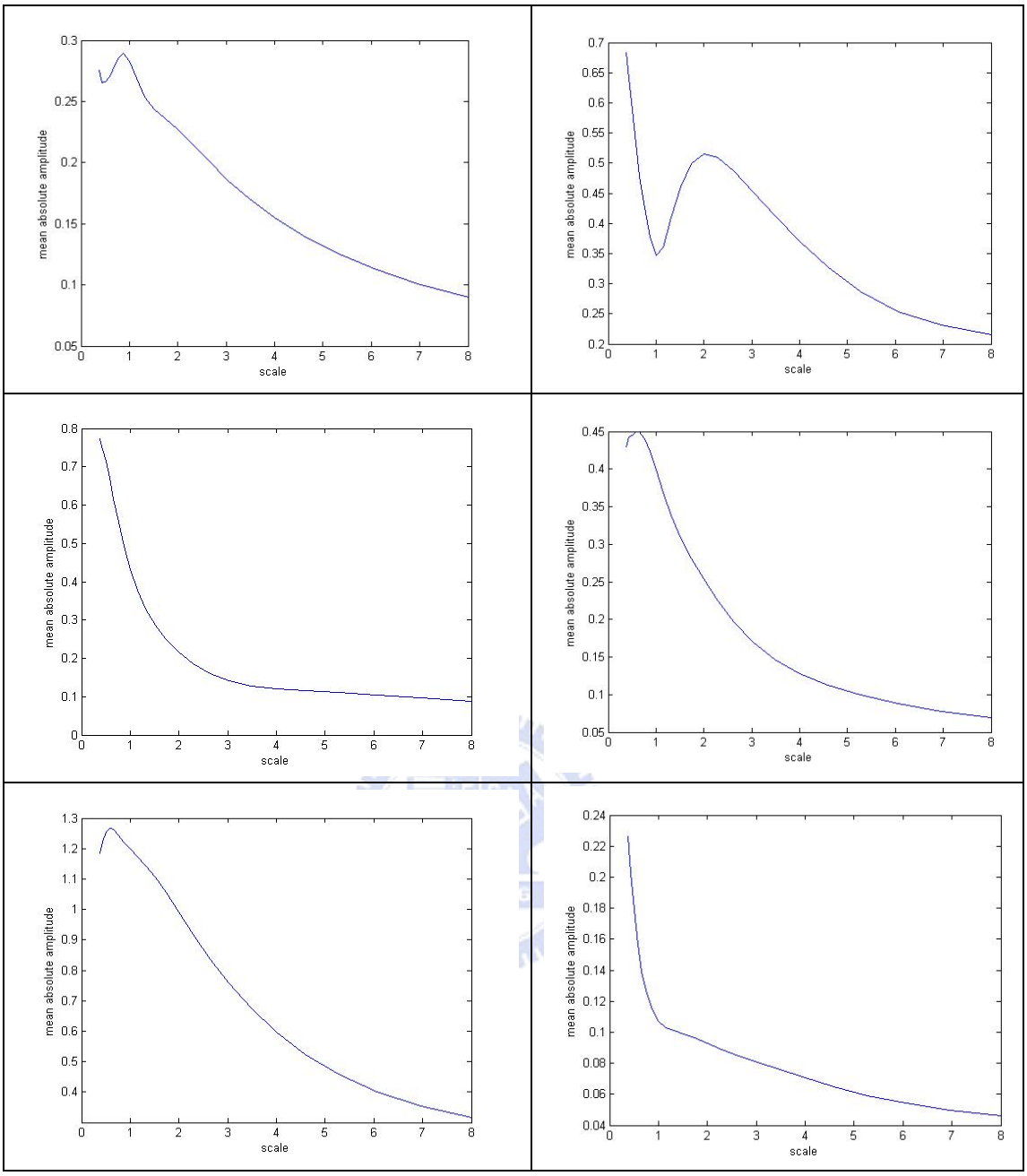


圖 3.13 單獨子音音框觀察

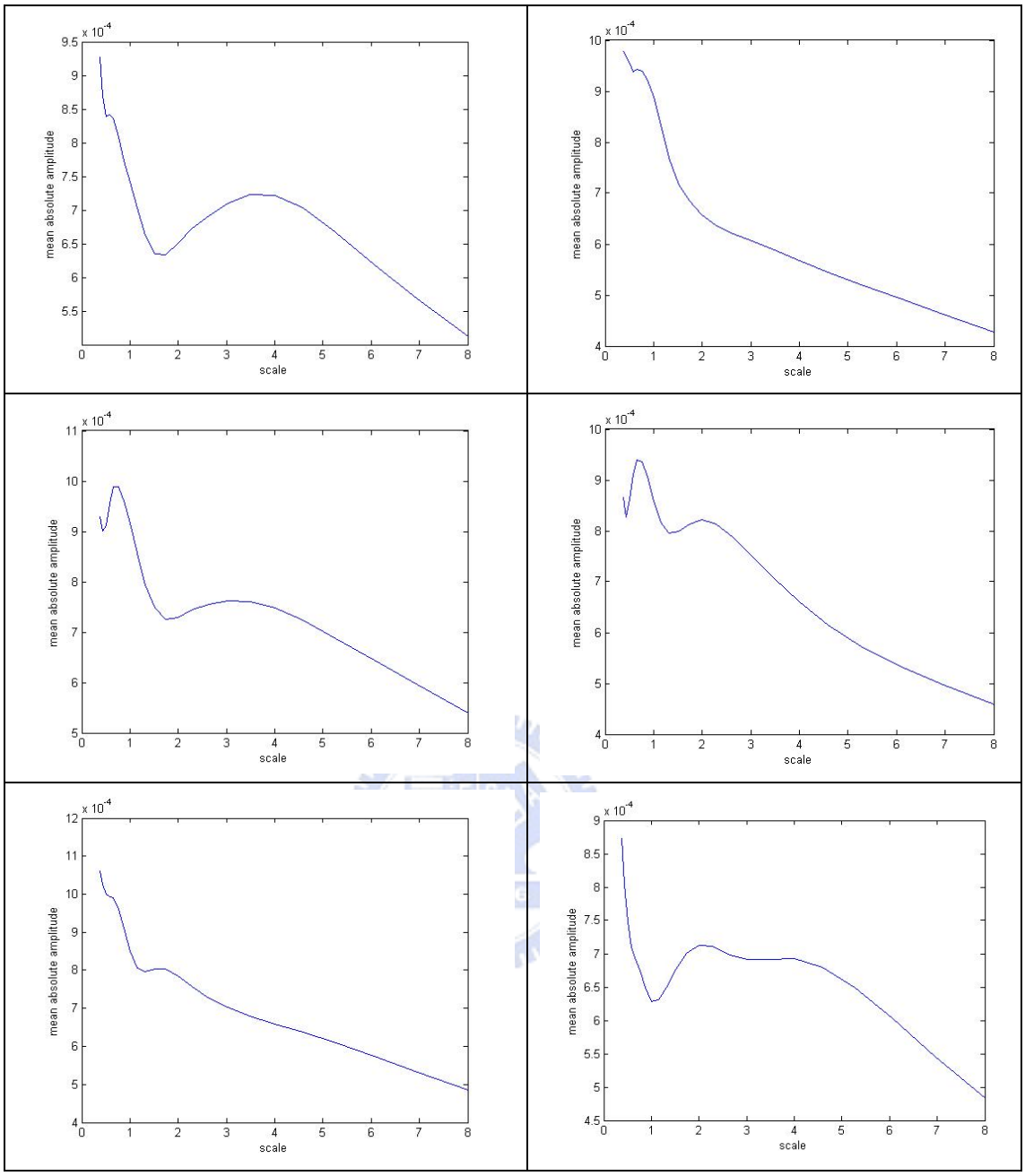


圖 3.14 單獨無聲音框觀察

· 3.4 語音分類與驗證

本章節討論不同聲音變化偵測器的決策策略(strategy)，並以不同的白噪音(white noise)語音訊號測試不同策略下的聲音變化偵測器，觀察其效能。

· 3.4.1 方法一 <以能量分類>

單就 23 維的尤拉距離(L-2 distance)做比較，來分出母音、子音、及 無聲。

主要作法:(1)將所有母音音框收集，做一平均的 23 維 scale 曲線當做範本(template)，子音和無聲也同上作法，大量收集，並以平均的 23 維 scale 曲線做出子音範本和無聲範本。(2)待分類的每一音框，抽取出 23 維 scale 向量後，分別與母音範本(voice template)、子音範本(unvoice template)、無聲範本(inactive template)做 L-2 距離比較，若與其中範本較相近則將音框歸為該類。

· 3.4.2 方法二 <以 23 維 scale 曲線趨勢分類>

單就 23 維的曲線趨勢做比較，來分出母音、子音、及 無聲。

主要作法:

(1) 將所有母音音框收集，做一平均的 23 維 scale 曲線當做範本(template)，子音和無聲也同上作法，大量收集，並以平均的 23 維 scale 曲線做出子音範本和無聲範本。(2)待分類的每一音框，抽取出 23 維 scale 向量後，分別與母音範本(voice template)、子音範本(unvoice template)、無聲範本(inactive template)做「相關係數」(correlation)比較，若與其中範本相關係數較高則將音框歸為該類。

- **3.4.3 方法三 <以曲線趨勢與能量分類>**

主要作法:

(1)求出母音範本、子音範本、無聲範本，作法同方法一及方法二之(1)。(2)待分類音框求出與範本相關性之前，先同時減去該範本的平均能量，再求彼此間相關性，若待分類音框與該範本相關性較高則將音框歸為該類。(式子說明，STi, 第 i 類的範本，Rk, 所接收到的第 k 個音框之 23 維 scale 向量)

$$P_i(R_k) = \frac{\sum[(ST_i - mean(ST_i)) \cdot (R_k - mean(ST_i))]}{\| (ST_i - mean(ST_i)) \| \cdot \| (R_k - mean(ST_i)) \|}$$
$$i = \{voice, unvoice, inactive\}$$
$$ST_i = template\ of\ \text{i}$$
$$\mathrm{R_k} = kth\ \text{number of frame} \qquad (3\text{-}1)$$

- **3.4.4 方法四 <以高斯訓練模型為分類>**

主要作法:

(1)將所有母音、子音、無聲的 23 維 scale 向量，以高斯混合模型(GMM)做訓練。

(2)待分類音框將其 scale 向量分別對三組高斯混合模型做比對，並得到一機率統計值，並以此值為依據判斷該音框所屬類別。

- **3.4.5 比較與驗證**

將前四小節所提出的四個方法，以不同語音及不同訊雜比(SNR)測試聲音變化偵測器的效能後，方法一和方法三和方法四都表現的比一般單一能量門檻(single threshold)聲音變化偵測器好。而由於母音及子音的種類繁多且變化性高，因此方法二僅用曲線趨勢相關性顯得分類判斷上較不確定，最後我們以方法四<高斯訓練模型為分類>做為本論文研究中聲音變化偵測器之策略。

以下圖示，分別顯示聽覺頻譜圖及時間軸上聲音變化偵測器偵測結果

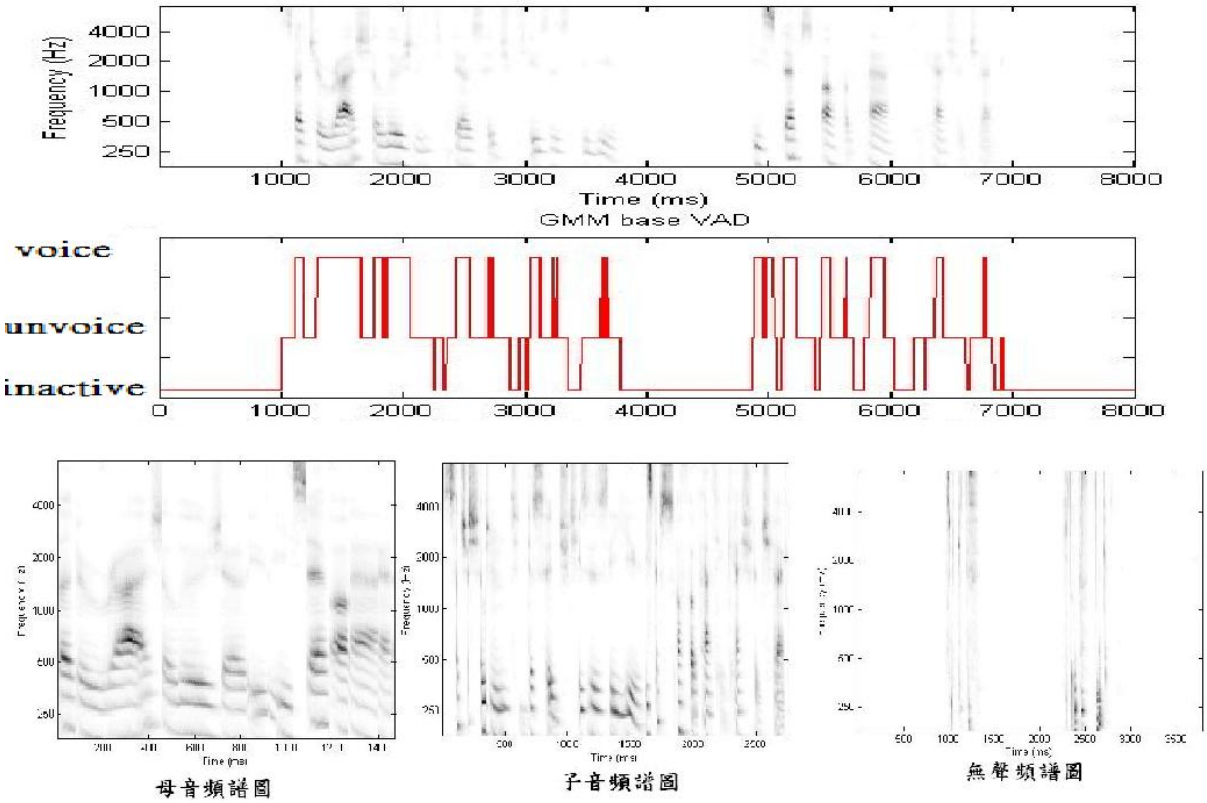


圖 3.15 VAD 偵測分類頻譜圖

圖 3.15 最上圖顯示原始語音頻譜圖，及所對應的音框分類，將母音、子音、無聲音框分三個等級(母音值最高、子音值中間、無聲圖最低)。圖 3.15 下圖，則顯示分類後的母音頻譜圖，子音頻譜圖，及無聲頻譜圖，由觀察中可發現母音部分偵測量好;子音部分由於音框大小(frame size)、語音開始(speech onset)、語音結束(speech offset)，本身判斷上即不顯著，因此會被歸於子音音框;無聲部分判斷極為良好，無聲頻譜圖中些許能量由於語者呼氣(breath)或錄音時剛開始的狀態，導致圖上有些許能量，是可接受之範圍。

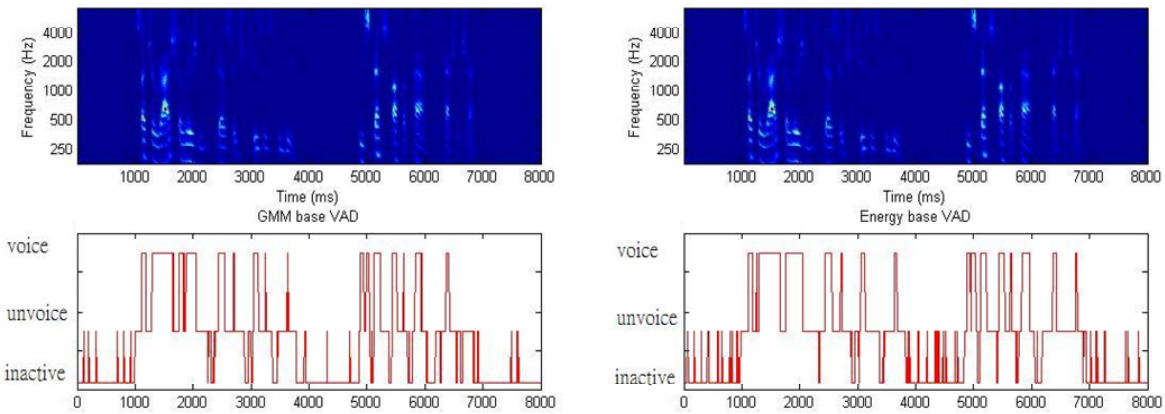


圖 3.16 VAD 比較圖

由圖 3.16 可發現,以同一句訊雜比(SNR)為 15db 的失真語音,左方以 23 維參數式之 VAD 與右方能量門檻式的 VAD，在左圖明顯較能偵測出有聲及無聲的差別，母音保留部分教完全;而右方能量方法明顯容易將有噪音之語音中的無聲音框誤判為子音音框。

第四章 高斯混合模型下的語音品質判斷

前一章節已說明如何將語音分為三類，本章節將說明如何判別品質好壞，並利用高斯混合模型(Gaussian Mixture Model，GMM)作判別語音品質之模型，並簡單概述所測試的語料。

4.1 評測模型

4.1.1 概念

模型式(model base、black box base，見 1.2.2)評測，假設建立在人腦對語音品質的判斷並沒有一一細分破壞語音品質的類型，而是根據個人自身經驗來告訴該語音的品質好壞，前階段參數抽取的部分，大多希望能抽取出代表人耳接收的語音訊號，而後一階段比對及評分，則是模擬人腦比對的過程，此比對過程不再細分語音品質遭破壞的類型，而是將參數與乾淨語料庫做比對。事實上，人腦在給予一語音品質時，我們可想像人腦本質就是人從小到大所聽過語音的「資料庫」，評分時我們參照資料庫(所聽過的語音)來給予一語音訊號評分。

評測模型現今已發展出多種，在這裡我們選擇高斯混合模型(Gaussian Mixture Model，GMM)以乾淨的語音總共 120 句，全部語音皆為標準英文發音，性別包含男性及女性，每個語音檔為八秒鐘，不過長也不過短，避免評分時給分者因為語音過長或過短影響評分，以上乾淨語料皆由國際電信聯盟標準化部門(ITU-T)出版之 Supplement p.23[32]中「original」資料夾中取得。

首先我們依照論文[22]設定，母音以 16 個 mixture 數，子音以 16 個 mixture 數，無聲以 2 個 mixture 數，以大量的母音、子音、及無聲音框轉成聽覺倒頻譜參數(Auditory Cepstral Coefficients)後，以高斯混合模型(GMM)訓練出三個類別之乾淨模型。

我們將待測語音用同樣的步驟，以 16ms 為一音框(frame)，經聲音變化偵測器(VAD)分類後，將該音框轉成聽覺倒頻譜參數(ACC)並丟入該類別(母音、子音、無聲)之乾淨模型，比對高斯混合模型(GMM)後，每一音框(frame)會從所屬類別產生一機率統計值，此值為機率密度函數的對數值(log-pdf)，並與機率值有正相關。

因此每一句 8sec 之語音檔，會有 500 個音框，並產生 500 個 log-pdf 值，在第五章我們會說明如何將 500 個值對應到平均意見分數(MOS)之評分。

· 4.1.2 國際電信聯盟標準化部門(ITU-T)之標準

PESQ(Perceptual Evaluation of Speech Quality)[18](2001)

PESQ 為 2001 年國際電信聯盟標準化部門所提出之語音品質測量方法，此測量方式為「侵入式客觀語音品質測量」(Intrusive-Objective Speech Quality Measurement)，測量適用於窄頻通訊(3.1khz handset telephony and narrow-band speech codec)，參考了 1998 年提出的 PSQM 的聽覺模型，運用了梅爾倒頻譜參數(mfcc)，以及參照英國電訊所 PAMS(Perceptual Analysis Measurement System)的時間對位法(time-alignment algorithm)，由於主要是用於語音編碼(speech codec)的評測，並著重於單傳輸失真(one-way speech distortion)及外加噪音(noise on speech quality)的影響，因此有需多因素無法有效測量，例如，回音(echo)，響度失真(loudness loss)等，又由於時間對位法的不準確性，使得語音訊號稍有時間上的位移(shift)，即無法準確評分。

PESQ，為目前「侵入式客觀語音品質測量」的標準，由於其評分效果與實際的平均意見分數(MOS)有高達百分之九十以上的相關性，因此在侵入式測量中仍常拿來使用。(下圖附錄 PESQ 之基本流程)

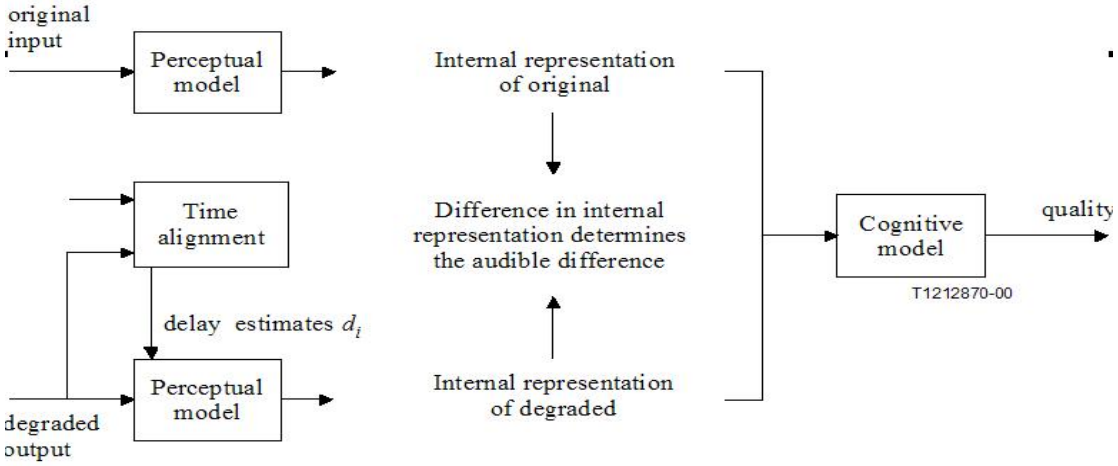


圖 4.1 PESQ 概觀流程圖(先將失真訊號及原始訊號做時間軸上對準，再以感知模型做轉換，最後用時、頻域的相異程度做為差距，最後階段則以此差距評分)

附表: PESQ 目前已知可測試的影響變因

經實驗證實可以由 PESQ 分數估得可以接受的準確度:
Test factors:
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
Environmental noise at the sending side
Effect of varying delay in listening only tests
Short-term time warping of audio signal
Long-term time warping of audio signal
Coding technologies
Waveform codecs, e.g. G.711; G.726;
G.727CELP 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
Application
Codec evaluation
Codec selection
Live network testing using digital or analogue connection to the network
Testing of emulated and prototype networks

表 4-1 PESQ 可測試之變因

經實驗證實由 PESQ 分數無法準確測量的變因:
Test factors Listening levels (See Note.) Loudness loss Effect of delay in conversational tests Talker echo Sidetone **Coding technologies**Replacement of continuous sections of speech making up more than 25% of active speech by silence (extreme temporal clipping) **Applications** In-service non-intrusive measurement devices Two-way communications performance

表 4-2 PESQ 無法測試之變因

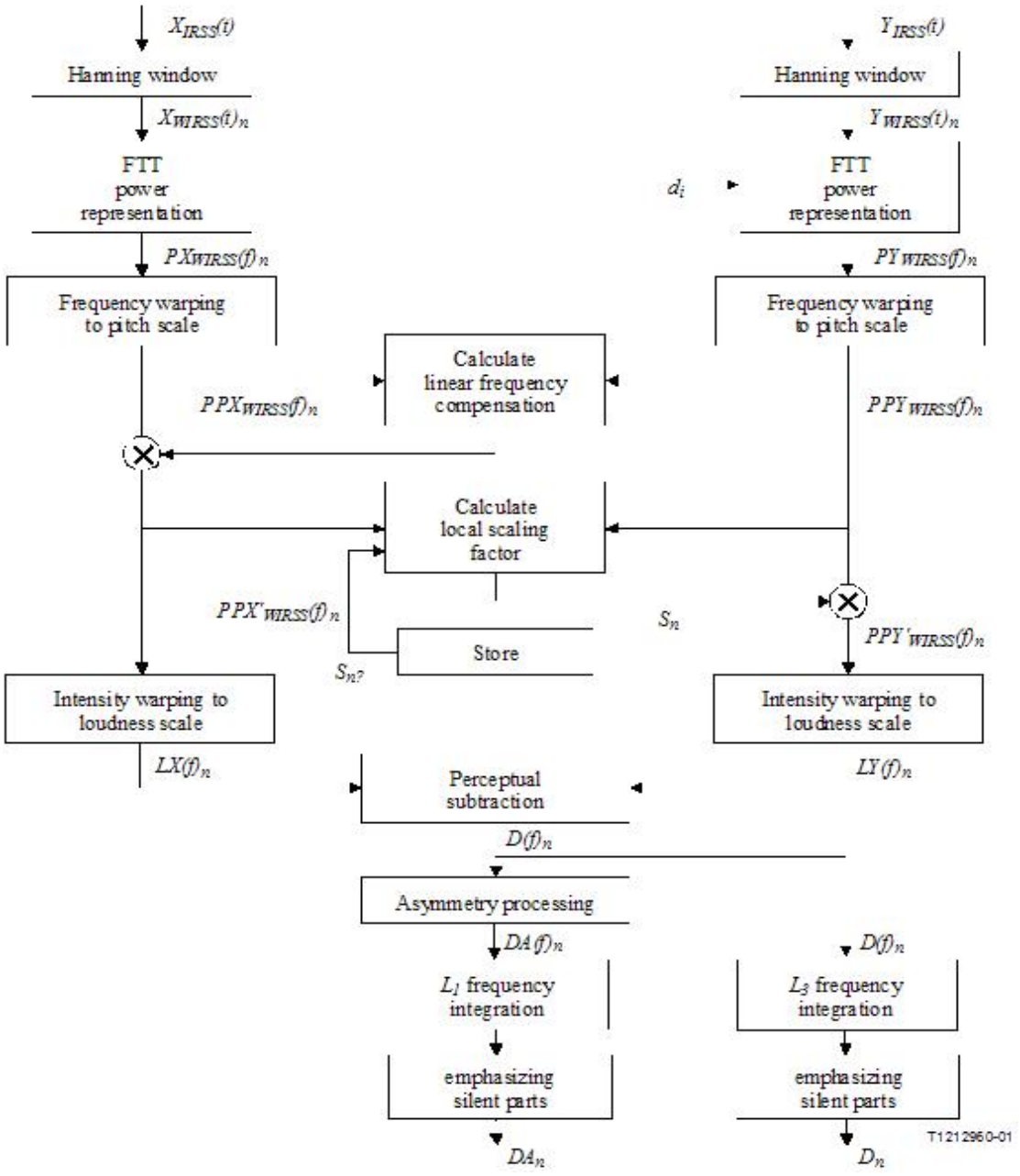


圖 4.2 PESQ 感知模型[18]

圖 4.2 為 PESQ 之感知模型，前階段先將對準的語音檔之第 n 個音框通過漢明窗(Hamming Window)，接著轉至頻域軸做頻濾扭曲(frquency warping)，下一步則將人耳基本的心理聲學反應，如響度門檻(loudness threshold)加入訊號中，最後將原始語音與失真語音的頻譜相減得到差距，Asymmetry processing 則將差距做一非線性轉換(由於差距值不能線性對應至語音品質)，最後階段則是做非語音部分的補償。得到的 DA_n 及 D_n ，必須對時間軸上多個音框做累計，最後才能求出整句語音檔的品質差距。

P.563[33](2004)

P.563 為 2004 年國際電信聯盟標準化部門所提出之語音品質測量方法,此測量方式為「非侵入式客觀語音品質測量」(Non-intrusive Objective Speech Quality Measurement),品質測量分三部份,一為特殊失真(distortion-specific)的測量,包含時間軸的不連續(temporal clipping)及噪音(noise)估計等;二為語音重建及對照模型(speech reconstruction and full-reference perceptual models),將失真訊號(degraded signal)做一個粗估的比對;三為音高同步(pitch synchronous)、口腔模型與線性預估係數分析(vocal track model & LPC)。將以上的三部分做綜合性的考量,並給予分類(classification)及權重(weighting),最後對應到平均意見分數(MOS)。

P.563,為目前「非侵入式客觀語音品質測量」的標準,其評分效果與實際的平均意見分數(MOS)達到百分之八十以上的相關性,因此在非侵入式測量中常拿來作為比較。

而由 PESQ 及 P.563 的演算法可看出,已有越來越多的感知現象及參數加入演算法中。

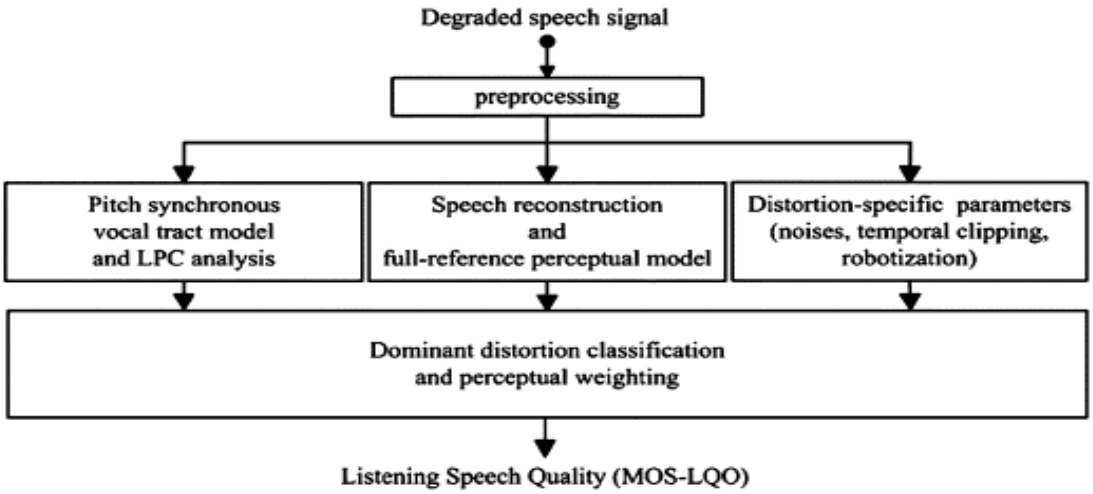


圖 4.3 P.563 概念流程圖

圖 4.4 [33]P.563 之評分流程,在將訊號做完前處理後,以三個機制判斷品質好壞,包含特殊失真的測量、以類乾淨語音做比對、及對口腔模型的估計。

4.2 評測之語料庫

本論文測試語料及訓練語料來自 ITU-T P.Supplement 23[32]，所有的語料盡量以簡單(simple)、簡短(short)、有語意(meaningful)為錄製標準，錄音內容多從報章或非技術性文章中擷取，更進一步地，句子與句子間並沒有明顯的語意關聯，過短或過長的語音會被刪除，所有的語音檔皆為 8 秒鐘(註:128000 16-bit sample)。

本語料庫中包含三個實驗，實驗一，以無線傳輸中之標準語音編碼為變因，例如:G.729，G.726，GSM…；實驗二，以環境背景音為測試變因，由於評分方式並非平均意見分數(MOS)，且主觀的實驗流程亦不同，因此實驗二在本論文中不討論；實驗三，以傳輸通道造成的削弱(Channel Degradations)為變因，包含隨機的位元錯誤(random bit error)等狀況。

語料庫含法文、德文、日文、英文等語系，為將語言的韻律等差異性排除，本論文僅使用英文語料，語料中包含男性語音及女性語音，實驗一共測試了 176 句，實驗三共測試了 200 句。除此之外，在訓練乾淨模型時，本論文僅用了 120 不同的原始語音檔。

實驗一　測試變因

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		
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 4-3 實驗一測試變因

實驗三　測試變因

Cond No.	Codec	Trans-codings	Noise Type	Error Type	Error Rate (%)	Input Characteristic	Q Value
1	G.729	1	Clean	–	–	A-Law, M. IRS	–
2	G.729	1	Clean	Random Frame	3	A-Law, M. IRS	–
3	G.729	1	Clean	Random Frame	5	A-Law, M. IRS	–
4	G.729	1	Clean	Bursty Frame	3	A-Law, M. IRS	–
5	G.729	1	Clean	Bursty Frame	5	A-Law, M. IRS	–
6	G.729	1	Vehicle	–	–	A-Law, M. IRS	–
7	G.729	1	Vehicle	Random Frame	3	A-Law, M. IRS	–
8	G.729	1	Vehicle	Random Frame	5	A-Law, M. IRS	–
9	G.729	1	Vehicle	Bursty Frame	3	A-Law, M. IRS	–
10	G.729	1	Vehicle	Bursty Frame	5	A-Law, M. IRS	–
11	G.729	1	Street	–	–	A-Law, M. IRS	–
12	G.729	1	Street	Random Frame	3	A-Law, M. IRS	–
13	G.729	1	Street	Random Frame	5	A-Law, M. IRS	–
14	G.729	1	Street	Bursty Frame	3	A-Law, M. IRS	–
15	G.729	1	Street	Bursty Frame	5	A-Law, M. IRS	–
16	G.729	1	Hoth	–	–	A-Law, M. IRS	–
17	G.729	1	Hoth	Random Frame	3	A-Law, M. IRS	–
18	G.729	1	Hoth	Random Frame	5	A-Law, M. IRS	–
19	G.729	1	Hoth	Bursty Frame	3	A-Law, M. IRS	–
20	G.729	1	Hoth	Bursty Frame	5	A-Law, M. IRS	–
21	G.729	2	Clean	–	–	A-Law, M. IRS	–
22	G.729	3	Clean	–	–	A-Law, M. IRS	–
23	G.729	2	Clean	Random Frame	3, 3	A-Law, M. IRS	–
24	G.729	3	Clean	Random Frame	3, 0, 3	A-Law, M. IRS	–
25	G.729	2	Clean	Bursty Frame	3, 3	A-Law, M. IRS	–
26	G.729	3	Clean	Bursty Frame	3, 0, 3	A-Law, M. IRS	–
27	G.729	2	Vehicle	Random Frame	3, 3	A-Law, M. IRS	–
28	G.729	2	Vehicle	Bursty Frame	3, 3	A-Law, M. IRS	–
29	G.729	1	Clean	Random Bit	1	A-Law, M. IRS	–
30	G.729	1	Clean	Random Bit	3	A-Law, M. IRS	–
31	G.729	1	Clean	Random Bit	5	A-Law, M. IRS	–
32	G.729	1	Clean	Random Bit	10	A-Law, M. IRS	–
33	G.729	1	Clean	Burst Frame/Random Bit	3, 1	A-Law, M. IRS	–
34	G.729	1	Clean	Burst Frame/Random Bit	3, 3	A-Law, M. IRS	–
35	G.729	1	Clean	Burst Frame/Random Bit	3, 5	A-Law, M. IRS	–
36	G.729	1	Clean	Burst Frame/Random Bit	3, 10	A-Law, M. IRS	–
37	G.726	1	Clean	–	–	A-Law, M. IRS	–
38	G.726	1	Vehicle	–	–	A-Law, M. IRS	–
39	G.726	1	Street	–	–	A-Law, M. IRS	–
40	G.726	1	Hoth	–	–	A-Law, M. IRS	–
41	MNRU	1	Clean	–	–	UPCM, M. IRS	10 dB
42	MNRU	1	Clean	–	–	UPCM, M. IRS	15 dB
43	MNRU	1	Clean	–	–	UPCM, M. IRS	20 dB
44	MNRU	1	Clean	–	–	UPCM, M. IRS	25 dB
45	MNRU	1	Clean	–	–	UPCM, M. IRS	30 dB
46	MNRU	1	Clean	–	–	UPCM, M. IRS	50 dB
47	Direct	–	Clean	–	–	UPCM, M. IRS	–
48	Direct	–	Vehicle	–	–	UPCM, M. IRS	–
49	Direct	–	Street	–	–	UPCM, M. IRS	–
50	Direct	–	Hoth	–	–	UPCM, M. IRS	–

表 4-4 實驗三測試變因

第五章 結果與討論

本章 5.1 節，先以侵入式測量做實驗，檢視有參照訊號測量下，以不同訊雜比的白噪音為變因，得到分數與 PESQ 求出分數之相關性。5.2 節及 5.3 節探討非侵入式測量，以第二章、第三章、及第四章為概念，在無參照訊號下做語音品質評分，測試變因為白噪音及 ITU-T 中含有實測品質分數之語音資料庫。所有結果將列於 5.4 節，並在 5.5 節給予結論。

5.1 侵入式測量

經由第二章介紹，已知侵入式語音品質測量需要原始訊號及失真訊號，由於侵入式語音評測較類似比較式的評測，最簡單的比較方式即為相減求取差距，此觀念的假設是基於當「人」比較兩語音檔時自然會尋求其中相同性與差異性，自然會找出話語中同個句子中的同個字，比較相似程度。

本章的實驗設計流程如圖(5.1)，當原始語音訊號及失真訊號同時輸入時，先兩者經由處理做音訊至聽覺頻譜圖的轉換，由聽覺頻譜圖轉換至 rate-scale domain(時域變化-頻域變化區域)，以互相對應的 rate-scale 圖比較差距，累計後評分。

當轉換至大腦區域時,得到的是一四維的表示式 STRF(見 2.2 節),四維分別包含「時域(Time)」、「頻域(Frequency)」、「時間變化(rate)」、「頻率變化(scale)」，聽覺頻譜圖(Auditory spectrogram)的維度為「時間」、「頻率」，若我們在 STRF 中選擇固定的時間點與頻率，將會得到該點的時間變化率與頻率變化率圖(即之前所說的 rate-scale 圖)。在本論文研究中我們假設人僅在時間軸上分辨相似性，因此我們將頻率軸做平均，四維的 STRF 縮減成三維的「時間(Time)」、「時間變化(Rate)」、「頻率變化(Frequency)」。評測時首先須將原始語音與失真語音的時間音框對準，才不會有對錯字對錯音的狀況。

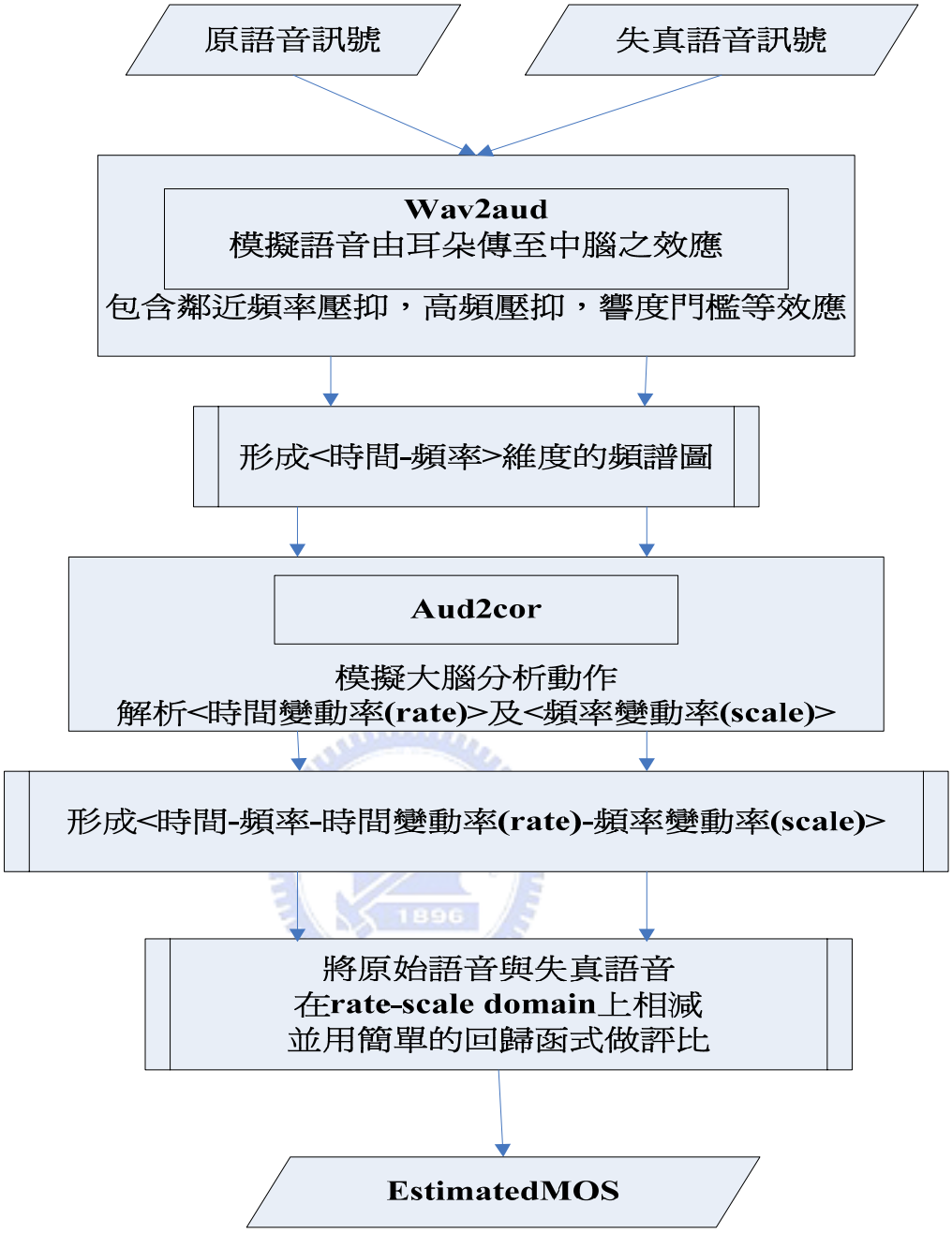


圖 5.1 侵入式語音品質測量流程圖

以 TIMIT 語料庫，區域一(dr=1)的乾淨原始語料，取 14 位女性及 14 位男性，每位語者取 6 組語句，並用白噪音(White noise)做測試，訊雜比範圍從[-10db~45db]，以 PESQ 做為此語料的評分(PESQ 之介紹詳見 4.1.2)，以失真訊號與原始訊號時間軸上對應之音框，個別求出不同 rate-scale 下的數值並相減，最後在時間軸上累計差異，並以二次多項式作回歸評分，結果將一併附在 5.4 節中。

‧ 5.2 非侵入式測量

在第四章中，我們已知本論文評分概念:每句話在經過處理後，形成聽覺頻譜圖，以時間軸上切割，將聽覺頻譜圖形成多個音框，每個音框由先前的聲音變化偵測器被予標註為母音(voice)、子音(unvoice)、及無聲(inactive)，以上分成三類之一個音框以對應的高斯混合模型(GMM)做比對，每一音框得一對數機率密度函數值(log-pdf)。

在做語音評測時，每個語音檔分為數個音框，每個音框都會有一對應之對數機率密度函數值，本章以這些值來做評分依據。

統計值之選擇:

平均值(mean):將一句語音中所有母音音框之對數機率密度函數取平均值，同樣地，所有子音音框之對數機率密度函數取平均，無聲音框亦取出一平均值。用來模擬失真語音中，三者類別與原始乾淨語音之模型的平均差距。

變異數(variance):將一句語音所有的母音音框之對數機率密度函數取變異數，所有的子音音框及無聲音框亦做同樣的動作求出變異數。求取各類別音框的變動性。

比重(weight):求取一句話中，母音所佔的百分比、子音所佔的百分比，及無聲所佔的百分比。

音高(pitch): 利用 AMDF(平均幅度差函數),求取一整句話中非無聲音框(Non-Inactive frames)對應的音高值，為避免性別的特性及說話的內容影響音高高低，在求取出音高後，做一階微分，並取變異數，來代表音框與音框間之變動性。另取偵測出之音高音框數佔有聲音框數的比例，當語音編碼將原有波形破壞時，諧波的失真造成偵測的音高音框數會減少，此時音高音框數佔有聲音框數的比例會降低。

5.3 回歸函式

本論文將用前節提出之各個統計值，作一簡單的回歸曲線，由值對應到 1 分至 5 分的平均意見分數，本論文以簡單的二次多項式(pure quadratic polynomial)做回歸曲線，變數與變數間無互相影響(no interaction terms)，每一變數皆做一二次多項式並將每個多項式相加，求得平均意見分數值，8 個變數如下:

(1) 所有母音音框之對數機率密度函數取平均值除以標準差，X_1

(2) 所有子音音框之對數機率密度函數取平均值除以標準差，X_2

(3) 所有無聲音框之對數機率密度函數取平均值除以標準差，X_3

(4) 所有母音音框之對數機率密度函數取變異數除以平均值，X_4

(5) 所有子音音框之對數機率密度函數取變異數除以平均值，X_5

(6) 所有無聲音框之對數機率密度函數取變異數除以平均值，X_6

(7) 所有音高音框微分後之變異數，X_7

(8) 音高音框數佔有聲音框數的比例，X_8

$$\hat{MOS} = \sum_i P_i(X_i)\ ,\quad i=1\sim 8\ 。$$

$$P_i(X_i) = \alpha_i X_i^{\,2} + \beta_i X_i + \gamma_i$$

α_i、β_i、γ_i 為所求出之回歸係數(regression coefficients)

5.4 實驗結果

1.**侵入式測量結果**:共 1764 句以白噪音(white noise)為變因，訊雜比範圍由(-10db)~(45db)。PESQ 評分與 NSLtool 評分彼此間的相關性高達 94.6%，也證明了在白噪音(white noise)為測試變因下，NSLtool 的感知訊號評分方式足以取代 ITU-T 所提出的標準。除此之外，NSLtool 的感知訊號評分結果在同一 db 狀態下有較低的標準差，此意味我們提出的評分方法有較強健(robust)的特性。而 NSLtool 中，使用大腦模擬階段來求取侵入式測量分數，當白噪音訊雜比為-3db 以下，此時差距的鑑別度已經縮小，以二次多項式做評分，語音品質分數差距多為小數點以後幾位之差別，這也說明模擬大腦機制的判斷下，對於品質「很好」與「非常好」以及「很差」與「非常差」的判斷已經非常小，而這也與人的品質判斷相符(當語音品質好壞達到極端狀況，人難以細分差異)。

2.**非侵入式測量結果:**以白噪音(white noise)為變因，測試 P.563，ACC，及 MFCC 參數下評分效能。

以白噪音(white noise)為變因之測試(評分由 PESQ 取得)			
評分方式	P.563	ACC	MFCC
與 PESQ 分數的相關性	83%	98.7%	98%

表 5-1　以白噪音(white noise)為變因之測試(評分由 PESQ 取得)

3.**非侵入式測量結果:**以 ITU-T p.23 supplement[32]實驗一，測試不同語音編碼下各類評分。

ITU-T P.23 實驗一			
評分方式	P.563	ACC	MFCC
與主觀分數的相關性	78.90%	58.5%	54%

表 5-2　ITU-T P.23 實驗一

4.**非侵入式測量結果:**以 ITU-T p.23 supplement[32]實驗三，測試相同語音編碼下，不同通道狀況及背景音下各類評分。

ITU-T P.23 實驗三			
評分方式	P.563	ACC	MFCC
與主觀分數的相關性	78.70%	60.6%	58.8%

表 5-3 ITU-T P.23 實驗三

▪ 5.5 結論

由以上實驗所得結果顯示，有參照訊號狀況以感知模型做測試，以白噪音為變因下，與 ITU 提出之標準可達到相當程度的匹配(百分之九十以上相關性);在非侵入式測量中，同樣使用白噪音為變因，以 ACC 為訊號參數甚至可與 PESQ 分數相關性達百分之98%，我們可說明倒頻譜參數對應至 PESQ 分數的相關性甚高。

侵入式語音測量中，音框的對準為重要的概念，在 PESQ 演算法中，由於一開始即須要做時間軸對準(Time-alignment)，若參照訊號與待測訊號有時間軸上的平移(time-shift)，則會造成極大誤判。本論文中侵入式測量可選擇將時間軸效應做平均，得到之分數與 PESQ 分數相關性亦可達到百分之九十。

參數 ACC 的評分皆比 MFCC 之評分在正確的主觀分數上更有相關性，而 ITU 所提出之標準在任何實驗中較能維持一致性,且相關性皆維持在 80%上下,而在以白噪音(white noise)為變因之測試下，使用 ACC 的評分效能勝過其餘兩者。在 ITU-T p.23 supplement 實驗一的狀況下，ACC 效能甚至高於 MFCC 的測量達到約 5%左右。

▪ 第六章 未來展望與方向

(1) 在聲音變化偵測器方面，本論文僅做簡單的參數特性觀察與試驗，事實上，現今的聲音變化偵測器，多數結合適應性門檻(Adaptive threshold)及多維判斷式(Muiti-decision)，本論文所提出之聲音變化偵測器先就單一因素觀察，未來若能結合時間軸上的適應性門檻，及多維的判斷式，相信在運用上可以更廣泛，並能克服時變性(time-variant)的噪音。

(2) 最後的語音評測測試中，以顯示評測效能以 ACC 參數會比 MFCC 參數好，但仍無法勝過標準的 p.563，由於本論文中僅以參數式的比對方式作評分，而真正影響品質的因素，可能需要抽取更多不同且能代表不同類型語音訊號破壞的參數，才能提升整體評分效能。

(3) 本論文僅以簡單的二次多項式做回歸曲線並評分，事實上，評分可能並非線性或是多項式的對應而已，現今多數作法皆用到大量的統計模型或資料探勘(Data mining)技術，甚至有用上類神經網路的方式來模擬回歸並評分的過程，以上方式不易看出物理特性及意義，因此本論文並未使用; 期待在本論文之後，有更多關於「資料與評分」之研究，並尋求更適合的評分公式。

參考文獻

[1] Antony W. Rix, John G. Beerends, Doh-Suk Kim, “Objective Assessment of Speech and Audio Quality—Technology and Applications”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 6, November 2006

[2] Chi, T., Ru, P., and Shamma, S., ”Multi-resolution Spectro-tempor Alanalysis of Complex Sounds”, J. Acoust. Soc. Am., 2005

[3] “Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs”, ITU-T Rec. P.830, 1996

[4] “Methods for Subjective Determination of Transmission Quality”, ITU-T Rec. P.800, 1996

[5] “Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems” , ITU-R Rec. BS.1534-1,2005

[6]“General Methods for the Subjective Assessment of Sound Quality”, ITU-R Rec. BS.1284-1, 2003

[7] “The E-model, a Computational Model for Use in Transmission Planning”, ITU-T Rec. G.107, 2005

[8] “Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems” , ITU-R, BS.1116, 1998

[9]“Wideband Extension to Recommendation for the Assessment of Wideband Telephone

Networks and Speech Codecs” , ITU-T P 862.2, 2005

[10] Weixia Li and Robert F. Kubichek, ”Output-based Objective Speech Quality Measurement Using Continuous Hidden Markov Models” 1-4 July 2003

[11] Tiago H.Falk, Wai-Yip Chan, “Nonintrusive Speech Quality Estimation Using Gaussian Mixture Models”, IEEE Signal Processing Letters, 2006

[12] R J B Reynolds, and A W Rix , “Quality VoIP-An Engineering Challenge”, BT Technology Journal, Volume 19, Number 2, 2001

[13] S.Quackenbush, T.Barnwell, and M.Clements, ”Objective Measure of Speech Quality”, Prentice-Hall, New York, 1988

[14] S. Wang, A. Sekey, and A. Gersho, “An Objective Measure for Predicting Subjective Quality of Speech Coders”, IEEE J. Sel. Areas Commun., vol. 10, no. 5, pp. 819–829, Jun. 1992

[15]J. Beerends and J. Stemerdink, “A perceptual Speech-quality Measure Based on a Psychoacoustic Sound Representation,” J. Audio Eng. Soc., vol. 42, no. 3, pp. 115–123, 1994

[16] S. Voran, “Objective Estimation of Perceived Speech Quality—Part I: Development of the Measuring Normalizing Block Technique,” IEEE Trans. Speech Audio Process., vol. 7, no. 4, pp. 371–382, Jul. 1999

[17] ——, “Objective Estimation of Perceived Speech Quality—Part II: Evaluation of the measuring normalizing block technique”, IEEE Trans. Speech Audio Process., vol. 7, no. 4, pp. 383–390, Jul. 1999

[18] “Perceptual evaluation of speech quality (PESQ)”, ITU-T Rec. P. 862, 2001

[19] O. Au and K. Lam, “A Novel Output-based Objective Speech Quality Measure for Wireless Communication,” in Proc. 4th Int. Conf. Signal Process., vol. 1, pp. 666–669, 1998

[20] P. Gray, M. Hollier, and R. Massara, “Non-intrusive Speech-quality Assessment Using Vocal-tract Models,” Proc. Inst. Elect. Eng. Vision, Image and Signal Processing, vol. 147, pp. 493–501, 2000

[21] H. Hermansky, “Perceptual Linear Prediction (PLP) Analysis of Speech”, J. Acoust. Soc. Amer., vol. 87, pp. 1738–1752, 1990.

[22] Tiago H.Falk, Wai-Yip Chan “Nonintrusive Speech Quality Estimation Using Gaussian Mixture Model”, IEEE 2006

[23] D. 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

[24] Weitzenfeld, A., M. Arbib, A. Alexander, “The Neural Simulation Language: A System for Brain Modeling”, MIT Press, MA. 2002.

[25] Chi, T., Gao, Y., Guyton, C. G., Ru, P., and Shamma, S. , ” Spectro-temporal Modulation Transfer Functions and Speech Intelligibility”, J. Acoust. Soc. Am.,1999

[26] H.Hermansky, “Perceptual Linear Prediction PLP analysis for speech”, Jounal of the Acoustic Society of America, 1990

[27] R. V. Cox and P. Kroon “Low Bit-rate Speech Coders for Multimedia Communications” IEEE Commun Mag, vol 34, pp34-41, Dec.1996

[28] R.L.Bouquine-Jeannes and G.Faucon, “Study of Voice Activity Detector and Its Influence in a Noise Reduction Syste”, Speech Commun, vol. 16, pp.245-254, 1995

[29] J.Sohn, N. S. Kim, and W.Sung, “A Statistical Model-base Voice Activity Detection”IEEE Signal Processing Lett., vol. 6, pp. 1-3, Jan.1999

[30] Y.D.Cho, K.Al-Naimi, and A.Kondoz, “Mixed Decision-base Noise Adaptation for Speech Enhancement”and not “A Statistical Model-based Voice Activity Detection” Electron. Lett., vol. 37, no.8, pp. 540-542, 2001

[31] Mark Marzinzik and Birger Kollmeier, “Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics”IEEE Transaction on Speech and Audio Processing, Vol 10, No. 2 ,Februrary 2002.

[32] ITU-T Rec. P. supplement 23, ITU-T Coded-Speech Database, Int. Tekecommun. Union, Geneva, Switzerland, Feb. 1998

[33] “Single Ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications”, vol.3, pp.1076-1079, 2004