The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 372-386

基於深層遞迴類神經網路之多通道電視回聲消除系統

Multi-Channel Television Echo Cancellation based on Deep Recurrent Neural Networks

黃宏 Huang Hung
國立台北科技大學電子工程學系
National Taipei University of Technology Department of Electronic Engineering
mainmemory1103@gmail.com

洪瑋嶸 Hung Wei-Jung
國立台北科技大學電子工程學系
National Taipei University of Technology Department of Electronic Engineering
waylong711022@gmail.com

廖元甫 Liao Yuan-Fu
國立台北科技大學電子工程學系
National Taipei University of Technology Department of Electronic Engineering
yfliao@ntut.edu.tw

摘要

本論文研究智慧型電視操作情境下之電視節目回聲消除，希望能在電視節目持續播放的情形下，仍能錄到說話者的清晰語音，並能應用在即時語音通訊與遠距語音辨認人機介面上。本論文的回聲消除系統演算法是以遞迴類神經網路(Recurrent Neural Network，RNN)演算法，再配上多通道麥克風做回聲消除，達到人聲增強，抑制噪音雜訊，提高語音清晰度。實驗分別實作單純電視節目聲、人聲混電視節目聲兩種實驗，再導入前五秒無人聲預訓練，後五秒有人聲之電視節目回聲消除模式實驗，實驗結果以回聲衰減量來判斷效能優劣。實驗顯示，以多通道深層遞迴類神經網路效能優於其他方法，透過多聲道 RNN 處理，的確能有效地濾除雜訊。

關鍵詞: 聲學回聲消除、適應性濾波器、類神經網路、遞迴類神經網路

一、簡介

聲控電視是非常人性化的功能，但礙於電視回聲與雜訊等問題影響，常會干擾使用者語音操作。因此一般需要加上聲學回聲消除系統，以適應性濾波器演算法[1-2]，在實際空間環境下，學習回聲路徑，預測並消除電視節目回聲，增強使用者語音質量。

聲學回聲主要是聲音經喇叭播出，由空間響應導致的一次或多次的反射聲到麥克風所引起，主流的回聲消除方法運用其架構如圖 1 所示，其使用適應性演算法自動地調整濾波器權重係數，使輸出信號能夠逼近所期望的信號。

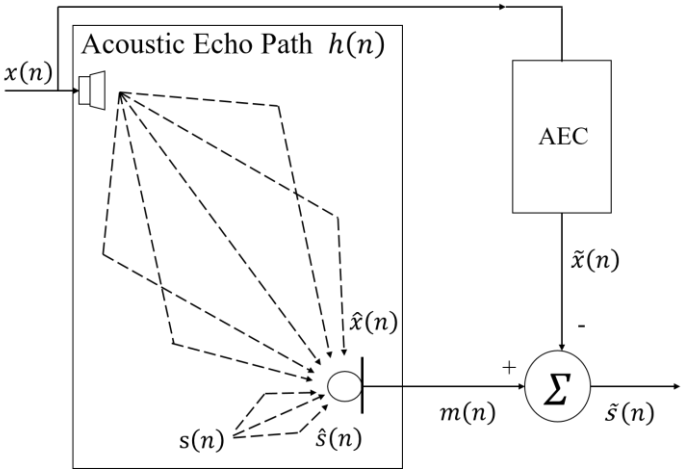

圖 1 典型聲學回聲消除系統架構圖

適應性濾波器演算法有線性與非線性消除兩種作法，傳統上線性方法常以正規化最小均方演算法(Normalized least mean squares，NLMS)[3-4]來實現，但是由於在室內背景雜訊眾多，NLMS 演算法未能有效解決非線性回聲，於是常進一步引進非線性適應性演算法，改善非線性回聲的部分，常用的非線性演算法包括輻射基底函數類神經網路(Radial Basis Function Neural Network，RBFNN)[3,5-6]與多層感知機(Multi-layer Perceptron Neural Network，MLP)[3, 5-6]演算法，其透過具適應性學習能力的類神經網路演算法，克服非線性的因素。不過 RBF 與 MLP 雖然相當有效，但因其輸入訊號的長度通常是有限的且不能太長，以免運算太久，因此通常只能處理較短時間的回聲，若回聲的影響時

間太久，其效果通常不好。

但在觀看電視時，因電視節目聲音通常開很大聲，且回聲經過重重反射，殘響時間通常很長。而且電視節目是動態持續的播放著，上一個時間點播放的聲音會影響下一個時間點的聲音。若我們想捕捉長期時間關聯的資訊，就必須使用含有回授功能的深層遞迴類神經網路[7]，其把上個時間點的輸出值存下來，並重新導入到輸入端，以便在時間上能夠抓取大長度的輸入訊號，使系統能有大量的歷史資料去學習電視節目回聲的路徑。此外我們並考慮空間資訊，改用多支麥克風組成的多通道系統[8]，以強化語音輸入訊號。因此，基於以上兩點考量，在本論文中我們將提出多通道深層遞迴類神經網路電視回聲消除系統，在以下章節將會再詳細介紹其架構與訓練方法。

二、相關研究

常用的回聲消除方法包括 NLMS、RBF 即 MLP，其中最小均方演算法(LMS)簡易實現、計算量小及穩定特色，受到不少人青睞且廣泛地運用，同時為了解決系統收斂緩慢的缺點，在使用上將輸入訊號的能量進行正規劃處理，也就是 NLMS 演算法，其採用可變步長的方法來穩定收斂過程。LMS 與 NLMS 的基本原理是計算輸入信號與參考信號的關聯性，下圖 2 為 NLMS 自適應濾波器架構圖。

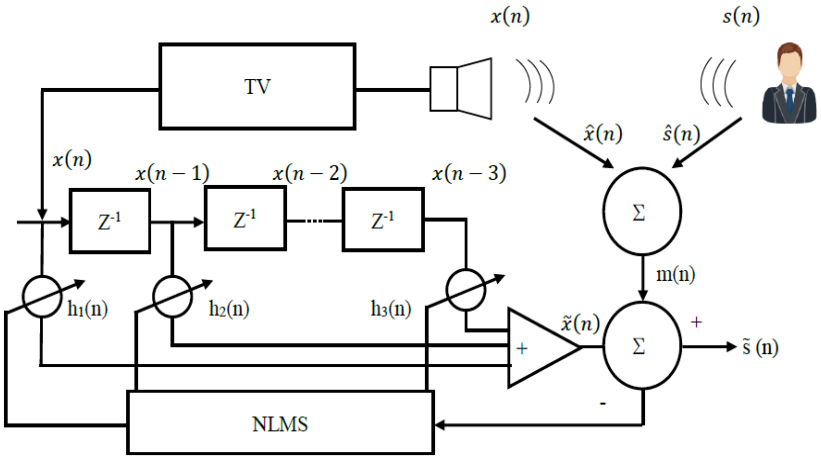

圖 2　NLMS 自適應濾波器架構圖

另外，輻射基底函數類神經網路(RBFNN)具有逼近任意的非線性函數的能力，而且具備一般化能力，對未知資料能有效地處理，再加上快速的學習收斂速度及低計算複雜度，已成功廣泛應用在非線性回聲消除，下圖 3 為 RBF 類神經網路用於回聲消除的架構圖。

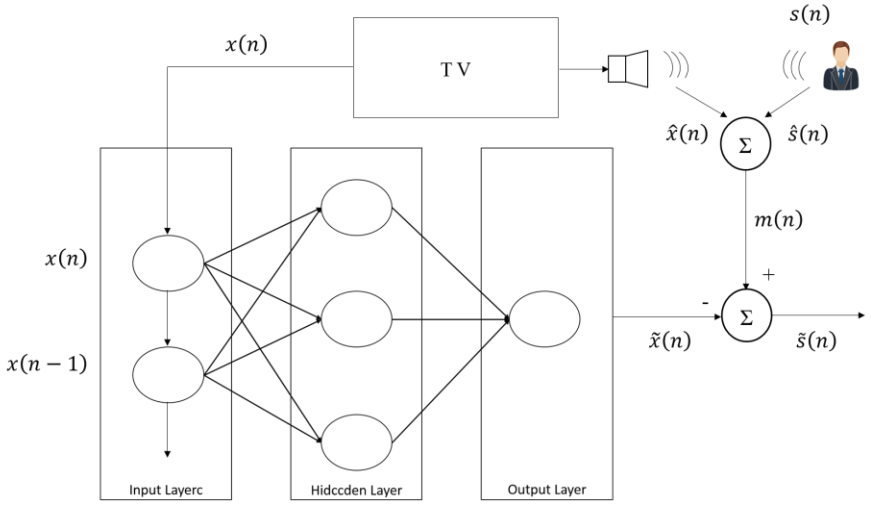

圖 3 RBF 神經網路架構圖

RBF 網路架構只由一層的隱藏層、輸入層及輸出層共三層所構成，其核心為高斯核函數所組織，訓練過程可看作在高維空間中找尋最佳逼近參考數據的解，主要由第一層輸入層的感知單元將神經網路與外界相聯接收訊息，並直接傳遞到隱藏層；接著第二層隱藏層則是藉著對輸入向量空間到隱藏空間之間的非線性映射變換，而第三層輸出層經由線性組合加權變成輸出。

而多層感知機類神經網路含有兩層以上的隱藏層，其中用到的激活函數為 Sigmoid 激活函數，核心演算法為反向傳播演算法，用於回聲消除的網路架構如下圖 4 所示:

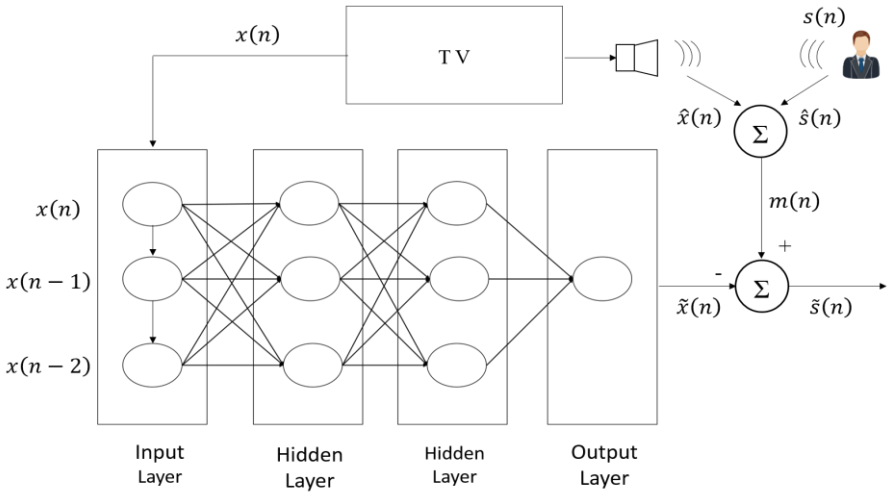

圖 4 三層多層感知機網路架構圖

主要網路架構由輸入層、多層隱藏層及輸出層組合而成，網路運作為輸入層把資訊傳送到隱藏層，再由最後的隱藏層傳輸到輸出層，加權總和出整個網路的輸出值，此時為前饋網路，而且權重值及偏壓值是固定值，其中層與層之間溝通是神經元透過權重及偏壓值相互的連結，得到整個網路的輸出值後，對比目標期望的輸出值算出誤差值，透過反向傳播演算法把誤差倒傳送回神經網路中，去做權重及偏壓值調整，此時為倒傳遞網路，得到整個網路輸出層的值後，與目標期望的輸出值相減獲得誤差值，再利用其誤差值所以定義出來的成本函數，透過反向傳播演算法來對其網路權重及偏壓值作更新，然後再代回網路求新誤差值，使誤差平方趨近極小值，讓網路輸出逼近於目標期望值。最後，由以上討論可知，NLMS，RBF 與 MLP 能看到的歷史訊號受到它的濾波器長度，或輸入神經元數目限制，通常是有限的且無法太長。

三、基於深層遞迴類神經網路之回音消除系統

因為電視節目聲會在室內反射造成殘響，而且殘響時間長度常常大於 0.5 秒，而傳統消除回聲系統都受限於輸入長度固定，如果輸入設定太長，會導致運算量過大，收斂太慢，無法有效增強使用者語音訊息，故我們改用深層遞迴類神經網路[8]，因其可以經由回授線路看到很久以前的資料以學習長時間的回聲路徑，並猜出下一時間點的聲音，而單

層的遞迴類神經網路就足以獲取的長時間資訊，因此多層遞回類神經網路就能看得更廣，抓取更大的資料。以下先以單聲道回聲消除的做法，然後在說明多聲道迴音消除的做法。

(一) 單聲道深層遞迴類神經網路回音消除

下圖 5 其運作原理為運用麥克風所收集到的語音 m(n)與透過電視音源線輸出取的電視節目聲 x(n)，利用深層遞迴式神經網路預測可能錄到的電視回聲 x̃(n)，相減後將回聲消除，此時系統的誤差訊號 s̃(n)為輸出得到的清晰語者聲音。

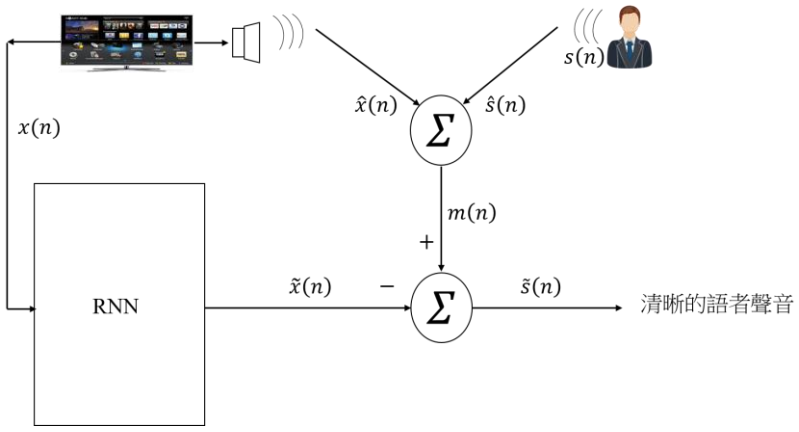

圖 5 單聲道深層遞迴類神經網路系統架構

其中本論文提出深層遞迴類神經網路(RNN)架構[6]除了與一般類神經網路一樣有輸入層、隱藏層以及輸出層外，還會多出兩個或兩個以上的遞迴層當回授功能，其運作原理主要透過回授線路，把上個時間點的隱藏層神經元輸出值記錄下來，並重新導入到隱藏層神經元輸入端；與輸入層輸入值整合在一起，當下一個時間點的隱藏層輸入，以此讓神經元下個時間輸入值與過去輸出值有關，簡單來說讓整個神經網路是有記憶性的，也可以說是把輸入層給放大了，其核心參數更新演算法還是反向傳播演算法，但不一樣的地方在於需要根據時間先後順序來做權重調整，所以權重調整會透過不同時間點的隱藏層訊息進行，先由最後時間點開始對於成本函數作偏微分，往前算出到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，使誤差平方趨近極小值，讓網路輸出逼近於目標期望值。下圖 6 為本論文中使用的深層 RNN 運作構造。

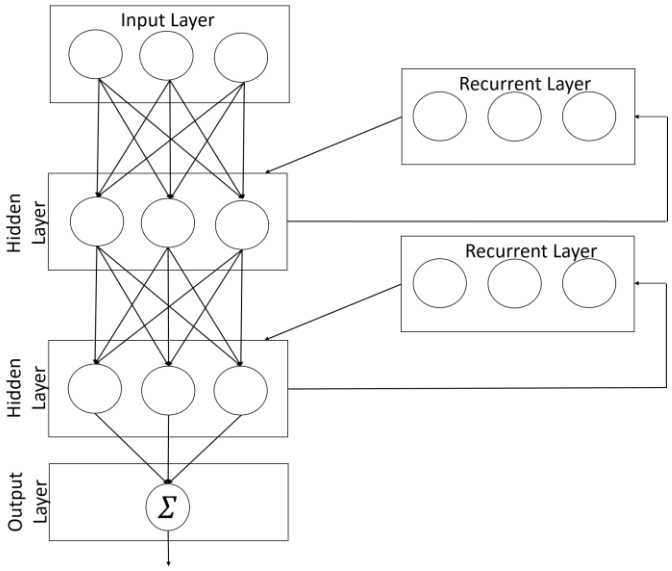

圖 6 深層遞迴網路架構圖

(二)多聲道深層遞迴類神經網路回音消除

為了提高收音品質以及降低收音角位的影響，我們改進圖 5，加上多隻麥克風放在不同位置，構成麥克風陣列，建立如圖 7 的多通道深層遞迴類神經網路之電視回聲消除系統架構，其運作原理為運用 Kinect for Xbox one 麥克風 4 通道收集到的語音 $m_i(n)$，再透過電視線輸出取得電視節目聲 x(n)，分別去做四次深層遞迴式神經網路，將回聲消除系統的誤差訊號加權平均以加強語音訊號。

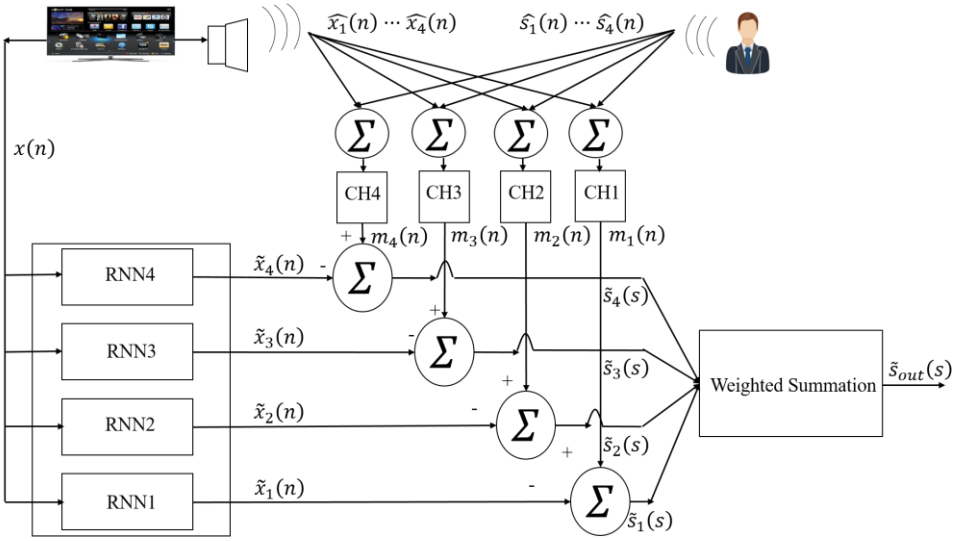

圖 7 多通道遞迴類神經網路電視回聲消除系統架構

其中 x(n)原始電視節目聲，並把它當成聲學回聲系統的參考信號，同時也是四個深層類神經網路的輸入。i 為麥克風聲道數，i=1,2,3,4。$m_i(n)$為 Kinect for Xbox one 麥克風陣列所錄製的電視節目聲及使用者說話聲的混合聲音。$\hat{x}_i(n)$為電視節目聲的回聲，i=1,2,3,4。s(n)為使用者說話聲，$\hat{s}_i(n)$為說話的回聲。$\tilde{x}_i(n)$為逼近 $\hat{x}(n)$的估計值，即為回聲抵消預測信號。$\tilde{s}_i(n)$為 $m_i(n)- \tilde{x}_i(n)$的誤差信號；即我們想要獲得的使用者說話聲。$\hat{s}_{out}(n)$為四個 RNN 回聲消除系統輸出的加權平均誤差信號。

四、實驗結果

本論文的實驗語料包含 8 個語者 TCC300 測試人聲以及 4 類 40 個電視節目聲，測試 NLMS、RBF、MLP 以及 RNN 等四種適應性濾波器演算法。其中輸入音框為 2048 的取樣點，RNN 演算法設定為兩層隱藏層，每層神經元為 100。在以下實驗中先測試(A)全部方法在單通道回聲消除的效果，取其中最好的方法在(B)測試多通道的情形。

(A)在單通道回音消除:

將進行三個實驗，實驗一單純先考慮只有電視節目音，而實驗二加入人聲混合電視節目聲。由於實驗二有可能會學習到人聲，除了消除電視節目回聲外，可能也把人聲給濾掉了，導致人聲訊號變形，因此再加入實驗三，實驗三為前五秒單純電視節目聲預訓練後五秒人聲混電視節目聲回聲消除模式，利用前五秒只有電視節目聲，可以預先學習環境響應，這樣不會學到使用者的人聲，而後五秒則把學習率調低，降低濾除人聲的可能性，達到消除電視回聲並增強使用者的語音訊號。在所有實驗中，我們透過回聲返回損失強化作為回聲消除後的評估，來比較四種演算法的好壞，並選出最好的演算法。

(B) 多通道回音消除實驗中:

依據之前實驗三的設定，進行多通道回音消除實驗，以在單通道表現最好的 RNN 依後面的實驗結果，製作多通道回聲消除，並與單通道 RNN 的回聲消除實驗比較。

(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，且為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 個音檔，從中隨機擷取十秒鐘片段節目聲，且每一類有 10 個音檔；所以共有 40 個測試背景電視節目聲。由於 TCC300 的語料檔案格式為.pcm 檔與所下載的電視節目聲音壓縮格式為 MP4 檔，加上兩者語音內容長短差異過大，導致無法直接作程式輸入測試的音檔，所以先以音頻處理工具作格式及時間上的整理。以下表 1 為語料格式整理。

表 1 語料格式設定整理

	測試人聲 (speech)	電視節目聲
聲音壓縮格式	.wav	.wav
取樣率	16K Hz	44.1K Hz
樣本大小	16 bit	16 bit
音檔長度	串音後原始長度	每個音檔取 20 分鐘

(二) 語料錄音實驗情境假設

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，並在 Kinect for Xbox one 位置左右兩旁平行放上兩顆主動式監聽喇叭，當成電視回聲背景音的來源，且在 Kinect for Xbox one 的正前方距離 2m 處也擺上主動式監聽喇叭播放出人聲，模擬使用者正在講話，如此一來；可以想像出，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，這樣可以用來當我們聲學回聲消除系統的語料了。實際上我們共錄製說話者在 90，60 與 30 度角位置的人聲，但目前本論文實驗回聲消除只有拿 90 度角的 TCC300 測試人聲錄音語料來使用，暫不考慮其他角度，其佈局擺設如下圖 8 及 9。

圖 8 90 度角擺設圖

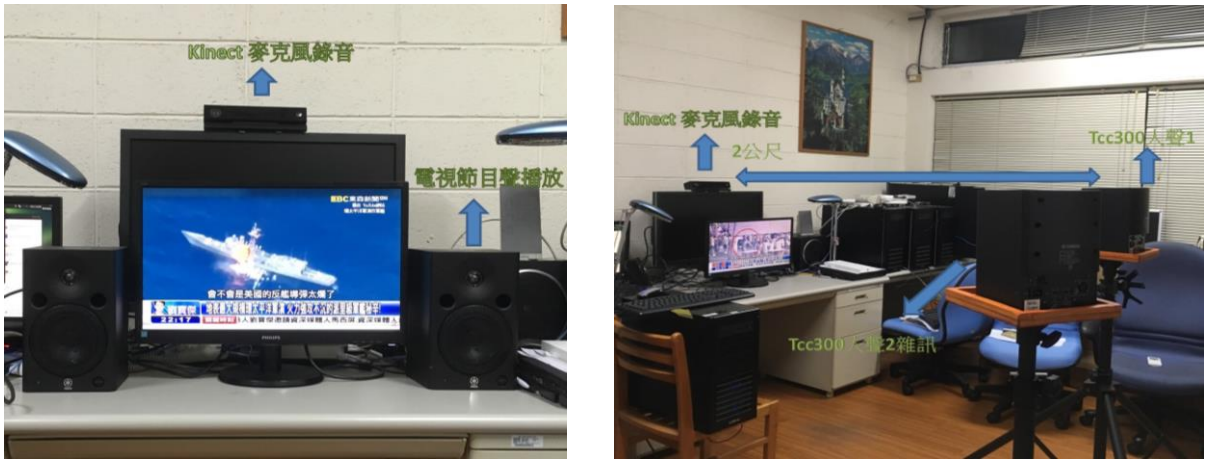

圖 9 實際電視喇叭、收音麥克風與實際模擬說話者之音源喇叭擺設

(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)[8]數值化在時域上的差異，作為回聲消除後的評估其方程式如下所示:

$$ERLE=10\log_{10}\frac{m^2(n)}{s^2(n)} \quad (1)$$

其中，m(n)為 TCC300 測試人聲混電視節目聲的錄音訊號聲。s̃(n)為誤差訊號；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，測試人聲與電視節目聲混合的比例皆保持一樣大聲，即模擬 SNR = 0dB 的情形。以下進行兩大類實驗，包刮(A)單聲道(B)多通道電視回聲消除實驗。

(A)單通道電視回聲消除實驗

1. 實驗一，純電視節目聲消除實驗:

演算法對電視節目聲每一類的音檔做回升消除，並做回聲返回損失強化性能比較評估，用來了解演算法在不同類別的電視節目聲表現如何，然後看四類性能比較評估做總平均觀察其對四類電視節目聲整體表現。 如下圖 10 所示。

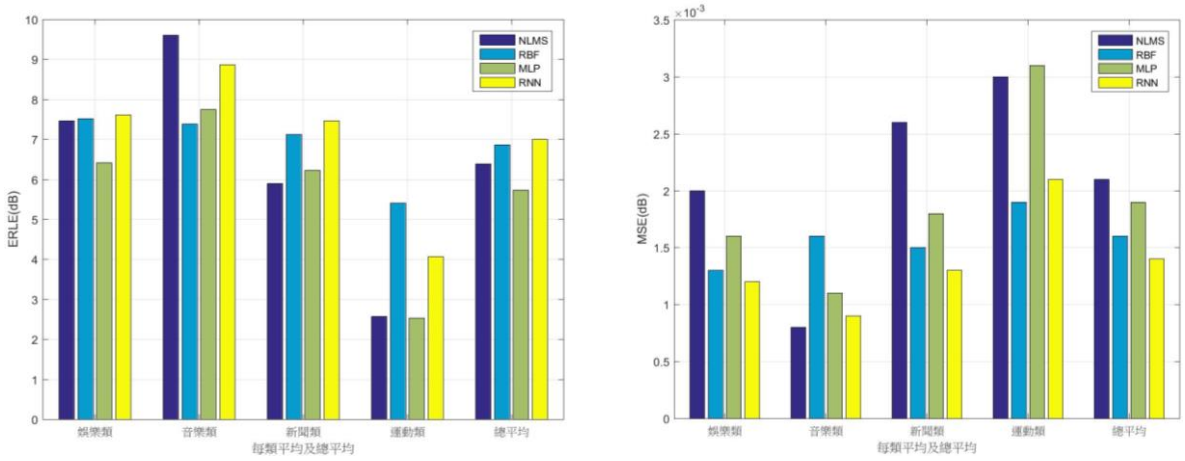

圖 10 實驗一 所有電視節目類型以 NLMS，RBF，MLP 與 RNN 回聲消除的 ERLE 與 MSE 結果圖

在實驗一純電視節目聲消除實驗上，整體來說音樂類最容易處理，新聞類最難處理。此外，RNN 在大部分情況下的表現都相當不錯，所以其總平均分數優於所有方法。

2. 實驗二，人聲混電視節目聲實驗:

以下是八個不同的測試使用者混電視節目聲的回聲消除實驗，可藉此觀測不同測試使用者在每一類電視節目聲的演算法平均效能，最後再看八個人的總平均，結果如下圖 11 所示，當電視回聲混合人聲時，所有方法都會變差，其中以 RBF 受到的影響最小，RNN 次之，且兩者差距不大，但效果都遠比 NLMS 與 MLP 要好。

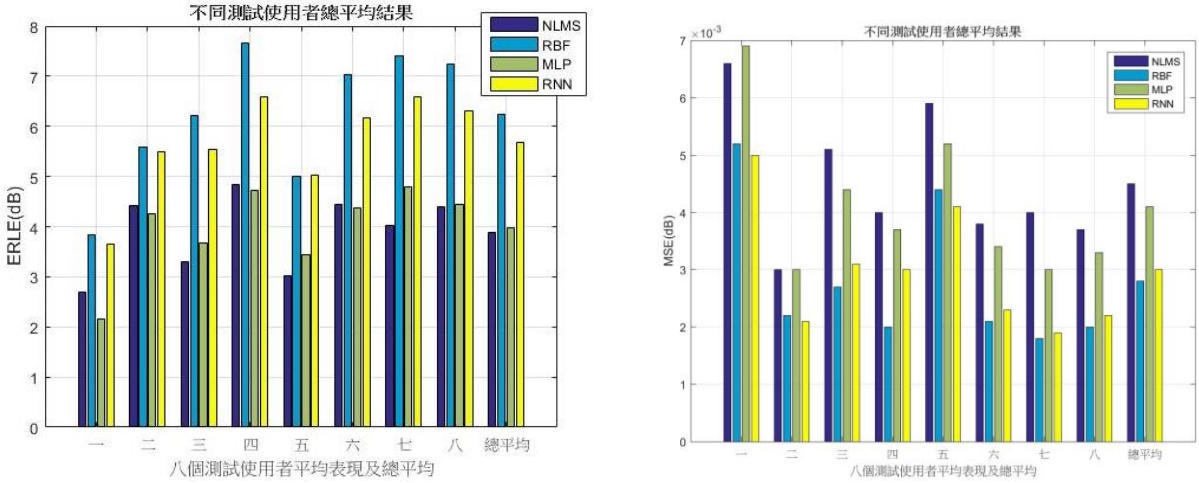

圖 11 實驗二 所有人聲混電視節目聲以 NLMS，RBF，MLP 與 RNN 回聲消除的 ERLE 與 MSE 結果圖

3. 實驗三，前 5 秒純電視節目聲預訓練加後五秒混人聲與電視節目回聲消除實驗:

結果如下圖 12 為前五秒單純背電視節目回聲，後五秒混有不同的測試使用者的聲音，測試演算法平均效能以及八個人的總平均。由圖 12 明顯得知此時 NLMS 表現最差，單聲道深層遞迴類神經網路回聲消除系統在總平均要好於 RBF 及 MLP。

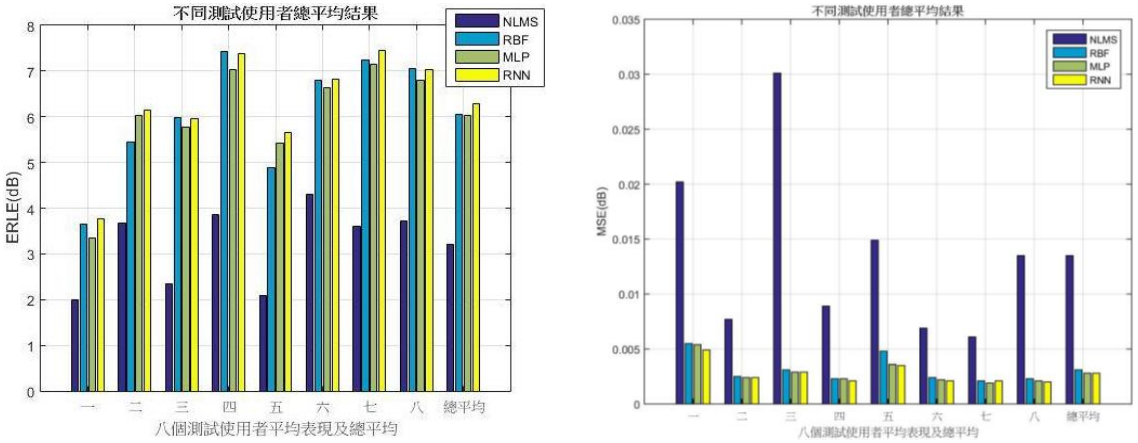

圖 12 實驗三 前五秒單純電視節目聲預訓練，後五秒混和人聲電視回聲消除模式，對所有不同測試使用者的回聲消除 ERLE 與 MSE 結果圖

(B)多通道電視回聲消除實驗

1. 實驗一，單通道 RNN 及多通道 RNN 的回聲消除實驗

結果如下圖 13 為前五秒單純背電視節目回聲，後五秒混有不同的測試使用者的聲音，測試單通道 RNN 及多通道 RNN 演算法平均效能以及八個人的總平均。由圖 13 得知，多通道深層遞迴類神經網路回聲消除系統相當好，且更優於單聲道深層遞迴類神經網路回聲消除系統。

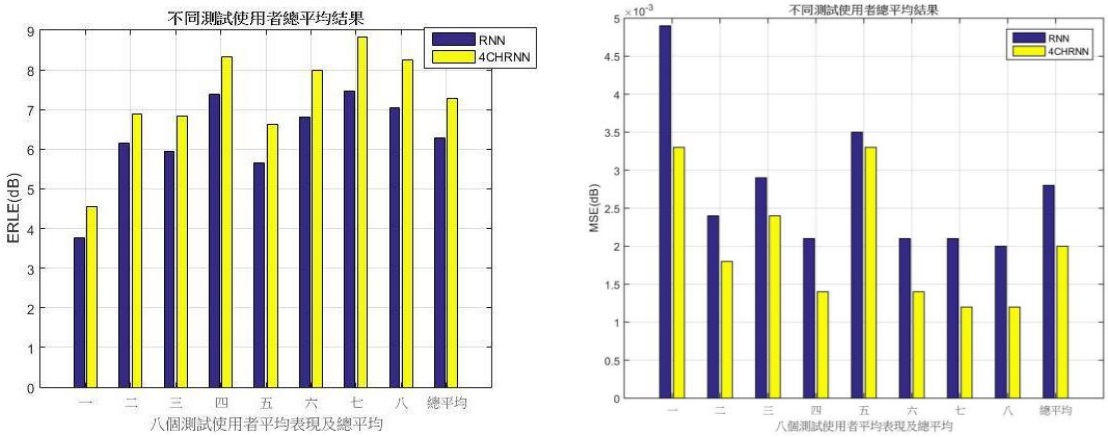

圖 13 實驗一 比較單通道 RNN 與多通道 RNN 電視回聲消除模式，對所有不同測試使用者的回聲消除 ERLE 與 MSE 結果圖

經由這些模擬實驗結果可以發現，在(A)單通道實驗一純電視節目聲消除上，對於某些種類的節目而言，有些演算法消除效能表現比較突出，例如：NLMS 在音樂類、RBF 在運動類等；也顯示了非線性濾波演算法對於比較複雜的環境具有較佳的消除回聲能力。而在實驗二人聲混電視節目聲消除上，可以發現其人聲會溢入到適應性濾波器中，所以學習出來的電視背景聲有些許的人聲，這會影響消除性能，但整體來說非線性濾波演算法對於混聲消除效能表現比較佳。最後在實驗三透過先在前五秒純電視節目聲時學習背景回聲路徑，接著後五秒人聲混電視節目聲直接用訓練後的系統作回聲消除，藉此來降低人聲溢入適應性濾波器的影響，結果顯示了此方法對於非線性濾波演算法有明顯的提升。而在(B)多通道實驗中加入多通道深層遞迴類神經網路，因其能錄取不同方位的聲音資訊，比單通道系統得到更多的資料，也就能更有效的學習電視節目回聲路徑，由實驗結果可以知道，多通道深層遞迴類神經網路回聲消除系統更好於單聲道深層遞迴類神經網路回聲消除系統。

五、結論

在本實驗中，利用了監聽式喇叭及 Kinect 等器材，實際錄音模擬智慧型電視操作情境下的回聲，作為電視節目回聲消除實驗的語料，接著分別先後導入了線性濾波算 NLMS 以及非線性濾波演算法如 RBF、MLP、RNN 等，作電視節目回聲消除實驗。

經由實驗結果可知，在大部分情況下，RNN 的表現相當穩定，尤其是多通道深層遞回類神經網路回聲消除的效果可以達到最佳。本論文實驗結果，可以為日後回聲消除研究提供參考，相信未來仍有許多可以改進的空間，例如：深層類神經網路增加層數以及每一層網路神經元數的調整等等，還可以多模擬一些情況，來了解對於回聲消除系統有什麼影響，像角度回聲的影響以及多人使用者情況下等等，這些都是日後可以加以考慮的因素。

致謝

本研究感謝教育部『大學以社教機構為基地之數位人文計畫』（A36 號）與科技部專題計畫（MOST 104-2221-E-027-079, 105-2221-E-027-119 and 103-2218-E-027-006-MY3）支持。

參考文獻

[1] 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,2007.

[2] 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.

[3] Paulo S. R. Diniz, Adaptive Filtering Algorithms and Practical Implementation 4th,New York：Springer Verlag,2012,pp.469-477.

[4] 張晉維,主動式噪音控制耳機之設計與實現,碩士論文,國立台灣科技大學電子工程研究所,2008.

[5] 諶愛文,基於 BP 和 RBF 神經網路的數據預測方法,碩士論文,中南大學,007.

[6] Liu Yong and Zhang Liyi, Implementation of BP and RBF neural network and their performance comparison, Master Thesis, University of Technology, 2007.

[7] 郭奕志,基於背景音消除法之智慧型電視的聲控技術研究,碩士論文,國立台北科技大學,2013.

[8] 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.