The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 122-136


# 基於遞迴類神經網路之麥克風嘯叫抑制系統

# Recurrent Neural Network-based Microphone Howling Suppression

林政陽　Cheng-Yang Lin, 廖元甫　Yuan-Fu Liao
國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
chengyang@speech.ntut.edu.tw, yfliao@mail.ntut.edu.tw

潘振銘 Chen-Ming Pan, 郭姿秀 Tzu-Hsiu Kuo
Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan
chenming@cht.com.tw gaga820402@cht.com.tw

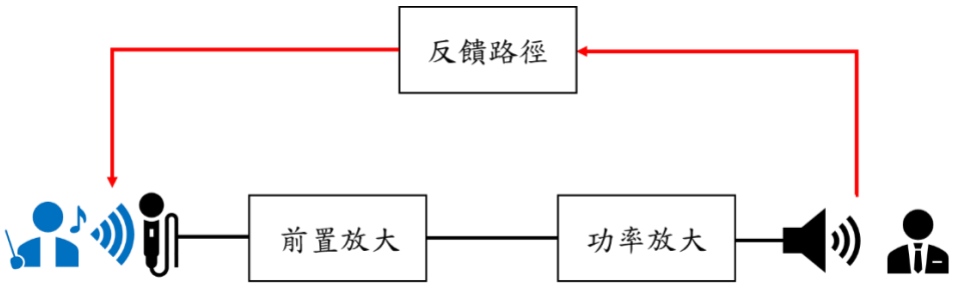
## 摘要

在使用卡拉 OK 系統唱歌時，常會因麥克風拿離喇叭太近，或是擴大機功率開太大，產生正回授而導致嘯叫，造成歌者跟聽眾都非常不舒服。一般處理麥克風嘯叫，常是利用移頻打斷共振，或是用帶阻濾波器做事後補救，但有可能會造成音質破壞。因此我們想改用適應性回授消除演算法，利用擴大機喇叭的輸入音源當參考訊號，來自動估算在不同空間環境、不同歌曲、不同訊雜比下，麥克風可能錄到的回授訊號，並在做訊號增益前先將其消除，以直接從源頭消除嘯叫發生的可能性。基於以上想法，在本論文中，實現了 normalized least mean square（NLMS）的嘯叫消除演算法，尤其是進一步考慮擴音系統的非線性失真，提出基於 recurrent neural network（RNN）的進階演算法。並在實驗時分別測試在時域或是頻域處理，與使用 NLMS 或是 RNN，對不同曲風、不同環境空間響應情況下，不同演算法的收斂速度、計算量需求與嘯叫抑制效果。由實驗結果發現：(1) 在時域實現收斂比較快，(2) 在頻率可實現計算量小於時域，(3) RNN 在收斂速度及突然變化的頻率消除上優於 NLMS。

關鍵詞： NLMS、遞迴類神經網路 RNN、適應性濾波器、麥克風嘯叫

## 一、簡介

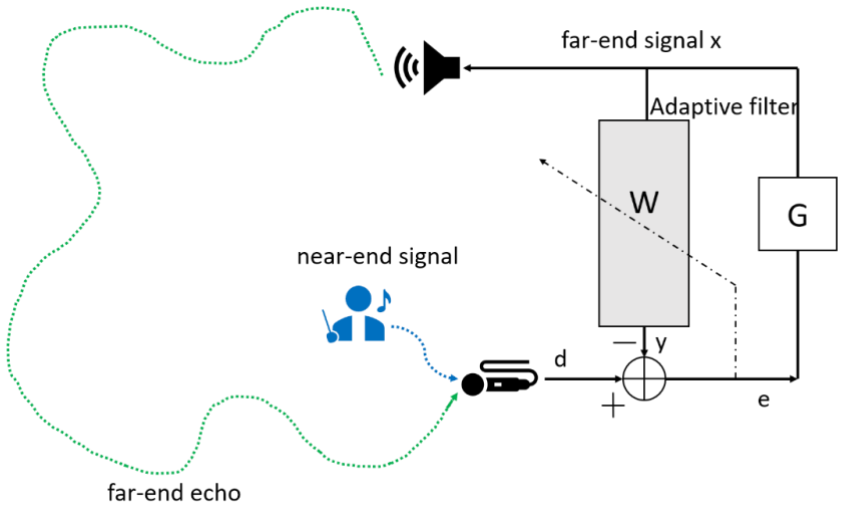
在使用麥克風唱卡拉 OK 的系統中，由於收音的喇叭與麥克風並未隔離在不同區域，當喇叭發出之聲音透過空間傳到麥克風，由於放大電路增益過高而導致正回授反饋如圖一，不斷將放出之聲音重複收入進而發生嘯叫。此種閉鎖迴路的嘯叫主要原因為整個電路與環境對某些共振頻率的增益過大，當提升喇叭通道之增益時，這些增益過大的共振頻率先達到聲學反饋所需的強度條件，若此頻率的反饋類型剛好為正反饋，則必定在此頻率上產生自激震盪現象，便是我們所說的嘯叫。




圖一、麥克風正回授嘯叫

為解決這件事情，可針對硬體上麥克風的指向性收音，將特定方向的聲音收進麥克風，盡量不讓除了近端人聲外的聲音收入，亦或者是增加麥克風之收音敏感度，可減少非必要收入麥克風的聲音;在演算法上有使用移頻打斷共振,將訊號頻率做些許升頻或降頻，破壞了嘯叫的發生條件，進而抑制了嘯叫，或是使用帶阻濾波器將會發生嘯叫的頻段做衰減，如果衰減這些過強的頻率就能抑制住嘯叫，但此兩者雖然只對小範圍的頻率做了調整,但仍會破壞原始訊號聲音,甚至是人耳可聽出的差別,且嘯叫仍是時不時的發生，因而目前在實際控制嘯叫發生的作法仍是治標不治本。

為了改善過去抑制嘯叫的缺點，這裡我們使用適應性濾波器來提前消除回聲以抑制嘯叫，首先我們先介紹基於 NLMS 之適應性濾波器演算法的回聲消除系統[1-2]如下圖二，利用擴大機喇叭的輸入音源作為參考訊號 x，自動估算在不同空間環境、不同歌曲、不同訊雜比下，麥克風可能錄到的回授訊號，再將預測出之期望信號與輸入的麥克風訊號 d 相減，使回聲訊號增益前就將其消除，直接從源頭消除嘯叫發生的可能性。




圖二、傳統聲學回聲消除系統架構圖

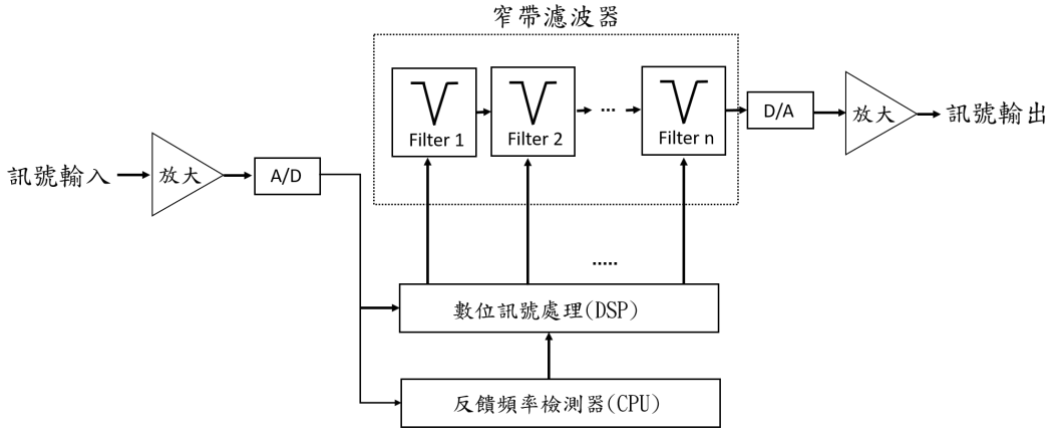
然而在使用卡拉 OK 麥克風時，時常因有能量非常大的擴大機喇叭撥出之聲音，與室內

複雜的空間環境響應，會有嚴重的非線性失真，使得線性演算法的 NLMS 無法有效的解決非線性問題，進而提出了基於 RNN 的進階演算法。由於 RNN 是擁有回授功能的遞迴類神經網路，能將上個過去的時間之輸出值儲存下來，再重新導回到輸入端，使得系統能夠抓取長度較長的時間輸入訊號，讓系統擁有龐大的過去資料來學習環境響應的路徑，改善非線性的部分。

此外嘯叫的發生都是即時且突然多變的，而時域的每一點調變一次，在面對突然變化的音樂或頻率變化，不確定是否仍能有效的收斂與消除，但在頻域做演算法可顧及到不同頻段的環境訊號，因此我們也針對此點做了時域、頻域演算法的比較，觀察時域與頻域在細膩度與收斂速度上是否有著明顯差異。另外也因 RNN 在計算量與 NLMS 有著明顯的差距，我們也觀察其計算量是否有與其效果成正比。因此基於上述考量，於本篇論文中我們將提出遞迴類神經網路麥克風抑制嘯叫系統，比較傳統時域、頻域 NLMS 和時域 RNN 在不同曲風、不同環境空間響應情況下，不同演算法的收斂速度、計算量需求與嘯叫抑制效果，在以之後章節將會介紹其模擬方法與架構。

## 二、相關研究

常用的回聲消除方法包括過去的帶阻濾波器、移頻，到近年主流的 NLMS 和基於預測誤差方法的適應性濾波器(Prediction Error Method-based Adaptive Feedback Cancellation，PEM AFC)。過去在使用移頻是使用一種可以改變聲音頻率的設備移頻器，其工作原理類似變調器，能夠將聲音訊號增加、減少 5Hz，破壞了嘯叫發生的條件，進而抑制了嘯叫，雖然對聲音的破壞很小，但其在演唱和樂器中就會有明顯差異，光 5Hz 的音調變化對人耳已經有明顯的感覺了；此外帶組濾波器至今在應用上仍是主流抑制嘯叫的工具，在音響系統中出現嘯叫是由於正反饋使音頻信號中的某些頻率點不斷被加強而造成的，因此將這些頻率點切除或進行大幅度衰減，就可以有效抑制聲反饋，其原理如圖三，主要是利用機器快速掃描尋找出發生嘯叫的頻段，並自動生成一組與這些嘯叫頻率相同的窄帶濾波器來切除嘯叫頻率，進而達到消除回聲而抑制嘯叫[3]。




圖三、帶組濾波器架構

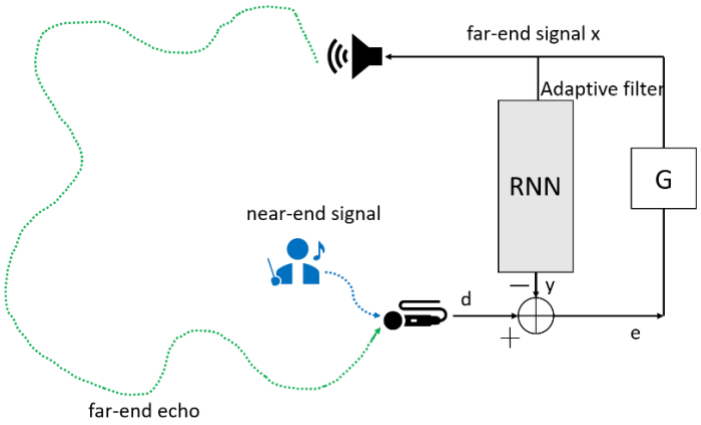
為了有別於過去方法，我們希望可以在嘯叫發生前就將嘯叫清除，因此使用了適應性濾波器，在嘯叫發生前先讓濾波器學習環境響應的路徑而改變其權重，穩定的消除多出來的回聲，其中最小均方演算法(LMS)是最易實現、且穩定及計算量小[4]，時至今日仍受到許多人喜愛且廣泛地運用，同時為了解決系統收斂緩慢的缺點。便將輸入訊號正規化，而演變成 NLMS 演算法，其採用可變步長的方法來穩定收斂過程。

另外由於閉鎖迴路的關係，近端訊號和揚聲器之間有嚴重的相關性問題，造成嘯叫更容易發生，因此後面有關於回聲消除的問題便有部分人著重在去相關性上面，有人在閉鎖迴路中的前向路徑加入全通濾波器來降低相關性[5]，此外也有論文提出了預測誤差方法適應性濾波器(PEM AFC)來減少近端訊號與揚聲器之間相關性[6]。而在 PEM 方法中，是利用反向近端訊號模型對麥克風和喇叭進行預濾波，然後將這些訊號送至自適應濾波演算法中，此便達到了降低相關性的目標。而對於近端語音訊號，通常使用線性預估(Linear Prediction，LP)來估計[7]，語音訊號由於時間相近的訊號點彼此有相關性，每個訊號點可由相近的訊號點藉由線性組合加以逼近或估測。故可藉由 LPC 估計去相關預訓練濾波，將語音訊號中分離成口腔與聲帶訊號，以去除喇叭輸出訊號與人聲訊號各自的相關性。

預測誤差方法次適應性濾波器能夠從期望訊號中去除相關分量，因此該反饋消除器對不相關信號和夾帶(entrainment)能夠使之不反應，但不幸地，當期望訊號是週期性號時，預測誤差方法自適應濾波器將會給出零，倘若是此種情形嘯叫便不能被消除了，因為嘯叫也是有週期性的，因此若發生此種情形或許需要回頭依靠傳統系統，因而也有人對此現象做出了嘯叫偵測器對此情況作出傳統和 PEM 發法的切換[8]。

## 三、適應性濾波器抑制嘯叫系統

由於麥克風與喇叭的閉鎖迴路會在室內反射造成回聲，而回聲時間會受房間大小影響，但傳統回聲消除系統會被輸入的長度固定而限制住，倘若輸入時間過長，將導致計算量過於龐大而收斂較慢，因此常無法有效地提取使用者長時間的訊號，故本論文改用遞迴類神經網路，其架構圖如圖四，將可以由回授路線看到過去較長時間的資料，以學習較長時間的環境響應路徑，並推測下一時間點的聲音，且僅用單層的 RNN 便足以獲取足夠的長時間資訊，因此多層的 RNN 就能看得更深更廣泛，抓取更多的訊息。下列先說明時域與頻域的 NLMS 做法，後面再介紹遞迴類神經網路的適應性濾波器消除演算法。




圖四、基於遞迴類神經網路之適應性濾波器抑制嘯叫系統架構圖

(一) NLMS

其模擬為利用麥克風所收集到的聲音 d 與透過喇叭輸出的聲音 x，利用 NLMS 預測可能錄到的環境響應，使其相減後將回聲消除，此時系統的誤差訊號 e 為輸出得到的清晰語者聲音。下列為 NLMS 演算法:

$$e(n) = d(n) - \hat{w}^H(n)u(n)$$

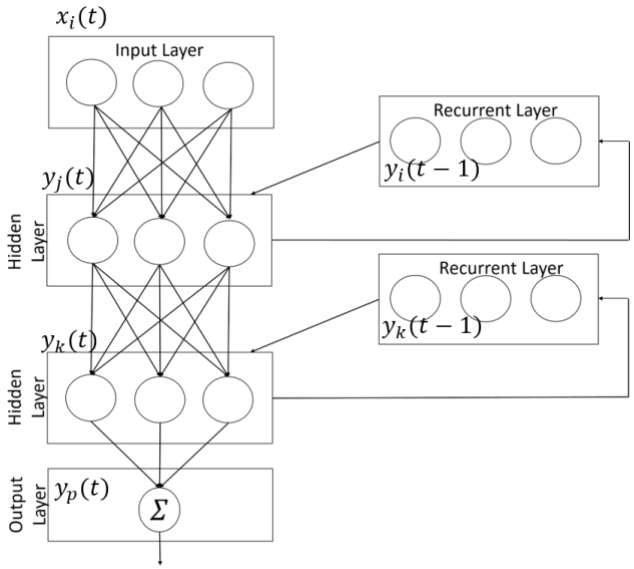
$$\hat{w}(n+1) = \hat{w}(n) + \frac{\tilde{\mu}}{\|u(n)\|^2} u(n)e^*(n)$$

在頻域 NLMS 演算法部分，我們每 512 點取一個音框，將此音框所有點由 FFT 轉為頻域，之後把每個音框的 512 點皆做一次 NLMS，運算完成後再將訊號由 Inverse FFT 轉回時域後做 overlap-add。此外顧慮到環境響應殘響長度，我們將取音框的數量增至每次取 8 個音框，讓演算法有足夠長度的時間序列去做學習。

(二)神經網路 RNN

這裡我們使用的架構為透過近端人聲與喇叭輸出的音檔結合後為麥克風收音 d，利用遞迴類神經網路預測麥克風收進的環境響應 y，相減後將環境響應消除，最後系統輸出誤差訊號 e 即為得到的剩餘近端人聲之聲音。

下圖五為本論文中使用的遞迴類神經網路運作結構。本論文之遞迴類神經網路架構為兩層的神經網路，輸入使用滑動視窗的方式每次抓取 1024 點、兩個隱藏層皆是 100 個節點，最後輸出為當前時間的點，此外擁有兩個的遞迴層。




圖五、深層遞迴網路架構圖

其想法為透過回授的方式，把過去的時間點之隱藏層節點輸出記錄起來，並重新輸入至隱藏層節點之輸入端，將輸入層與輸入值此兩個資訊連結在一起，以之作為下一個時間點的隱藏層輸入，讓當前的節點保有過去的輸出資訊，使得整個網路架構有記憶的特性。此外 RNN 的核心演算法為反向傳遞演算法(Backpropagation)，以下為反向傳遞演算法的推導:

最初輸入$x_i(t)$透過權重$v_{ji}$傳遞到第一個隱藏層，再加上第一層的隱藏層的偏壓值$b_j$，經過轉換函數f(·)，產生在第一個隱藏層第一個時間點的神經元輸出$y_j(t)$，如下式(3.1)為第一層隱藏層輸出，n 為輸入個數。

$$y_j(t) = f\left(\sum_i^n v_{ji}x_i(t) + b_j\right) \quad (3.1)$$

此時的第一層隱藏層輸出會輸入到第一層的回饋層，回授連結權重$r_{1,jm}$再到第一層隱藏層，也就是上一個時間點(t-1)神經元的轉換狀態，下式(3.2)為結合輸入$x_i(t)$所產生的新的輸出方程式。

$$y_j(t) = f\left(\sum_i^n v_{ji}x_i(t) + \sum_i^m r_{1,jm}y_i(t-1) + b_j\right) \quad (3.2)$$

由上層隱藏層輸出$y_j(t)$會輸入第二層的隱藏層，權重$w_{kj}$連接第二層隱藏層，加上第二層隱藏層偏壓值 bk，經轉換函數f(·)轉換，產出在第二個隱藏層這一個時間點的神經元輸出$y_k(t)$，其輸出如下方程式(3.3)。

$$y_k(t) = f\left(\sum_j^n w_{kj}y_j(t) + b_k\right) \quad (3.3)$$

此時的輸出一樣會輸入到第二層回饋層，透過回授連接權重$r_{2,kg}$第二層隱藏層，也是上一個個時間點(t-1)神經元轉換狀態，下式(3.4)為結合輸入$y_j(t)$所產生的新的輸出方程式。

$$y_k(t) = f\left(\sum_j^n w_{kj} y_j(t) + \sum_j^g r_{2,kg} y_j(t-1) + b_k\right) \quad (3.4)$$

最後再由權重 upk 連接第二層隱藏層輸出$y_k(t)$到輸出層$y_p(t)$，此時$y_p(t)$就是深層遞迴式網路最終輸出。

有別於一般 RNN 的反向傳播演算法(Backpropagation Through Time，BPTT)，我們會根據時間前後順序來調整權重值，因此調整權重會經由不同時間點的隱藏層資訊進行，由最後時間點對成本函數( cost function )作偏微分，往前估算出一開始時間點的偏微分值，直到整個權重作出調整完後，再代回網路求新誤差值，使 MSE 接近最小值，讓輸出接近期望的值。

## 四、實驗結果

本論文使用之音檔為自行錄製真實人聲(A Cappella)與人聲同一曲目的伴唱音檔，比較時域 NLMS、頻域 NLMS、時域 RNN 等三個演算法。

為避免麥克風一打開收音就接收大量能量，以致適應性濾波器尚未收斂便引發不可收拾的嘯叫，因此將測試音檔皆編輯成:第一段(音樂)、第二段(音樂)、第三段(人聲)、第四段(人聲)、第五段(人聲)、第六段(音樂)之長度，而第一段(音樂)為純伴唱音樂且未將濾波器消除結果輸出至喇叭，為單純學習環境響應、調整濾波器權重；第二段(音樂)才將濾波器消除結果加入喇叭輸出，到第三段(人聲)才正式將近端人聲加入系統之中。此模擬類似為 google 之 Google Home 與 apple 之 HomePod，在開機時給予一個提示音，使系統學習環境響應的情況，避免每次開機皆是不同環境的情形。

### (一)音檔說明

表一、音檔格式

| | 位元率 | 聲道數量 | 取樣頻率 | 長度(秒) | 曲風 |
|---|---|---|---|---|---|
| 眉飛色舞 | 16-bit | 單聲道 | 16000 Hz | 33.5 | 快歌 |
| 天黑黑 | 16-bit | 單聲道 | 16000 Hz | 33.5 | 慢歌 |

### (二)環境響應說明(房間大小、回音長度)

以下環境響應為 RIR-Generator 生成之環境響應音檔[9]，我們將此音檔摺積輸出音檔以之來模擬真實環境下麥克風所接收到的環境響應回聲。以下表二為環境響應參數設置，皆為聲速(Sound velocity): 340 (m/s)、取樣頻率(Sample frequency) = 16000 (sample/s)。

表二、環境響應參數設置(單位:公尺)

| | 大房間(禮堂大小) | 中房間(會議室大小) | 小房間(車內大小) |
|---|---|---|---|
| Source position | [10.0 4.0 2.0] | [2.5 1.0 1.2] | [1.7 0.2 0.2] |
| Receiver position | [6.0 4.0 1.5] | [2.5 2.0 1.6] | [1.7 0.8 0.8] |
| Room dimensions | [15.5 8.5 6.0] | [5.0 4.0 3.0] | [2.3 1.7 1.2] |
| Beta(殘響長度)(s) | 1.8 | 0.45 | 0.1 |

(三)抑制評估-MSE

回聲消除效果除了主觀由耳朵聽聲音之外，我們將使用平均誤差值(Mean Squared Error，MSE)來數值化其回聲消除的效果上，其方程式如下所示:

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(s(n) - e(n))^2$$

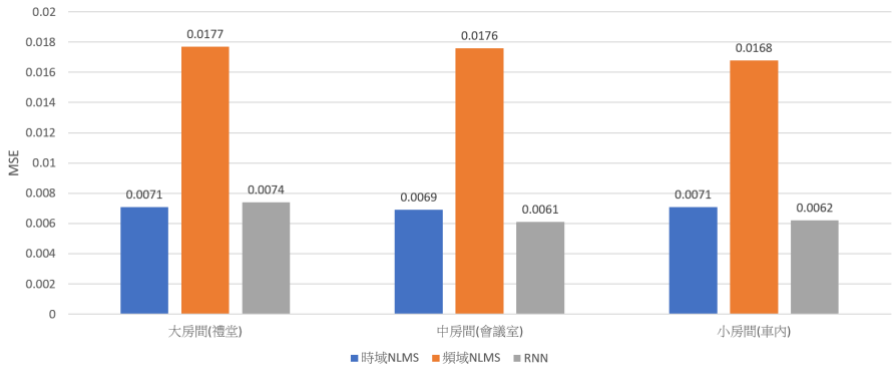
其中，s(n)為近端人聲之原始音檔。e(n)為剩餘訊號:即經回聲消除系統消除環境響應後得到的剩餘人聲。借由原始的近端人聲 s(n)與經回聲消除的剩餘人聲 e(n)，兩者相相減取平方，當 MSE 值愈小，表示消除效果愈好，代表也得到愈乾淨的近端人聲。

(四)實驗結果

每個實驗皆含 (時域 NLMS、頻域 NLMS、RNN)

1.實驗一，同一首歌_不同環境(房間大小、回音程度)

下圖六中，在不同環境下每個演算法的效果皆是穩定的，整理來看是時域會好過頻域，不過以演算法來說 RNN 與 NLMS 彼此的效果卻是相差不多的。

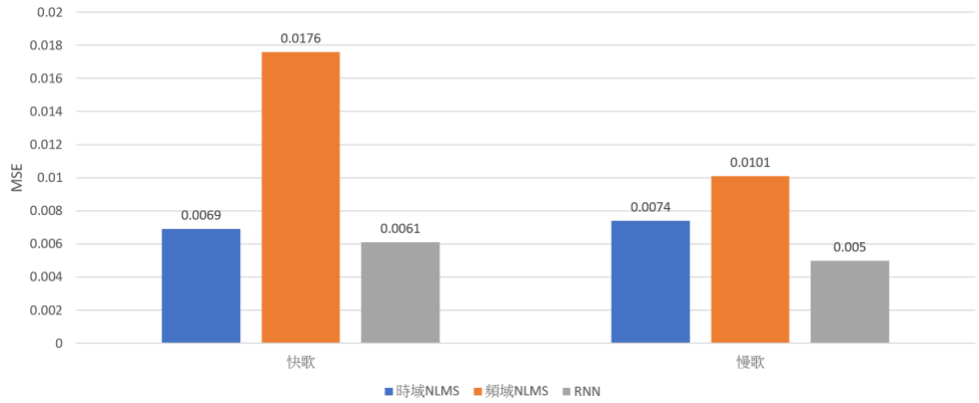



圖六、實驗一比較圖

2.實驗二，同一環境_不同首歌

下圖七中，在不同首歌下慢歌比快歌效果來的更好，這或許跟嘯叫發生頻段有關，在快

歌中的某些頻段剛好與嘯叫發生頻段符合，因此快歌在消除效果上沒慢歌來的好。

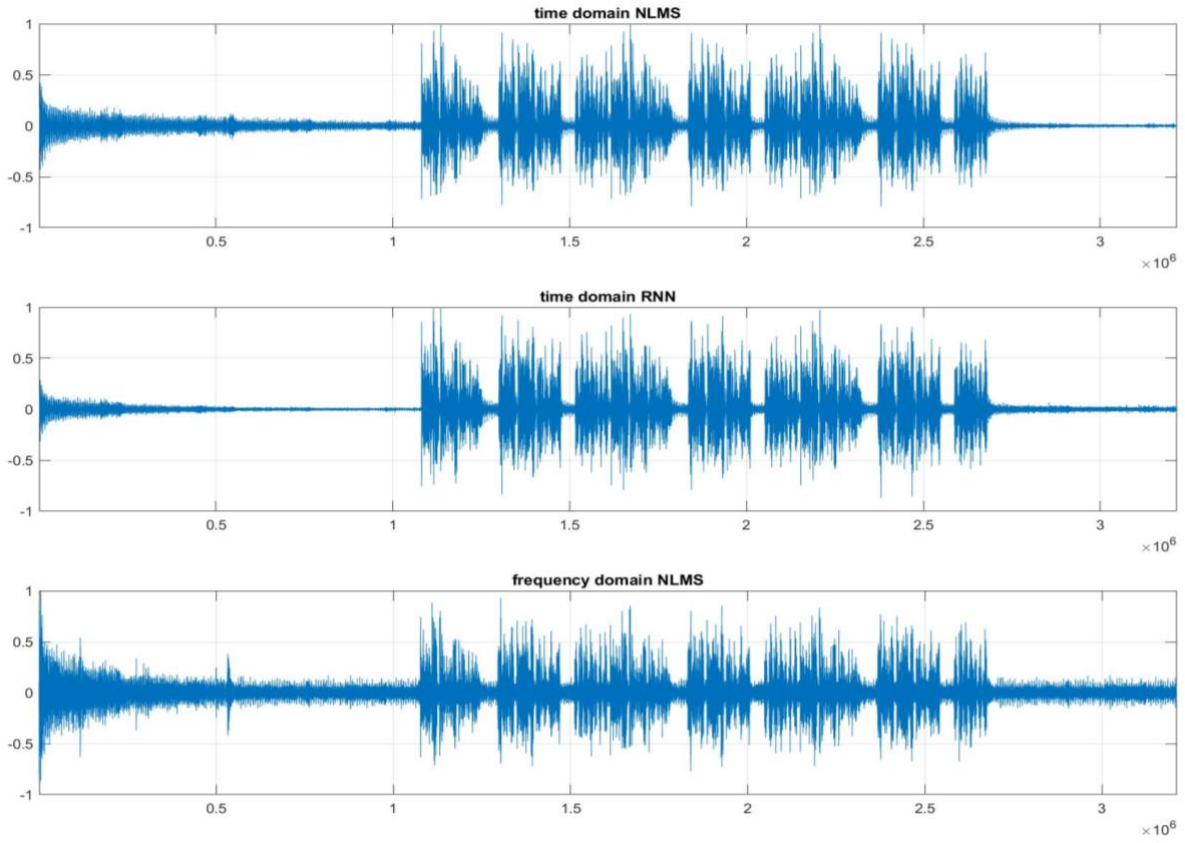



圖七、實驗二比較圖

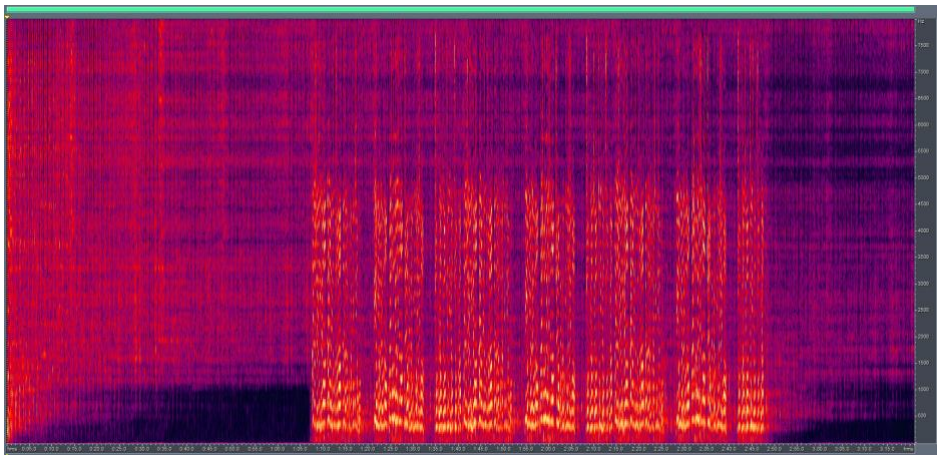
3.時域 NLMS、頻域 NLMS、時域 RNN 比較(同一首歌、同一環境)

下圖八、九、十、十一，為時域 NLMS、時域 RNN、頻域 NLMS 的時域圖、頻譜圖比較，比較下來時域 NLMS 與時域 RNN 效果相差不多，而頻域 NLMS 由於收斂比較慢的關係，所以效果略遜色於其他兩者，接下來會比較第一、二段與最後一段的效果與收斂速度。

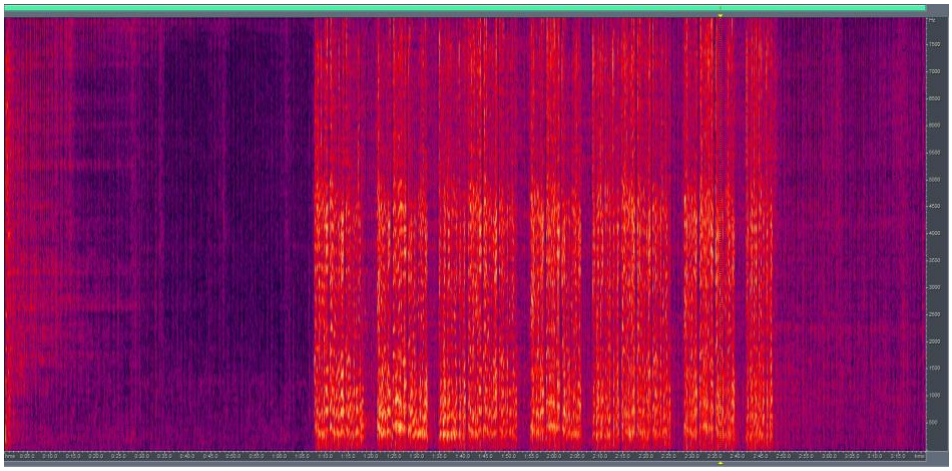
.



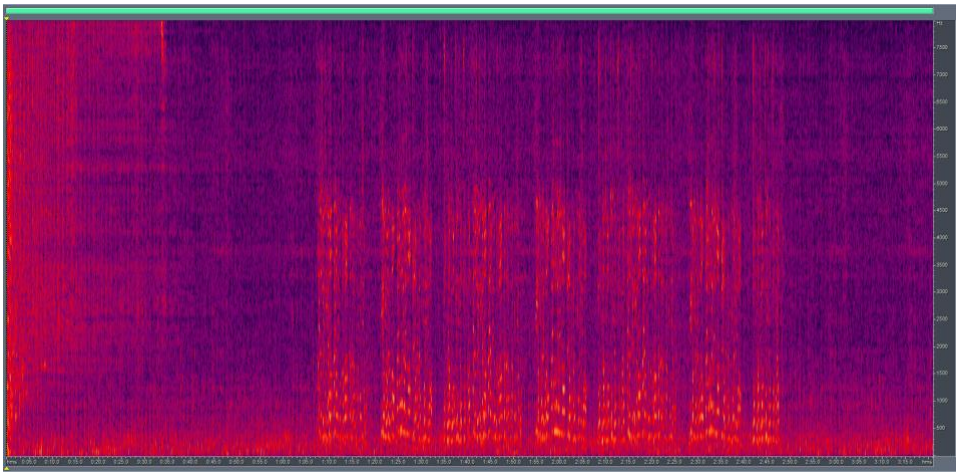

圖八、時域 NLMS、時域 RNN、頻域 NLMS 時域比較圖



圖九、時域 NLMS 剩餘人聲頻域圖



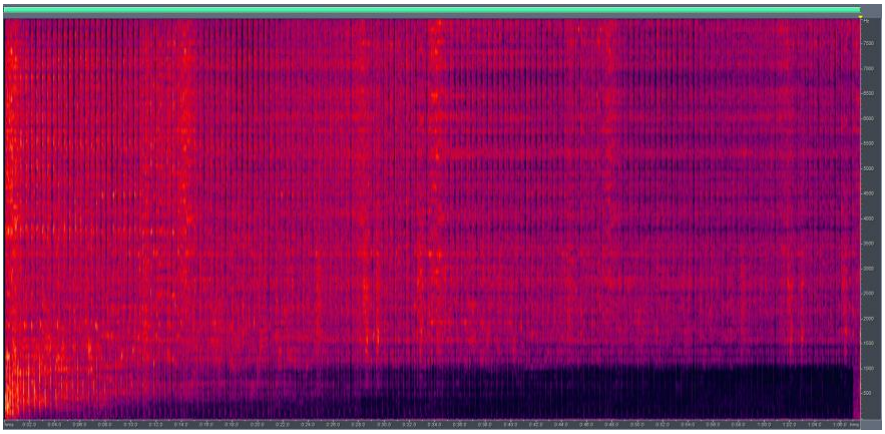
圖十、時域 RNN 剩餘人聲頻域圖



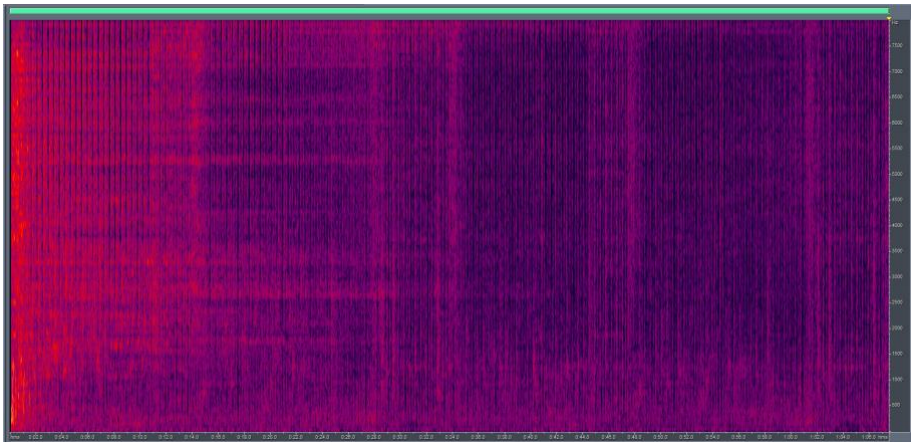
圖十一、頻域 NLMS 剩餘人聲頻域圖

由圖十二、十三、十四可知，第一段、第二段收斂過程中，時域 RNN 的消除效果較好且較快，接著是時域 NLMS，但從演算複雜度來看，時域是每一點都計算一次，也就是說第一、二段 67 秒的歌曲中便演算了 NLMS 1072 千次，每一點都調整一次權重；而頻域由於是每 256 點才取一次 512 點數做運算，512 點每一點都用一樣的調整量，頻域的

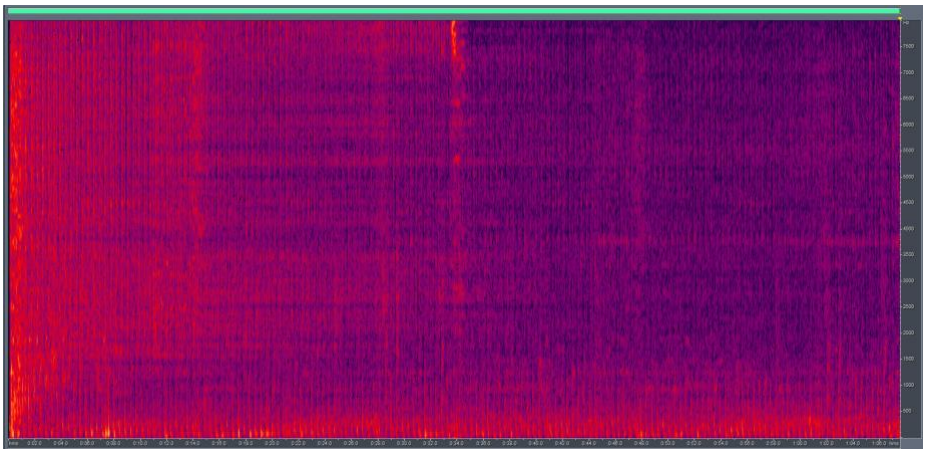
收斂速度會相較慢的。



圖十二、時域 NLMS 第一、二段頻域圖



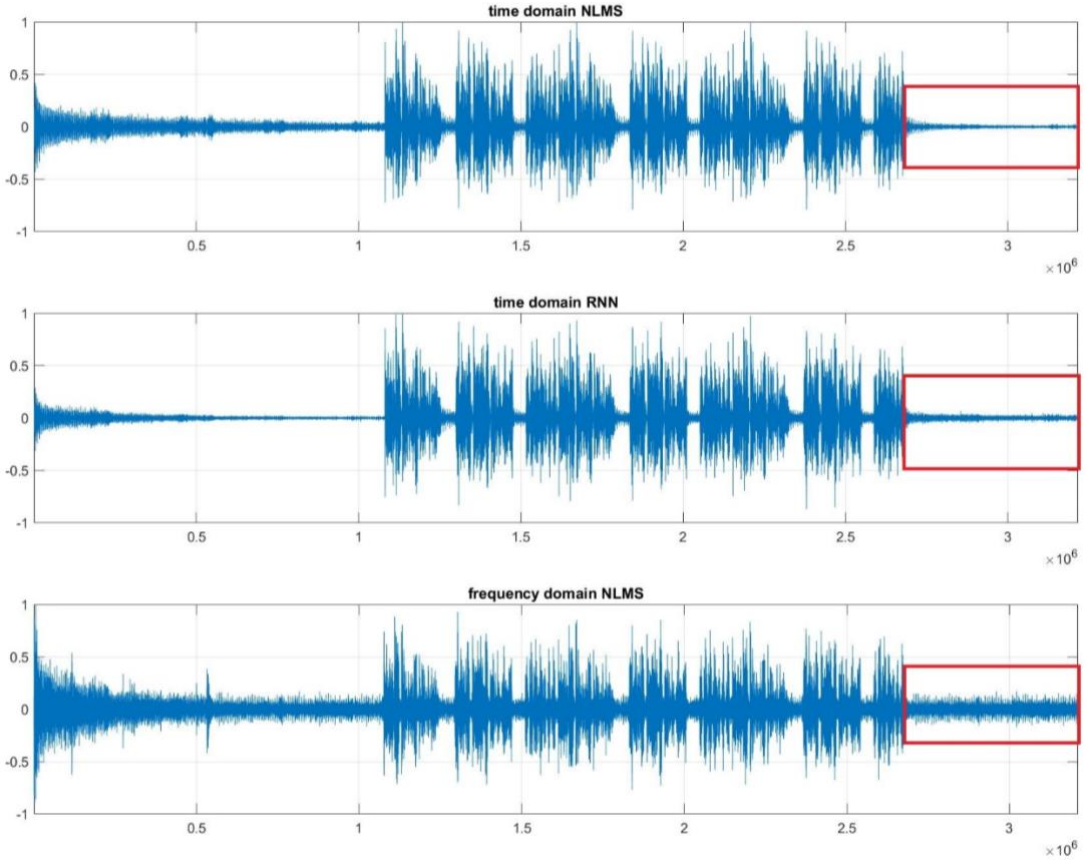
圖十三、時域 RNN 第一、二段頻域圖



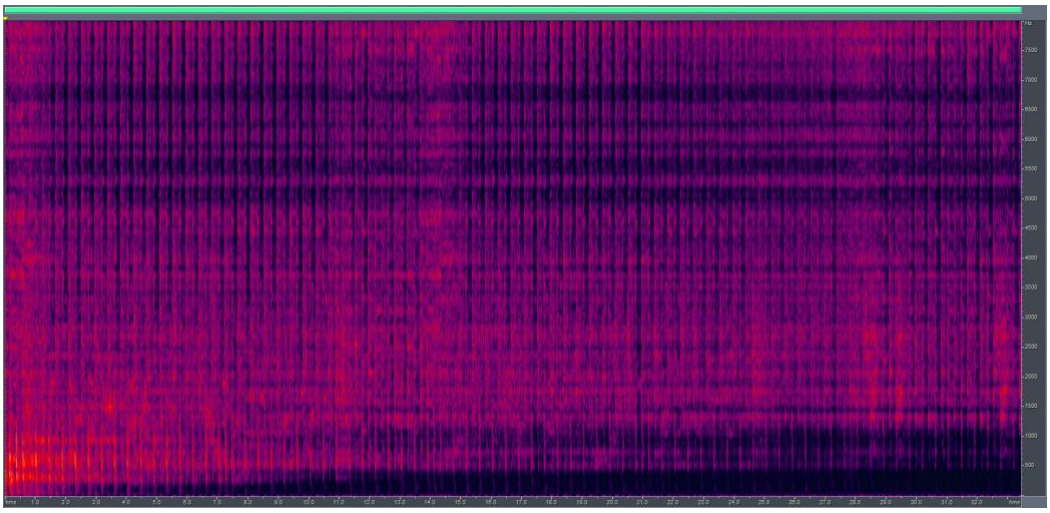
圖十四、頻域 NLMS 第一、二段頻域圖

經過中間第三、四、五段有加入人聲後，回到最後一段僅有音樂的第六段，如圖十五紅框部分所示，此時彼此都已經收斂得差不多了，但時域部分此時已經演算了 3216 千次的運算量了，而頻域演算法才做了 12.56 千次左右的計算與 FFT，彼此在 NLMS 計算量

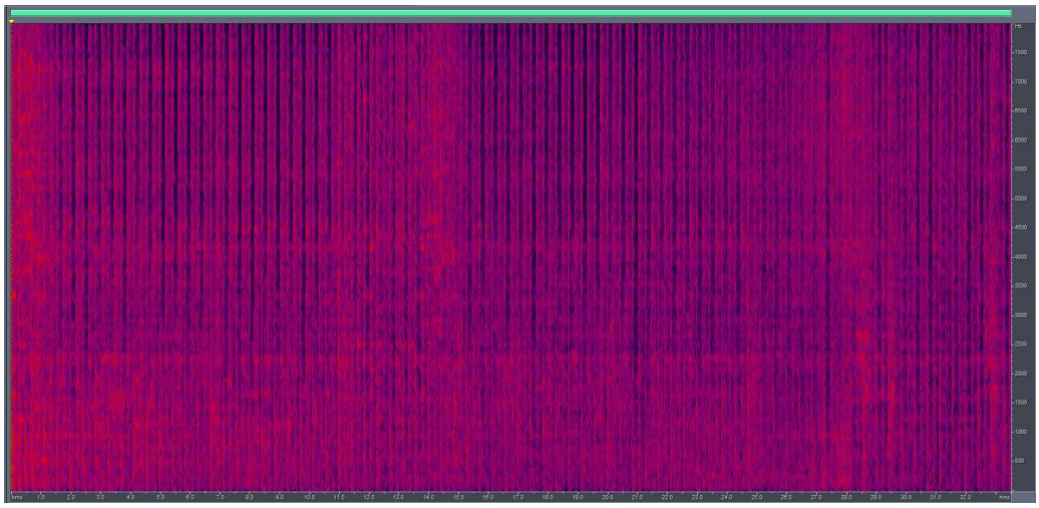
仍相差了 256 倍，而時域 RNN 是其中演算最複雜的。此外從圖十六、十七、十八比較可得知，頻域 NLMS 在 500 Hz 的地方仍有些許沒消乾淨，而時域 NLMS 在音檔突然變換的情形下，就需要一段時間才能重新收斂，但頻域 NLMS 因為有照顧到每個頻段的關係，所以突然的變化，仍可以穩定消除回聲；而 RNN 雖然也為時域運算，但因為其為非線性的演算法，在這種情形下便仍有良好的效果。在做更久的演算下，時域或許仍因每一點都運算一次而比頻域效果來的更好，但由於彼此都收斂的情形下，效果並不會相差太多，但三者的演算量與穩定度在日後的應用上，需要在效果與計算量上做取捨了。



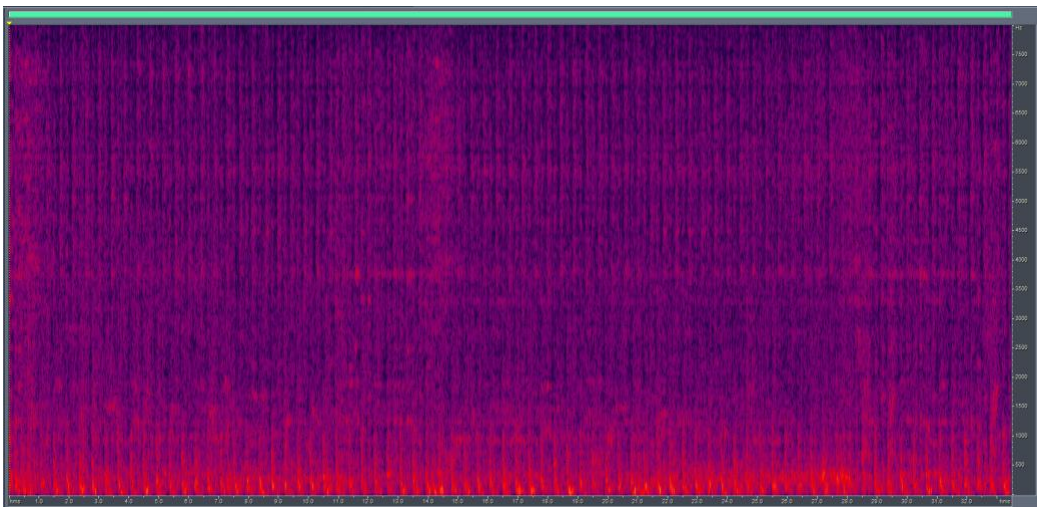

圖十五、時域 NLMS、時域 RNN、頻域 NLMS 第六段時域圖



圖十六、時域 NLMS 第六段頻域圖



圖十七、時域 RNN 第六段頻域圖



圖十八、頻域 NLMS 第六段頻域圖

## 五、結論

在本實驗中，模擬了真實環境中麥克風在不同空間下的收音與環境響應情形，分別模擬了大房間、中房間、小房間的環境響應，語料部分也分了不同曲風的歌曲來做實驗，接著分別測試了線性的時域 NLMS、頻域 NLMS 以及非線性濾波演算法 RNN，作麥克風嘯叫抑制實驗。在時域 NLMS 與頻域 NLMS 上，在一開始時域的由於是每一點就做一次，其收斂效果會比頻域來的更快，甚至消除效果更好，但在兩者都演算了一段時間後，彼此都已達到了收斂，效果其實是差不多的，但頻域在某些突如的高頻或低頻放面會比時域來的效果更好，計算複雜度上，頻域的摺積比時域來的簡單，且 256 點才做一次運算，但每次都是一次調整 512 點，因此計算量是相差不多的。此外時域 RNN 上由於非線性與有時間的記憶，在某些部分消除的效果其實是最好的，但由於其運算量龐大，日後若有應用，在這三者之間，計算量與效果好壞的取捨便端開使用的情況。


## Acknowledgements

This work was partly supported by Taiwan Ministry of Science and Technology MOST contract No. 107-2911-I-027-501, 108-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067 and partly supported by Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan contract No. TL-108-D301.


## 參考文獻 [References]


[1]. 胡立宁. 自适应回声消除算法的研究与实现. MS thesis. 吉林大学, 2007.

[2]. Stenger, A., L. Trautmann, and R. Rabenstein. "Nonliner Acoustic Echo Cancellstion with 2nd Order Adaptive Volterra Filters, IEEE Int." Conf. on Acoustics, Speech & Signal Procrssing (ICASSP). 1999.

[3]. 杜鷁、吳樂華，電聲技術與音響系統/國防工業出版社/ 2015-05-01

[4]. Tyagi, Ranbeer, Roop Singh, and Rahul Tiwari. "The performance study of NLMS algorithm for acoustic echo cancellation." 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC). IEEE, 2017.

[5]. Boukis, Christos, Danilo P. Mandic, and Anthony G. Constantinides. "Toward bias minimization in acoustic feedback cancellation systems." The Journal of the Acoustical Society of America 121.3 (2007): 1529-1537.

[6]. Ngo, Kim, et al. "Prediction-error-method-based adaptive feedback cancellation in hearing aids using pitch estimation." 2010 18th European Signal Processing Conference. IEEE, 2010.

[7]. Van Waterschoot, Toon, and Marc Moonen. "Adaptive feedback cancellation for audio applications." Signal Processing 89.11 (2009): 2185-2201.

[8]. Kashima, Kakeru, et al. "Adaptive feedback canceller with howling detection for hearing

aids." 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 2015.

[9]. Habets, Emanuel AP. "Room impulse response generator." Technische Universiteit Eindhoven, Tech. Rep 2.2.4 (2006): 1. Available: https://github.com/ehabets/RIR-Generator [Accessed: Jul. 15, 2019]