The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 37-51

基於次頻道遞迴類神經網路之麥克風陣列電視回聲消除系統

洪瑋嶸 Wei-Jung Hung[a]、蘇世安 Shih-An Su[a]、廖元甫 Yuan-Fu Liao[a]
國立臺北科技大學電子工程學系[a]
waylong711022@gmail.com, susan1101415057@gmail.com, yfliao@ntut.edu.tw

摘要

本論文研究聲控智慧電視情況下之電視節目回聲消除，希望在電視節目持續播放的干擾下，仍能收到清晰的語者的語音指令。因此本論文先使用最小方差無失真波束形成器(Minimum Variance Distortionless Response Beamformer，MVDR)，指向語者聲源位置。接著以頻域遞迴類神經網路(Recurrent Neural Network，RNN)學習房間響應路徑，清除電視回聲，最後加上頻譜消去法(Spectral Subtraction，SS)做後處理，將殘餘的回聲進一步的濾除掉。實驗針對不同電視節目類型、語者，人聲電視聲訊雜比與語者角度的組合作情境模擬，並以回聲衰減量(Echo Return Loss Enhancement，ERLE)來判斷電視回聲消除效能。實驗顯示，我們提出的方法，在不同情境下皆有良好的電視回聲消除表現，平均 ERLE 結果為 11.75dB，優於傳統的 NLMS 的 5.78dB，且處理速度比一般時域 RNN 快 15 倍，的確能有效地濾除電視回聲雜訊。

關鍵詞：頻域 RNN、MVDR、聲學回聲消除、適應性濾波器、遞迴類神經網路

一、簡介

聲控電視是非常人性化的功能，但是通常使用者在執行語音操控時，會受到電視節目的回聲與周遭背景雜訊影響，干擾使用者的語音操作。這是一種聲學回聲消除(Acoustic Echo Cancellation，AEC)[1][2]的問題，針對這個問題，傳統上大都先用麥克風陣列以 MVDR 做波束成型(beamforming)，然後以最小均方演算法(Normalized least mean squares，NLMS)[3]適應性濾波器，學習減低房間響應的影響，最後再用頻譜消去法來做後處理，進一步消除噪音的部分，如下圖一-A 所示。

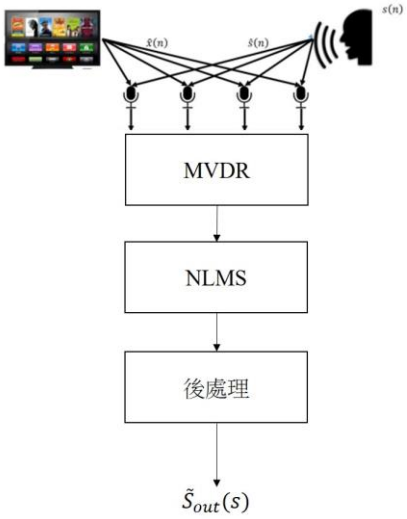
然而，在觀看電視節時，因電視節目是持續播放且聲音通常開很大聲，回聲經過重重反射後，殘響持續時間通常很久。而且，電視節目的原始聲音訊號，跟經過喇叭播放出來後，再被麥克風陣列收錄進來的聲音訊號間，通常呈非線性關係。但是傳統 NLMS 濾波器基本上是線性系統，只會考慮輸入跟輸出訊號間的線性關係，無法處理非線性訊號損耗。此外，因為 NLMS 是前饋式系統，若系統要考慮很長的殘響時間，會需要很長會需要很長的系統參數，所以常會導致運算過久，不容易收斂。

尤其是當 NLMS 在 time domain 運作，以 raw sample 為處理單位的情況下，這個問題會特別嚴重。例如在 16 kHz 取樣頻率下，若要涵蓋 0.25 秒的電視回音，就需要一個有 4096 個 taps 的適應性濾波器。

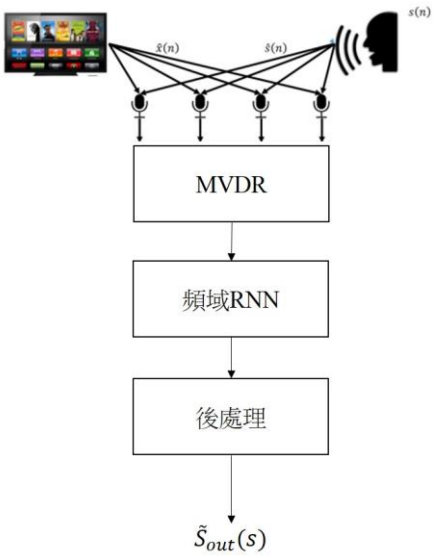
因此，為了解決這問題，所以我們改用含有回授功能的深度遞迴類神經網路(deep recurrent neural networks, RNNs)[4]，因為 RNN 能把輸出再回饋回來當輸入參數，因此能以較少的參數，記憶較長的狀態歷史資訊，幫助處理電視回音問題將系統改成而且，RNNs 的輸出實際上是一非線性函數，因此可以用來估計喇叭，麥克風陣列與房間殘響間的非線性關係。如下圖一-B 所示。

此外，在一前置實驗中，我們發現若在時域實現 RNNs 電視回音消除架構，有時候在某些頻帶常常消除得不夠好，尤其是在高頻部分。這是因為電視節目回聲是經過很多反射回來的聲音疊加起來，可以視為原始聲音跟多重路徑房間響應的卷積疊加的結果，往往在不同的頻帶上會受到不同程度的干擾。如果我們只使用時域 RNNs 架構，常會無法照顧不同頻帶的變化，導致無法有效消除雜訊。所以我們在此論文中，進一步提出頻域 RNN 電視回聲消除系統，就可以針對不同頻帶的訊號，應用不同的調整參數。

最後，在以下章節，將詳細介紹我們所提出的 RNNs 電視回音消除架構與其訓練方法。



圖一-A、
傳統 NLMS 電視回聲消除系統架構圖



圖一-B、
頻域 RNN 電視回聲消除系統架構圖

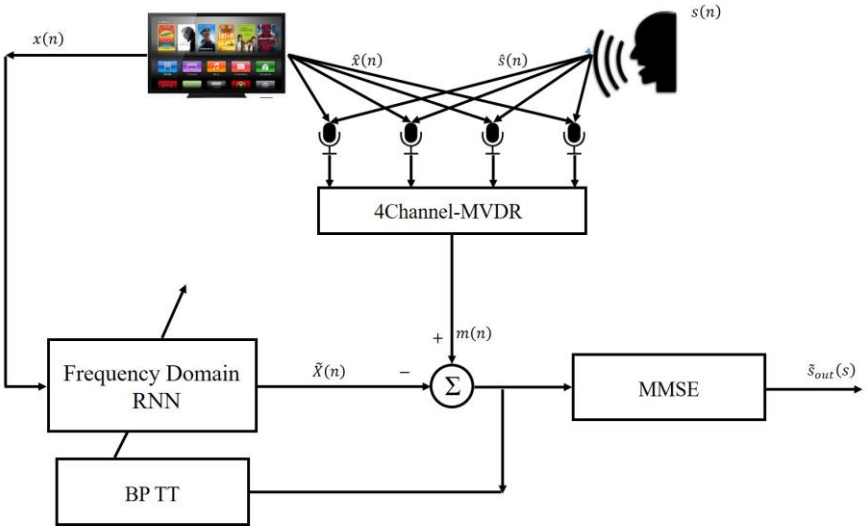
二、基於次頻帶 RNN 之麥克風陣列電視回聲消除系統

本論文使用多通道麥克風陣列為系統的輸入[5]，為了在收音時，能夠指向使用者聲源，先使用最小變異無失真響應波束形成器(Minimum Variance Distortionless Response Beamformer,MVDR)[5][6]，利用空間資訊進行指向性處理。再用有非線性處理能力及能補捉長時間資料的遞迴類神經網路 RNN[4]，作為核心來做回聲消除，而且為了讓 RNN 在回聲路徑上能夠預估得更細微，我們把時域訊號轉換到頻域，並將訊號分成不同頻帶進行不同處理。最後則使用 Ephraim's Minimum Mean Square Error log- Shor time Spectral Amplitude（MMSE-log-STSA）[9]頻譜消去法做後處理，進一步利用時間資訊，消除殘餘的回聲雜訊。

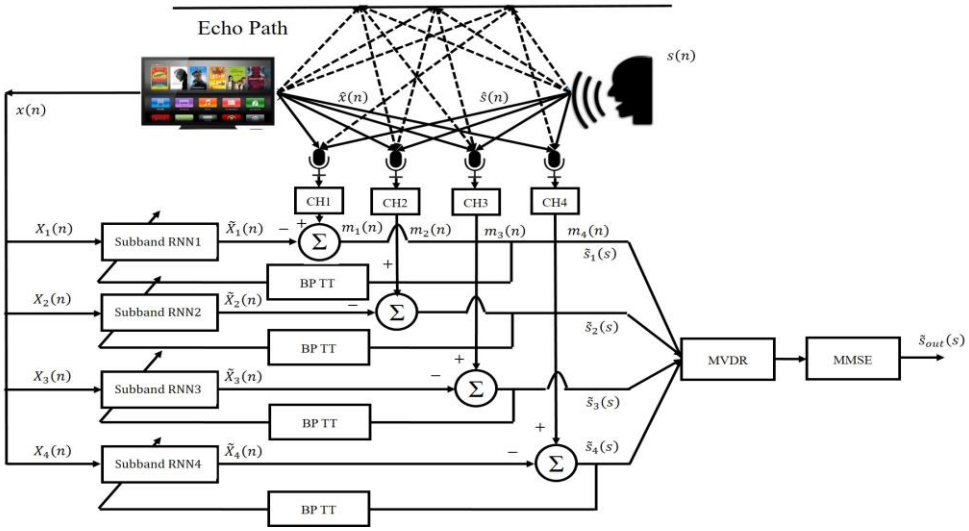
此外，我們考慮兩種不同的 MVDR 與 RNNs 組合架構，分別如下圖二-A 與二-B 所示。其中下圖二-A，是先經過 MVDR 做指向性處理，再用頻域 RNN 做預估與抵消。下圖二-B 則是先將每個通道麥克風收到的聲音先經過 RNN 消除回聲，讓每個麥克風的誤差達到最小之後，再做 MVDR 的動作會不會比較好。

以下進一步詳細此兩架構中的各子模組，包括（一）MVDR，（二）頻域 RNNs 電

視回聲消除架構與（三）MMSE-log-STSA 後處理方法。



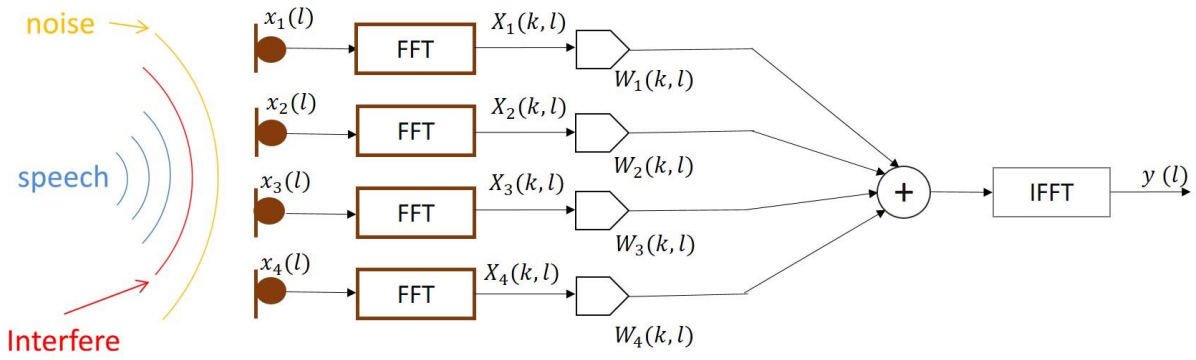
圖二-A、MVDR+頻域 RNN 之電視回聲消除架構圖



圖二-B、頻域 RNN+MVDR 電視回聲消除架構圖

（一） MVDR 最小變異失真響應

假設麥克風陣列會收到目標語音，以及來自其他方向的干擾雜訊與非目標語音信號。我們使用 MVDR Beamformer[7][8]，。因為在每支麥克風在不同角度的關係，使得語音到達每支麥克風都會有延遲。如果能透過使用者聲源方位，計算聲音傳至麥克風的時間差[5]，並透過最小輸出方差準則找到頻域濾波器係數的最佳解，進行 filter-and-sum 補償，就可將使用者說話聲增強。如下圖三所示。



圖三、MVDR 架構圖[7]

其中麥克風接收聲音訊號$x_m(n)$中含有目標聲源訊號$s(n)$跟干擾加噪音訊號$v(n)$。通過對時域麥克風輸入訊號應用快速傅里葉變換（FFT），則麥克風頻域信號為{$X_m(k,l)$，m=1,2…,4}。其訊號成分可表示為

$$X_m(k,l) = e_m(k,\theta)S(k,l) + V_m(k,l) \tag{2.1}$$

其中k是頻率索引，l是輸入的短時段索引；$e_m(k,\theta)$表示第 m 個麥克風對目標來源的到達方向（DOA），θ是目標來源，$S(k,l)$和$V_m(k,l)$分別表示變換後的語音信號和乾擾加噪聲信號。假設語音與干擾加噪聲信號不相關，則$X_m(k,l)$的空間頻譜相關矩陣為

$$\boldsymbol{R_{XX}}(k,l) = \sigma_S^2(k)\boldsymbol{e}(k,\theta)\boldsymbol{e}^H(k,\theta) + \boldsymbol{R_{VV}}(k,l) \tag{2.2}$$

其中H是表示共軛轉置，$\boldsymbol{R_{XX}}(k,l)$和$\boldsymbol{R_{VV}}(k,l)$是空間譜相關矩陣，$\sigma_s^2(k)$是語音功率譜密度，波束形成器的輸出訊號$Y(k,l)$可以寫成下式。此外$Y(k,l)$訊號可以再用反傅立葉轉換，獲得最後轉檔或時域的麥克風輸出訊號。

$$Y(k,l) = \boldsymbol{W}^H(k,l)\boldsymbol{X}(k,l) \tag{2.3}$$

而為了找到最佳權重向量，MVDR 波束形成器將每個頻率的總輸出信號功率最小化，同時限制波束形成器的權重大小，讓麥克風指向的方向的響應為 1，即依下式尋找

最佳解。

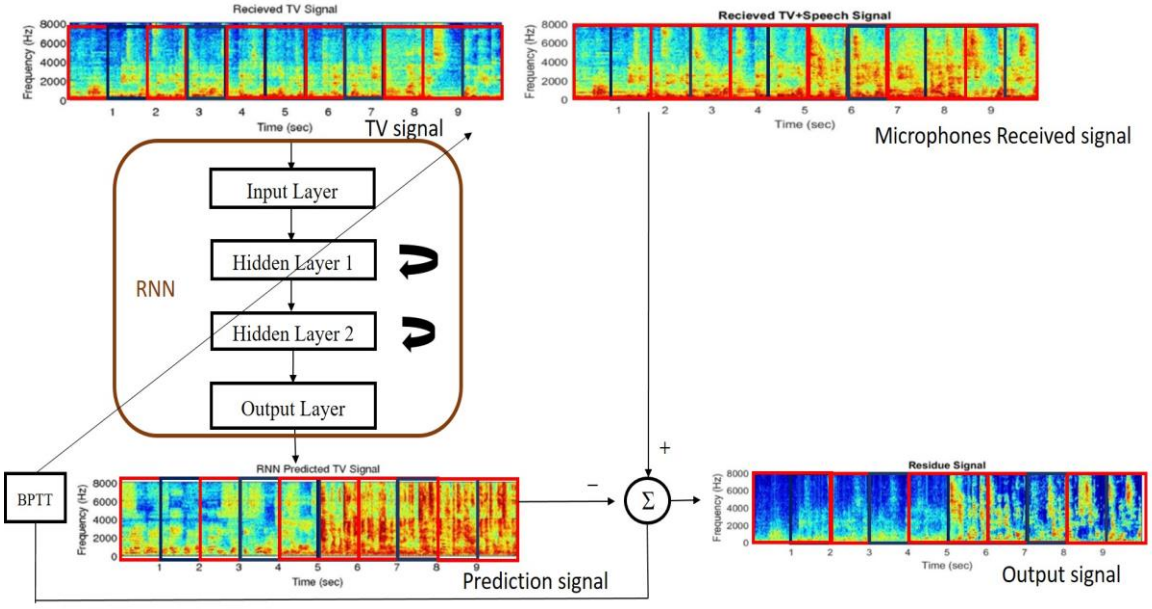
$$\boldsymbol{W}_{MVDR} = \underset{w}{\operatorname{argmin}}\, \boldsymbol{W}^{H}\boldsymbol{R}_{XX}\boldsymbol{W} \text{ ，subject to } \boldsymbol{W}^{H}\boldsymbol{e} = 1 \tag{2.4}$$

其中$a(\omega, k)$為拓撲向量，這裡假設雜訊與目標訊號無關聯且平均為 0，將 $W^{H}(\omega)$對 $X(\omega)$做矩陣處理。則在無失真準則無雜訊的情況下，對任意$R(\omega)$之 MVDR 最佳解為

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{e}}{\mathbf{e}^{H}\mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{e}} \tag{2.5}$$

（二） 頻域 RNN 適應性電視回聲消除

RNN 適應性電視回聲消除的做法，主要是先將電視原始訊號輸入 RNN，讓 RNN 預測麥克風收到的電視回聲訊號。再與真正收到的人聲加電視回聲訊號相減後，消掉電視回聲，得到較不受電視回聲影響的人聲訊號。不過我們進一步將整個運作從時域轉到頻域，變成次頻帶 RNN 適應性電視回聲消除法。此頻域 RNN 架構與一般時域 RNN 相同。不同的是將輸入做改變，原本時域 RNN 的輸入是一整段的時域語音訊號，而頻域 RNN 做法是將訊號切成音框，將音框從時域用 FFT 轉到頻域，所以輸入就改為這一小段音框的頻譜訊號。下圖四為次頻帶 RNN 適應性電視回聲消除流程圖。



圖四、頻域 RNN 架構圖

其中，因麥克風陣列取樣頻率為 16Khz，我們將 FFT 設為 256，而且由於 FFT 轉換對實數訊號具有對稱性，所以 RNN 的輸入只需要 128 點的資料。學習完的頻域輸出訊號可用 IFFT 反轉為時域訊號，再使用 Overlap-Add 一一疊加起來，就能還原成一整段的時域輸出訊號

本論文所提出的頻域 RNN 電視回聲消除法做法如下。首先將原始的電視聲經過 RNN 計算。公式如下(假設只有一層隱藏層的情況下)，其中$W_{xh}(t)$為隱藏層權重值，$W_{hy}(t)$為輸出層權重值：

$$\begin{aligned}\mathbf{h}_t &= f(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}) \\ \mathbf{y}_t &= g(\mathbf{W}_{hy}\mathbf{h}_t),\end{aligned} \tag{2.6}$$

$\boldsymbol{y}_t$是預測出來的訊號，並與麥克風收到人聲和電視聲的混合訊號相減，得到某一時間點的誤差訊號。再以均方差計算成本函數，公式如下（其中$l_t(j)$為麥克風收到人聲加電視聲的訊號）：

$$E = c\sum_{t=1}^{T}\|\mathbf{l}_t - \mathbf{y}_t\|^2 = c\sum_{t=1}^{T}\sum_{j=1}^{L}(l_t(j) - y_t(j))^2 \tag{2.7}$$

最後則利用反向傳播算法(back propagation through time，BPTT)調整 RNN 權重。BPTT 先由最後時間點，開始對於成本函數作偏微分，再往前算到一開始時間點的偏微分值。RNN 的調適公式如下：

$$w_{xh}^{new}(i,j) = w_{xh}(i,j) - \gamma\sum_{t=1}^{T}\frac{\partial E}{\partial u_t(i)}\frac{\partial u_t(i)}{\partial w_{xh}(i,j)} \tag{2.8}$$

（三） MMSE log-STSA 最小均方誤差短時譜幅度估計

後處理的部分使用 MMSE[10]來做更進一步的的消雜訊動作。首先利用語音活性檢測(voice activity detection，VAD)[11]，對輸入信號的一個區塊提取特徵然後對這個區塊進行分類，切出語音訊號，估計背景雜訊頻譜。然後再利用頻譜消去法，去掉背景雜訊，讓輸入訊號透過頻譜相減達到降噪的效果。

三、實驗結果

本論文的實驗語料，包含 8 個語者的 TCC300 語料，以及考慮 4 種電視節目類型，每一類選 10 個電視節目聲。SNR 設定則為人聲與電視聲一樣(0dB)、人聲比電視聲小(-6dB)、人聲比電視聲大(+6dB)三種。其中單通道實驗取靠近麥克風中心點最近的第二聲道做為測試。多通道則使用 4 個麥克風。頻域處理音框大小設為 256，使用漢明窗，RNN 演算法設定為兩層隱藏層，每層神經元為 100。

（一） 語料說明

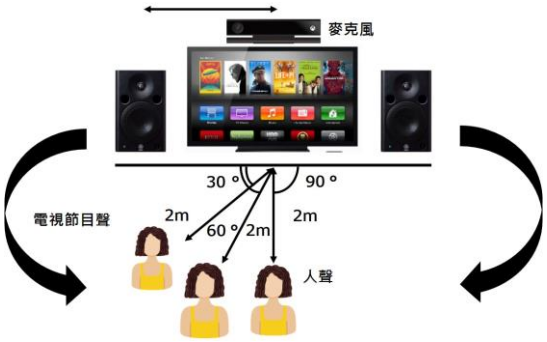
測試人聲錄音語料從 TCC300 語料庫中，選擇了 4 男 4 女的音檔，且隨機擷取十秒鐘片段說話聲；而所回錄的電視節目聲為四大類，每類為有 10 個長度大約半小時的節目，從中隨機擷取十秒鐘片段節目聲。所以每人共有 40 個測試電視節目聲。以下表 1 為語料的格式設定。

表 1　語料格式設定

	電視節目聲	語者人聲
Source	Youtube	TCC300
Format	Wave PCM	Wave PCM
Sample Rate	44.1K Hz	16K Hz
Bit Resolution	16-bit	16-bit
Vocal Track	stereo	mono

（二） 實驗情境

為了能夠模擬真實的聲學回聲消除系統情形，我們在一個類似客廳的房間模擬遠距離收音情境。首先，把 Kinect for Xbox one 充當接收端麥克風陣列。並在螢幕位置左右兩旁平行放上兩顆主動式監聽喇叭，播放電視節目的聲音。另外，在螢幕的正前方距離 2m 處，擺上另一顆主動式監聽喇叭播放人聲，模擬使用者正在講話。如此一來，當播放出人聲時，影片節目聲也同時混進疊加其中，一起被麥克風陣列所接收，並錄音起來當作我們聲學回聲消除系統的語料。實際上我們共錄製說話者在 90，60 與 30 度角位置的人聲，整體實驗環境擺設如下圖五、六。



圖五、角度擺設圖

圖六、實際電視喇叭與收音麥克風與實際模擬說話者之音源喇叭擺設

（三） 回聲消除效能評估

回聲消除成效除了主觀的由耳朵聽取聲音外，還可以用平均誤差值(Mean Squared Error，MSE)，與回聲返回損失強化(Echo Return Loss Enhancement，ERLE)[5]數值化。ERLE 評估函式式如下式所示：

$$\text{ERLE} = 10\log_{10}\frac{m^2(n)}{\tilde{s}^2(n)} \tag{3.1}$$

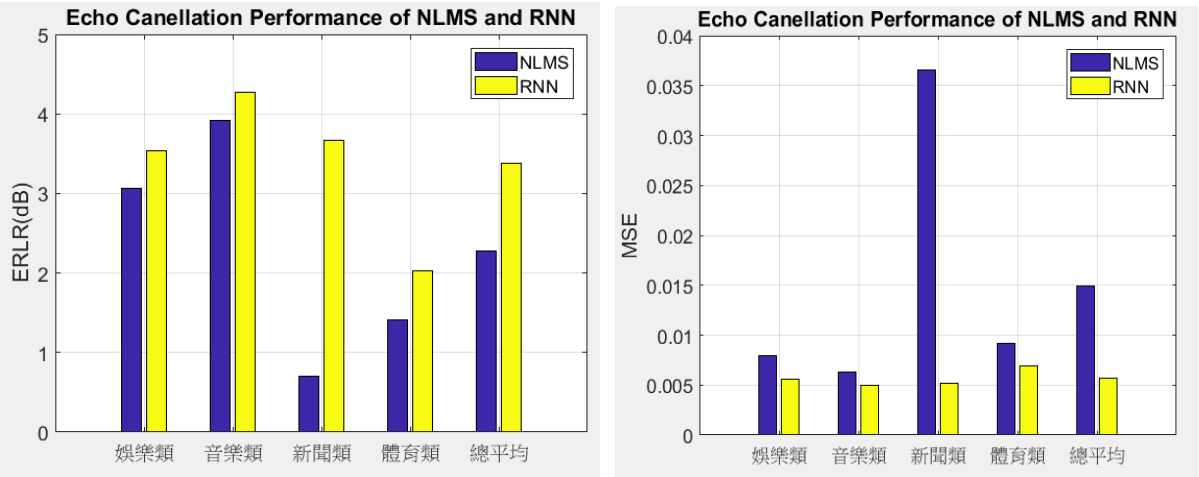
其中，m(n)為 TCC300 測試人聲混電視節目聲的錄音訊號聲。s̃(n)為消除回聲後的估計語音訊號；即經回聲消除系統消除影片節目聲後得到的 TCC300 測試人聲。藉由原始的訊號聲 m(n)與經回聲消除後得到的 TCC300 測試人聲 s̃(n)兩者相互去比較，當分子 s̃(n)越小時，此時 ERLE 值就愈大，代表消除性能愈好；也表示得到愈清晰的 TCC300 測試人聲。

（四） 實驗結果

在以下實驗中，先讓系統藉由前 5 秒純電視回聲的適應學習，再用來預測到下五秒的電視回聲，測試人聲與電視節目聲混合的比例先保持一樣大聲，即模擬 SNR = 0dB 的情形。實驗結果主要以後五秒為判斷系統好壞準則。以下介紹五種實驗：

- 實驗一：時域NLMS與時域RNN電視回聲消除實驗。
- 實驗二：時域與頻域RNN電視回聲消除實驗。
- 實驗三：MVDR加頻域RNN電視回聲消除實驗。
- 實驗四：MVDR加頻域RNN在不同角度電視回聲消除實驗。
- 實驗五：MVDR加頻域RNN再加上MMSE後處理電視回聲消除實驗。

1. 實驗一，時域 NLMS 與 RNN 電視回聲消除實驗

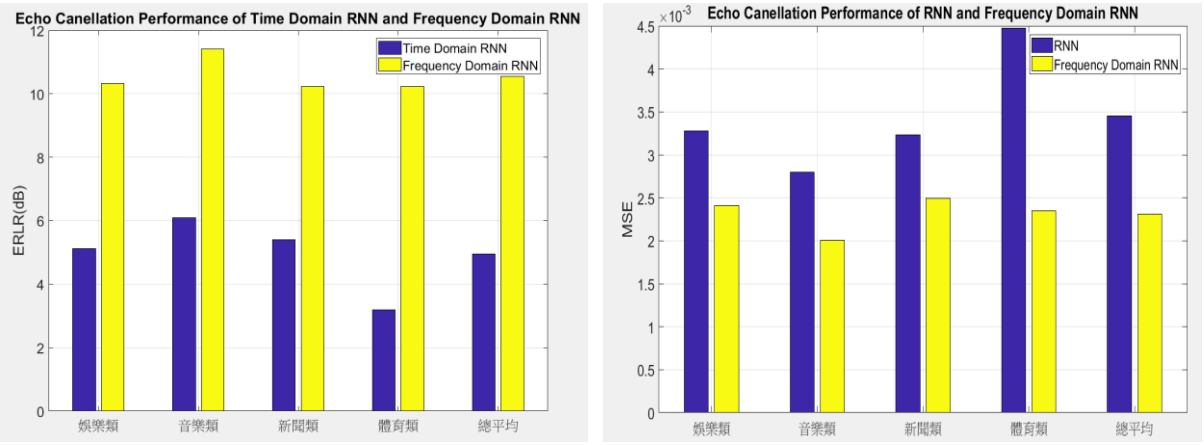


圖七、實驗一 NLMS 與時域 RNN 電視回聲消除系統的 ERLE 與 MSE 結果圖

實驗結果若以如圖七新聞類來為例，左圖的 RNN 在此類型的 ERLE 值有 3.673 遠高於 NLMS 的 0.698，右圖的 RNN 的 MSE 只有 0.00514，而 NLMS 有 0.03654，所以 RNN 效果明顯優於 NLMS。

2. 實驗二，時域與頻域RNN電視回聲消除實驗

實驗二考慮到利用多通道來增強人聲，所以先假設說話者在正前方，用麥克風陣列收集的 4 個聲道音訊，分別做 4 次的深層遞迴式神經網路，在加總增強語音，並比較 RNN 在時域與頻域消除電視回聲的效能。結果如下圖八所示：

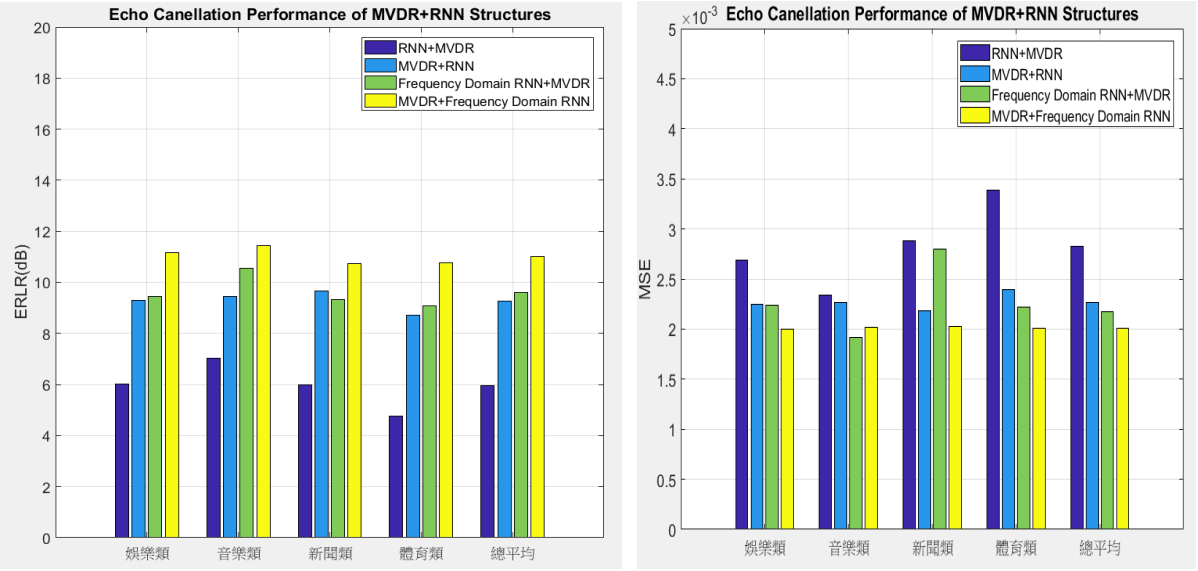


圖八、實驗二 RNN 與頻域 RNN 電視回聲消除系統的 ERLE 與 MSE 結果圖

從圖八的實驗結果，可以看出多通道頻域 RNN 在各種類型的電視節目都優越於多通道時域 RNN。

3. 實驗三，MVDR 加 RNN 電視回聲消除實驗

實驗三則實驗如圖二所示之兩種不同 MVDR 與 RNN 整合架構，一為測試收音後經過 MVDR 先抑制電視回聲，再經過頻域 RNN 過濾電視回聲，另一為收音後先由頻域 RNN 消除回聲，再用 MVDR 增強人聲。實驗比較加入 MVDR 後的回聲消除能否更為優越，跟兩種不同整合架構，那一個更突出。實驗結果如圖九所示：

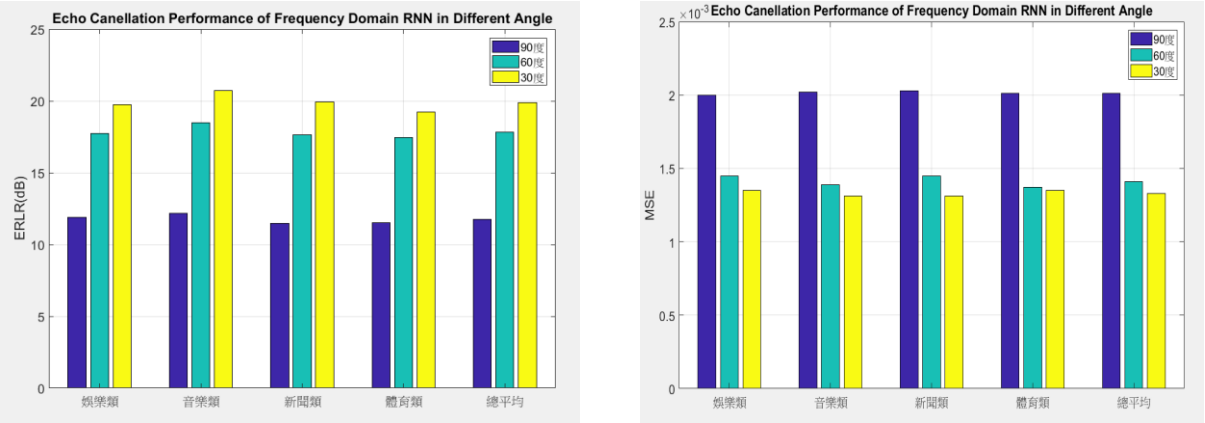


圖九、實驗三 各種 MVDR+RNN 電視回聲消除系統的 ERLE 與 MSE 結果圖

由圖九可以看出先用 MVDR 抑制電視回聲，再經過頻域 RNN 過濾電視回聲，在各種類型電視節目的效能，都比先用頻域 RNN 做消除電視回聲，再用 MVDR 增強人聲的效果好。

4. 實驗四，MVDR 加頻域 RNN 在不同角度電視回聲消除實驗

由實驗三可得知，MVDR+頻域 RNN 在各種回聲消除實現比較中，有最好的效能，我們就以 MVDR+頻域 RNN 測試使用者在不同角度的回聲消除效能，以便得知我們提出的系統能否可以在使用者在不同位置時，一樣能接受到乾淨的語音。實驗測試使用者在 90 度、60 度、30 度情形，實驗結果如圖十所示

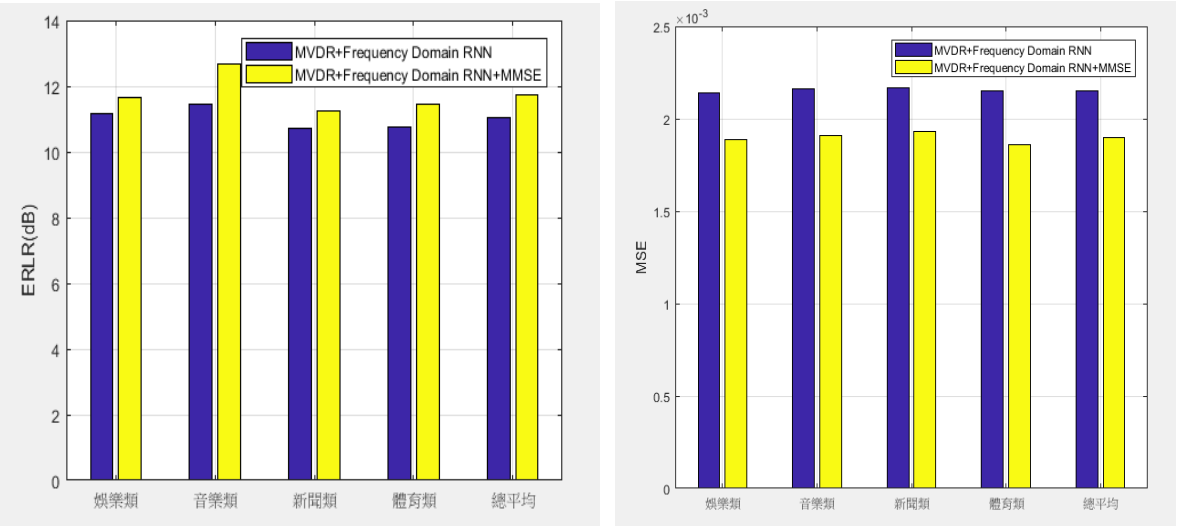


圖十、實驗四 頻域 RNN 在不同角度的電視回聲消除 ERLE 與 MSE 比較圖

由圖十的 ERLE 和 MSE 比較圖，可以得知不管使用者在哪一個方位，系統都能輸出較為乾淨人聲語音，也代表它能指向使用者位置做語音增強。而且，從圖十我們可以發現到在 30 度和 60 度時效果會比原本的正前方還要好。此實驗可證實我們提出的方法，能在使用者在不同角度時依然有很好的效率。

5. 實驗五，MVDR 加頻域 RNN 再加上 MMSE 後處理電視回聲消除實驗

由實驗三可得知，MVDR+頻域 RNN 在各種回聲消除實現比較中，有最好的效能，所以我們直接以說話者在正前面的 MVDR+頻域 RNN 與 MVDR+頻域 RNN 多加一個後處理的做比較。後處的方法使用 MMSE-log-STSA 方法，看能不能更進一步濾除掉剩餘的電視節目回聲。實驗結果如圖十一所示：



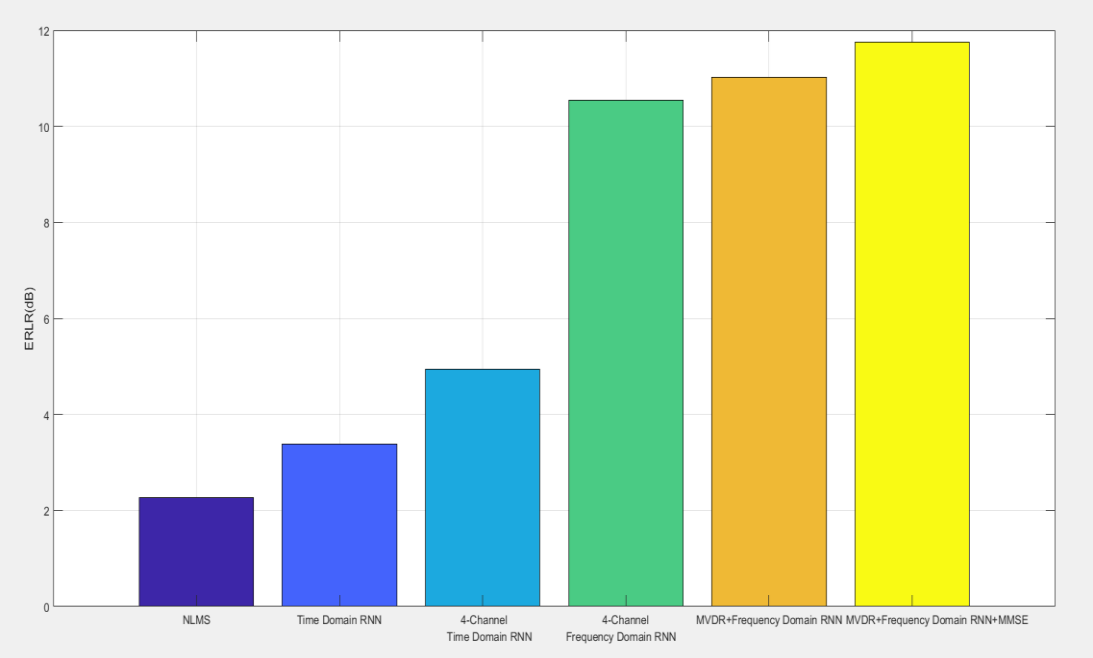
圖十一、實驗五 加入後處理後的電視回聲消除系統的 ERLE 與 MSE 結果圖

由圖十一左邊可以看到，對各種類型的電視節目的 ERLE 值，使用 MVDR+頻域 RNN＋MMSE 的後處理總平均為 11.754，比沒經後處理的 MVDR+頻域 RNN 的 11.023 還要好。

五、結論

本實驗中利用監聽式喇叭及 Kinect 等器材，實際錄音模擬智慧型電視操作情境，作為電視節目回聲消除實驗的語料。先後導入了 MVDR 做麥克風陣列，利用波束成型，將人聲增強，也更進一步導入非線性濾波 RNN，在頻域處理不同的頻帶的電視回聲，最後再加上 MMSE 做後處理。

本論文實驗結果總結如圖十二所示，從此電視回聲消除實驗結果，我們發現用 RNN 的效果比 NLMS 好，使用多通道的麥克風也比單通道的好，而且頻域 RNN 效果比時域 RNN 的好，還有先將收到的人聲加電視回聲做 MVDR，再做頻域 RNN 的動作會比先做 RNN 再做 MVDR 的效果還要好。最後將 MVDR 加上頻域 RNN 在做後處理的部分效果更好。



圖十二、實驗結果 ERLE 比較圖

參考文獻

[1] 莊世昌，用於網路電話之多頻帶聲學回聲消除研究，碩士論文，國立臺灣科技大學資訊工程研究所，臺北，2014.

[2] 胡立寧，自適應回聲消除算法的研究與實現，碩士論文，吉林大學，中國，2007. 適應性濾波器

[3] A. Stenger, 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.

[4] RNN 介紹：http://www.360doc.com/content/16 /0302/19/2459_538881000.shtml , 2016, March.

[5] 陶柏戎，運用多個聯網麥克風進行室內環境語音語音樂之增強:波束成形方法開發與評估，碩士論文，國立清華大學電機工程研究所，新竹，2016.

[6] 劉淵翰,語音強化與立體聲回聲消除於智慧型電視之應用,碩士論文,國立交通大學, 2013.

[7] J. Capon, High-resolution frequency-wavenumber spectrum analysis, Proc. IEEE vol.57, no.8, 8,August 1969.

[8] Shengkui Zhao, Douglas L. Jones,Suiyang KhooandZhihong Man, Frequency-domain beamformers using conjugate gradient techniques for speech enhancement , 2014.

[9] Simit Shah and Roma Patel,MMSE STSA Based Techniques for Single channel, Electronics and Communication Department, Parul institute of engineering and technology, Vadodara, Gujarat india, 2015.

[10] Simit Shah and Roma Patel,MMSE STSA Based Techniques for Single channel, Electronics and Communication Department, Parul institute of engineering and technology, Vadodara, Gujarat india,2015.

[11] Improving Single Frequency Filtering based Voice Activity Detection (VAD) using Spectral Subtraction based Noise Cancellation ,Department of Electronics and Communications NMAM Institute of Technology, Karnataka State, India,2016