國立政治大學資訊科學系

Department of Computer Science

National Chengchi University

碩士論文

Master's Thesis

大型網路語音會談中回音消除方法

Echo Cancellation In Large-Scale VoIP Conferencing

研 究 生：祁立誠

指導教授：連耀南

中華民國九十八年十二月

November 2009

大型網路語音會談中回音消除方法

Echo Cancellation In Large-Scale VoIP Conferencing

研 究 生：祁立誠　　Student：Li-Chen Chi

指導教授：連耀南　　Advisor：Yao-Nan Lien

國立政治大學

資訊科學系

碩士論文

A Thesis

submitted to Department of Computer Science

National Chengchi University

in partial fulfillment of the Requirements

for the degree of

Master

In

Computer Science

中華民國九十八年十二月

November 2009

大型網路語音會談中回音消除方法

摘要

隨著網路技術的發展，目前網路電話(VoIP)已有逐漸取代傳統電話的趨勢。尤其能夠允許多人同時在線上進行會談是其最大的優勢之一。但在多人參與網路會談時，因為聲音在空間中傳遞或反射等因素，使得由喇叭發出的聲音再次被麥克風收回，造成回音的產生。會談中只要有一位使用者的裝置發生回音時，回音訊號就會在與會者之間擴散，使得所有使用者均會受到影響，進而嚴重影響網路通話的進行。此狀況在參與會談人數越多時，發生機率越高，且對通話品質影響越嚴重。

傳統電話在一對一通話時，通常使用遠端回音消除機制(Near End Echo Canceller)，由接收端在接收聲音後先暫存在記憶體中再播放，再將麥克風擷取的聲音與事先暫存的訊號反向後混合，以抵銷回音。網路會談的環境下，由於沒有標準的聽筒設備，使得回音發生的時間難以預估。且多人參與的網路會談中，由於收聽者所聽到的聲音可能混合多個使用者說話的聲音與回音，使得回音訊號難以偵測。另外，由於網路傳輸的特性，回音訊號到達的時間與順序都難以預估，這使得回音消除機制在多人網路回談中經常失效。

本研究提出藉由語音動態偵測(Voice Activity Detection-VAD)的方式分辨回音訊號，藉由本研究所提出的語音能量 VAD 判定機制，能夠有效區別正常語音與回音的差異，即可有效的消除回音，同時發揮靜音抑制（Slience Suppression）的效果，阻擋不含語音內容的封包，降低網路頻寬耗用。本研究以自行開發的 VoIP 軟體進行實地測試實驗，實驗中顯示，我們的方法能消除 85%以上的回音。

Echo Cancellation In Large-Scale VoIP Conferencing

Abstract

With the prosperous development of Internet technology, traditional phone service is being replaced gradually by Voice-over-IP (VoIP) technology. One of the critical problem that is yet to be improved is the echo problem. Due to the difference in working environment, conventional echo cancellation technology may not work well on VoIP system. The echo problem is becoming more critical as the number of participants in a talk session increases. As long as one user fails to depress echos, every other participant in the conference will be infected. The more participant, the higher probability of echo infection.

We propose an energy based Voice Activity Detection (VAD) mechnism that effectively differentiate echo from speech signal. Our VAD algrouthm records a user's speech volume, and based on this information to determine whether the frame is echo or not. By applying this mechnism to network conference, we can filter out echo frames and suppress slience at same time to save bandwidth consumption. We experimented on a self-developed VoIP software platform, the experiment result shows that our method can eliminate more than 85% of the echo.

誌謝辭

『受人點滴，當銘記在心』—當我的論文逐漸成型的同時，我一再提醒自己：僅靠我個人的努力，是無法順利完成研究，取得學位的。我首先必須感謝的是我的指導教授連耀南教授，老師以無比的耐心與專業，帶著我這個菜鳥研究生踏入資訊科學的研究殿堂，在我跌跌撞撞的摸索過程中一路陪伴，總能在最關鍵的時刻從細節中抓出問題，為我指引方向。我還必須感謝張宏慶老師與蔡子傑老師的專業建議與提醒，讓我得以補足許多研究上的缺陷，並且對我的報告提出中肯的建議，使我受益良多。當然，我還必須感謝系上的每一位老師的辛苦付出，因為有老師們安排充實的課程，才能讓我在研究所期間充分學習計算機科學領域的相關知識。

除了老師，在研究所相處時間最多的當屬我的同門師兄師弟—也就是同一間實驗室的同學們。在研究方面，因為有你們的切磋，我才能思考的更周全；在課業方面，你們的合作讓我吸收課程的精要，順利取得學分；在生活方面，學長姐與學弟妹互相幫助扶持，使實驗室充滿溫情。我們的實驗室隨時充滿歡笑，都是你們的功勞！小 P 學長與怡萱學姐帶著剛成為政大新鮮人的我適應環境，育晟一路陪伴著我完成研究，若不是他的激勵，我絕對無法如期完成論文。山高，小 M 與筱慈讓實驗室充滿活力，且為我打點口試所需的瑣事。而智賢與啟禎學長則在專業知識方面惠我良多。此外其他實驗室的同學—文卿，東諺，國淵，文捷以及眾多學弟妹們，你們的陪伴讓我有勇氣踏出每一步。當然，最值得我感謝的還是我的父母—謝謝你們！因為有你們從小到大的拉拔與關懷，才有今天的我。

謹以此文獻給我的雙親與所有家人，謝謝你們的關心與照顧。

祁立誠 (卡比) November 2009

目錄

圖目錄

表目錄

第一章 簡介

1.1 多人/大型網路語音會談

近年來，為響應全球減碳運動，並節省差旅費，許多公司內部召開之大規模跨國會談均採用網路會談(conference)方式進行。通常在網路會談進行時，並不會特別採購新的硬體設備與租用網路服務，而是採用現行的終端設備(如 PC，PDA 或 Laptop 筆記型電腦等等)，而網路頻寬也多採取現有 ISP 公司所提供的服務(例如：ADSL，通常為下行頻寬 2~3Mbps，上行頻寬 384Kbps 左右)。目前常見的網路會談都是在此架構上執行多人的 VoIP (Voice Over Internet Protocol) 應用。VoIP 的運作方式是將使用者的聲音訊號包裝為訊框(Frame)，再透過網路封包(Packet)送至對話的對象，並且以相反的程式取出資料，將數位訊號轉為類比聲音。在使用上，VoIP 對一般操作者而言與傳統電話並無不同，但在進行大型網路語音會談時，由於上線人數增加，隨之所必須面對的挑戰也更為艱鉅。本研究的目的在於維持一定語音品質下，盡可能提高網路會談的與會人數。

1.2 網路會談的常見問題

當參與會談人數增加時，許多網路會談的問題就顯得更為嚴重。例如頻寬需求量，連線人數限制，封包遺失率或者聲音品質等等因素，以下提出幾個常見的問題：

- 總頻寬需求與 stream 數量問題：當使用者增加時，stream 數目隨會談成員增多呈指數成長。若與會人數為 N，則總共需要存在的 stream 數量為 C(N,2)，當連線人數上升至某個程度時，此數量即相當可觀。
- 無線網路的連線品質問題：若參與會談的某些使用者以無線網路傳輸聲音封包，需克服無線網路連線品質不佳的問題，包括低穩定性，低頻寬，delay 時間變動大(jitter 很大)，封包遺失率高，無線電波高雜訊及頻道競爭所造成的信號干擾等等問題。
- echo/noise epidemic 問題：當參與會談人數增加時，產生回音或噪音的機率也隨之增加，使得會談過程免於回音/噪音干擾的機率大幅降低。且回音會有遞迴的效應，即回音本身還會產生回音，如此反覆產生回音。

- time sequence disorder 的問題：亦即聲音不依照發送順序到達接收端，而發生先發言卻較慢被聽到的狀況。例如 A 成員比 B 成員更早發言，但 B 成員的聲音卻比 A 成員的更早被收到。

諸如以上問題，都是有可能在大型網路語音會談中所發生的，而針對以上種種問題，目前分別都有提出一些在應用程式端或網路架構上的解決方案，整體的概觀可由表 1 與表 2 說明，△代表這種解決方法能有效改善該問題，▽則代表此方法對此問題有負面影響，◇代表此方法有機會改善此問題，但不排除加重此問題的可能性：

表 1：會議通話常見問題及解法(Client Side Solutions)

VoIP Conferencing Issues and Solutions		Key Issues							
		Bandwidth Consumption	No. Of Streams	Echo/Noise Epidemic	Time Sequence Disorder	Delay	Packet Loss	Jitter	Voice Quality
Client Side Solutions	Silence Suppression	△	△	△					
	Echo Cancellation			△					△
	Adaptive Data Rate	◇							△
	Adaptive Packet Payload	◇				◇			
	Redundant Voice Streams (without Piggyback)	▽	▽				△		△
	Redundant Voice Streams (with Piggyback)	▽					△		△

表 2：會議通話常見問題及解法(Network Side Solutions)

VoIP Conferencing Issues and Solutions		Key Issues			
		Bandwidth Consumption	No. Of Streams	Node Stress	Delay
Network Side Solutions	Full Mesh Broadcasting	▽	▽	▽	△
	P2P Multicasting	△	△	△	▽

針對網路會談所發生的種種問題，在使用者軟體(client)端與網路(network side)端各自有一些解決方案，以下分別進行分析。

Client Side Solutions

1. Silence Suppression：此法偵測聲音訊號中實際是否包含語音資料並阻擋之。當 Silence Suppression 偵測到一個聲音訊框中並未包含任何有意義的語音資料時，那就阻擋這個訊框傳送至網路上。在網路會談中同一時間通常只有一位或少量發話者需要上傳資料，其餘使用者僅需接收，使用 Silence Suppression 擋下非發話者之封包即可節省頻寬之耗用並減少連線數量。同時 Silence Suppression 也能避免一些不必要的雜音或回音混入會談聲音中，提高會談的品質。

2. Echo cancellation：此法用來消除與會者麥克風收到喇叭聲音所造成的回音(acoustic echo)以及傳統 PSTN 電話線路中，在線路混合(hybrid circuit)時因為阻抗不匹配所造成的回音，消除回音與可提高會談之品質。

3. Adaptive data rate：採用可變動的 data rate (例如變動取樣頻率或壓縮比)。採用此方法，在網路資源允許時(有足夠頻寬可使用時)，提高取樣頻率或降低壓縮比，即能提昇會談音質，而當網路資源不夠時，可降低 data rate，得到較低品質但資訊量較少的語音封包。

4. Adaptive packet payload：每個網路封包內通常可裝載數個聲音訊框(frame)。若將一個封包盡量裝滿，雖可提高網路的使用效率，但卻可能增加聲音的延遲(delay time)。反之，若為了節省時間而將每個訊框裝載在一個單獨的封包中時，聲音即時性較好，但卻降低封頻寬利用率(增加 overhead)。因此可以採用可變動的方式，根據網路可用頻寬調整每個封包內裝載的訊框數量，當頻寬足夠時，可降低 packet payload，反之則將多個訊框放於同一個封包內以節省頻寬。

5. Redudant voice streams (without piggyback) ：網路語音通訊通常採用 UDP 封包傳送語音資料，因此不能保證封包必然到達，一旦封包遺失，可能對聲音品質造成影響。

此方法將同一語音封包以多個網路串流(stream)重複傳輸數次以降低封包遺失率(packet loss rate)，提昇音質。但此方法必然會提高頻寬使用量與串流連線數量，對網路造成較大負擔。

6. Redudant voice streams (with piggyback) ：將同一個語音封包傳輸多次減少遺失率，提昇音質。但此方法不建立額外的串流連線，而是將前一個語音封包附加於下一個封包內一併傳輸，如此網路頻寬使用量雖會提高，但卻不用增加串流數量。

Network Side Solutions

為了解決網路會談的相關問題，在網路傳輸架構方面，有別於傳統需要伺服器協助轉送資料的 centralized server 架構，相關研究提出了兩種不需伺服端的傳輸架構(serverless)。這兩種方法分別為：

1. Full Mesh Broadcasting：參與會談的節點之間直接建立連線，每位使用者將語音封包直接傳送給會談中的所有與會者。此方法需要耗費相當大的網路頻寬與連線數量，對於每一個節點而言，負擔相當的沈重，但由於資料不需其它節點轉送，故此法最大的優點為延遲時間較短。

2. P2P Multicasting：此方法根據網路會談中每個節點的資源，每位發話者建立一個用來傳輸聲音資料的 multicasting tree，用以廣播封包到所有其他成員。此方法使得發話者僅需將資料傳送給下一個轉送點，對於每個與會者而言需要建立的連線數量通常也較 Full Mesh 架構少，整體節點負擔較輕。但其最大缺點為封包必須經過多個轉送點，延遲時間必然增加。

1.3 回音現象

本研究考慮的問題為如何改善 echo epidemic 對於通話品質所造成的影響，同時作到 Silience Suppress 降低總頻寬使用量。目前針對 echo epidemic 的問題，最常使用的解決方案為 Acoustic Echo Cancellation (AEC) ，以下分別針對回音現象，現有回音解

決方案與本研究所提出的解決方案一一提出說明。

聲音由音源發出後，若被反射回音源處即產生回音。若第三者直接接收到音源發出之聲音後，又接收到反射的回音，則接收到兩次以上的相同聲音。回音的情形在全雙工(full-duplex)語音通訊系統中經常發生。由於全雙工系統能夠同時傳送與接收聲音訊號，因此收聽者(listener)本身也同時為說話者(talker)。而聲音可在固體與空氣中傳導，如此可能使收聽者的收音裝置收到自己的播放裝置所發出的聲音，回音的現象就此產生。

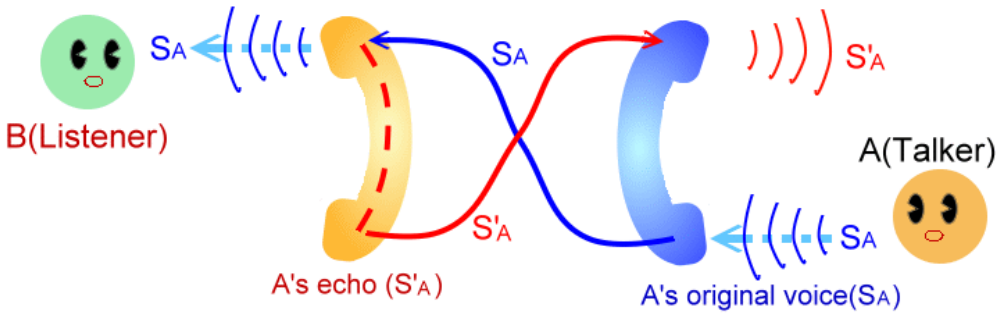

圖 1：基本回音現象

參照圖 1，以傳統電話(PSTN)系統為例，若說話者(A)說出的聲音，傳至收聽者(B)後，經由電話聽筒的傳遞，再次被麥克風收回而由說話者 A 的聽筒播放出 A 的聲音，此即一般所稱呼的『聲學上的回音』(acoustic echo)[15]。除此之外，PSTN 系統還可能因為訊號由於類比訊號混合時的阻抗不匹配導致信號反射的回音，此種回音則稱為 Network Echo。但此種回音僅發生於純類比系統中，在數位通訊系統中並不會發生[4]。

另一種由電話所造成的回音狀況常常發生於電視或廣播電台的 Call-in 節目中。當觀眾打電話與主持人溝通時，如圖 2 所示，若講電話的觀眾的電話話筒收到電視發出自己的聲音，則回音也會產生，使得主持人與所有電視觀眾都會聽到不斷反覆重疊的回音。

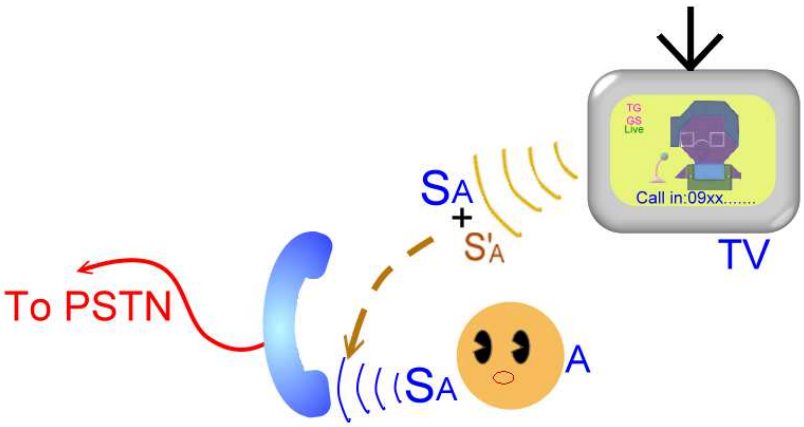

圖 2：Call in 節目中出現的回音狀況

1.4 回音消除基本原理

若要消除前述的由聲音上傳輸造成的回音，最簡單的方式就是加入回音消除機制(echo cancellation)。對傳統電話而言，由於有標準的聽筒裝置，因此很容易預估回音的強度與傳遞所需要的時間，建構此回音消除機制並不困難。

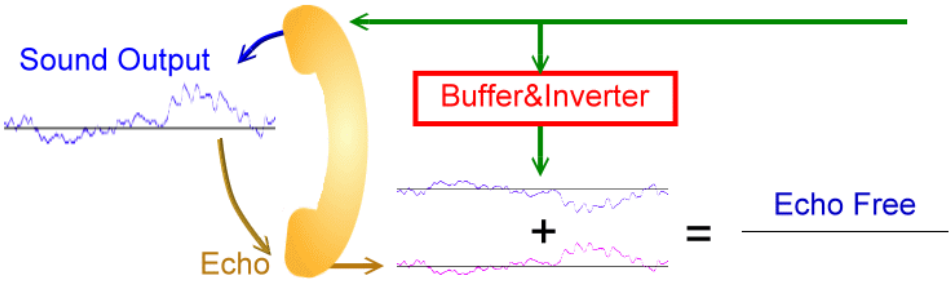

圖 3：傳統電話中的基本回音消除機制

如圖 3，聲音在由聽筒播放之前，先存入暫存器之中，並且加以反向(inverter)，當話筒收回聲音訊號之後，再將先前存下的訊號經過適當衰減後與此相加，即可消除由話筒收回的回音[2]。

1.5 VoIP 中的回音

在VoIP通話進行時，很多使用者並不會特別配備專用的設備，而是採用電腦原有的喇叭與麥克風進行網路通話。如此一來聲音由喇叭播放後，在空氣中傳輸，經過空間反射後再次被麥克風收回，同樣會產生回音，此種現象的發生將可能影響通話的品質，嚴重時甚至影響會談的進行。例如圖4中說話者(A)的聲音傳送至收聽者(B)的喇叭播放後，

經過牆壁反射，又被麥克風收回，返回至A的喇叭播放，造成回音。聲音自喇叭至麥克風的路徑稱為”Accoustic path”。

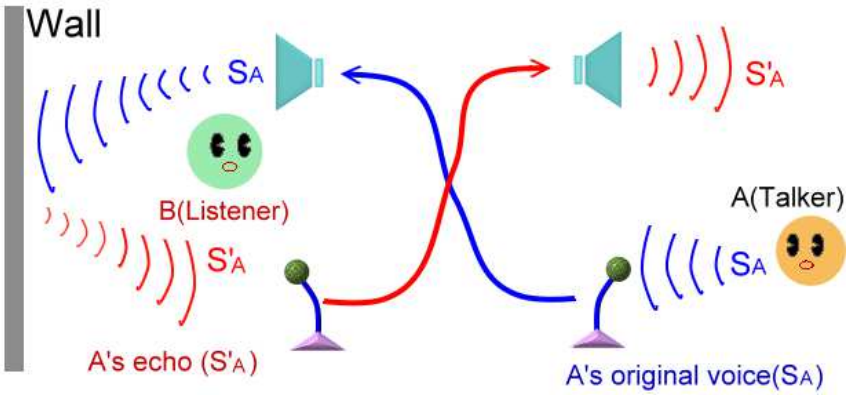

圖 4：因為聲音的傳輸或反射而造成的回音

由於 VoIP 沒有標準的聲音播放與擷取設備，回音在 Acoustic path 上的傳輸延遲時間變動極大，再加上多人參與會談時，回音的條件與特徵可能變得難以預估，因此 VoIP 系統中的回音問題並不容易徹底解決。

在兩人以上的 VoIP 會談中，每部參與會談的裝置(通常為個人電腦)通常都具有回音消除機制，使得即使麥克風收到喇叭所播放的聲音時，也能夠有效消除回音，不至於影響會談之進行。

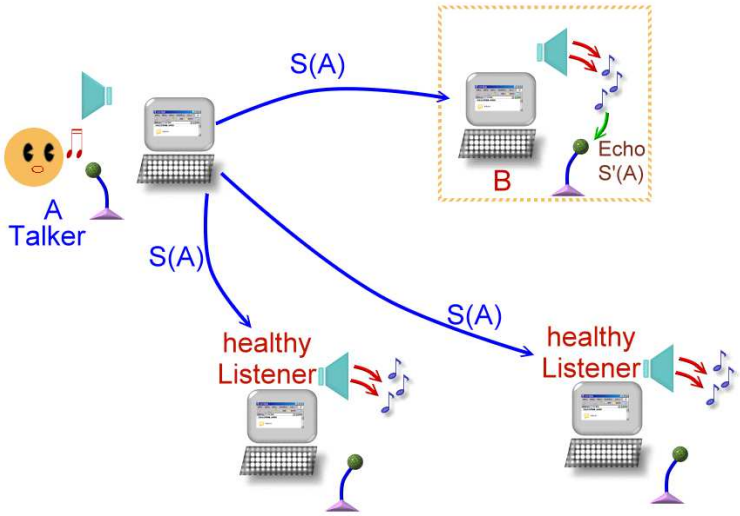

圖 5：回音消除機制正常時的多人網路會談

如圖 5 為多人網路會談的狀況，其中一位與會者 B 的喇叭與麥克風距離太近，使得由喇叭放出的訊號被麥克風收回。當回音消除機制正常啟動時，此訊號在裝置 B 就被順

利消除，而不至於對網路會談造成任何影響。

然而電腦軟硬體都有失效的機會，使得消除機制未必能夠正常運作，多人會談中只要有一個參與者的回音消除機制失效，成為回音的產生者而產生回音時，則可能對整個會談造成影響，導致所有使用者的通話都受到干擾。雖然每部參與會談裝置的回音消機制故障率很低，但當參與會談使用者大量增加時，則回音發生率相對提高。若單一設備音效卡之回音消除機制故障機率為 x，則當 n 個使用者參與會談時，沒有回音存在的機率即為：$(1-x)^n$。單一使用者故障與整體產生障礙的分析如圖 6：

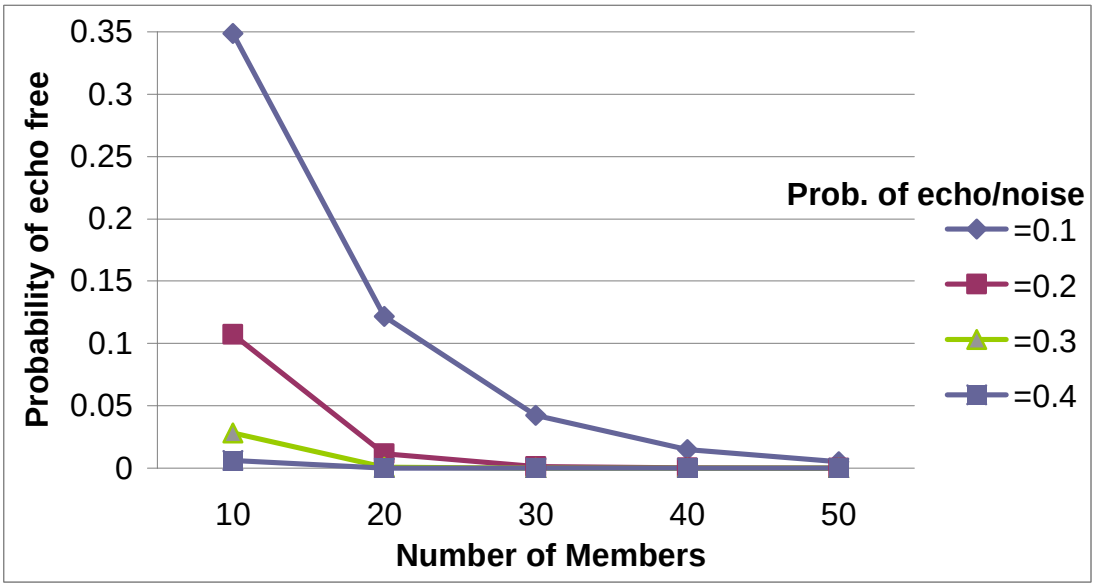

圖 6：單一使用者故障與整體故障率

若單一使用者故障率為 10%，則十人參與會談時，沒有回音存在的機率僅為 $(1-0.1)^{10}$=34.9%，由此可見參與人數越多時，免於回音干擾，順利進行會談的機率越低。根據我們利用 Skype 於台美兩地所進行的五人以下多方語音會談實驗所測得的結果顯示，回音消除機制部份失效(因回音效應而產生的噪音)的機率很高，全部失效(產生與原始語音相同的回音)的機率則較低。

1.5.1 單一回音產生者

如同前述網路語音會談中，有其中一部裝置(B)的喇叭發出的聲音被麥克風收音，且該裝置的回音消除機制並未正常工作，這將使得 B 成為回產生者(即 Echo Generator)，此時狀況如圖 7。

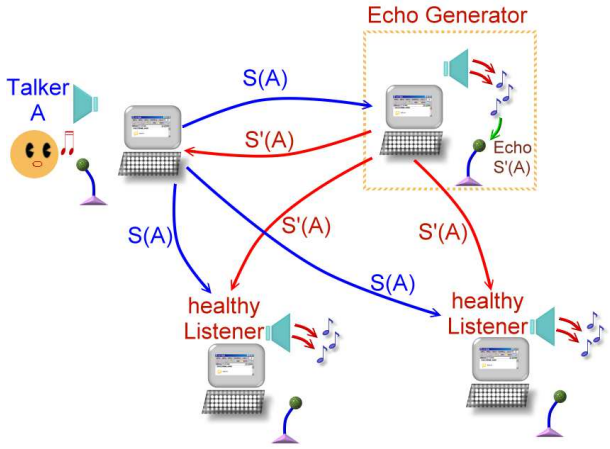

圖 7：一個 Echo Generator 的回音影響

圖 7 中 A 為發話者，其說話的聲音訊號為 S(A)，B 為回音消除機制故障之 PC(即回音產生者)，其造成的回音為 S'(A)。此時，當發話者說話後，回音將會從以下兩種路徑干擾參與通話者：

- Talker -> Echo Generator -> Talker：發話者聽到自己的 Echo，即 S'(A)
- Talker -> Echo Generator -> healthy Listener：其他收聽者聽到發話者的聲音以及 Echo，即 S(A) + S'(A)

因此，只要有任何一個裝置的回音消除機制故障，則其產生的回音就會干擾發話者以及其他所有收聽者，造成整個網路會談充滿回音的聲音訊號，嚴重影響通話品質。

1.5.2 多個回音產生者

而若在一個會談中有超過一個以上的 Echo Generator 存在時，問題會變得更為複雜，例如當同時有兩個 Echo Generator 時，回音訊號的反射如圖 8：

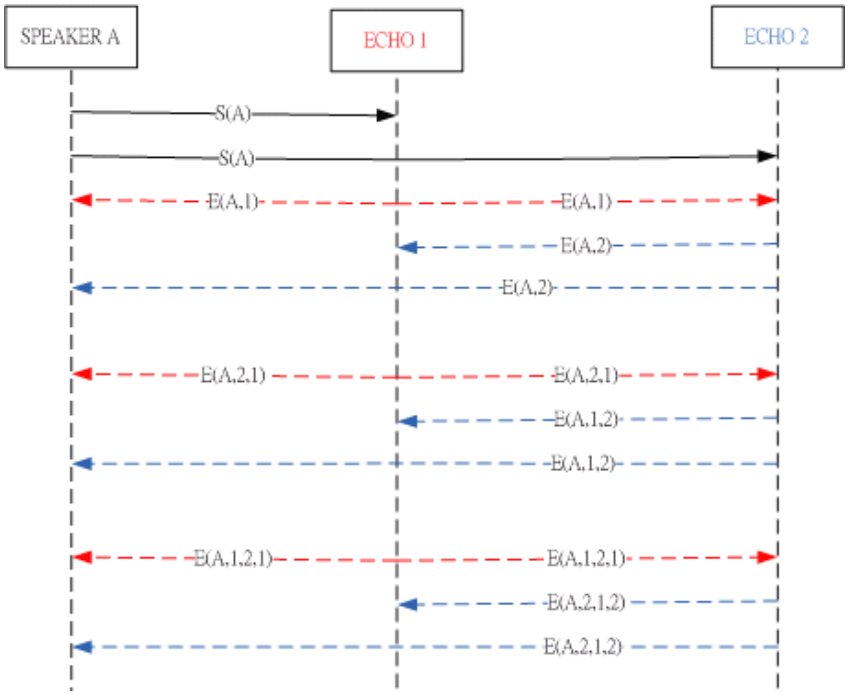

圖 8：會談中有兩個 Echo Generator 存在時的回音傳遞情形

當說話者(Speaker A)說出一段語音 S(A)後，傳遞給兩位 Echo Generator，則會由 Echo Generator1 與 Echo Generator2 分別傳回一次回音，接著，由 Echo Generator1 傳給 Echo Generator2 的回音會再次產生回音，傳給發話者與 Echo Generator1，如此一來，只要任一發話者說出一段語音，兩個 Echo Generator 將不斷重複產生回音，直到聲音衰減至無法收音為止，此時的回音傳輸路徑為：

- Speaker A -> Echo Generator A & Echo Generator B
- Echo Generator A -(回音)-> Speaker A & Echo Generator B
- Echo Generator B -(回音)-> Speaker A & Echo Generator A
- Echo Generator A -(回音的回音)-> Speaker A & Echo Generator B
- Echo Generator B -(回音的回音)-> Speaker A & Echo Generator A
- ……..Loop，直到聲音能量衰減至無法收音

在此狀況下，比起只有一位回音產生者的狀況更加嚴重，只要有一位使者說一句話，所有與會者就會不斷聽到重複遞迴的回音訊號，使得會談嚴重受影響，甚至難以進行。

1.5.2 Proximity Problem 造成的回音

所謂的 Proximity Problem 為電腦之間距離太近時，所發生一種聲音傳遞上造成的干擾問題[7]。若兩部電腦距離太接近時，則使用者的聲音（假設只有一個人說話）及兩部電腦的喇叭聲音，同時被兩部電腦的麥克風收回，此狀況發生的狀況如圖 9 所示。則此時兩部電腦的喇叭將會同時放出包含發話者以及不斷重複的回音，且如此混亂的聲音將再次被雙方的麥克風收音，導致此現象不斷循環。如此一來，將會造成複雜且混亂的回音加上互相收音所造成的雜音。就如雪崩效應一般，收回的聲音被放大後播出，接著又立刻被收音且播放，不斷反覆而造成類似震盪的刺耳噪音。

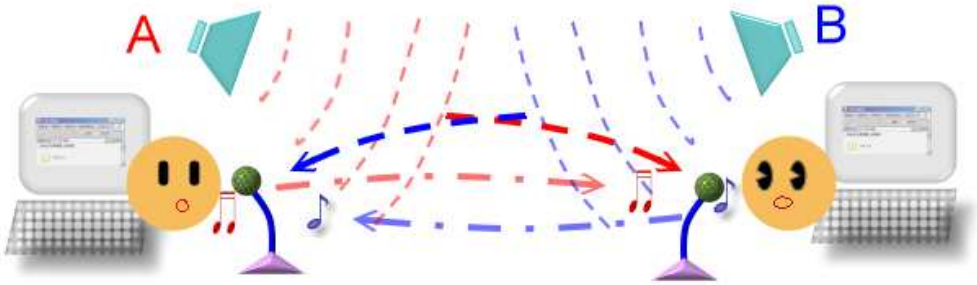

圖 9：Proximity Problem 現象

1.6 名詞定義

a. **Echo Generator**：即回音消除機制產生故障，而使得會談中其他使用者聽到回音產生的裝置，亦即此裝置為 Compromized Node。反之，功能正常者則稱為 Healthy Listener。

b. **Direct Echo**：由 Echo Generator 所產生，使得 Talker 聽到自己的聲音的回音。

c. **Indirect Echo**：網路會談中其他 listener 聽到 Talker 聲音的回音。

d. **Infected Conference**：因為有 Echo Generator 參與，而造成含有 Indirect Echo 亂竄的網路語音會談。

1.7 研究動機與目的

由於目前各種針對回音消除的方法，大多僅針對一對一的通話所產生的回音進行消除，且考慮的僅有聽筒到話筒間固定長度且短距離造成的回音，針對距離不固定，甚且

參與會談人數增加時所造成的回音狀況均沒有加以歸納分析，以至於常常無法正確的消除回音。雖然現在的電腦音效卡中均有內建回音消除演算機制，VoIP 軟體中也有加入類似的判斷與修正機制，但卻常常發生失效的狀況，更常見的是部份失效，亦即回音消除不完全而產生噪音，且當多人進行大型網路語音會談時，經常由於一個使用者的設備出現問題，而造成整個會談的通話品質受到嚴重干擾，以至於難以溝通。

此外，當與會成員沒有在說話時，傳統的會談程式仍然會持續將空白的語音封包送至網路上，如此將造成網路資源的浪費。基於以上理由，本研究希望分析在大型網路語音會談中，回音消除機制失效的成因與語音的特性，以此設計消除回音與 Slience Suppression 的方法，抑制會談中不必要的回音與靜音，確保語音會談正常進行的方法，同時節省網路頻寬。

第二章 背景與相關研究

本章針對目前已經存在的回音消除機制相關作法與其歷史進行說明，探討它們的實作原理與尚須改進的部份。

2.1 回音消除技術演進

在 1950 年代以前，電話系統中並沒有回音消除機制存在。當時的通訊系統為了消除回音對通訊造成的干擾，而採用回音抑制(echo suppression)的方式降低回音。由於當時的通話系統均為一對一通話，沒有多方通話技術出現，因此人與人的溝通是以半雙工(half-duplex)方式進行，亦即同一時間僅其中一方在說話，另一方是收聽者[20]。此回音抑制機制會判定電話的那一方是說話者，則在說話期間就保留正常的語音，將收聽者回傳的訊號視為回音進行衰減或阻止其傳輸，以達到回音抑制的目的。此方法雖有效消除半雙工通話中的回音，但回音抑制機制的判定速度往往跟不上發話者切換的速度，如此可能導致使用者開始講話時的語音遺失，或者雙方同時講話時無法判定那一方是說話者而造成誤判產生回音[21]。

直到 1970 年代，隨著半導體的進步，市場上才逐漸出現回音消除機制的產品。此時的技術開始採用前述的以訊號暫存與相減的方式消除回音[11]，以此方式取代先前回音抑制機制的信號開關。直到 90 年代，隨著 DSP 數位訊號處器普及，回音消除機制才逐漸整合於電話交換機內，並且能更精準的消除延遲時間不固定的回音。

2.2 回音消除原理

如同前述，在 VoIP 系統中，不像傳統電話，由於沒有標準的聽筒裝置，因此回音的延遲時間與音量/失真狀況相當難以預估。所以，針對此狀況需要加入更多的判斷機制來掌握回音的狀況，並即時的進行消除。例如 Perry P. He 等人提出的"Network Echo

Cancellers: Requirements, Applications and Solutions"[4]一文中所提出的解決方案，此系統與傳統電話的回音消除機制類似，在受話端加入回音消除裝置，但此裝置包含了訊號即時比對的裝置：

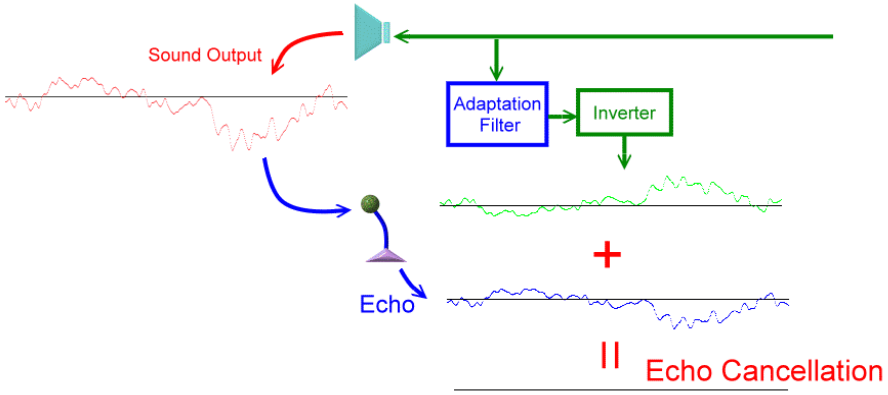

圖 10：VoIP 系統中回音消除機制

圖 10 為 VoIP 中所使用的回音消除機制示意，其中包含的主要元件為 Echo Inverter 與 Adaptation Filiter，其功能分別為：

a. Echo Inverter：此部份包含了暫存記憶體(buffer)與訊號反向的運算機制。目的為將接收到的語音訊號暫存，若經由演算判斷回音發生時，則取出暫存器中的訊號做反，相並與回音訊號相減，以達到抵銷的目的。

b. Adaptation filter：此部份包含預期回音傳回的時間，判斷回音是否發生與決定相減訊號的衰減程度等等。透過 LMS/NLMS 等演算法[9]，將可能有回音的訊號做誤差還原後，逐一與原始訊號做比較，判斷回音是否發生。若回音發生時，則由此部份決定回音傳遞時間與衰減幅度等等因素，將前述 Echo Inverter 中原始訊號取出，經過處理後相減以消除回音。

通常，此部份需要不斷進行暫存，濾波器參數運算調整，比對，訊號處理與相減等等複雜的計算工作，因此在目前絕大多數的系統中，都是由 DSP (Digital Signal Processor-數位訊號處理器) 進行運算[8]，以避免如此大量的計算影響 VoIP 等通訊軟體的正常工作。

2.2.1 回音消除演算法

在VoIP系統中最常使用的回音消除演算法為NLMS (Normalized Least-Mean-Square

Algorithm，正規化最小平方誤差演算法) 或LMS (Least Mean Square Algorithm，最小平方誤差演算法) 。此演算法為1960年由B. Widrow等人所提出[17][18]，最初為用來作為信號誤差修正用，後來被DSP系統用來實做回音消除裝置。例如U. I. Choudhry等人提出的"A Highly Adaptive Acoustic Echo CancellationSolution for VoIP Conferencing Systems"[1]一文中，就採用LMS演算法。由於經過空間傳遞或反射後的回音訊號，無論在頻率或強度上均會與原本的聲音有所誤差，因此需要將此誤差值以計算的方式試圖還原，才能與再次收回的聲音訊號做相似度比對，找出正確的回音訊號，並以相減的方式消除。此文中所使用的LMS演算法算式如下：

$$a_k(i+1) = a_k(i) + 2\beta \frac{e(i)y(i-k)}{N\sigma^2} \cdots\cdots(1)$$

其中：

- a_k為濾波器中的第k個係數
- i為取樣的編號
- N為濾波器係數的數量
- β為適應步階常數(adaptation step)，用來調整收斂時間與調適品質
- e(i)為訊框i中剩餘的回音訊號
- y為Listener的語音訊號
- σ^2為參考信號強度

在上面所有參數中，β收斂速度的決定參數，若β的值很小，則代表收斂速度跟著變慢(跟不上訊號變化)，但卻能得到較精確的結果。相反地，β值大則表示速度雖然快，但一方面可能較不精確，且可能得到發散的結果。

其中N為濾波器係數的數量，針對每一個聲音訊框中的每個濾波器係數a_k，都需要經過上式運算。而N的數量取決於回音路徑的長度：例如在8Khz取樣時，以表三中所列之傳輸時間而言，5.88~14.69ms的延遲時間至少需要48~118個係數才能夠有效偵測回音。若需要偵測的delay時間越長，則需要的係數量就越多，同時需要花費更多的運算量來偵測回音訊號。

2.3 回音消除方法分類

回音消除機制根據實做的使用者端不同，可分為以下兩種方式：

2.3.1 Listener Echo Cancellation

這是『由回音產生者消除回音』的方法，亦即由麥克風擷取聲音(發話者)一方自己進行回音消除的動作。若此方法使用在一對一的傳輸狀況下則稱為 Near End Echo Cancellation。

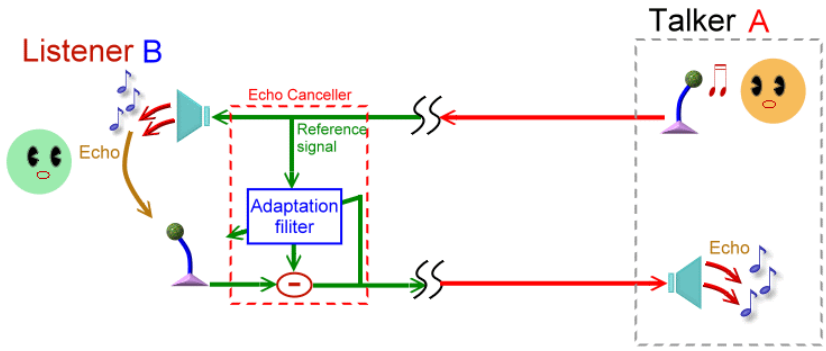

圖 11：Listener Echo Cancellation 機制

如圖 11，此方法是由產生回音的收聽者 B 負責消除回音。當收聽收到來自發話者 A 的訊號時，會先經過暫存。當麥克風收到回音時，就將此包含回音的訊號與原本暫存反向後的訊號相減，以去除回音。

2.3.2 Listener Echo Cancellation 失效原因

在 VoIP 系統中，雖然每部參與會談的裝置的都有 Listener Echo Cancellection，但卻常因某些因素造成此機制失效，這些原因包含：

- VoIP 沒有標準的話筒與聽筒，延遲時間難以計算，且回音的音量也難以預估，不一定能正確將回音訊號抵銷。
- 根據使用者所使用的收/放音設備不同，聲音傳輸距離造成的回音延遲時間差異相當大。可能的聲音傳輸距離與其所需時間範圍如表 3 所列(以音速在空氣中速度為 340.29 m / sec 計算)：

表 3：Acoustic 延遲時間

Echo 路徑	距離	Delay Time
手機聽筒---話筒	10cm	0.29ms
一般電話聽筒---話筒	15cm	0.44ms
喇叭---麥克風	50cm	1.47ms
喇叭---牆壁反射---麥克風	2~5m	5.88ms~14.69ms

- 由麥克風擷取的回音可能經過牆壁反射而造成相位相反，此時若再與反向過的訊號相加，則不但不能減少回音，反而使回音更為嚴重。另外，由於環境(包含空間，材質，距離等等)因素，由麥克風收到的回音訊號可能會在相位與頻率上產生嚴重失真，而難以判定回音是否產生。

2.3.3 Talker Echo Cancellation

與前述相反，這是『由回音接收者消除回音』的方法。由聽到回音的收聽端在播放聲音之前，先將聲音中的回音去除再播放。若此方法使用在一對一的傳輸狀況下則稱為 Far End Echo Cancellation。[5][6]

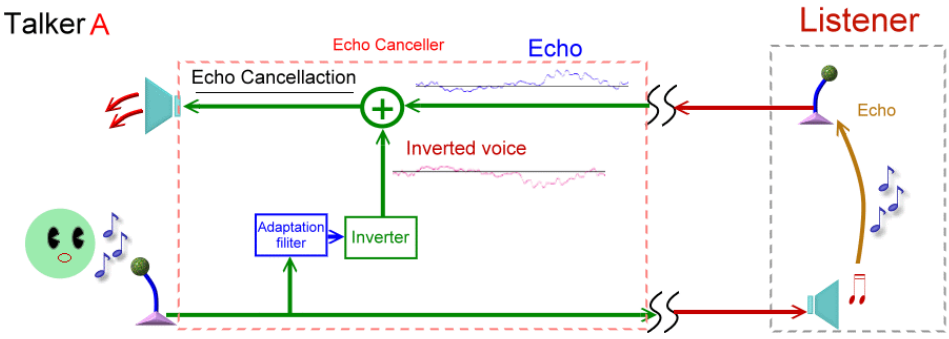

圖 12：Talker Echo Cancellation 機制

如圖 12，此方法是由說話者(A)在送出聲音之前，先將自己的聲音存下，當回音訊號透過收聽端傳回後，由說話者 A 取出先前暫存的訊號，比對回音是否存在，一旦發現回音存在，就將訊號經過衰減後相減。

2.3.4 Talker Echo Cancellection 的挑戰

如前述，Listener Echo Cancellection 有許多導致失效的因素。但同樣的，若要設計

一套由 Talkerer 端主動消除回音的機制也會有相當多挑戰存在，尤其在多人網路會談中，某些問題將變得比一對一時更加複雜，這些問題列舉如下：

1. 回音時間難以預估

在語音會談中，若使用者之間距離相當遙遠(例如跨國 VoIP 會談)，則回音到達的時間可能難以預估，其傳輸途徑如圖 13：

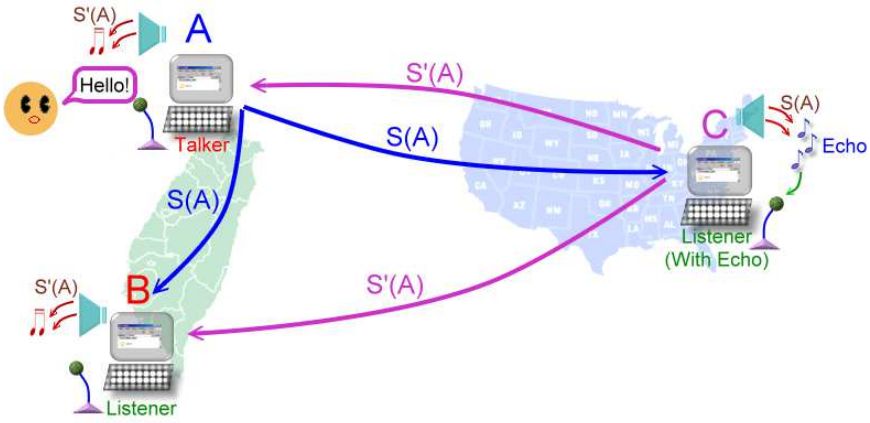

圖 13：遠距離 VoIP 會談時的傳輸時間難以預估

如圖 13，發話者 A 說出 S(A)聲音後，若在遠方的另一使用者 C 為 Echo Generator，則 A 聽到回音訊號 S'(A)的延遲時間為：

$$T_{echo} = 2\left(T_e + T_d + T_t + T_p\right) + T_f \cdots\cdots(2)$$

其中：

T_{echo} 為說話者聽到回音所需的總延遲時間。

T_e 為語音訊框編碼(encode)時間。

T_d 為語音訊框解碼(decode)時間。

T_t 為封包傳輸時間(transmition delay)。

T_p 為傳輸延遲(propagation delay)。

T_f 為聲音在空間中傳遞的時間。

在這些時間中，T_c 與 T_t 為固定值，很容易事先預估。而如同前述，由於 VoIP 沒有固定的聽筒與話筒，故 T_f 的傳遞時間並不容易預估。另外，說話者與收聽者之間距離越遠，所需要的網路傳輸時間 T_p 也隨之增加，若再考慮網路擁塞時封包在 router 中排隊(queueing)等待傳送的時間，則更加難以預估。因此，當說話者與使用者之間的距離越

遠或經過的網路 hop 數量越多，Listener 端的回音就越難以消除。

2. 說話人數與聲音到達順序

在一對一的 VoIP 中，通常都可假設絕大部份時間的講話者只有一位，即一方為說話者，另一方必定為收聽者，即使有兩者同時說話，也只有短暫時間。不過一旦與會人數很多時，多人同時說話的機率隨之增加。

假設參與會談人數為 N 個時，同時說話的組合有 2^N 個組合：假設每個使用者都可能有『說話』或『不說話』兩種狀態，則會有非常多組合產生，相對的特徵值就會有非常多組。在此情形下，多個人聲混合後特徵值不明顯，聲音會彼此調合因而失去自己的特性，這將使得前述演算法難以進行聲音特徵比對，判斷回音是否存在。

此外，在封包傳輸時，有可能因為網路擁塞或路徑選擇等因素，造成封包不依照順序到達(即發生 time sequence disorder 的問題)。這將導致回音和正常語音可能會同時出現，甚至可能先聽到回音，再聽到正常語音。此狀況發生時，也會讓回音判定演算法失效。

3. 運算時間之限制

對於 VoIP 而言，使用者能忍受的 mouth-to-ear delay time(聲音由說話者傳送到收聽者的總延遲時間)不可超過 300ms，而在 VoIP 系統中，除了傳輸延遲時間外，還多了編解碼所需時間，使得容許的回音消除運算時間被大幅壓縮。

聲音送出時必須經過類比數位轉換(A/D)，訊框編碼(Encoding)與封包化(Packetize)的過程(如圖 14)。

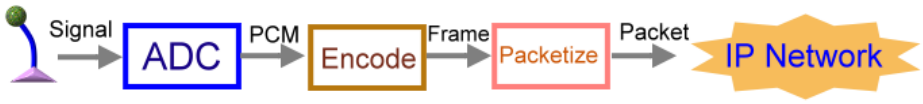

圖 14：聲音訊號由麥克風傳至網路的必經過程

相反的，在接收端也必須經過解封包，訊框解碼與數位類比轉換(D/A)的過程，圖 15 為由網路收取封包後，還原為聲音的過程。

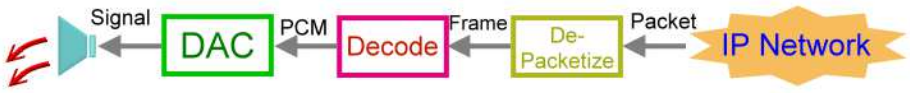

圖 15：由封包還原為聲音的過程

在這些必須消耗的時間以外，留給 Talker 端執行回音消除運算的時間就極為短暫，對程式實作的挑戰相當大。

4. Mixer 造成的困擾

在 P2P Multicast 的 VoIP 系統中，可能會有某些裝置擔任 mixer 負責將來自數個使用者發出的聲音先進行混音後再轉送[3]。

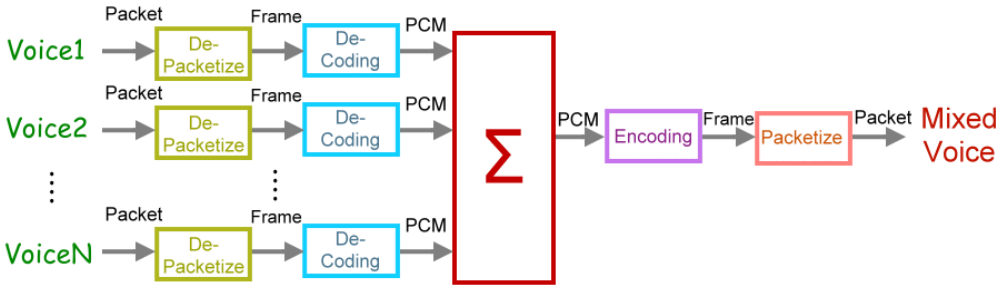

圖 16：mixer 的混音程序

此時，若使用者收到的是混合後的聲音，則不但回音分析的複雜度大幅提昇，而且要消除混雜於其他聲音之內的回音亦極為困難。

5. 使用者說話行為造成誤判

在多人會談時，使用者們有些說話行為會造成『很像回音，但卻不是回音』的情況產生。以下兩個狀況就是經常發生的情形。

例 1，在選舉造勢場合中的『一呼百應』:

台上主持人:『大家說對不對？』

台下的眾人:『對！』

例 2，軍隊或團體活動時的『呼口號』:

主席:『復興中華文化！』

眾人:『復興中華文化！』

以上這種說話的行為，對於程式而言都與回音的狀況非常類似(具有類似特徵的語音在短時間內重複出現)，但實際上卻是正常的會談內容，此時就很容易發生誤判。

2.4 一對一 VoIP 回音消除機制

若將前述之回音消除機制應用於 VoIP 之中，如 G. Periakarruppan 等人提出的 "Packet based echo cancellation for VoIP networks" [11]一文中所設計的 PBEC 系統(Packet Based Echo Canceller)，此為一個獨立於網路中的系統，其作用方式為：由 VoIP Gateway 抓取 VoIP 的語音封包，經過封包拆解，取出其中語音資料後，採用前述 Echo Canceller 方法消除回音，接著再將資料重新包回封包內，重新送入 VoIP Gateway 內，傳送至通話的對象。其網路架構如圖 17：

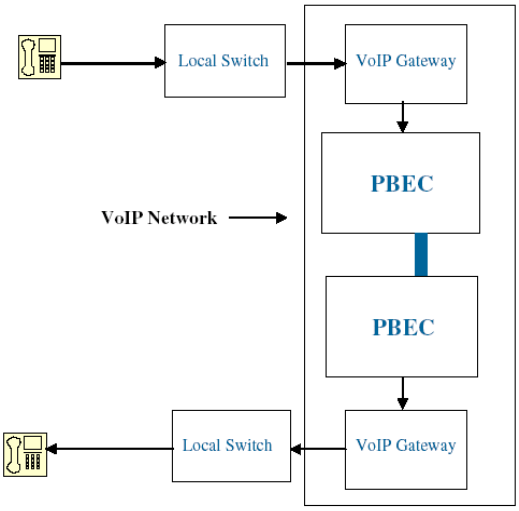

圖 17：Packet-Based 回音消除系統

此架構中，其 PBEC 內部採用 Listener Echo Cancellation，其採用的回音判斷消除機制為 NLMS 演算法(Normalized Least-Mean-Square Algorithm，即正規化最小平方誤差演算法)，此演算法被許多數位系統應用於濾波器中用以消除噪音、提供頻譜修正組成、回音消除或進行信號檢測之計算。

2.5 總結

回音對於語音通訊而言是影響通訊品質的一大障礙，因此過去數十年來回音消除系統一直是備受矚目的研究議題。相關研究提出了從類比電路至數位運算等各種方法用於解決電話系統中的回音問題，目前在傳統電話一對一的溝通上，回音消除系統幾乎都能穩定的發揮作用，有效消除回音。但隨著網路會談的流行，回音又成為新的挑

戰：VoIP 沒有像傳統電話一般的標準話筒，因此難以預估回音延遲時間。而當與會人數增加且網路品質不穩定時，回音的狀況更加複雜。

雖然目前的 VoIP 軟體與電腦音效裝置均提供回音消除機制，但卻經常失效(或回音消除不完全)而導致回音產生。在多人會談中，只要一位使用者的回音消除機制故障，就會導致所有與會者受到干擾，嚴重影響通話品質。本研究的目標即為在不增加額外硬體設備與運算負擔的前提下，提出一套回音消除機制的方法，並可兼具靜音消除的功能，大幅降低網路會談的頻寬需求。

第三章 MET VAD 靜音及回音消除機制

3.1 需求分析及研究目標

如同前述，雖目前用於進行 VoIP 的裝置均內建有 Listener 端回音消除機制，但卻常因為各種原因而失效造成 Indirect Echo 產生。

因此，本研究希望找出一套適用於大型網路語音會談中消除回音的方法，當網路會談進行中，因為與會者裝置的 Listener 回音消除機制失效而產生 Indirect Echo 時，能有效判別回音訊號與正常語音的差異並消除之，以避免因為回音的影響而造成 Infected Conference，而影響聲音品質。此判定機制除了針對回音以外，同時能分辨訊號中不包含語音的部份，阻止會談程式將此部份語音封包傳送至網路，藉以節省頻寬的使用。

3.2 解決方法

本研究提出的解決方法為在每一部參與語音會談裝置加入一套判定麥克風收到的聲音是否為正常說話語音的機制 (若該裝置原有 Listener 端回音消除機制，則置於其後)。當沒有回音產生或回音消除機制正常運作，送入正常的使用者說話聲音訊號時，就將此聲音以正常方式送至網路中。反之，若因為有回音存在或回音消除機制故障或使用者沒有說話，導致麥克風聲到回音或是靜音訊號時，就由此機制阻擋訊框，不送入網路，以避免此回音訊號成為 Indirect Echo，同時節省網路頻寬的耗用。

在本研究提出一套最大能量追蹤 VAD (Maximum Energy Tracking VAD - MET VAD) 語音動態偵測作為此判定機制，由此 VAD 負責擋下非正常語音的回音以及靜音訊號。

3.3 VAD 語音動態偵測

語音動態偵測 VAD (Voice Activity Detection) 的目的為找出聲音訊號中，實際含有語音內容的區段。

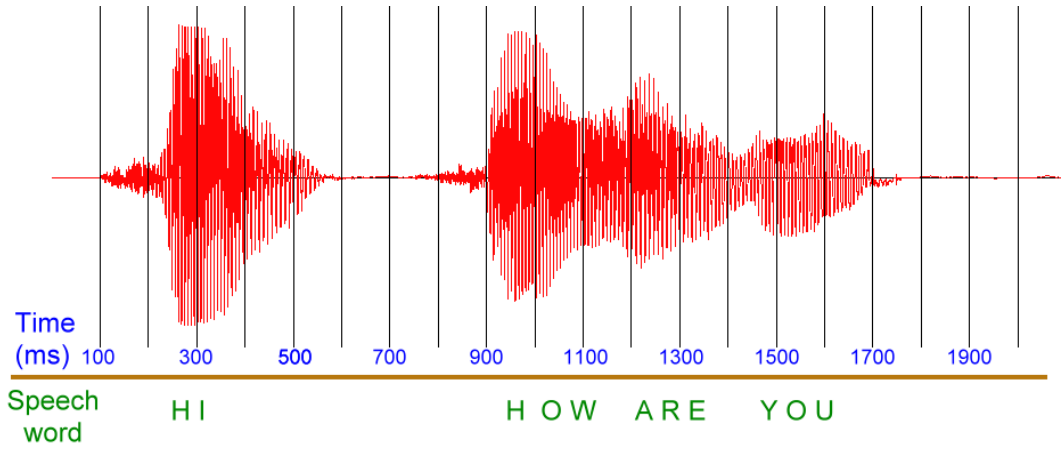

圖 18：一段語音的時域信號圖

由於人與人之間的溝通會話是一種語音存在與不存在不斷交替的訊號，圖 18 即為一段語音的振幅分佈情形。人在說話時，語句與語句之間並不會緊密的連接在一起，而會有中斷的間隔存在。而 VAD 就是根據聲音訊號之中的某些特性區分語音(speech，即實際有說話內容的部份)或非語音(non-speech，沒有說話內容)區段。

過去 VAD 技術用於 VoIP 的目的是用來決定一個訊框之中是否包含了有意義的語音資料[12]：假設 VAD 偵測到一個聲音訊框中並不包含任何有意義的語音資料，那麼就不將這個訊框傳送至網路上，以節省頻寬使用量。

一個有效的 VAD 方法是利用語音的頻率特性做判斷，但因回音之頻域特性與正常語音類似，故不適用於回音消除。另一法為根據能量大小作為判斷依據。其技術上的作法為：定義每一個聲音訊框的能量值(energy)，同時設定一個臨界值(threshold)作為判定用的依據，其判定的演算法如下(其中，Ej 為能量值，Er 為臨界值)：

IF (Ej> Er)

THEN Frame is **ACTIVE**

ELSE Frame is **INACTIVE**

當一個聲音能量的能量值超過臨界值時，即將此訊框視為語音。反之，即視為非語音。而本研究將以能量作為判斷為正常語音或回音的依據。

3.4 系統架構

本研究提出以 VAD 的方式作為正常語音與否的判定機制。此系統架構如圖 19 所示。

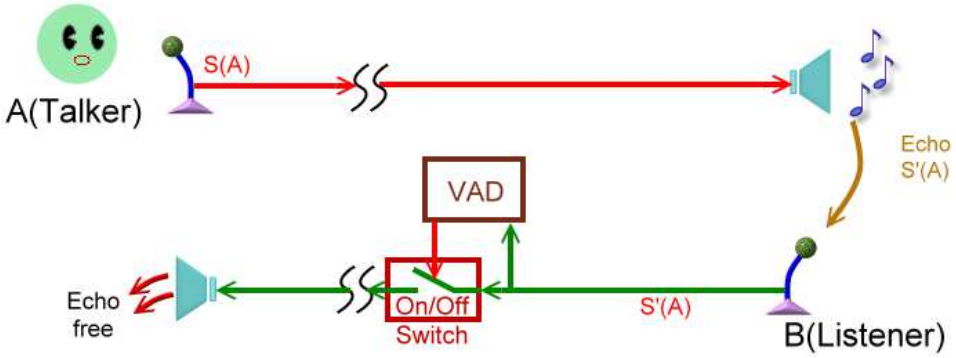

圖 19：以 VAD 判定正常語音與否之架構

圖 19 中，由收聽端(B)麥克風收到的聲音訊框，在應用程式送出之前，先由 VAD 進行判定，若發現該訊框中為回音或不包含語音，則 VAD 負責擋下此訊框(開關切為 Off)。反之，若該訊框內包含有意義的語音，則 VAD 會允許其傳送。

若使用者的終端設備原內建有回音消除機制，此架構也能與現有之 Listener 回音消除機制搭配共同運作，當原有回音消除機制失效時，VAD 仍然能有效阻擋回音與靜音訊框。如同前述，在一個 Listener 回音消除機制失效的收聽端，回音訊號被麥克風收回後，會傳送給其他網路會談的與會者播放，造成 Infected Conference。

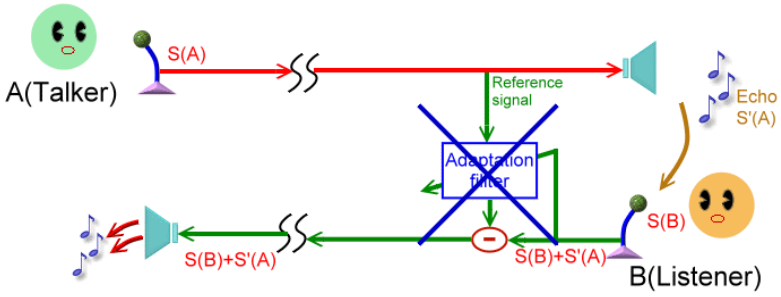

圖 20：Listener 端回音消除機制失效的狀況

如圖 20，收聽者(B) 所造成的 Direct Echo 會傳給原始說話者(A)。而為了避免這種狀況，本研究所設計的系統即在原本的回音消除機制之後加入由 VAD 控制的訊號開關(switch)，控制是否讓聲音通過。

Case1：Listener Echo Cancellation 正常運作時

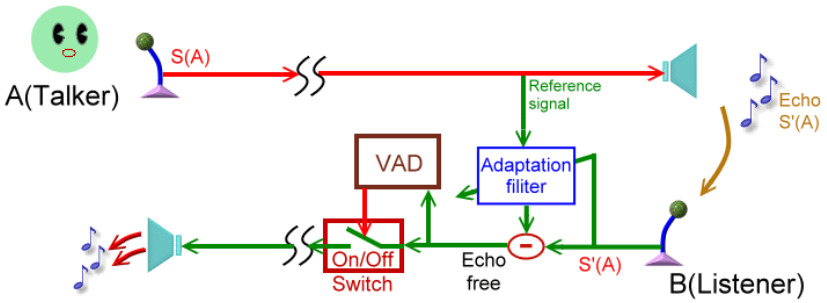

圖 21：加入 VAD 機制之系統

如圖 21，當回音消除機制正常運作時，收聽端的回音正常被消除，此 VAD 機制並不會對原系統造成任何影響。

Case2：Listener Echo Cancellation 失效/且收聽者未說話時

若回音消除機制失效，造成收聽者 B 的回音無法正常被消除時，此 VAD 機制即發揮功能：

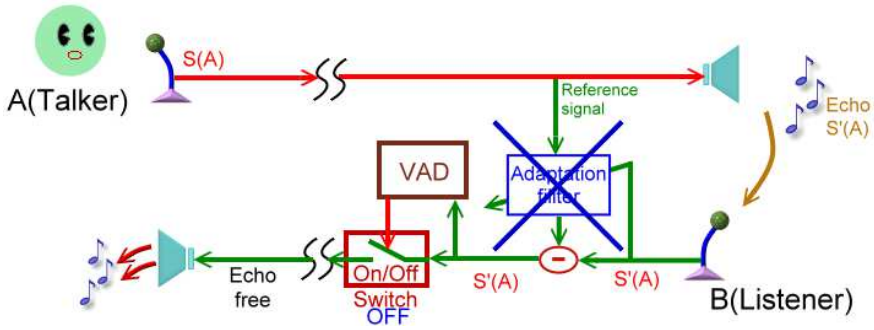

圖 22：加入 VAD 機制，且回音消除機制失效時的狀況

如圖 22，Listener Echo Cancellation 失效，導致回音透過 B 麥克風收音，且無法正確被消除。此時 VAD 機制即可檢測出此回音訊號，並將其擋下。

Case3：Listener Echo Cancellation 失效/且收聽者說話時

若 Listener 端回音消除機制失效，導致收聽者端麥克風收入回音，但同時此收聽者也在說話的狀況如圖 23：

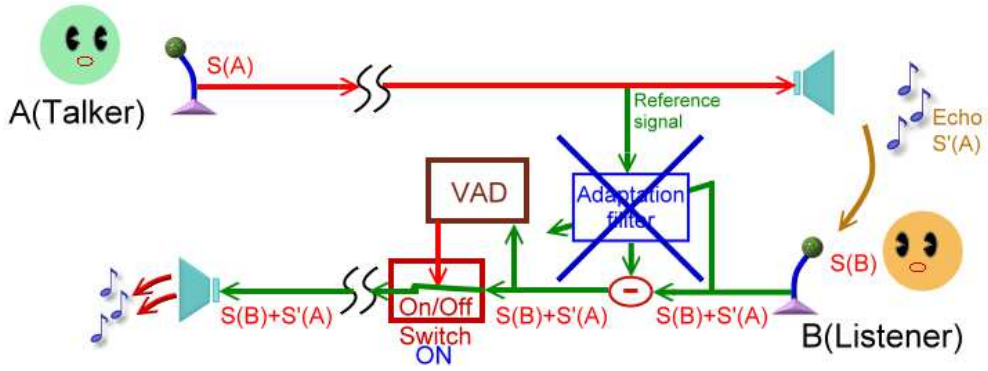

圖 23：回音消除機制失效，同時收入回音與說話聲音時的狀況

如圖 23，B 端的 Listener Echo Cancellation 失效，導致回音之產生，但同時 B 也在說話，亦即 B 的正常語音 S(B)與說話端 A 的回音 S(A)夾雜在一起被送出時，VAD 機制能判斷出語音訊號存在，並不會將此聲音擋下，仍然會送出此聲音訊號，而使溝通能正常進行，但回音並未被消除。

如上述，藉由加入 VAD 語音判定機制，當聲音傳送至網路以前先判斷為有效語音或是回音訊號，以避免 Listener 端回音消除機制失效時導致 Infected Conference。

3.5 細部設計

以下將針對本研究所提出的以 VAD 方式偵測並消除回音解決方案的細部設計方式。由正常語音與回音之間的差異，設計適當的判定演算法有效判定回音與否，並且以 VAD 方式控制聲音通過與否。

傳統的 VAD 為了判斷人聲與否，常用的方法為時域(time domain)與頻域(frequency domain)兩種特性判斷方法。由於人類的聲音必然集中在特定的頻率範圍，因此採用頻域判定通常能得到比較好的判定結果。但為了分析頻域數值，輸入的每一個訊框都必須先轉換為頻域數值，此部份的運算量相當可觀。在本研究中，根據實驗可以發現能夠精確判斷人聲與否的頻域 VAD 方法並不適用於回音判斷，其原因為無論語音或回音，聲音的頻率範圍均相同(來源都是說話聲音)，在此狀況下，採用運算複雜度較低的時域判斷反而能夠有較好的判定結果。

時域分析時通常採用振幅(amplitude)作為能量計算的依據，平均振幅高的訊框，其能量值也較高。另外相關研究也提出另一種採用過零量偵測(Zero Crossing Detection)為判斷依據的 WFD 演算法(Weak Fricatives Detector)[13]。此演算法假設訊框內容為語音

時，其訊號振幅越過零點的次數為每個訊框 15~60 次之間(訊框長度為 30ms 時)。當 WFD 演算法計算輸入訊框過零量在此範圍內時，即將此訊框視為語音，否則視為非語音。本研究的實驗結果顯示，採用聲音振幅計算能量具有較佳的辨識能力。

3.5.1 聲音能量紀錄

在使用者進行網路會談的實際環境中，由於喇叭音量會被使用者調至適當大小，由喇叭發出再次被麥克風收回的回音通常會經過空間上的傳遞與反射，造成能量的衰減，在進入麥克風時，其音量比正常使用者說話的聲音，通常較小。

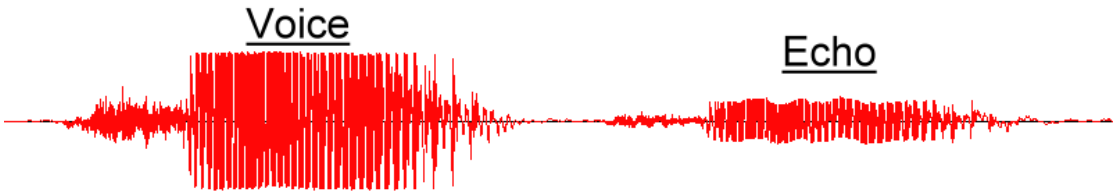

圖 24：正常語音與回音的振幅差異

圖 24 即為在一段聲音訊號中，正常語音與回音的音量比較。一般的會談中，使用者必然面對麥克風直接說話，因而麥克風得到的音量(即振幅)必然較大，相反地，回音由於經過傳遞時的衰減，因此振幅較小。由於具有此種信號特性上的差異，即可做為時域能量 VAD (Time Domain Energy-Based VAD) 的判定依據。

在時域(Time-Domain)訊號分析中對於聲音訊框的能量值(Energy)定義如下式[14]：

$$E = \sum_{k=0}^{N} S^2[k] \cdots\cdots (3)$$

其中：

E 為訊框的能量值。

N 為訊框的取樣總數。

S [k]則為第 k 個取樣的振幅。

亦即將一個聲音訊框中每一個取樣的振幅平方值加總後，即為該訊框的能量值。而本研究的設計中，每當由麥克風收到的聲音經過取樣後，都先經過上式計算出該訊框的能量值，以作為後續 VAD 演算法的判定依據。

3.5.2 LED VAD 演算法

LED VAD 即為『線性能量偵測』(Linear Energy-Based Detector)，如同前述，VAD 機制會定義一個臨界值作為判定語音訊號與否的依據，而此演算法就是用以定義並更新此臨界值的方法[13][16]。

LED VAD 演算法對於臨界值的定義如下：

初始臨界值(Initial Threshold)為第一個訊框的能量值：由於假設第一個訊框的內容必然為非語音，因此用此訊框的能量值作為背景雜訊的初始值。

接下來收到每個訊框號，對臨界值更新如下式：

$$E_{dnew} = (1-p)\cdot E_{dold} + p\cdot E \cdots\cdots(4)$$

其中：

E_{dnew} 為每次更新後的臨界值。

E_{dold} 為前一次的臨界值。

E 為最近一次的訊框能量。

而上式中的 p 則為可調整的參數，可根據不同聲音環境或需求做調整。當 p 越大，則臨界值更新越快速，亦即新到達的 Frame 能量影響越大。相反的，p 值小時，臨界值更新速度慢，通常會將此 p 值固定為 0.2。使用 LED VAD 演算法，臨界值會隨著每個新到達的訊框能量而有所變化，整體來說，會成為追隨著能量值變化的趨勢：

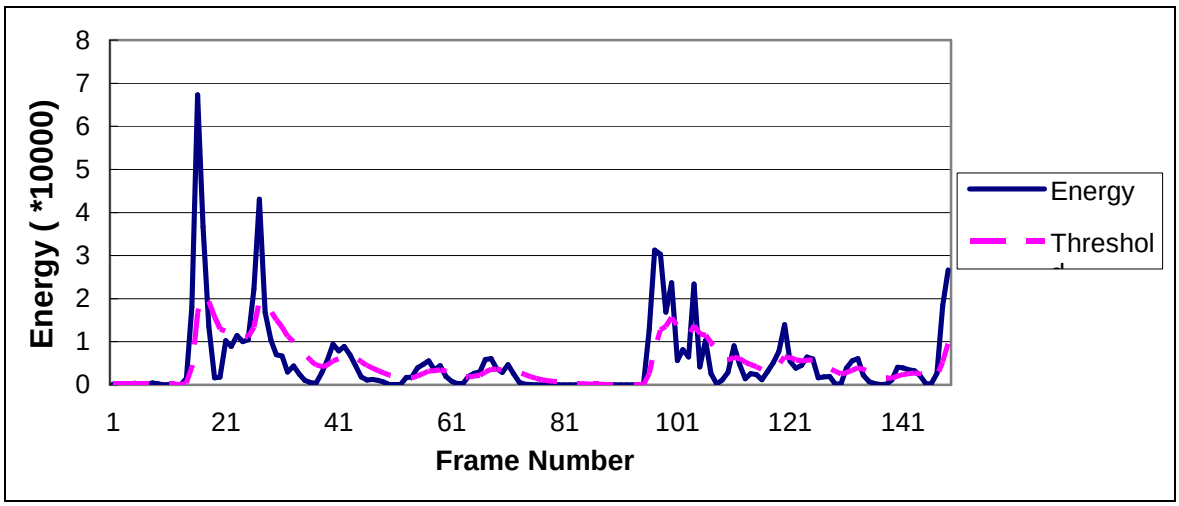

圖 25：聲音能量與臨界值之間的變化關係

圖 25 為一段語音的能量變化與 LED VAD 演算法求得的臨界值變化關係(p=0.2)。如訊框中包含語音，能量會上升而超過臨界值，即將此訊框標視為語音。反之，若能量下降至臨界值以下則視為非語音訊框。因此 LED VAD 演算法能夠在聲音具有語音與非語音的交替狀況下，找出其中真正具有語音內容的部份。

但若以 LED VAD 演算法進行回音與否的判斷，會產生一些問題：由於 LED VAD 演算法的臨界值變化隨時跟著聲音能量(即振幅大小)變動，因此除非使用者非常頻繁的說話，否則當正常語音中斷時，回音的訊號極容易被誤判為正常語音訊號，降低回音消除之效率。

3.5.3 MET VAD 演算法

為了改善前述 LED VAD 演算法所遇到的問題，必須針對語音會談的回音判定演算法重新進行設計。本研究針對多人參與的網路語音會談進行實驗與分析，大多數情況下會有一些共同的特點：

- 根據實驗，回音的能量值大部分都在正常語音能量的 10%~15%以下。
- 當參與會談人數越多時，平均每位使用者發言的機會就越少。
- 每一位使用者會在上線後的短時間內發言，以告知其他使用者(例如打招呼)，因此會談程式在啟動後短時間內應該會收到一段正常語音訊號。

根據以上的特點，本研究設計了一套專門針對回音偵測的 VAD 判定演算法，稱為『最大能量追蹤 VAD』(Maximum Energy Tracking VAD - MET VAD)。此方法可以確保使用者於持續一段時間未說話時，仍然不會將回音誤判為正常語音。MET VAD 演算法之實際步驟如下：

a. 開始進行判斷之前，可由程式根據先前的設定值，作為初始臨界值。若無先前之臨界值，則以第一個訊框能量值的 10 倍作為初始臨界(與 LED VAD 演算法相同，假設第一個訊框內容必然為非語音)。
b. 每個訊框輸入後，判斷其能量是否大於臨界，若是則視為語音，否則為非語音。
c. 若輸入訊框能量之 15%超過目前之臨界值，則更新臨界值，以此訊框能量之 10%作為新的臨界值。
d. 為了避免因為雜訊導致臨界值設定過高，當超過 100 個訊框沒有收到語音訊框時，

則將臨界值降低為 95%。

根據以上原則，此判定機制會抓取輸入的最高能量值，並取此值的 10%作為臨界值，由於絕大多數回音訊號都低於此能量，故可以依據每一位使用者的說話能量，過濾掉回音。

為了避免因為雜訊或其他因素導致臨界值設定過高，故加入了自動降低臨界值的機制：當長時間沒有收到正常語音時，就略微調低臨界值，避免使用者因為意外而無法送出聲音，其中，降低臨界值的時間長度可隨著與會人數增加而增加(如假設中所述，人數越多，平均說話頻率越少)。若使用者未說話的時間太長，則臨界值將不斷下降，甚至有機會降至趨近於零。此時回音消除機制將失效，而回音與靜音訊框將可能被視為正常語音而送出。因為根據實驗，麥克風可能接收到的雜訊強度是難以預知的，若因為臨界過高而造成使用者聲音完全無法送出，對於會談造成的影響性遠比回音來的嚴重，故 MET VAD 選擇寧可放過回音而避免切斷正常語音。圖 26 為 MET VAD 演算法之判定流程圖，其中 t 為臨界值決定參數，在此定為 10%。

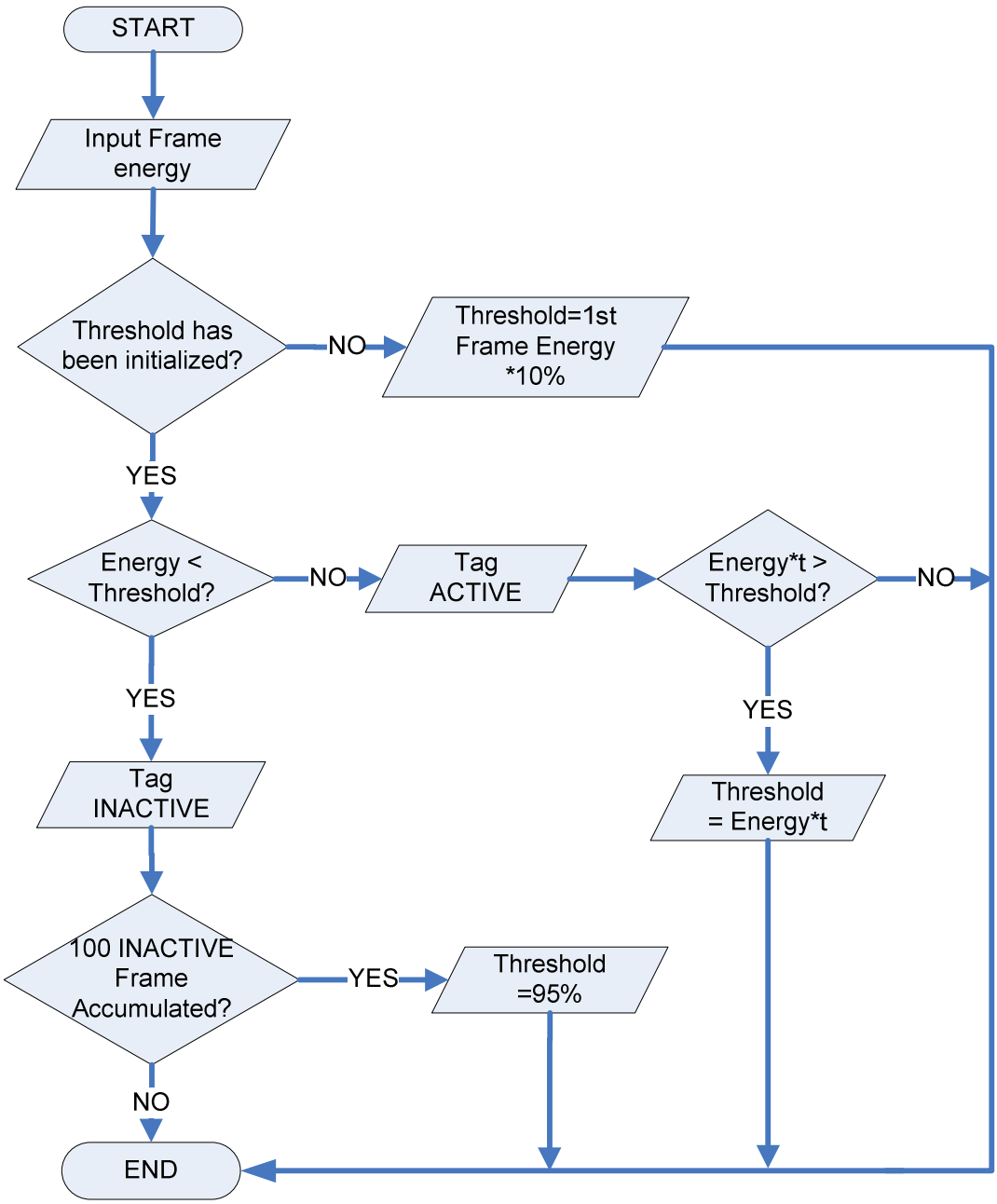

圖 26：MET VAD 演算法流程圖

在此演算法中，每一個輸入的訊框所需要的運算僅有能量值的計算與臨界更新，並沒有任何迴圈存在。當訊框之取樣數為 n 時，整體的時間複雜度僅為線性時間 O(n)。對於目前電腦而言，每秒 8000 次取樣的運算量負擔相當輕微。而此流程之虛擬碼 (pseudo code)如下：

```
NewFrameArrival(){
        IF(Threshold not been Initialzed)
            Threshold = 1st FrameEnergy*10%;
         ELSE IF(FrameEnergy > Threshold){
              TagACTIVE();
              IF(FrameEnergy*t > Threshold)
                   Threshold=FrameEnergy*t;
               }
          ELSE IF(FrameEnergy < Threshold)
              TagINACTIVE();
     IF(100 INACTIVE Frame Accumulated)
        Threshold=Threshold*0.95
}
```

第四章 效能分析

4.1 實驗目的

我們針對本研究所提出的方法進行實際驗證，首先以樣本資料對 MET VAD 與現有的其他 VAD 進行量化評比，其次，於真實的網路會談程式中進行質化驗證。

4.2 實驗設計

以下簡單的描述各項實驗的目的，概略作法及評比參數。

4.2.1 以聲音樣本評比各種 VAD 演算法

此實驗以完全相同的樣本聲音資料作為輸入，針對 Time Domain 常用的 VAD 演算法與本研究提出的 MET VAD 演算法比較回音辨識率。樣本資料為包含 Indirect Echo 之語音訊號，其特性於實驗前經過分析統計，作為量化評比各種演算法之依據。VAD 演算法會將輸入的聲音以訊框為單位，將判定為語音的訊框保留，丟棄其餘封包。本實驗可根據誤判率與回音能量殘留指標評比演算法效能。

4.2.2 以網路會談實測 MET VAD 之效能

本實驗將 MET VAD 演算法實際結合在一個簡單的網路會談程式中，並且刻意讓麥克風接收喇叭產生的回音(或由軟體製造回音訊號)。藉由使用者間實際進行交談，將 Echo Generator 端的麥克風實際錄下的聲音與經過 VAD 演算法濾除回音之後的聲音做比較，藉由誤判率與使用者實際試用觀感作為效能評估依據。

4.2.3 Proximity Problem 的回音消除測試

本實驗測試 Skype 對於 Proximity Problem 造成的回音消除之能力，並且與 MET VAD 演算法做比較，以驗證 MET VAD 是否能有效消除 Proximity Problem 造成的回音。

4.3 評估指標

本研究使用誤判率及 MOS 作為評量指標。

4.3.1 誤判率

誤判率可分為兩種：False Positive (把回音當成正常語音的誤判)與 False Negative (將正常語音當成回音的誤判)。若 False Positive 高，則通話聲音中剩餘的回音將干擾會談，若 False Negative 高，則正常的語音會被 VAD 刪除，使用者溝通可能受阻。

4.3.2 MOS

由於聲音品質對於使用者而言是因個人主觀感受而異的。每個人在進行語音會談時，對於語音失真與回音程度的可接受範圍均不相同。實際將聲音資料經過 VAD 處理後，由使用者聆聽並評估其對回音的過濾能力以及對語音資訊的破壞程度，作為評估 VAD 效能的依據。

一般常用來衡量網路會談通話品質的指標為MOS (Mean Opinion Score) [21]，表4為其評分的參考依據與意義：

表 4：Mean Opinion Score

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

分數由高至低分別代表品質最好至最壞，分數的計算是由特定的發話者與聽話者在特定的環境下，透過收集測試者在各種不同情景下的主觀感受，再根據建議的分析法則得出該語音的品質。本研究之實驗藉由使用者實際在包含回音消除機制的網路會談中進行溝通，並且根據其主觀感受給予MOS評分。

4.4 實驗一：以聲音樣本評比各種 VAD

本實驗事先製作了一段聲音樣本資料作為測試樣本，其中包含大部分的回音與部份的語音，並且事先分析語音的位置，長度與能量等資訊。

4.4.1 實驗目標

本實驗以事先分析的語音樣本測試三種不同 VAD 的回音消除能力。實驗結果根據

False Positive (把回音當成正常語音的誤判)與 False Negative (將正常語音當成回音的誤判)兩種誤判率評估個別 VAD 對於非語音的過濾能力以及對於語音訊號的誤判機率。

4.4.2 實驗環境

本實驗採用的樣本聲音資料如表 5：

表 5：實驗一使用的樣本聲音資料

聲音取樣率	8000 Hz
取樣格式	Mono PCM
取樣位元數	8 bits (256 Levels)
聲音長度	15 秒
訊框長度	30ms
總訊框量	500
語音插入位置	1.5 秒 (第 50 個訊框)
語音插入長度	1.2 秒 (共 40 訊框)
回音能量峰值:語音能量峰值	22%

圖 27 為本實驗輸入聲音之波形：

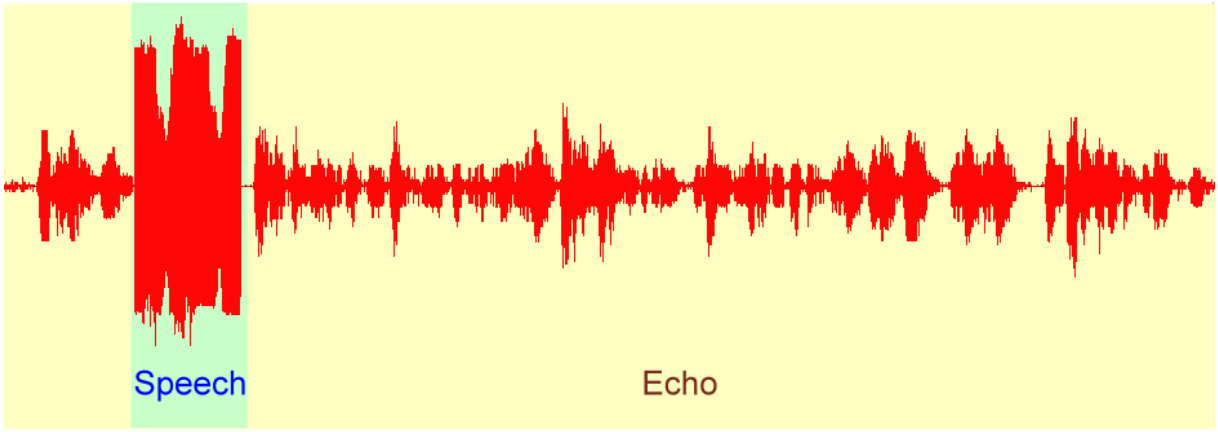

圖 27：實驗一輸入之聲音波形

本實驗使用 LED VAD，WFD VAD 與 MET VAD 三種時域能量 VAD 演算法判定以上的輸入樣本訊號，本實驗中三種演算法的設定資料如表 6。

表 6：實驗一的演算法參數設定

演算法	LED VAD	WFD (Weak Fricatives Detector)	MET VAD
初始臨界	第一個訊框的能量	———	第一個訊框的能量的十倍
參數設定	p 分別為 0.05，0.1，0.15，0.2，0.25，0.3(新訊框能量做為臨界更新值的比率)	過零量分別為 10~65，15~60，20~55，25~50，30~45，35~40 次時，視為語音。	a. t 分別為 0.05，0.1，0.15，0.2，0.25，0.3(語音峰值能量作為臨界更新值的比率) b. 連續 100 個非語音訊框，臨界下降為 95%。

4.4.3 實驗流程

本實驗將樣本聲音資料輸入至三種不同 Time Domain VAD 演算法中，並且分別以三種演算法針對每一個 Frame 判定為語音或非語音。最後統計每一種演算法將回音部份誤判為語音的比例，作為演算法濾除回音效能的依據。

4.4.4 實驗結果分析

a. LED VAD 演算法

以 LED VAD 演算法分析此段聲音訊號所得到的能量與臨界值變化如圖 28~圖 33，其中調整參數分別為 0.05，0.1，0.15，0.2，0.25，0.3：

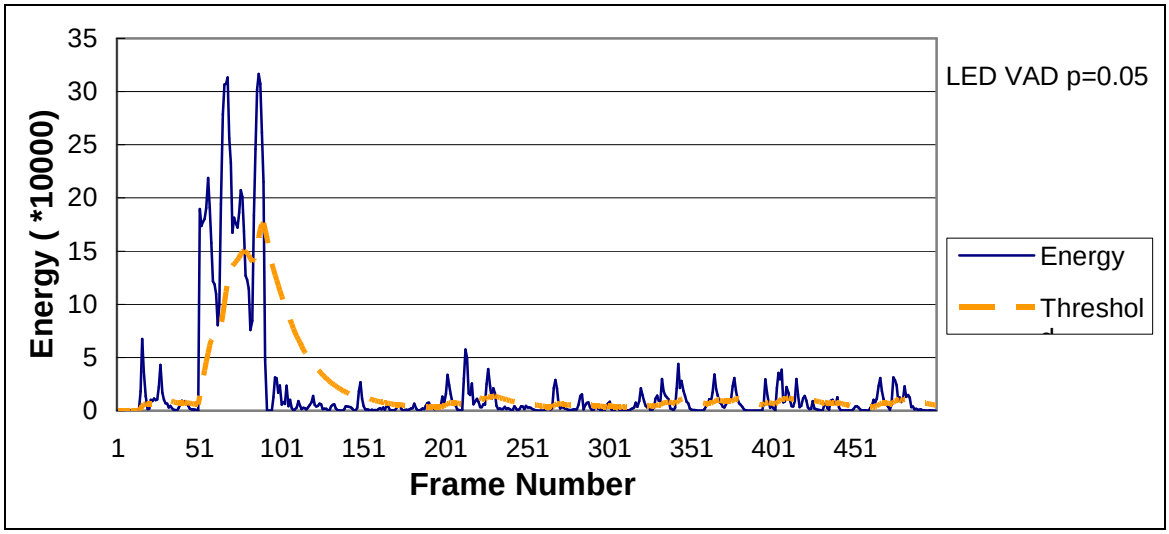

圖 28：以 LED VAD 演算法針對樣本資料的分析結果(p=0.05)

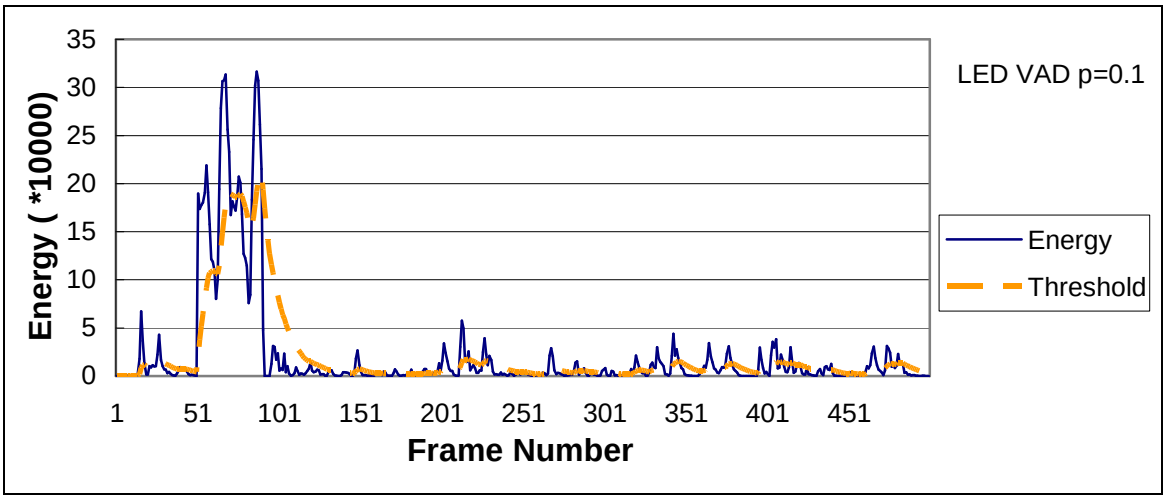

圖 29：以 LED VAD 演算法針對樣本資料的分析結果(p=0.1)

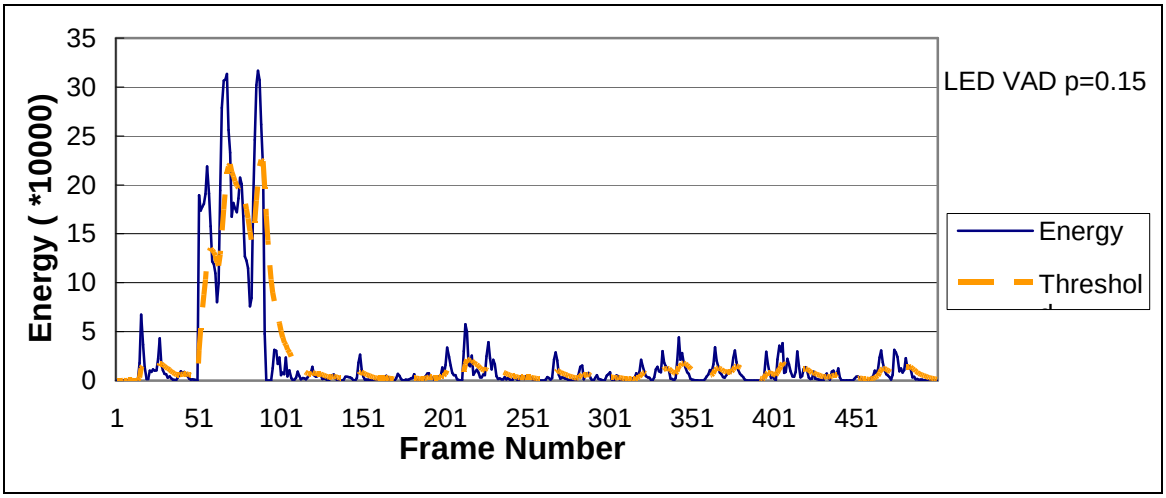

圖 30：以 LED VAD 演算法針對樣本資料的分析結果(p=0.15)

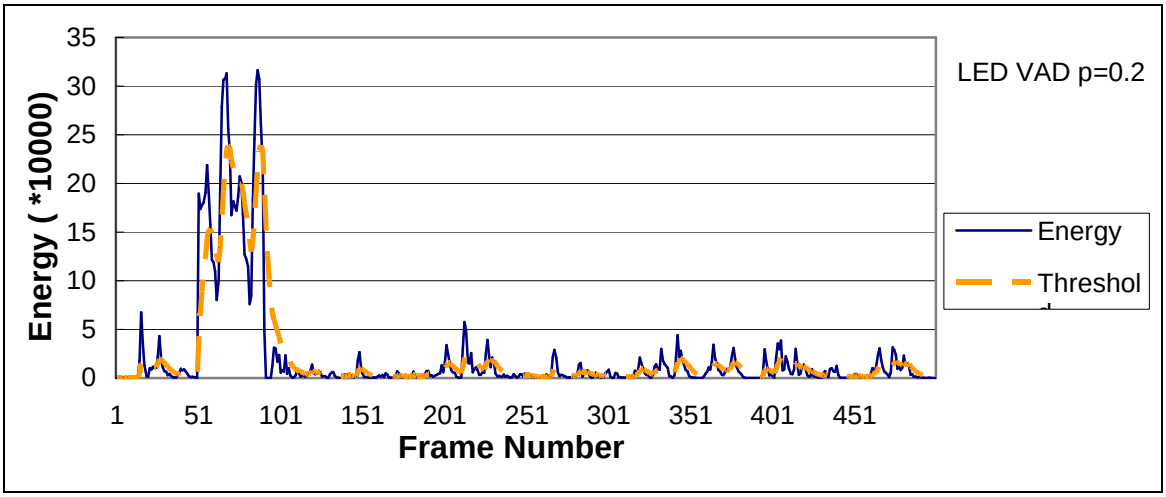

圖 31：以 LED VAD 演算法針對樣本資料的分析結果(p=0.2)

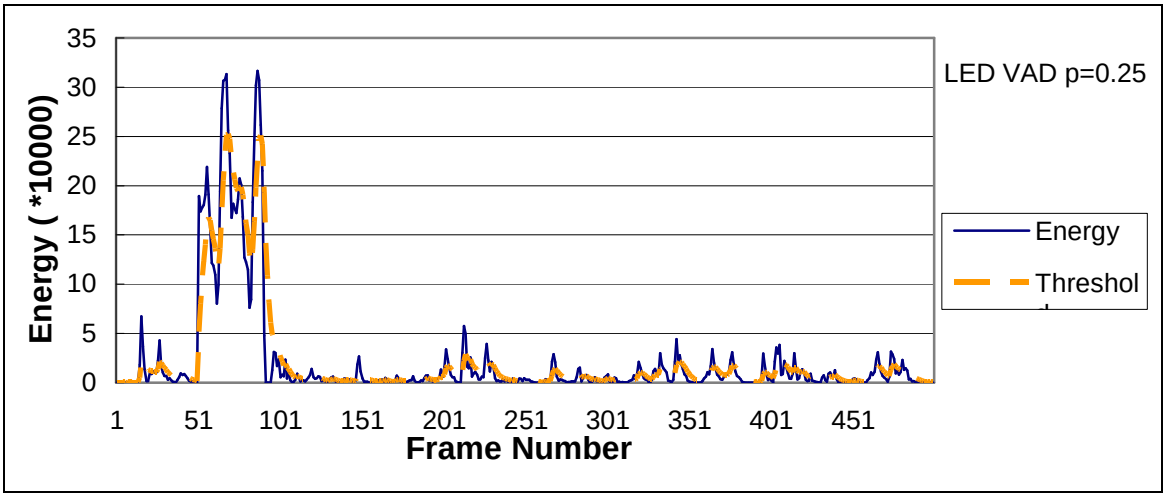

圖 32：以 LED VAD 演算法針對樣本資料的分析結果(p=0.25)

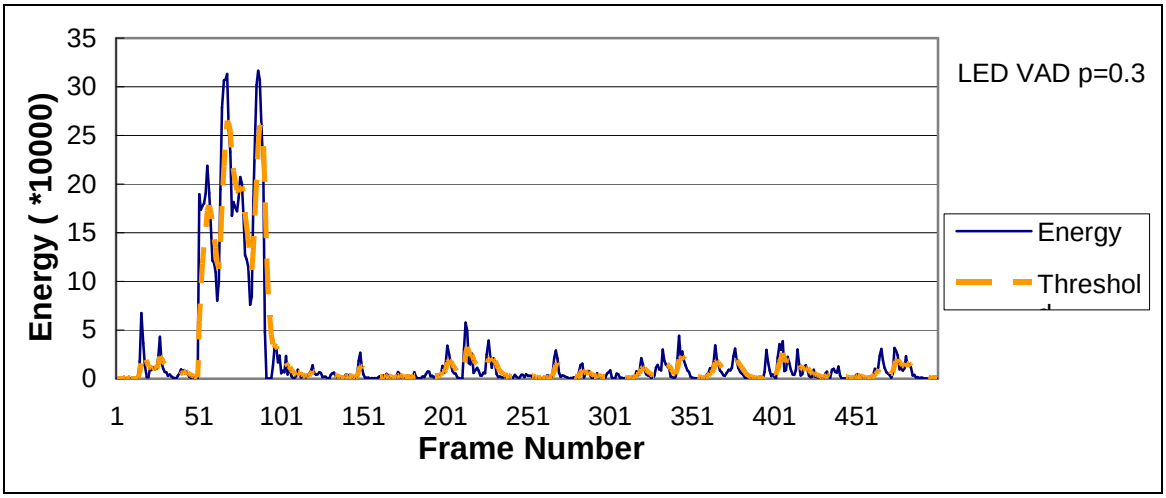

圖 33：以 LED VAD 演算法針對樣本資料的分析結果(p=0.3)

表 7：LED VAD 演算法的誤判率

參數 p	p=0.05	p=0.1	p=0.15	p=0.2	p=0.25	p=0.3
將回音視為語音機率 (False Positive) (%)	28.48	31.3	33.7	33.91	33.04	33.48
將語音視為非語音機率 (False Negative) (%)	12.5	30	42.5	45	47.5	50

LED VAD 在能量值(Energy) 超過臨界值(Threshold)時，會將該訊框將被視為語音，否則視為非語音。圖 28~圖 33 中可看出 LED VAD 演算法的臨界值會隨著輸入能量的大

小而變化臨界值，因此當聲音能量下降時，臨界值也會跟著下降，使得回音仍然會被判定為語音訊框。若採用不同的調整參數 p，其結果如表 7。若將 p 值降低，臨界值變化速度跟著變緩慢，雖在語音資料少時誤差小，但一旦語音資料交錯出現時，每段語音開始時均可能被誤判，造成語音訊框損失 s。

b. WFD 演算法

以語音樣本的過零量統計如圖 34：

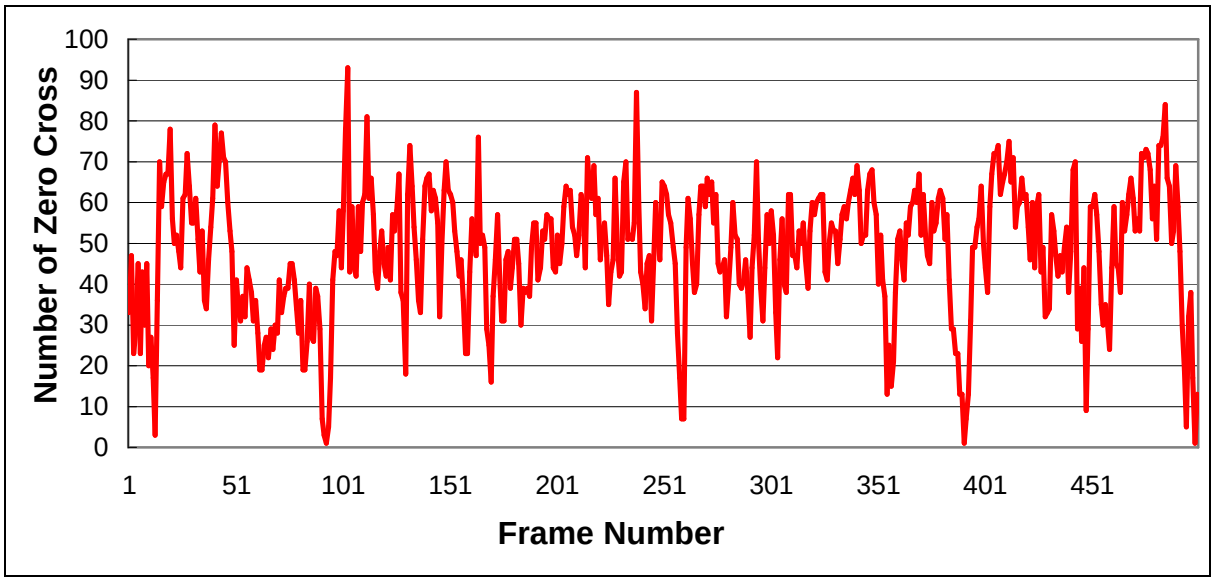

圖 34：以 WFD VAD 演算法分析得到的越零次數統計

WFD 演算法根據過零量的統計，當一個訊框內的過零量落在一個『識別區間』(Classification Zone)範圍內，則判定為語音，否則判定為非語音。表 8 與圖 35 為過零量識別區間分別為 10~65，15~60，20~55，25~50，30~45，35~40 時的辨識率。其中包含將回音訊框當作正常語音訊框以及將語音訊框當作非語音的誤判率。

表 8：WFD VAD 演算法不同過零量下的語音誤判率

過零量識別區間	10~65	15~60	20~55	25~50	30~45	35~40
將回音視為語音機率 (False Positive) (%)	84.13	67.39	53.26	36.30	20.65	5.65
將語音視為非語音機率 (False Negative) (%)	0	0	10	17.5	47.5	77.5

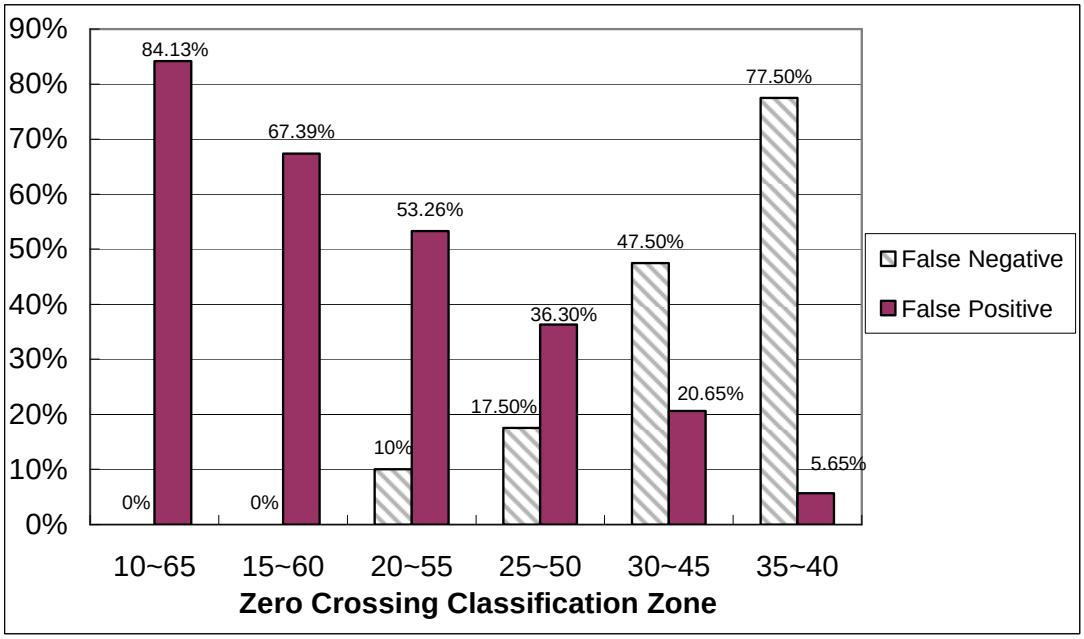

圖 35：以 WFD VAD 演算法分析得到的語音誤判率長條圖

由於無論是正常語音或是回音，都是實際語音，僅有能量的差異，其過零量並沒有顯著的差異。雖然如表 8 所列，將識別區間(過零量範圍)縮小可降低將回音訊框當成正常語音的誤判率，但同時也會遺漏掉更多正常語音訊框(將語音視為非語音之機率同時提高)，因此顯然無法有效從過零量分析訊框是否為回音。

c. MET VAD 演算法

圖 36~圖 41 為 MET 演算法分析樣本資料得到的臨界值，其中調整參數 t 分別為 0.05，0.1，0.15，0.2，0.25，0.3。

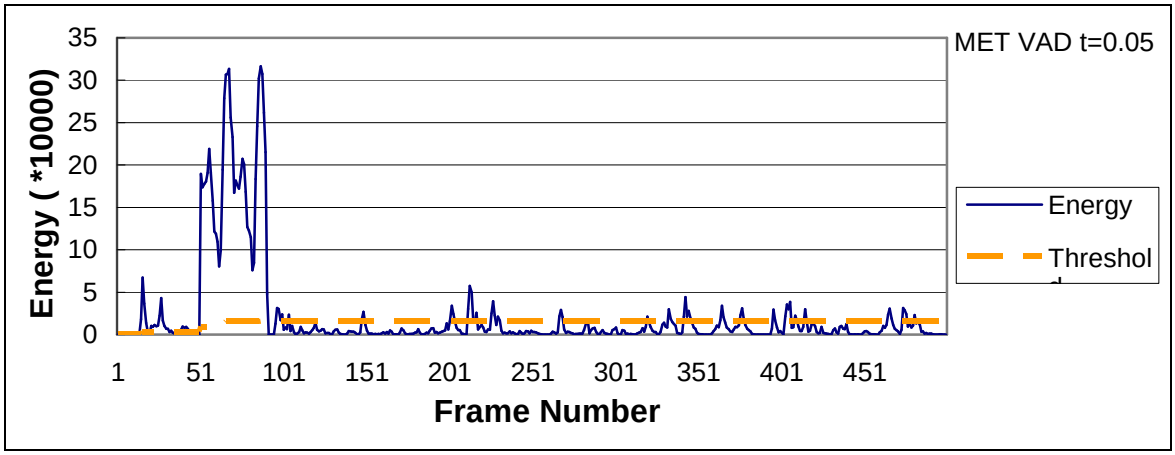

圖 36：以 MET VAD 演算法針對樣本資料的分析結果(t=0.05)

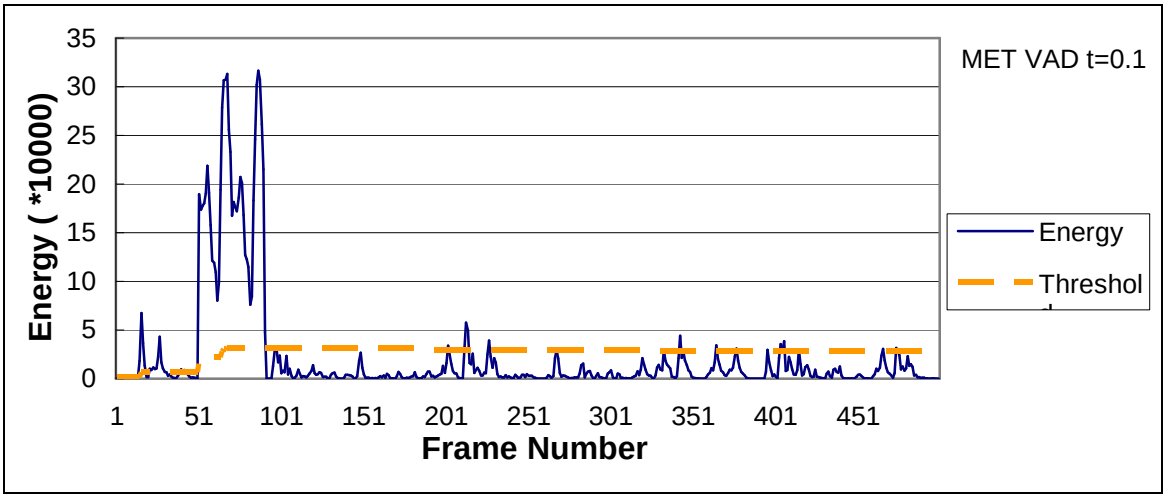

圖 37：以 MET VAD 演算法針對樣本資料的分析結果(t=0.1)

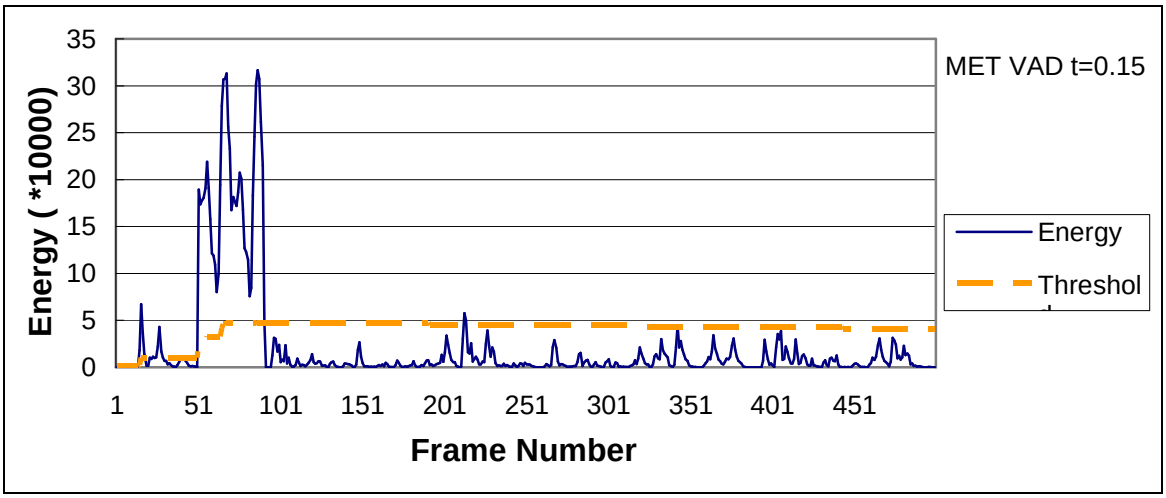

圖 38：以 MET VAD 演算法針對樣本資料的分析結果(t=0.15)

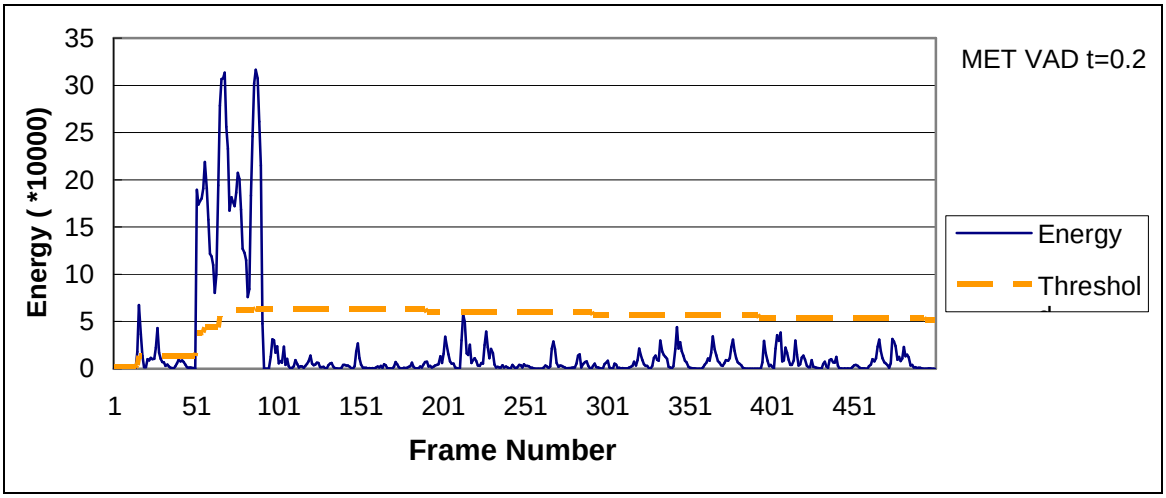

圖 39：以 MET VAD 演算法針對樣本資料的分析結果(t=0.2)

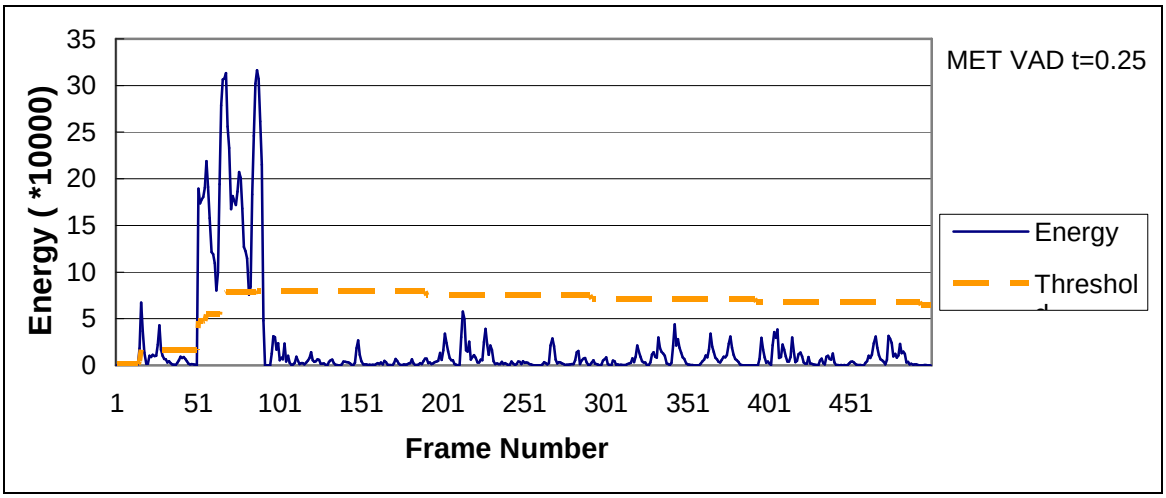

圖 40：以 MET VAD 演算法針對樣本資料的分析結果(t=0.25)

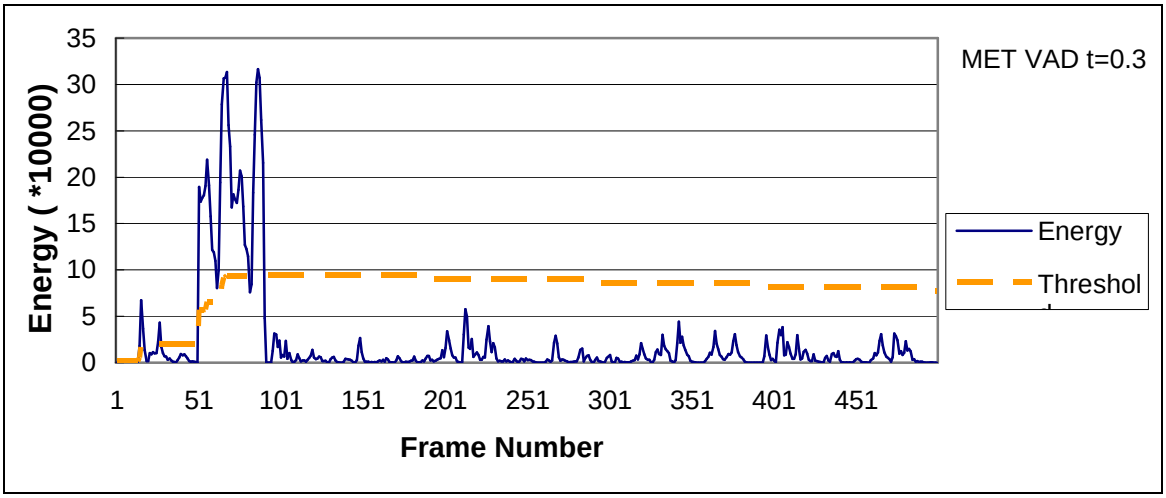

圖 41：以 MET VAD 演算法針對樣本資料的分析結果(t=0.3)

表 9：MET VAD 演算法的誤判率

參數 t	t=0.05	t=0.1	t=0.15	t=0.2	t=0.25	t=0.3
將回音視為語音機率 (False Positive) (%)	16.52	7.83	3.26	1.74	1.3	1.3
將語音視為非語音機率 (False Negative) (%)	0	0	0	0	2.5	5

如圖 36~圖 41 可以看出 MET VAD 演算法只要一開始有語音訊號出現時，即可紀錄此語音訊號峰值作為臨界值的設定依據，隨後即作為辨認之依據。由於在會談過程中，正常語音與回音的能量變化幅度通常不大，因此所紀錄下的語音峰值用以辨識回音之效果令人滿意。除了在輸入正常語音訊號之前，因為還沒得到正確的臨界值而導致較高的誤判率之外，一旦使用者開始說話，辨識能力隨即大幅提高。表 9 分別針對不同的調整參數 t 做測試，亦即調整語音峰值能量作為臨界更新值的比率，若 t 值越大，則代表以更大的語音峰值能量比率作為臨界更新值，在回音能量較大的狀況下，調大此參數可降低誤判回音為正常語音機率，但相對的也可能將能量較小的語音訊框誤判為回音，使用時應根據實際情況調整。

同時，為避免因為雜訊因素導致臨界值設定過高，無法送出正常語音，本演算法會定期略微降低臨界值。當使用者再次說話時，又會再次紀錄語音峰值，而將臨界值提高。如此，只要使用者不斷對著麥克風發話，臨界值就會保持在正常範圍內，而不會將回音

誤判為語音。

e.綜合比較

針對樣本資料,三種演算法在不同設定參數之下的誤判率統計於表 10。在 LED VAD 方面，隨著 p 值減少，臨界值之更新速度增加，在語音封包密集時可減少語音資料遺失之機率，但相反地在語音封包稀疏時，可能會遺失較多語音封包。而 WFD VAD 對回音的辨識率極差，很明顯的不適用於回音消除，若將辨識標準降低，可能會將大部分回音視為正常語音，而提高辨識標準則會增加封鎖正常語音封包的機率。而 MET VAD 則因依據使用者的說話音量作為辨識臨界值，因此能有效區分正常語音與否，而且可在不同的回音音量下調整參數，提高辨識精確度。

表 10：實驗一結果

演算法	參數	False Positive 誤判率	False Negative 誤判率
LED VAD	p=0.05	28.48%	12.5%
	p=0.1	31.3%	30%
	p=0.15	33.7%	42.5%
	p=0.2	33.91%	45%
	p=0.25	33.04%	47.5%
	p=0.3	33.48%	50%
WFD VAD	辨識區間=10~65	84.13%	0%
	辨識區間=15~60	67.39%	0%
	辨識區間=20~55	53.26%	10%
	辨識區間=25~50	36.30%	17.5%
	辨識區間=30~45	20.65%	47.5%
	辨識區間=35~40	5.65%	77.5%
MET VAD	t=0.05	16.52%	0%
	t=0.1	7.83%	0%
	t=0.15	3.26%	0%
	t=0.2	1.74%	0%
	t=0.25	1.3%	2.5%
	t=0.3	1.3%	5%

4.5 實驗二：網路會談實測

本實驗將 MET VAD 實際應用於 VoIP 會談中，測試其針對回音消除的的效果。在本實驗中，我們也針對市面上流行的會談軟體 Skype 進行測試，試圖在通話中刻意產生回音，以找出該程式所採用的回音消除方法之缺陷。

4.5.1 實驗環境

本實驗所使用的聲音格式如表 11 所示：

表 11：實驗二的音訊參數

聲音取樣率	8000 Hz
取樣格式	Mono PCM
取樣位元數	8 bits (256Levels)
聲音長度	15 秒
訊框長度	30ms

本實驗分別將 LED VAD，WFD VAD 與 MET VAD 結合於自行撰寫的網路會談程式中作為 VAD 判斷演算法，其參數設定如表 12。同時也使用 Skype 4.0 進行一對一會談實際測試。實驗進行時，為了使麥克風收回喇叭播放的回音，其中一位使用者刻意將喇叭音量放大，且將麥克風靠近喇叭，試圖收回喇叭之播放聲音，造成回音。

表 12：實驗二的演算法參數設定

演算法	**LED VAD**	**WFD VAD**	**MET VAD**
初始臨界	第一個訊框的能量	———	第一個訊框的能量的十倍
參數設定	p=0.2(新到達訊框能量之 0.2 倍為臨界更新值)	過零量為 20~55 次時，視為語音。	a.t=0.1(以語音峰值能量之 10%作為臨界更新值) b.連續 100 個非語音訊框，臨界下降為 95%。

4.5.2 實驗流程

a. 在沒有加入回音消除機制的 VoIP 會談中，故意讓其中一方的麥克風收到喇叭放出的

聲音，產生 Direct Echo 且紀錄此聲音的原始波形。同時將此包含回音的聲音經過 LED VAD，WFD VAD 與 MET VAD 過濾，紀錄過濾後的波形，比較過濾前後之差異。此步驟之目的在於評估 VAD 演算法將回音當作正常語音的誤判率(False Positive)。

b. 在正常的 VoIP 會談中，加入 VAD 過濾機制，比較過濾前後是否對正常語音造成影響。此步驟之目的在於評估 VAD 演算法將正常語音誤判為回音的誤判率(False Negative)。

c. 兩位使用者以 Skype 實際進行會談，其中一位使用者刻意以不同距離，不同擺放位置讓麥克風收進喇叭放出的聲音，試圖製造回音，驗證 Skype 回音消除機制失效的可能性。

4.5.3 實驗結果分析

a. Flase Positive 誤判率測試

圖 42 為一段包含回音的聲音波形(由 Echo Generator 端麥克風收錄後，未經任何處理)，回音的音量振幅約為正常聲音的 10%。

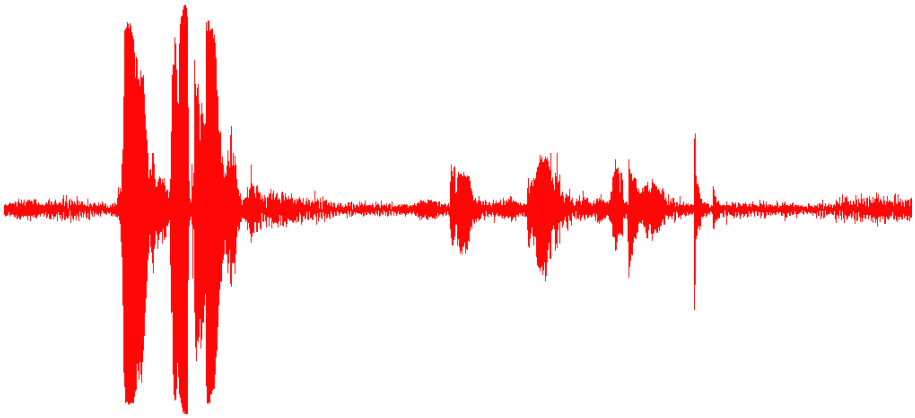

圖 42：含有回音的一段聲音波形

將此聲音透過 LED VAD 過濾後，得到的波形如圖 43：

圖 43：含有回音的聲音波形經過 LED VAD 過濾結果

此段含有回音的聲音，由 LED VAD 過濾後，僅能夠消除掉最初的小部份回音，超過 85%的回音會被保留下來，使得通話品質嚴重受影響。再加上 LED VAD 將能量小的混音音節誤刪機率相當高，故整體的通話品質不佳，MOS 只能達到 2(難以溝通)。

同樣將此段聲音經過 WFD VAD 過濾後，得到的波形如圖 44：

圖 44：含有回音的聲音波形經過 WFD VAD 過濾結果

由圖 44 可看出，由於無論是否為回音，都是使用者的說話聲音，因此回音與非回音訊框的過零量並無明顯差異。採用 WFD VAD 針對回音過濾的效果並不理想，至少 90%的回音被留下，正常語音誤刪率雖不高，但也無法消除回音封包，因此 MOS 僅有 2~3 之間。

同樣的將此段包含回音的聲音使用 MET VAD 過濾後的結果如圖 45：

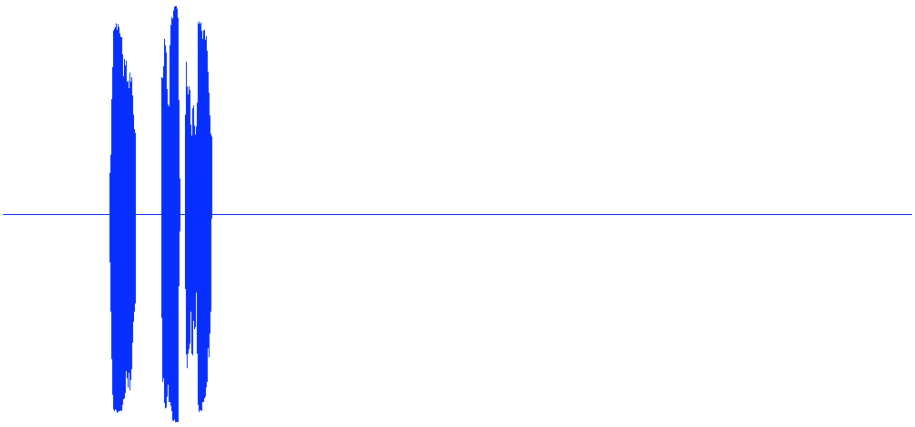

圖 45：含有回音的聲音波形經過 MET VAD 過濾結果

此段聲音在 MET VAD 過濾後，幾乎能夠消除所有非語音訊框，至少 85%以上的回音訊框能成功被消除，僅有與正常語音夾雜的回音無法消除，因此能有效提昇通話品質，使得 MOS 達到 3 以上。除了回音以外，MET VAD 也能夠有效的將音量過低的靜音訊框消除，節省傳輸頻寬同時降低背景雜訊對通話造成的干擾。

b. Flase Negative 誤判率測試

圖 46 為一段不包含回音的正常語音聲音波形：

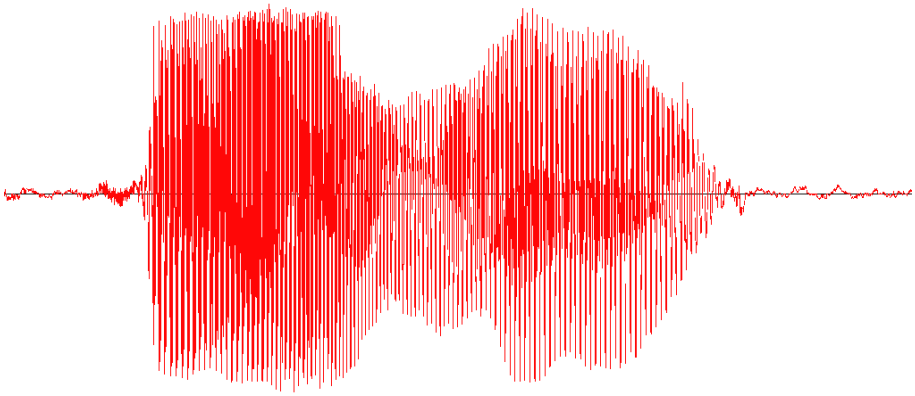

圖 46：不包含回音的語音聲音波形

將此段聲音波形分別輸入 LED VAD 過濾後，其結果輸出波形如圖 47：

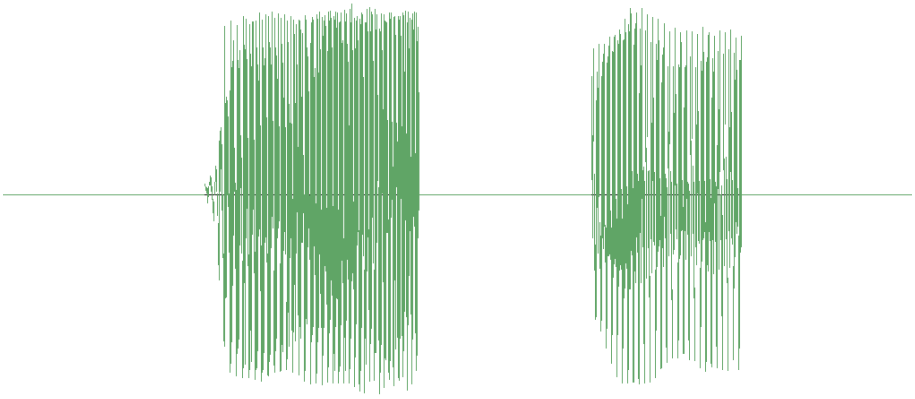

圖 47：不含回音的波形經過 LED VAD 過濾結果

由圖 46 中可發現，LED VAD 很容易將能量較微弱的聲音誤判為非語音訊框而誤刪(例如英文中的無聲子音音節)，最嚴重時誤刪率可達到 40%，如此將可能嚴重影響溝通品質。

同樣將此段聲音經過 WFD VAD 過濾後，得到的波形如圖 48：

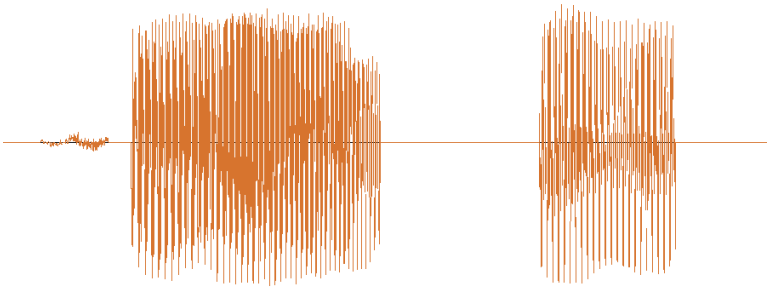

圖 48：不含回音的波形經過 WFD VAD 過濾結果

圖 48 可看出 WFD VAD 對於語音的誤刪率較低，但同樣在音節轉折與結尾處容易將語音訊框視為非語音而誤刪，最嚴重可能誤刪近 30%的語音訊框。

同樣將此段正常語音輸入 MET VAD 過濾，其輸出結果如圖 49：

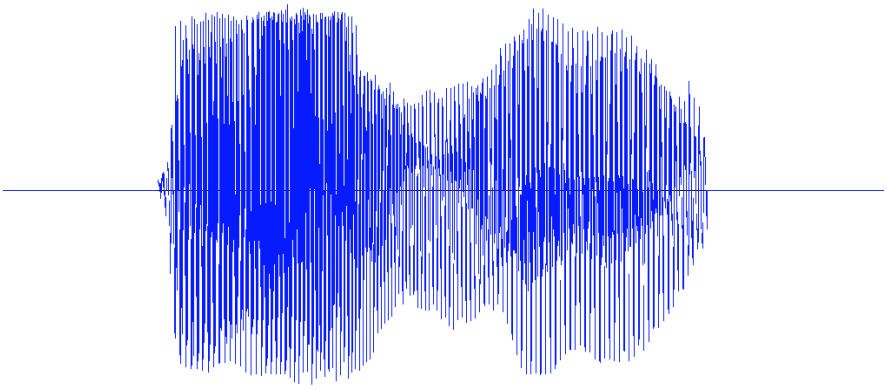

圖 49：不含回音的波形經過 MET VAD 過濾結果

MET VAD 會預先紀錄使用者的說話音量，因此並不容易將語音訊框誤刪，在正常溝通狀況下，誤刪率約在 15%以下，並不會對溝通造成太嚴重的影響。

c. Skype 通話測試

本實驗採用目前主流的 VoIP 程式 Skype 進行測試。實驗時刻意讓其中一位會談參與者的麥克風收入喇叭聲音，試圖製造回音，以驗證 Skype 之回音消除機制效能。

由實驗結果發現：在 Skype 會談中，無論是 Direct 或 Indirect Echo 都可能出現。且只要有一個 Echo Generator 存在，所有使用者都會聽到回音。此外，根據封包擷取的實驗結果，每個 Skype 與會者接收到的均為經過混音的聲音封包(即代表有混音傳輸的節點存在)，因此每個接收端並沒有所有發話者的原始聲音訊號作為判定回音的依據，故推測 Skype 所使用的為 Listener Echo Cancellation，在送出麥克風的擷取訊號之前，就先做回音相減。

整體而言，Skype 發生回音的機率不高，且由於目前市面上電腦音效卡均有內建回音消除 DSP (筆記型電腦的麥克風與喇叭位置可預知，更容易預估回音在 Acoustic path 所耗時間)，因此回音情況相當罕見。但是一旦回音發生，則所有參與會談使用者的通話品質都會受到影響。

根據本實驗，可以證明在 VoIP 中，若有 Echo Generator 存在，則加入 MET VAD 可有效預防因為回音消除機制失效而造成的 Infected Conference，而相對的若在沒有回音的會談中，即使加入 MET VAD 也不至於對正常交談聲音造成破壞。另外，根據 Skype 測試可得知：即使目前主流的 VoIP 程式，約有 10%的機率無法完全消除回音(尤其是 Indirect Echo，完全無法消除)，因此以 VAD 作為回音消除機制對於多人的網路會談系統

而言，確實有其必要性。本實驗測試結果如表 13 所列，其中雖仍以 Skype 產生回音機率最低，消除效果最佳，但由於其為商業軟體，使用的回音消除方法與程式碼均不對外公開。而 MET VAD 則能夠達到效能接近的回音消除效果，且完全公開演算法，提供相關研究與設計者使用。

表 13：實驗二結果

演算法	LED VAD	WFD VAD	MET VAD	Skype
False Positive Test (% of False Negative)	>85%	>90%	<15%	<10%
False Negative Test (% of False Positive)	>40%	>30%	<15%	<5%

4.6 實驗三：Proximity Problem 的回音消除測試

本實驗針對 Proximity Problem 這種狀況下所產生的回音與雜音，測試 MET VAD 的過濾能力。

4.6.1 實驗目標

此部份實驗將分別測試 MET VAD 與 Skype 對於 Proximity Problem 造成的回音之消除能力，目的為驗證 MET VAD 能夠適用於特殊的回音情況，維持正常會談進行。

4.6.2 實驗環境

本實驗採用與 Skype 與自製的 VoIP 程式搭配 MET VAD 分別進行測試。測試時以一對一會談，將兩部電腦距離拉近至 50 公分以內。自製的 VoIP 程式所採用的聲音格式如表 14 所示：

表 14：實驗三的音訊參數

聲音取樣率	8000 Hz
取樣格式	Mono PCM
取樣位元數	8 bits (256Levels)
聲音長度	15 秒
訊框長度	30ms

MET VAD 的參數設定如下：

初始臨界：以第一個訊框能量的 10 倍作為初始臨界。

調整參數：t=0.1(語音峰值能量的 10%作為臨界更新值)。

臨界下降頻率與幅度：連續 100 個非語音訊框(3 秒)，臨界下降為 95%。

4.6.3 實驗結果分析

a. 測試結果- Skype

使用 Skype 通話時，當兩部電腦距離靠近，互相收到對方喇叭發出的聲音時，當使用者對著麥克風說話後 3 秒之內，即產生 Proximity Problem，造成持續不斷的雜音與回音，其聲音波形如圖 50：

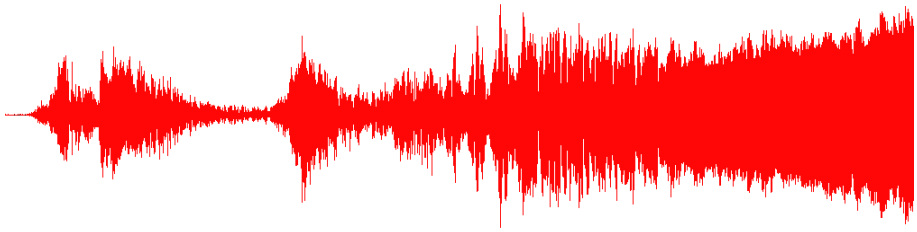

圖 50：Skype 通話時，Proximity Problem 所造成的回音波形

由此證明，Skype 之回音消除機制雖能夠克服大部分回音造成的干擾，但卻無法消除 Proximity Problem 產生的回音。因此若電腦之間距離過於接近或音量過大時，可能造成嚴重的回音問題，嚴重影響會談進行。

b. 測試結果- MET VAD

相對的，若使用 MET VAD 過濾，則能夠有效抑制 Proximity Problem 所產生的回音，即使將電腦間距離靠近，也不至於發生回音，此時聲音波形如圖 51：

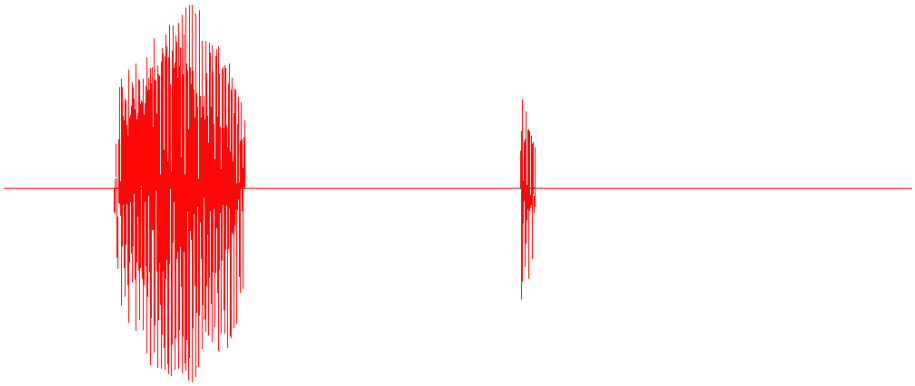

圖 51：MET VAD 有效抑制 Proximity Problem 的結果

使用 MET VAD 過濾後，即使電腦之間距離非常接近，能夠互相收到對方喇叭的聲音，也可在 Proximity Problem 開始擴散之前，根據能量差異判斷聲音為非正常語音，而有效抑制，僅有在使用者說話後 0.5 秒出現 100ms 左右的雜音，不至於產生刺耳的噪音。

表 15：實驗三 MET VAD 消除 Proximity Problem 之效能

測試環境	**Skype**	**MET VAD**
Proximity Problem	會產生(影響通話)	不會產生
回音情形	說話後 3 秒內產生嚴重噪音	說話後 0.5 秒後出現 100ms 雜音

第五章 結論與未來研究方向

本研究針對目前的網路會談中所產生的回音狀況，就其成因，影響以及現有解決方法等各種角度做了分析。我們分析許多造成現有回音消除機制失效的原因，並指出其發生的狀況與時機，證明在大型語音會談中要徹底消除回音是極為困難，且隨著與會人數增加，回音發生機率也呈指數成長。因此，本研究提出一種以最大語音能量紀錄為基礎的語音動態偵測(VAD)演算法(MET VAD)，根據回音的特性，使用最少的運算成本，將回音減至最低，以避免影響會談聲音品質。本研究提出的 MET VAD 演算法，比傳統的 VAD 演算法，更能有效的降低回音誤判率，在我們的實驗中，能夠將誤判率降低至 15% 以下，有效改善通話品質。我們的實驗中將 MET VAD 演算法與主流的語音會談軟體進行實測比較，證實 MET VAD 演算法針對回音消除效能接近 Skype。若 MET VAD 與現有之回音消除方法配合，將能夠達到互補的作用，且針對 Skype 等軟體所無法克服之 Proximity Problem 所產生的回音，能夠藉由 MET VAD 有效抑制。

除了回音以外，MET VAD 同時能夠有效發揮靜音抑制（Slience Suppression）的效果，阻擋語音會談中不含語音內容的封包。根據實驗，MET VAD 能有效阻擋 85%以上不含語音的靜音封包，且將語音誤刪的機率低於 15%，有效降低網路頻寬耗用而不影響通話品質。

大型語音會談的另一個影響品質的因素是環境噪音，本研究未來將以環境噪音的消除為主要研究目標。

參考文獻

[1] U.I. Choudhry, JongWon Kim, and Hong Kook Kim, "A Highly Adaptive Acoustic Echo Cancellation Solution for VoIP Conferencing Systems", IEEE International Conference on Computer Systems and Applications, 2006, pp. 433-436.

[2] G. S. Fang, "Voice Channel Echo Cancellation", IEEE Communications Magazine, Vol. 21, Issue 9, Dec. 1983, pp.11-14.

[3] Xiaohui Gu, Zhen Wen, Philip S. Yu, and Zon-Yin Shae, "peerTalk: A Peer-to-Peer Multi-Party Voice-Over-IP System", Parallel and Distributed Systems, IEEE Trans. on Publication, Vol. 19, No. 4, April 2008.

[4] Perry P. He, Roman A. Dyba, and Lucio F.C. Pessoa, "Network Echo Cancellers: Requirements, Applications and Solutions", AnalogZONE, 2004.

[5] Brant M. Helf, "Far end echo cancellation method and apparatus”, U.S. Patent 4,995,030, Feb. 19, 1991.

[6] M. Hiraguchi, "Full duplex modem having two echo cancellers for a near end echo and a far end echo", U.S. Patent 4,935,919, 19 Jun. 19, 1990.

[7] Yao-Nan Lien, Li-Cheng Chi and Yuh-Sheng Shaw, "A Walkie-Talkie-Like Emergency Communication System for Catastrophic Natural Disasters", Proc. of 10th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN09), Dec. 14-16, 2009.

[8] P. Marques, F. Sousa, and J. Leitao, "A DSP Based Long Distance Echo Canceller using Short Length Centered Adaptive Filters", Proc. of ICASSP, 1997.

[9] B. S. Nollet, and D. L. Jones, "Nonlinear Echo Cancellation For Hands-Free Speakerphones", Proc. of NSIP'97, Michigan USA, Sep. 1997.

[10] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceller with two echo path models", IEEE Trans. on Commun., Vol. COM-25, No. 6, June 1977, pp. 589-595.

[11] G. Periakarruppan, and H. A. Abdul-Rashid, "Packet based echo cancellation for VoIP networks", Computers and Electrical Engineering, Vol. 33, No. 2, 2007, pp. 139-148.

[12] Petr Pollak, Pavel Sovka, and Jan Uhlir, "Noise Sup-pression System for a Car", Proc. of the Third European Conference on Speech, Communication and Technology -EUROSPEECH'93, Berlin, Sep. 1993, pp. 1073-1076.
[13] R. V. Prasad, A. Sangwan, H. S. Jamadagni, and M. C. Chiranth, "Comparison of voice activity detection algorithms for voip", Proc. of IEEE Symposium on Computer and Communications, July 2002, pp. 530-535.
[14] R. V. Prasad, R. Muralishhankar, S. Vijay, H. N. Shankar, P. Pawelczak, and I. Miemegeers, "Voice activity detection for VoIP-an information theoretic approach", Proc. of IEEE Global Telecommunications Conference, 2006, pp. 1-6.
[15] I. Rassameeroj, and S. Tangwongsan, "Echo Cancellation in Voice over IP", Proc of 5th International Conference on Information Technology and Applications (ICITA), 2008, pp. 570-575.
[16] P. Renevey, and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions", Proc. of European Conference on Speech Communication and Technology (ISCA EUROSPEECH '01), Sep. 2001, pp. 1887-1890.
[17] B. Widrow, and M. E. Hoff, "Adaptive switching circuits", IRE WESCON Convention Record part 4, 1960, pp. 96-104.
[18] B. Widrow, and SD Stearns, "Adaptive Signal Processing", Prentice-Hall, Nglewood Cliffs, NJ, 1985.
[19] Echo cancellation, http://en.wikipedia.org/wiki/Echo_cancellation, Retrieved at November 11, 2009.
[20]Echo suppressor, http://en.wikipedia.org/wiki/Echo_suppressor, Retrieved at November 11, 2009.
[21] Mean Opinion Score, http://en.wikipedia.org/wiki/Mean_Opinion_Score, Retrieved at July 09, 2009.