元智大學

電機工程學系

畢業專題論文

基於位置資訊在分散式麥克風中之語音強化

Distributed Microphone Speech Enhancement Based On Location Information.

專題生: 謝維　　林伯翰

指導老師 : 方士豪 教授

摘要

因應車載語音助理、智慧家居的需求，語音提取技術的提升在其中勢必扮演不可或缺的腳色。考慮到多通道麥克風能夠提供更多的時空間資訊，因此本專題將使用分散式麥克風系統進行除噪。此外由於深度學習的快速發展，在聲音訊號處理也有突出的表現，因此本專題選用深度降噪自動編碼器(Deep Denoising Autoencoder,DDAE)做為神經網路的基礎架構開發出位置相關深度降噪自動編碼器(Location Associated Deep Denoising Autoencoder,LA-DDAE)。LA-DDAE與DDAE的差異在於輸入層除了聲音訊號之外，也會加入位置資訊進行綜合訓練，其目的為充分利用多通道系統提供的時空間資訊，藉此強化神經網路架構。

本專題所提出的LA-DDAE語音降噪系統在加入特定的位置資訊共同訓練後，能夠在頻域上有效消除語音訊號中的噪聲。在實驗上，我們使用台灣地區噪音下華語語句聽辨測驗（Taiwan Mandarin hearing in noise test, TMHINT）語料庫。實驗數據顯示，在模擬狀態下LA3-DDAE與LA15-DDAE能有效地提升語音品質，在真實場域中LA3-DDAE在語音品質方面也有較佳的表現。

目錄

圖目錄

表目錄

第一章、緒論

1.1 研究動機與目的

科技日新月異的進步，語音提取技術對人類的生活扮演至關重要的腳色，舉凡車載語音智能助理、智能家居等都需要使用到該項技術。因此提升免持設備的語音提取能力能夠大幅降低人與機器互動的阻礙，並且增加事務處理的效率。

但在實際的環境中是充滿噪聲的，例如在車用系統上，收音設備會收錄到引擎聲、街道噪音等噪音。又如家庭使用智慧家電，收音設備也會收錄到空間回聲、冷氣馬達聲、街道傳出的警笛聲等。這些干擾訊號與主要目標聲源同時被收錄進麥克風中，導致語音品質(quality)與語音可理解度(intelligibility)被大幅降低，更進一步造成人與機器的互動效率下降。語音增強的目的在於將受損的語音訊號進行降噪與還原，藉此提升語音品質與語音可理解度。

語音增強系統可分為單麥克風技術與分散式麥克風技術。單麥克風除噪技術的硬體成本相對較低，但因為沒有空間資訊的輔助除噪的效果有限，因此本專題選用分散式麥克風環境進行降噪。另外由於深度學習的快速發展，其成果在生醫、影像、音訊等訊號處理皆有突出的表現，因此本研究選用了深度降噪自動編碼器(Deep Denoising Autoencoder,DDAE)進行降噪。有別於以往的文獻，我們在神經網路的輸入端除了將音訊資料輸入，也會把各種位置資訊加入神經網路中綜合訓練，將分散式麥克風的優點最大化。不過輸入端從單麥克風技術開發成分散式麥克風技術並且加入位置資訊會使輸入層大小倍數增大，同時意味著神經網路模型會變得更複雜，耗費的資源也會更多，因此在輸入端要加入何種位置資訊，才可以讓模型效能增加的同時盡可能地減少資源消耗，會是本專題的挑戰與探討要點。

1.2 文獻回顧

語音強化的麥克風配置大致可以分為，麥克風陣列、單通道麥克風、分散式麥克風，考慮到車載語音智能助理、智能家具等應用，我們的實驗配置是模仿論文[2][3]中，使用分散式麥克風進行實驗的配置。

DDAE 為一種深度神經網路，在 2013 年首次被提出，與許多模型架構相同，DDAE 同樣需要給他一組”label”或稱”target”作為他訓練用的答案，其使用的”target”為語者的乾淨語音，藉由乾淨語音作為帶噪語音還原的目標，訓練神經層中的每一個參數，而研究指出 DDAE 相較於傳統的維那濾波器、MMSE 等方法有更好的語音降噪效果，因此我們選用 DDAE 作為我們模型的基礎架構。

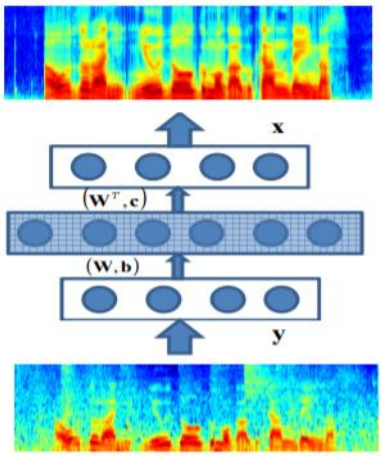


圖 1.1 DDAE 模型架構圖[1]

在模型架構設計上，主要受到兩篇論文的啟發，在論文[4]中，作者將靜態的病例資料轉為參數，送入模型進行訓練，藉此提升模型預測的準確度。而在論文[5]中，作者則是將人的頻率當作輸入的維度，送入 DDAE 模型第三層，提供模型訓練，設計出了 SaDAE。而 SaDAE 相較於 DDAE，在語音品質有更好的表現，約提升 7%。因此，我們將位置資訊轉為輸入的數值提供模型進行訓練，設計出 LA-DDAE(Location Associated Deep Denoising Autoencoder)，希望藉由位置資訊的加入，讓分散式麥克風中訊噪比較高的麥克風通道獲得較高的權重，提升模效能。

第二章、神經網路架構

2.1 深度降噪自動編碼器(Deep Denoising Autoencoder,DDAE)

深度降噪自動編碼器為深度神經網路(Deep Neural Networks, DNN)的一種變形，其目的為針對聲學領域設計的模型，本專題使用的深度神經網路皆參考深度降噪自動編碼器的架構，以下將介紹深度降噪自動編碼器的架構。

深度降噪自動編碼器是基於自動編碼器（Autoencoder, AE）的神經網路技巧，自動編碼器為輸入輸出相同的神經網路模型，但位於最中間層的隱藏層大小，通常較模型輸入輸出的維度小，以達到編碼壓縮的目的。而深度降噪自動編碼器是一種自動編碼器架構下，以降低噪聲為目的的深度神經網路模型，因此隱藏層會以最佳模型參數為目標設定隱藏層大小。與自動編碼器不同，降噪自動編碼器的輸入為帶噪語音資料，輸出為乾淨語音，換言之，深度降噪自動編碼器在自動編碼器的架構中加入了雜訊成份。由於模型架構與神經網路相似，增加了編解碼的結構，在訓練上，我們仍先準備帶噪、乾淨語音訓練資料集。

將$\hat{x}_i$定義為模型預測結果，y_i定義為帶噪語音，W定義為模型權重，b、c定義為偏差(bias)，σ定義為激活函數(activation function)，則 DDAE 模型可以表示成下式:

$$h(y_i) = \sigma(W_i y_i + b) \tag{2.1}$$

$$\hat{x}_i = W_{output} h(y_i) + c \tag{2.2}$$

其中式(3.1)表示輸入層與隱藏層還有隱藏層與隱藏層之間的關係，式(3.2) 表示最後一個隱藏層與輸出層的關係。而本專題實驗中採用的激活函數為 relu，因此可以表示為:

$$\sigma = f(z) = \max(0, z) \tag{2.3}$$

最後當我們在訓練模型時，損失函數(loss funtion)為判斷 DDAE 模型是否在每一個期(epoch)都有持續增強的重要指標，因此我們在每一個期執行完畢後都會計算損失函數，而損失函數選用的是均方誤差（Mean Square Error, MSE），將x_i定義乾淨語音，則均方誤差可以表示成:

$$L = \sum_i ||x_i - \hat{x}_i||_2^2 \tag{2.4}$$

在過去的實驗中指出深度降噪自動編碼器在降噪方面有顯著的效果，在輸入神經層加入位置資訊後也能強化模型的效能，使得在輸出層的訊號有更好的品質。

2.2 位置相關深度降噪自動編碼器(Location Associated Deep Denoising Autoencoder,LA-DDAE)

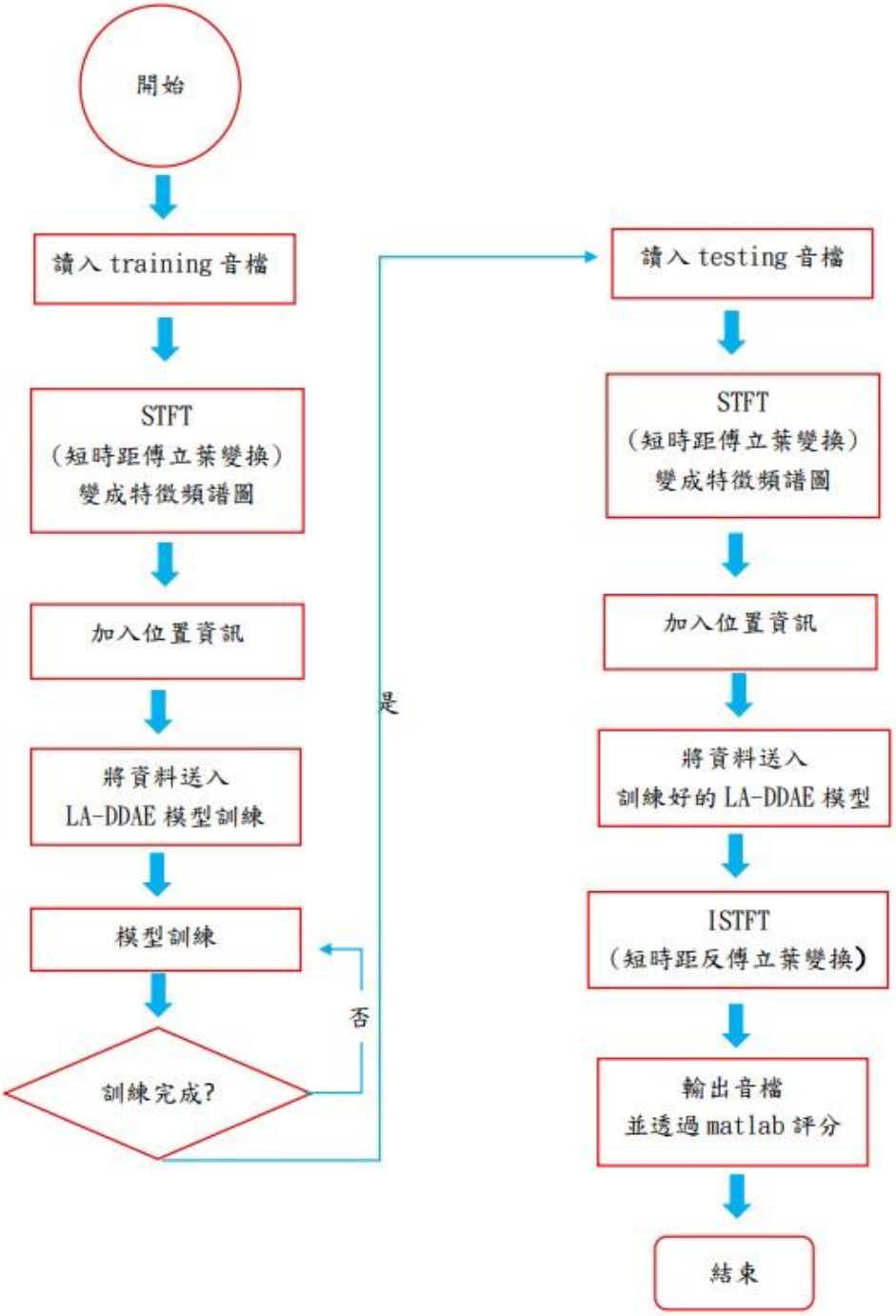


圖 2.1 LA-DDAE 流程圖

如圖 2.1 所示，LA-DDAE 一開始會讀入帶噪、乾淨的訓練資料集並且對其進行短時距傅立葉變換(Short Time Fourier Transform, STFT)藉此獲得時頻資

訊矩陣，接著我們將位置資訊加入時頻資訊矩陣中，對於測試資料集我們也進行同樣的動作。接著將經過處理的訓練集矩陣做為輸入送進 LA-DDAE 模型訓練，訓練一個期(epoch)後，再將測試集矩陣做為輸入送進 LA-DDAE 模型，接著我們會使用損失函數計算出 LA-DDAE 模型對訓練、測試資料集的準確程度，藉此判斷模型效能是否在每一個期都持續上升。之後重複上述動作直到程式完成執行指定的期。接著我們將帶噪測試集矩陣送入 LA-DDAE 模型並且對其結果做短時距反傅立葉變換，藉此得到降噪過後的音檔，最後我們將音檔評分並判斷 LA-DDAE 模型效能。

時頻資訊矩陣在程式中的表示法為特徵值*音框長，所以我們將實驗中得到的位置資訊當成新的特徵值加入，因此在加入位置資訊後的時頻資訊矩陣可以表示成(特徵值+位置資訊維度)*音框長。實驗中加入的位置資訊維度總共有 4 種，分別為 3 維度、6 維度、12 維度與 15 維度，3 維度是加入聲源座標，6 維度是加入聲源座標與聲源位置和通道位置的歐氏距離，12 維度是加入聲源座標與通道座標，15 維度是加入聲源座標、通道座標、聲源位置和通道位置的歐氏距離。為了清楚辨別加入的是何種位置資訊，我們將加入的位置資訊維度放入模型名稱中，舉例來說，若是加入的位置資訊是 3 維度，那麼模型名稱會是 LA3-DDAE。

第三章、實驗配置與資料庫

3.1 實驗配置

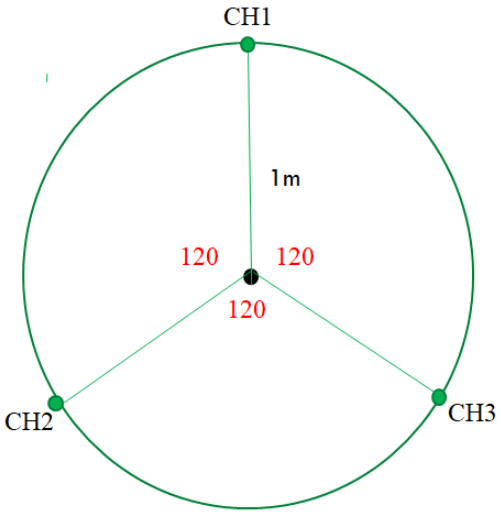


圖 3.1 分散式麥克風配置

本專題使用分散式麥克風系統進行資料收集，其配置如圖 3.1 所示，實驗場域為一個半徑一公尺，並且每隔 120 度設置一個通道麥克風，也就是圖中的 CH1、CH2、CH3，除此之外我們額外使用一隻手持麥克風錄製無雜訊的音訊。在錄製的時候我們在實驗室設置一個安靜、密閉的空間，並且按照上述設置架設三個通道麥克風。本專題總共進行四次實驗，實驗一、實驗二的目的為使用極端位置與少量資料測試 LA-DDAE 是否有效，實驗三、實驗四則是使用隨機位置與大量資料驗證 LA-DDAE 確實可行。

實驗一與實驗二使用同一組資料集進行訓練，資料集總共錄製 100 句音檔，聲源位置分別在 CH1 旁錄製 34 句，CH2 旁、CH3 旁錄製 33 句。由於聲源位置與通道麥克風近乎重疊，因此在加入位置資訊時，聲源位置與對應到的通道麥克風將視為同一位置。實驗一、實驗二不同的是我們將訓練與測試資料集的內容做改變，實驗一訓練資料集從三個聲源位置各取出 30 個音檔，並且對其混入 8 種不同的噪音與 8 種不同的訊雜比(Signal-to-noise Ratio)，因此我們總共獲得 90*8*8 個訓練資料，而測試資料集我們將剩下的 10 個音檔混入 4 種不同的噪音與 3 種不同的訊雜比，因此我們總共獲得 10*4*3 個測試資料，實驗一配置的目的為，檢驗加入最極端的位置資訊是否能幫助 LA-DDAE 模型訓練。實驗二訓練資料集

選用在 CH1、CH2 旁錄製的資料混入 8 種不同的噪音與 8 種不同的訊雜比，因此我們總共獲得 67*8*8 個訓練資料，而測試資料集將選用在 CH3 旁錄製的資料混入 4 種不同的噪音與 3 種不同的訊雜比，因此我們總共獲得 33*4*3 個測試資料。實驗二配置的目的為，檢驗 LA-DDAE 模型在遇到訓練資料集中沒有出現的位置資訊依然能夠有較佳的語音增強效果。

實驗三的通道資料皆為模擬出來的音檔，因此在實驗初步我們在實驗室設置一個安靜、密閉的空間錄製 320 句毫無噪音的音檔，之後使用 matlab 模擬在真實場域得到的音檔。模擬的過程中我們隨機在圓內產出 320 個位置，使得每一句音檔都能夠有不同的聲源位置，藉此模擬出人在空間中到處走動的情形。訓練資料集使用 300 句音檔混入 8 種不同的噪音與 8 種不同的訊雜比，因此我們總共獲得 300*8*8 個訓練資料，測試資料集將選用 20 句音檔混入 4 種不同的噪音與 6 種不同的訊雜比，因此我們總共獲得 20*4*6 個測試資料。實驗三配置的目的為，檢驗在模擬環境下加入不重複且隨機的位置資訊對 LA-DDAE 模型訓練是否有幫助。

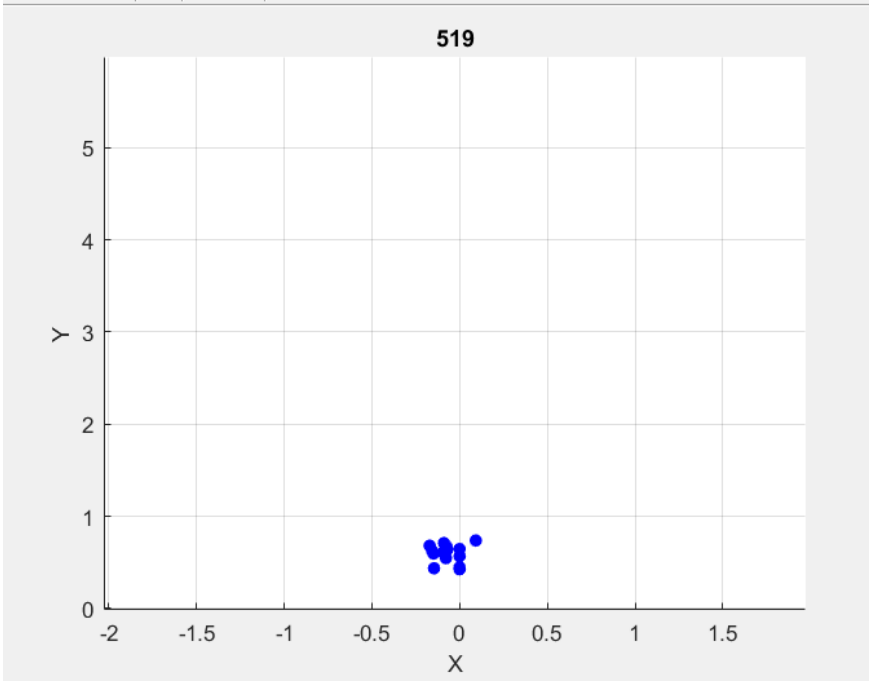


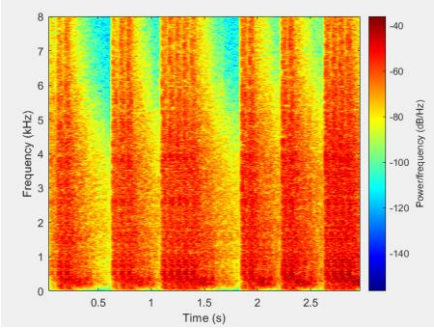
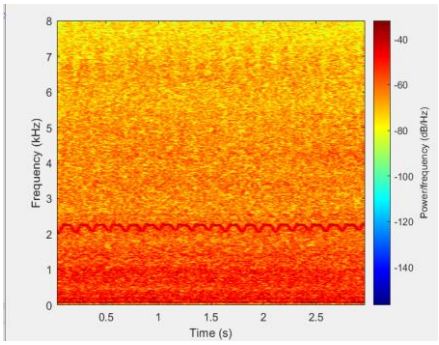
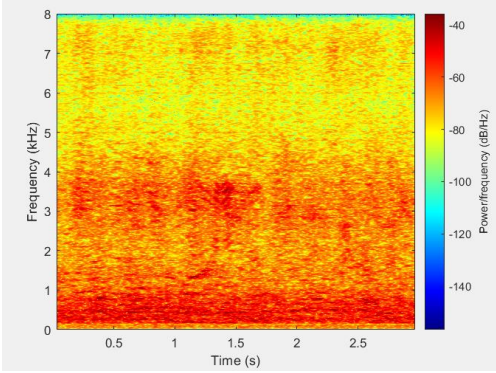
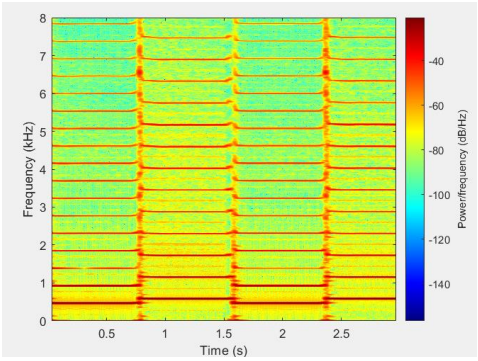
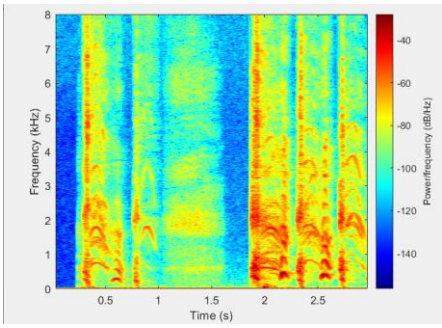
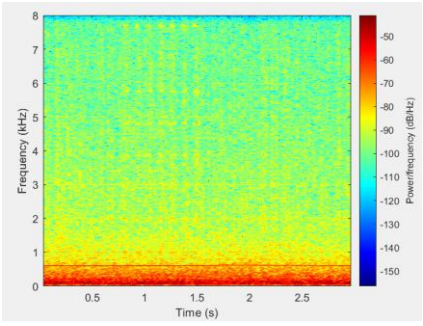
圖 3.2 毫米波雷達點雲座標圖

實驗四在真實的實驗場域中進行實驗，與實驗三相同，我們藉由在空間中隨機站在 320 個位置錄製資料，模擬出人在空間中到處走動的情形。訓練資料集使

用 300 句音檔混入 8 種不同的噪音與 8 種不同的訊雜比，因此我們總共獲得 300*8*8 個訓練資料，測試資料集將選用 20 句音檔混入 4 種不同的噪音與 6 種不同的訊雜比，因此我們總共獲得 20*4*6 個測試資料。此外在實驗四中我們在天花板架設毫米波雷達進行室內定位，如圖 3.2 所示，我們將點雲座標算出並且計算其平均藉此獲得聲源位置資訊。實驗四配置的目的為，檢驗在真實場域中加入不重複且隨機的位置資訊對 DDAE 模型訓練是否有幫助，實驗四與實驗三的最大差異在於錄製的資料有摺積性噪音。

3.2 資料庫建構

我們使用台灣地區噪音下漢語語音聽辨測試（Taiwan Mandarin Chinese version of Hearing in Noise Test, TMHINT)錄製實驗中所需的原始音檔，作為錄製者要說的句子內容，一句話有十個字，約在 3 秒鐘左右可以唸完。由於使用完整的資料集訓練需要數十個小時的時間才能將模型訓練完成，因此在實驗一與實驗二中，我們使用 100 句作為資料集訓練模型，在驗證想法有可行性後，實驗三與實驗四中，我們使用完整的 320 句作為資料集，進行模型的訓練與結果的驗證。

機槍噪聲	駕駛艙噪聲
	
自助餐內部吵雜聲	救護車噪聲
	
咳嗽聲	電腦風扇噪聲
	

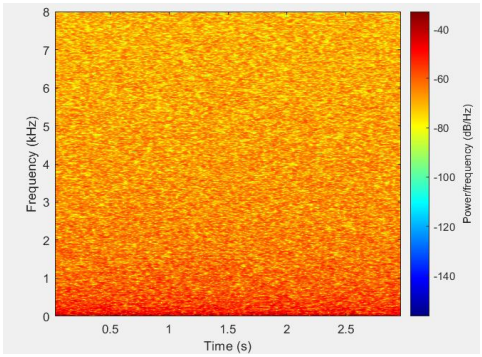
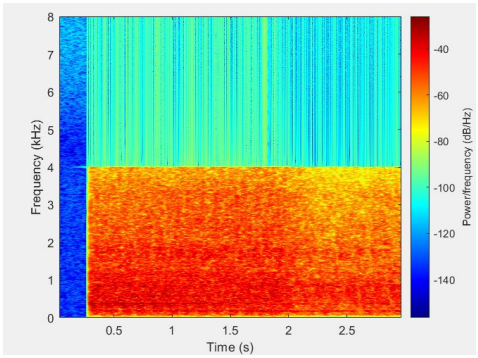
粉紅噪聲	發動機噪聲
	

圖 3.3 訓練噪聲時頻圖

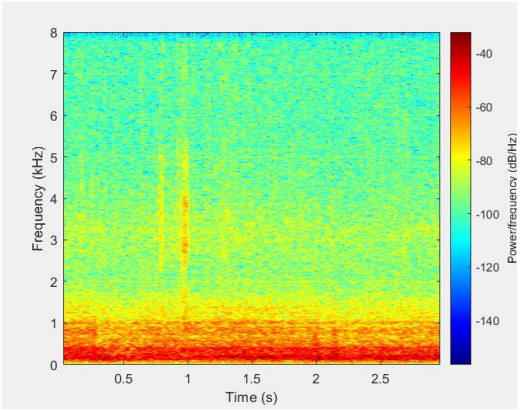
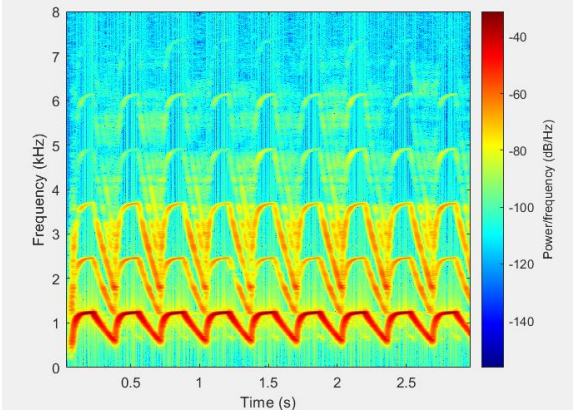
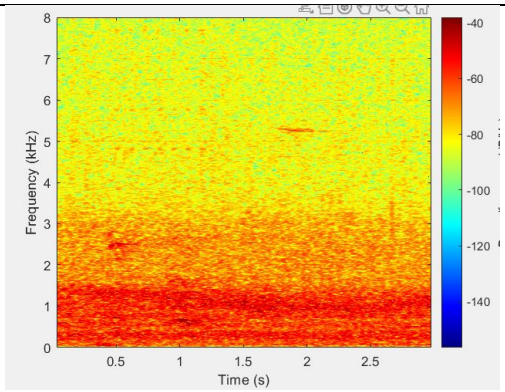
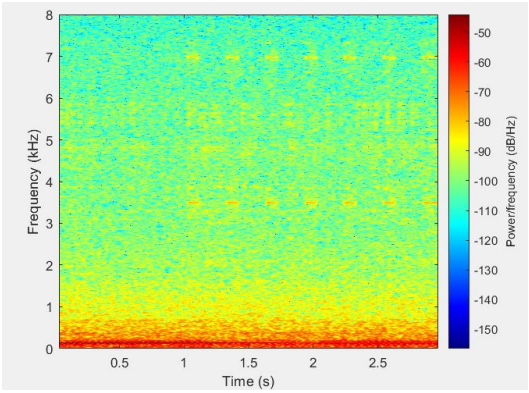
汽車噪聲	警笛噪聲
	
街區噪聲	冷氣機噪聲
	

圖 3.4 測試噪聲時頻圖

如圖 3.3、圖 3.4 所示，我們使用了 8 種噪聲建構訓練集資料，4 種噪聲建構測試集的資料，並且利用不同的訊號雜訊比(Signal-to-noise ratio, SNR)，與原始音檔混合，使資料集能夠更完整。訊號雜訊比可由以下公式表示:

$$SNR(dB) = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) = 20\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) \quad (4.1)$$

訓練集的部分我們總共產生了 8 種不同的訊號雜訊比，分別是(-5dB -2dB 1dB 4dB 7dB 10dB 13dB 16dB)，以 3dB 為等差數列遞增。測試集的部分，實驗一與實驗二中我們使用了(-5dB 0dB 5dB)，實驗三與實驗四中使用(-10dB -5dB 0dB 5dB 10dB 15dB)，皆以 5dB 為等差數列遞增。

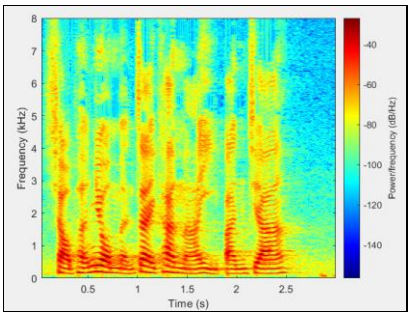


圖 3.5 乾淨測試語音時頻圖

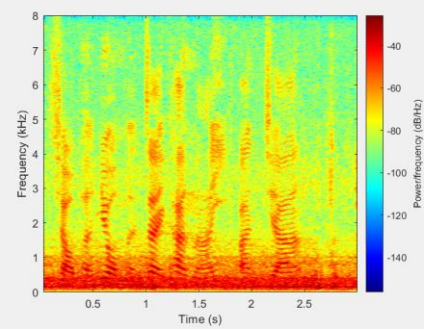
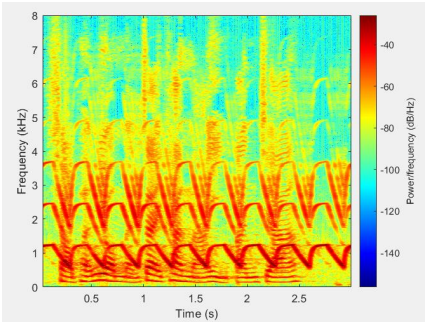
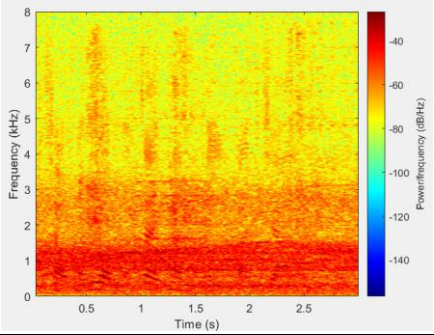
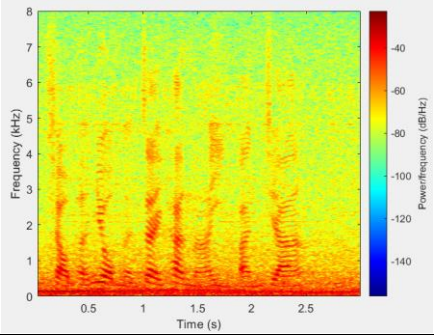
汽車噪聲+乾淨測試語音	警笛噪聲+乾淨測試語音
	
街區噪聲+乾淨測試語音	水冷噪聲+乾淨測試語音
	

圖 3.6 測試語音混入-5dB 噪聲之時頻圖。

由圖 3.5、圖 3.6 可以看出，原本乾淨的語音在加入噪聲後，頻譜特徵明顯遭受到破壞，而我們目標就是藉由 LA-DDAE 模型，消除這些噪聲，讓音訊恢復乾淨清晰的狀態。

第四章、結果與討論

4.1 評估指標

在這章的一開始會先介紹我們用來評估效能的指標，再利用圖表與表格進行比較。評估指標有三個，第一個語音品質指標(Perceptual Evaluation of Speech Quality，PESQ)是由國際電信聯盟電信標準化部門(ITU Telecommunication Standardization Sector, ITU-T）所訂定的指標，其目的是以客觀的評估方式取代以往的陪審團制度，運作的原理是將「原始的輸入訊號」與「經過處理的音檔」進行比較後，產生出評分的結果。而我們所使用的 PESQ 指標是[ITU-T P.862]中所制定的標準，其數值範圍是由[-0.5,4.5]，數值越高表示其語音品質越好，越接近乾淨的原始音檔，聆聽感受也會越佳。

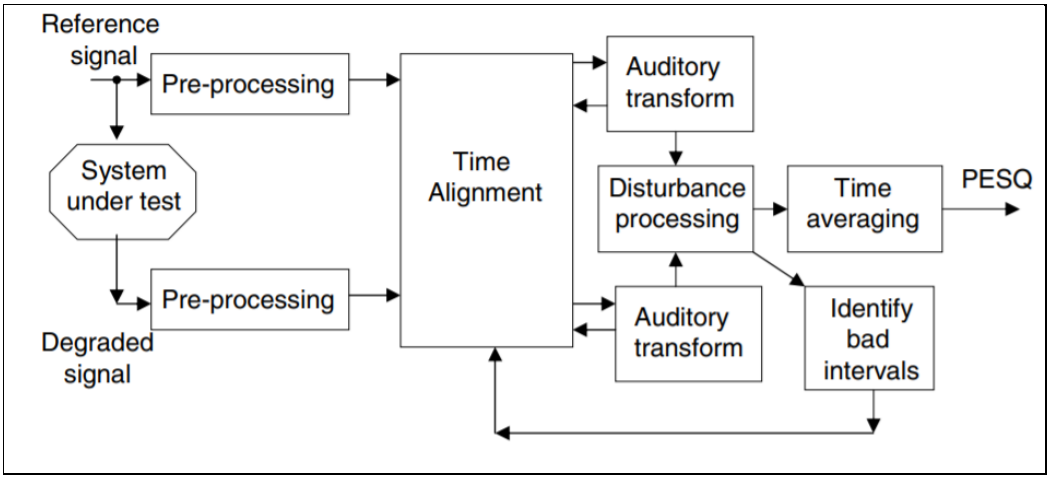


圖 4.1 Block diagram of the PESQ measure computation[6]

第二個指標稱為短時客觀可懂度(Short-Time Objective Intelligibility, STOI)[7]，是用來評估語音可理解度的指標，範圍介於[0,1]，數值越大表示該語句中可被理解的單詞數越高。其運作方式先是將帶噪語音與經過處理的語音，利用漢明窗 50%的重疊(overlapping)與 STFT 將音檔轉為時頻訊號，接著利用三分之一倍頻進行分析，經處理後可得第 j 個頻段第 n 個時間的短時能量頻譜，最後可得以下公式:

$$\text{STOI} = \frac{1}{JN}\sum_{j,n} d_{j,n} \tag{4.1}$$

第三個指標為訊號雜訊比改善(signal-to-noise ratio improvement, SNRI)，此指標為原始帶噪音檔的訊號雜訊比與經處理後降噪音檔的訊號雜訊比比較所產生的指標，數值越大代表經 LA-DDAE 模型後的音檔去除背景噪聲的效果越好，因為 SNRI 是比較兩者之間的訊號雜訊比值的差異，因此若原始帶噪音檔的訊號雜訊比值很高，訊號雜訊比改善的值會變小。

$$\text{SNRI} = SNR_{enchance} - SNR_{origin} \tag{4.2}$$

4.2 實驗一

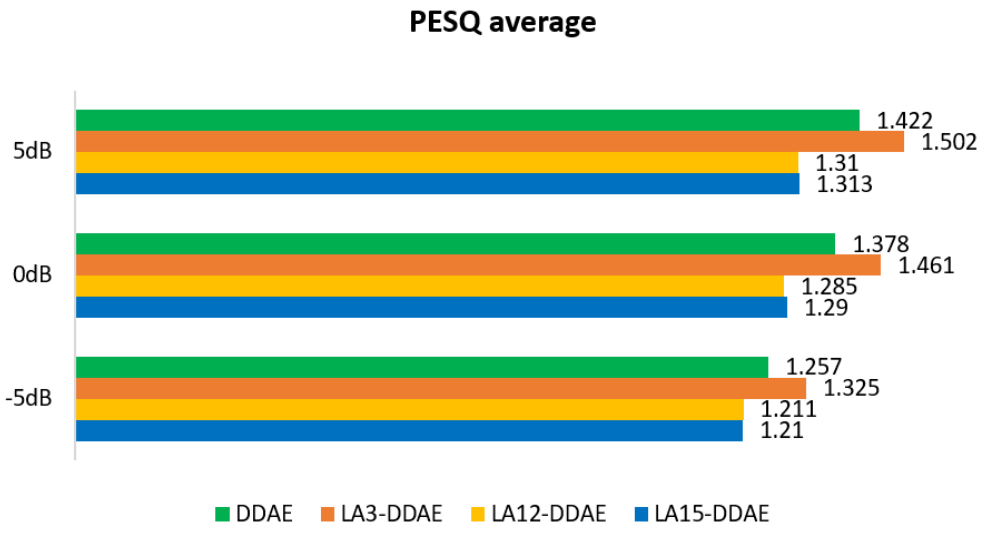


圖 4.2 實驗一 4 種噪聲平均的 PESQ

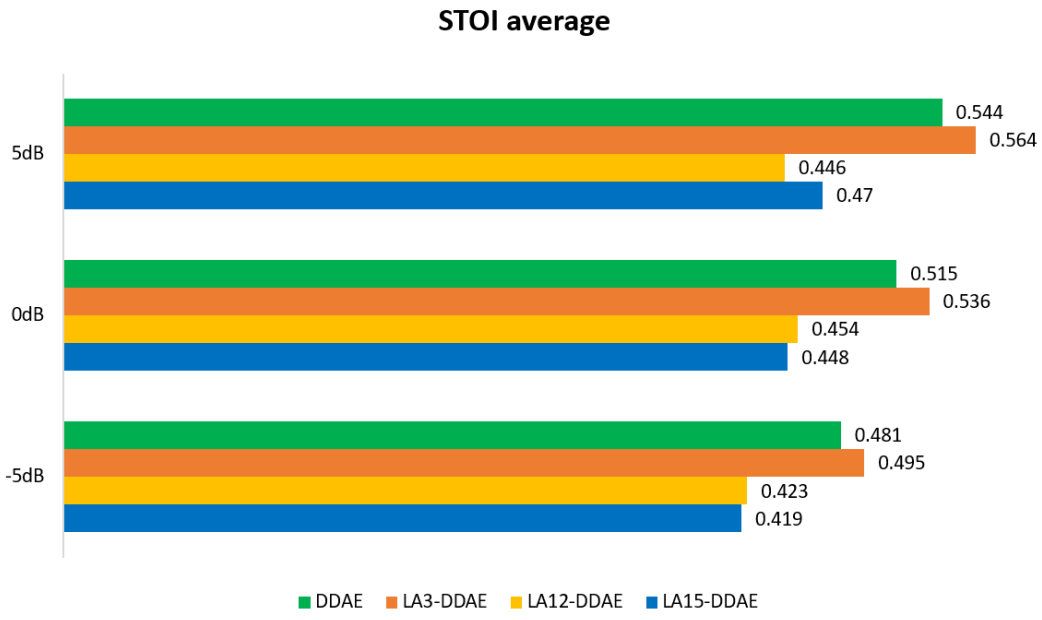


圖4.3 實驗一4種噪聲平均的STOI

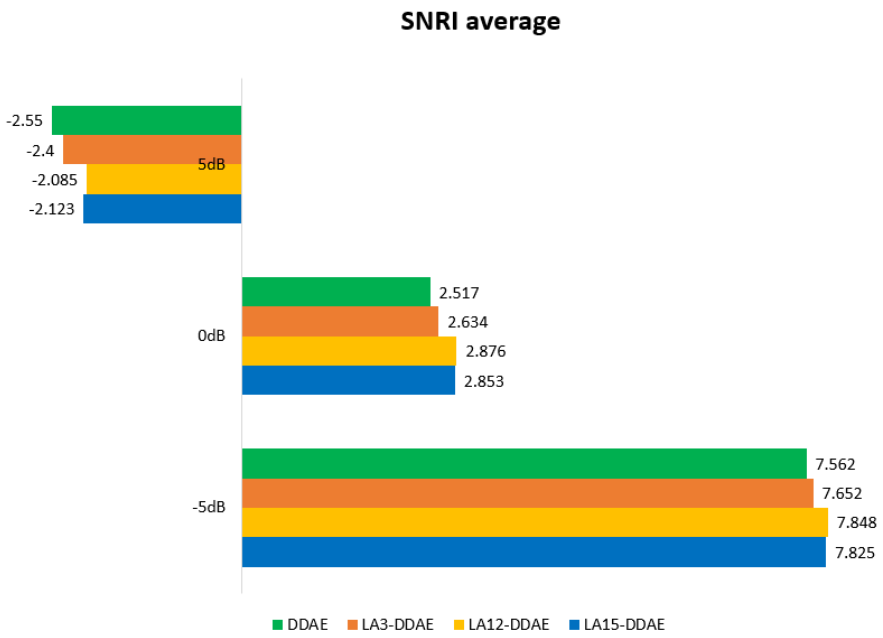


圖 4.4 實驗一 4 種噪聲平均的 SNRI

由圖4.2中可以發現，當考慮極端位置的狀況下，且測試音檔的位置資訊在訓練音檔的位置資訊出現過時，加入3維的位置資訊對模型的訓練是有幫助的，但是加入12維與15維的位置資訊不但沒有提升PESQ的效能，反而使數值變的更低。我們將有提升效能的LA3-DDAE模型與DDAE進行比較，在不同的SNR下PESQ的數值皆可提升5~6%不等。

由圖4.3中可以發現，STOI同樣是由LA3-DDAE的表現最好，與DDAE相比，約可以提升3~4%的數值。而由圖4.4可得知，當原始帶噪語音的SNR在0dB、-5dB時，經模型處理後可以有效的提升SNR的數值。

4.3 實驗二

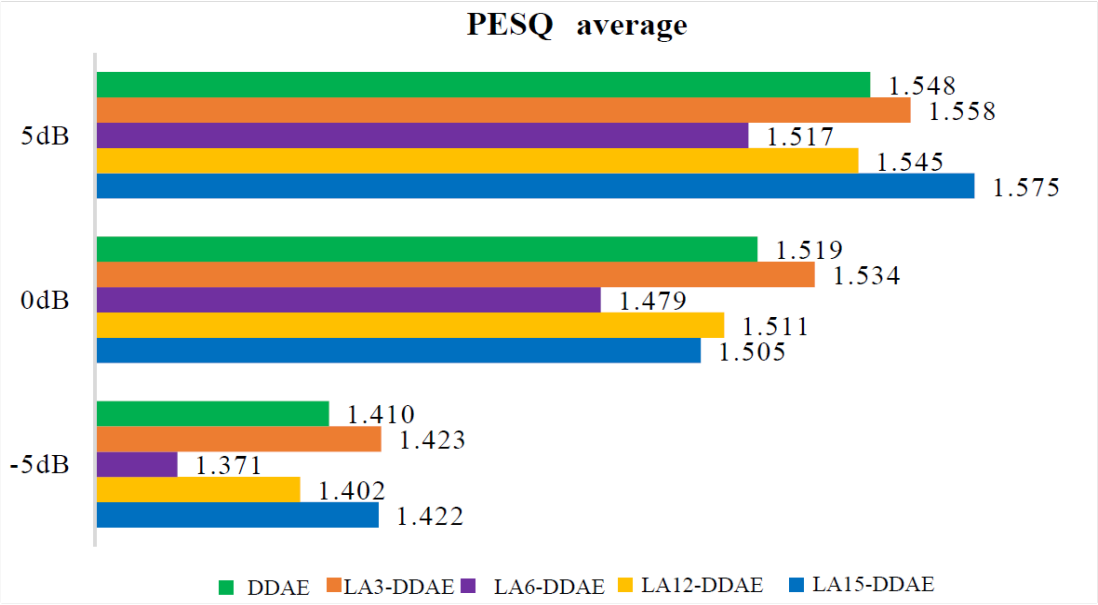


圖 4.5 實驗二 4 種噪聲平均的 PESQ

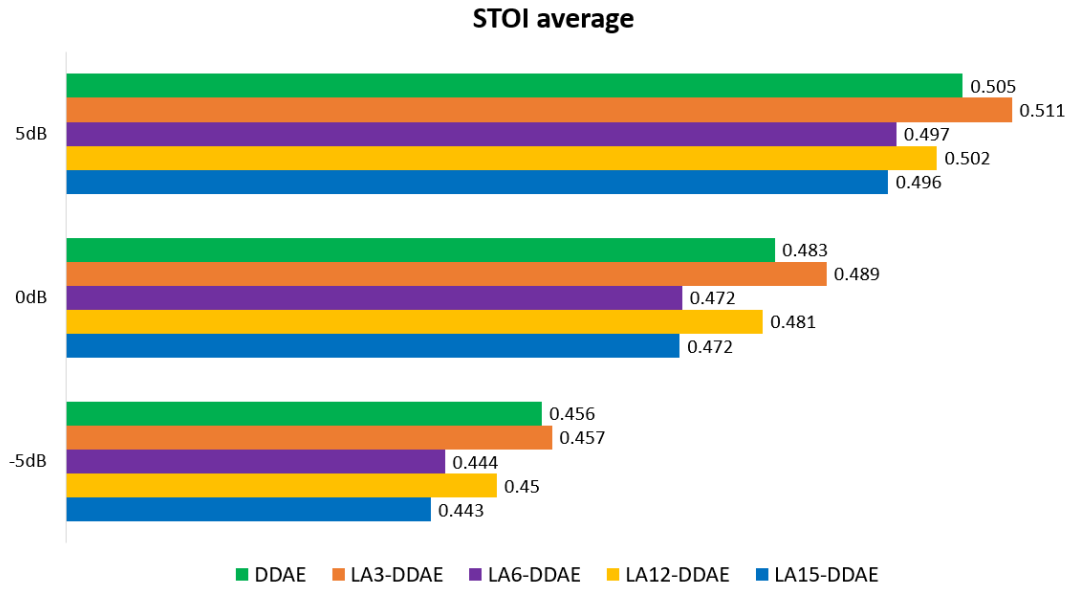


圖 4.6 實驗二 4 種噪聲平均的 STOI

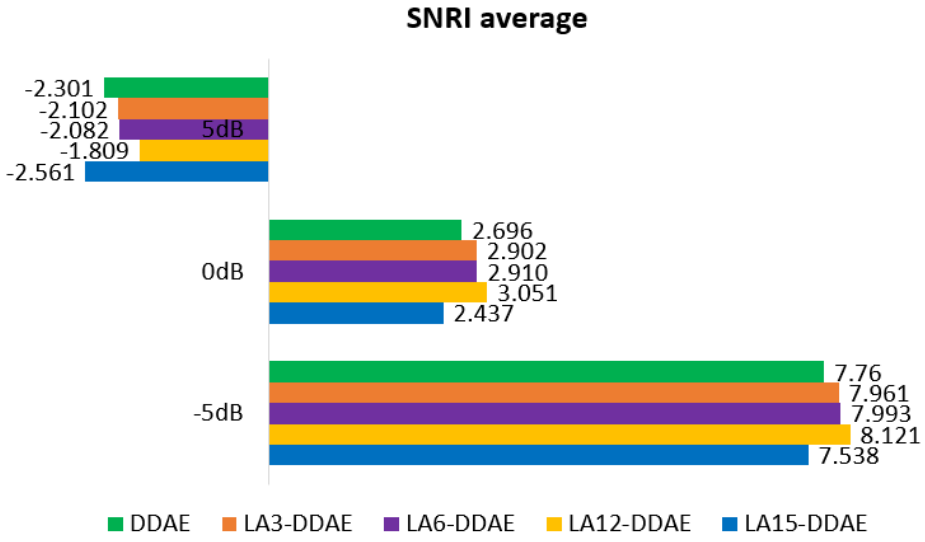


圖4.7 實驗二4種噪聲平均的SNRI

由圖4.5中可以發現，當考慮極端位置的狀況下，且測試音檔的位置資訊沒有在訓練音檔的位置資訊出現過時，加入3維的位置資訊與15維的位置資訊對模型的訓練是有幫助的，但是加入15維位置資訊的模型在SNR為0dB時表現的比未加入位置資訊的模型來的差。而加入6維與12維的位置資訊不但沒有提升PESQ的效能，反而使數值變的更低。與實驗一的結果相比，可以發現LA3-DDAE的PESQ數值相較於DDAE提升約1%左右，與實驗一提升的6%相比降低許多，因此得出測試資料位置資訊是否在訓練資料出現過，會影響模型的效能。

由圖4.6可以發現，LA3-DDAE的STOI與DDAE相比提升1%的數值，而由4.7可以得出與實驗一相同的結果，當原始帶噪語音的SNR在0dB、-5dB時，經模型處理後可以有效的提升SNR的數值。

4.4 實驗三

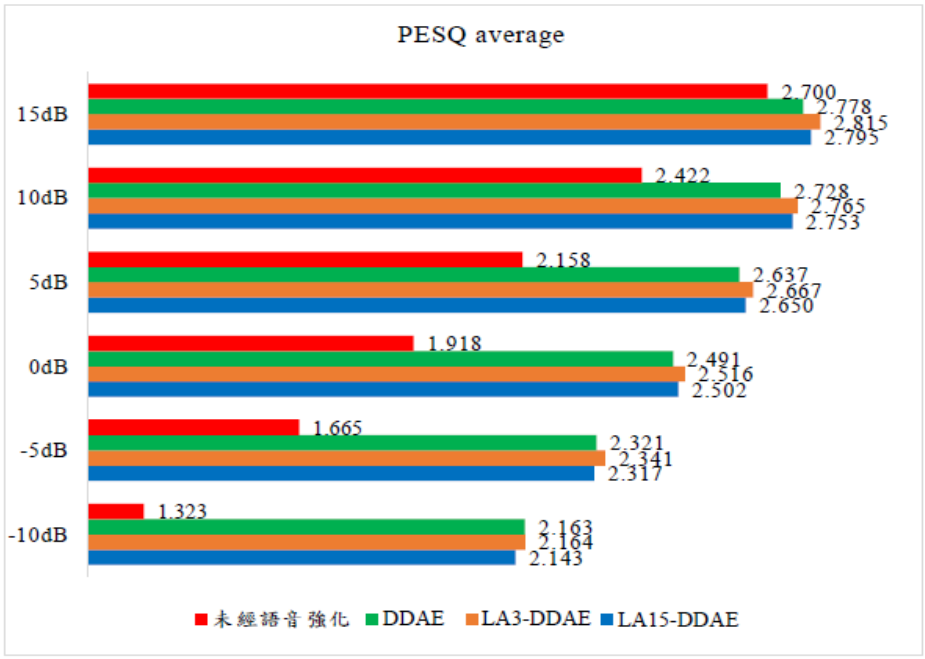


圖4.8 實驗三4種噪聲平均的PESQ

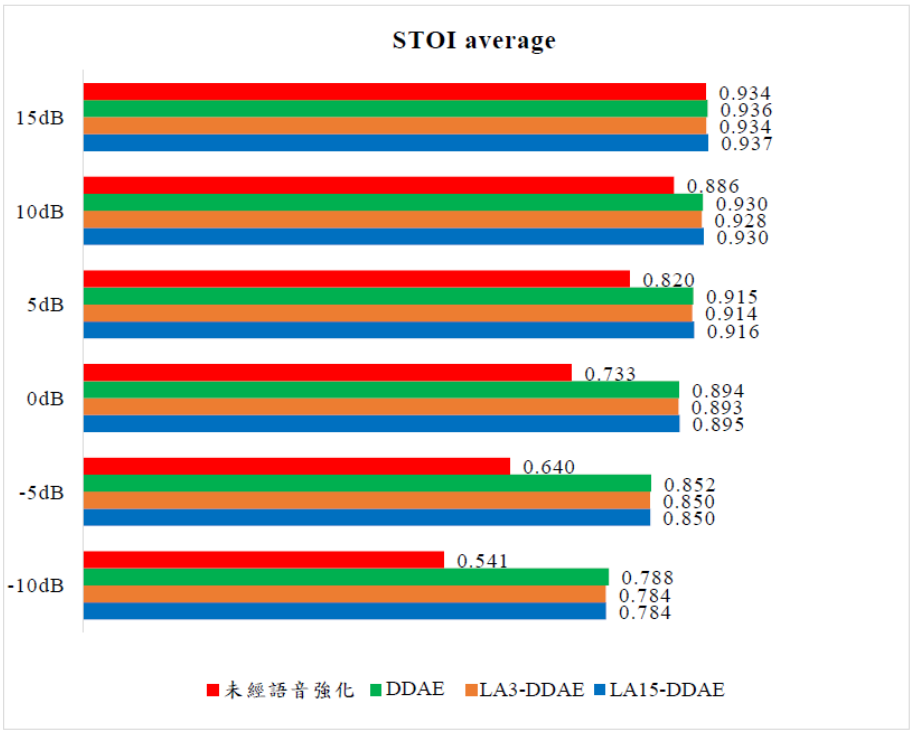


圖4.9 實驗三4種噪聲平均的STOI

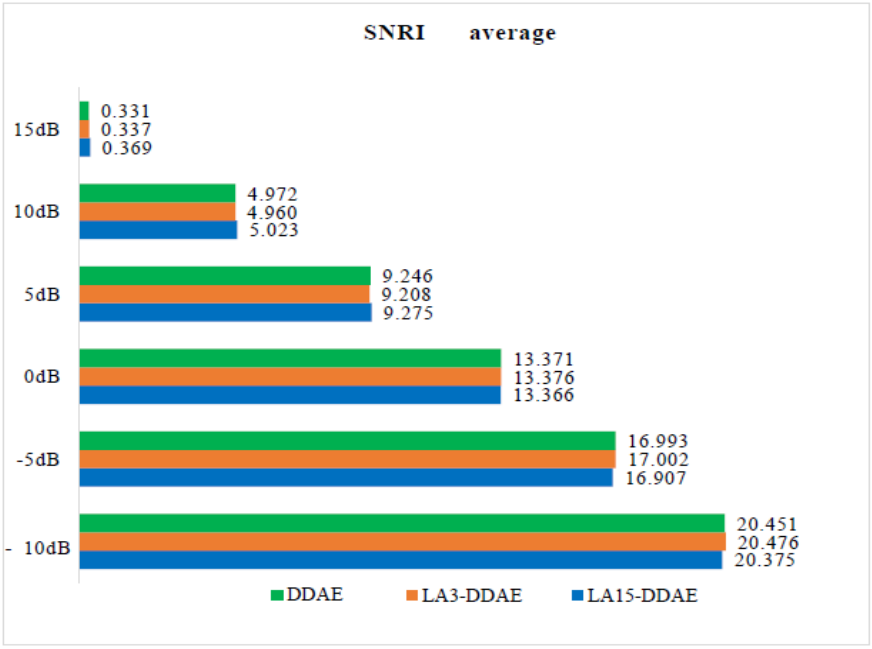


圖4.10 實驗三4種噪聲平均的SNRI

由圖 4.8 可以發現，當音檔經過語音強化後，PESQ 的數值比未經語音強化提升了許多，分別觀察 LA3-DDAE、LA15-DDAE 與 DDAE 的 PESQ，可以發現當 SNR 為 0dB 以上時，加入 3 維與 15 維的位置資訊皆能提升模型效能，但是在 SNR 小於 0dB 時，只有 LA3-DDAE 才能提升效能。與 DDAE 相比，經 LA3-DDAE 處理後的 PESQ 約可再提升 1~2%的數值。

觀察圖 4.9 可以發現，經模型強化後的音檔 STOI 數值明顯提高許多，但是加入位置資訊與否對 STOI 不會造成明顯的影響，差距皆在 0.5%以內。觀察圖 4.10 可以發現，當原始音檔的 SNR 值越低時，SNRI 的數值越高，表示原始音檔雜訊比越低，經語音強化後降噪的效果會越明顯，而當原始帶噪音檔的 SNR 為 15 時幾乎難以達到降噪的效果。

將實驗三與實驗一、實驗二的結果相比，可以發現當我們用 matlab 模擬音檔的結果來進行模型訓練時，PESQ、STOI 的數值皆高出許多，且由 SNRI 的結果發現當原始帶噪音檔的 SNR 為 15 時在處理後也能些微提升 SNR，而主要造成這樣的結果是因為實驗三的音檔沒有摺積性噪聲，這個部分在結論的地方會再提到。

4.5 實驗四

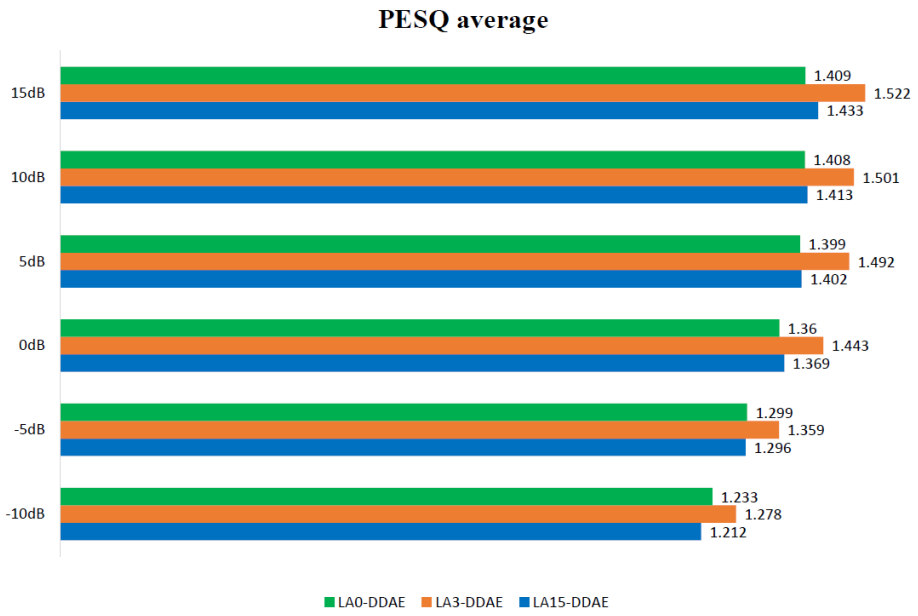


圖 4.11 實驗四種噪聲平均的 PESQ

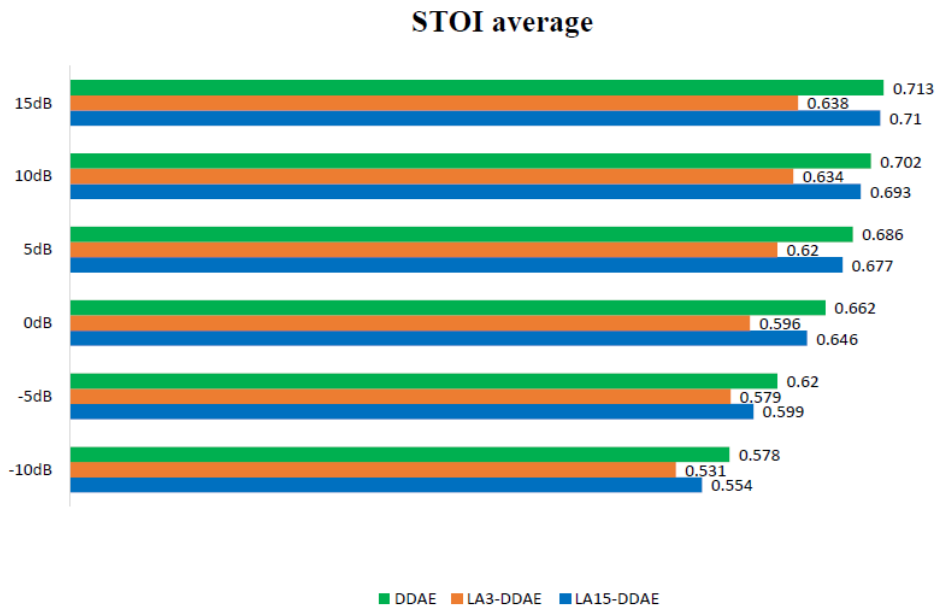


圖 4.12 實驗四 4 種噪聲平均的 STOI

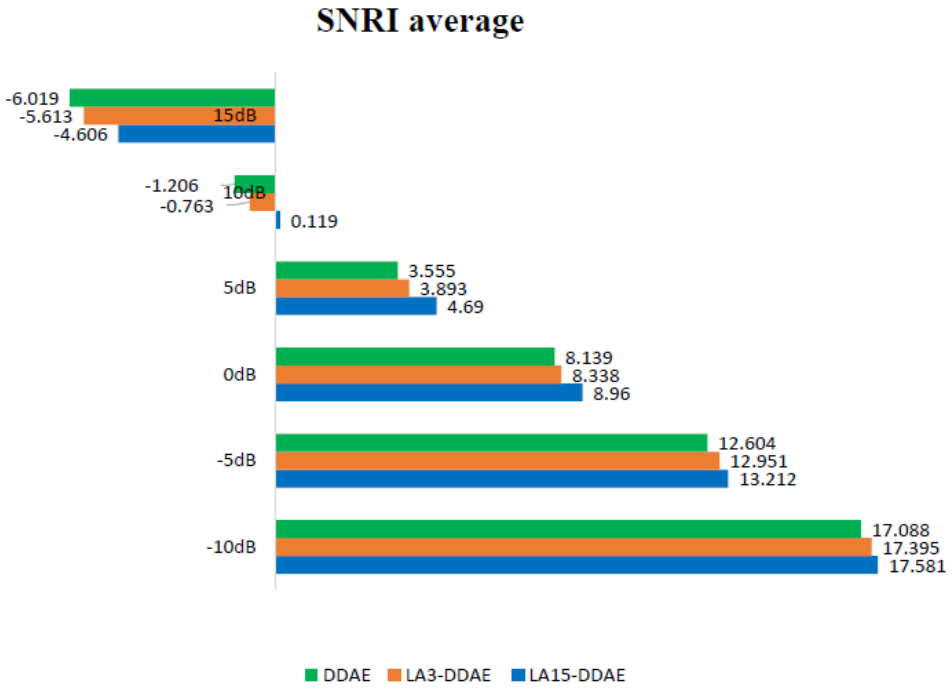


圖 4.13 實驗四 4 種噪聲平均的 SNRI

由圖 4.11 可以得知，LA3-DDAE、LA15-DDAE 與 DDAE 相比皆可以提升 PSEQ 的數值，但是由 LA3-DDAE 提升的最為明顯，在不同的 SNR 下提升 3~8% 不等。

STOI 的部分則與前面的結果相違背，LA3-DDAE 的數值比 DDAE 來的低，我們推測有可能是因為毫米波雷達版在錄製資料的時候有誤差，導致位置資訊不夠準確，因此模型訓練的結果比較差勁。

由圖 4.13 則可以得知，在使用完整的資料集下，相較於實驗一、實驗二能夠更有效的去除噪聲，當原始帶噪音檔的 SNR 為 5dB 以下時，經模型處理依然能提升 SNR 的數值。

第五章、結論與未來展望

下圖為實驗三中SNR是-5dB的警笛噪聲時頻圖，分別為圖5.1、圖5.2、圖5.3。可以觀察出，當帶噪語音經過DDAE模型後，雜訊被消除了許多，接著進一步比較圖7.2與圖7.3，可以發現當我們的模型未加入位置資訊時，時頻圖上還是會殘留著些許警笛噪聲的特徵，而當加入3維的位置資訊後，時頻圖上警笛噪聲的特徵幾乎被完全消除。

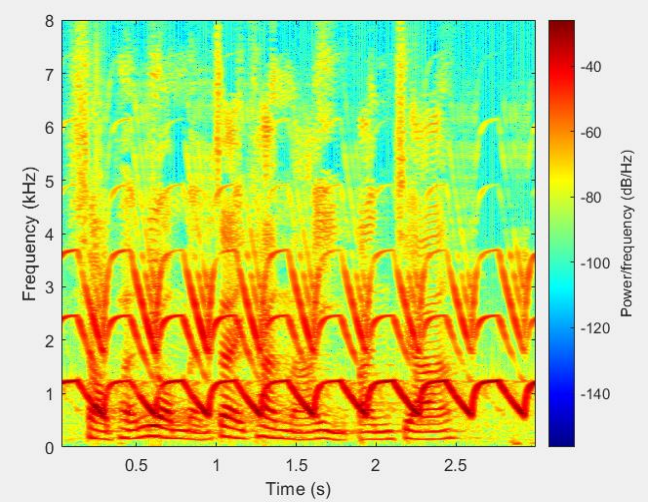


圖 5.1 未降噪音檔時頻圖

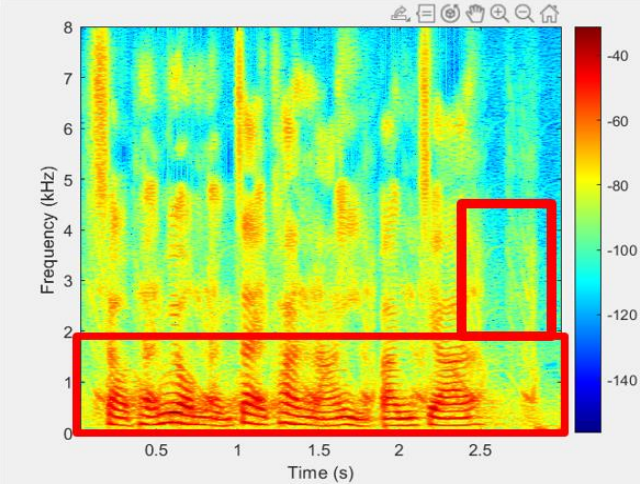


圖 5.2 經 DDAE 之時頻圖

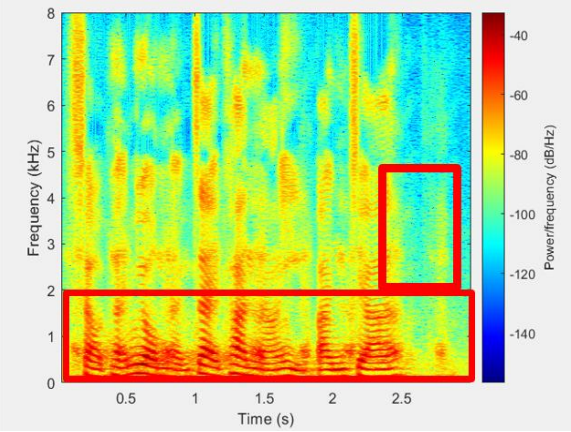


圖 5.3 經 LA3-DDAE 之時頻圖

表 5.1 圖 5.1 至圖 5.3 之 PESQ 比較表

	PESQ	PESQ 提升百分比
未降噪音檔	1.485	
DDAE	2.131	43.5%
LA3-DDAE	2.247	51.3%

綜合以上實驗得出以下結論，我們推測DDAE模型對於加成性噪聲(Additive Noise)有良好的除噪效果，但是對摺積性噪聲(Convolution Noise)，也就是麥克風錄製資料時，自身產生的通道雜訊的除噪效果較不明顯，且摺積性噪聲會嚴重影響模型的訓練，造成PESQ的數值普遍都比實驗三，matlab模擬出來的實驗結果低。

在所有實驗中LA3-DDAE的表現最為穩定，在所有實驗PESQ數值皆比沒有加入位置資訊的模型來的高。STOI的部分則是在實驗一、實驗二中表現得最為優異，在實驗四STOI異常低落，推測是毫米波雷達版錄製資料的誤差導致模型沒辦法訓練出較為精確的空間分布概念。

針對摺積性噪聲所造成的影響，我們規劃在未來實驗時，可以先使用一次濾波器，消除摺積性噪聲，再進行加成性噪聲混入與模型訓練，而在模型的維度選擇上也許可以再加入房間的大小以及空間的反響等維度，增加模型的複雜度與廣度，藉此提升模型的效能。

第六章、參考文獻

[1] Xugang Lu, Yu Tsao, Shigeki Matsuda, Chiori Hori, Speech Enhancement Based on Deep Denoising Autoencoder , 2013

[2] 梁又友，以深度學習技術實現分散式多麥克風語音增強系統，元智大學碩士論文, 2020 年

[3] Wang, S., Liang, Y., Hung, J., Tsao, Y., Wang, H., & Fang, S. (2019). Distributed Microphone Speech Enhancement based on Deep Learning. *ArXiv, abs/1911.08153*.

[4] Fang, S., Wang, C., Chen, J., Tsao, Y., & Lin, F. (2019). Combining acoustic signals and medical records to improve pathological voice classification. APSIPA Transactions on Signal and Information Processing, 8, E14. doi:10.1017/ATSIP.2019.7

[5] Chuang, Fu-Kai, Syu-Siang Wang, Jeih-Weih Hung, Yu Tsao and Shih-Hau Fang. “Speaker-Aware Deep Denoising Autoencoder with Embedded Speaker Identity for Speech Enhancement.” *INTERSPEECH* (2019).

[6] Loizou P.C. (2011) Speech Quality Assessment. In: Lin W., Tao D., Kacprzyk J., Li Z., Izquierdo E., Wang H. (eds) Multimedia Analysis, Processing and Communications. Studies in Computational Intelligence, vol 346. Springer, Berlin, Heidelberg.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214-4217, doi: 10.1109/ICASSP.2010.5495701.

第七章、專題歷程及分工

謝維:

專題研究可以分為神經網路與資料處理兩個面向，在神經網路方面，我與伯翰透過指導教授獲得一個深度神經網路在單麥克風下的模型架構，自此我將獲得的模型架構開發成適用分散式麥克風情境，期間藉由實驗室學長預錄的分散式麥克風資料集進行模型訓練，並且在當中不斷的除錯，最終獲得文中三個實驗的原始模型，接著為了加入位置資訊我和伯翰嘗試了許多不同的函式與方法，最終得到專題實驗中使用的程式碼。資料處理方面，我主要負責收集資料、製作帶噪語音、評分音檔、模擬空間中音檔、使用毫米波版獲取位置資訊，收集資料是在學長的指導下架設麥克風並且錄製實驗所需的音檔，製作帶噪語音與評分音檔在實驗室前輩的努力下已經有初步程式，但在開發期間仍然遇到許多困難，首先我在理解程式相關參數後，花費許多時間成功將程式修復，接著由於獲得的程式只能對單一音檔進行操作，因此我透過網路資源與過去所學將程式完全自動化，其功能為一次處理所有資料並將程式結果分類，模擬空間中音檔則是從 github 上抓取程式，並且在理解參數內容後與伯翰共同模擬出實驗三的音檔，使用毫米波版獲取位置資訊則是在實驗室的學長幫助下設定雷達環境，而我跟伯翰則是共同將雷達獲取的點雲轉成座標資訊。

在此我想先感謝專題組員伯翰，因為伯翰的協助與配合使專題製作更佳順利。在製作專題中，開發 DDAE 模型、模擬空間中音檔是我與伯翰共同負責的，兩者皆需理解程式中的參數並且加入其他語法以達到我們所需的程式目的，所以過程中我們會各自上網查詢資料，之後透過通訊軟體達到資訊與進度同步。在實驗的資料處理部分，首先我會將錄製的資料交給伯翰剪輯，之後將剪輯好的音檔進行混音並交給伯翰放入模型中訓練，最後再將模型中產出的音檔做評分，過程中我們藉由通訊軟體或雲端同步資訊與傳輸音檔。

在本份報告中我所負責的實驗部分在 2.1、2.2 節提到詳細模型架構，在 3.1 節提到收集資料過程、模擬音檔方法與毫米波雷達室內定位，3.2 節提到製作帶噪語音方法，第 4、5 章提到音檔評分方法與結果展示。在專題中遇到許多困難，個人方面開發 DDAE 模型與資料處理程式都遇到在課堂中不曾遇到的問題，因此我藉由文獻探討與開放資源解決語法和模型設計的相關問題，團隊合作方面，

一開始我們在同步資訊上遇到一些障礙，經常有不能明確指出問題的情形，不過在一年多的默契訓練下，我們的溝通障礙減少並且都能夠明確闡述自己的觀點。

林伯翰:

在專題中我主要負責的工作是，程式撰寫、進行 DDAE 的模型訓練與參數調整、圖表製作，與音檔剪輯，起初我們一開始要加入位置資訊的時候並不清楚要怎麼加，在看完[4][5]論文後，對靜態資訊串在模型中有了初步的概念，我們將位置資訊以歐幾里得空間座標加在時頻特徵後面,但是一開始的方式在王緒翔博士評估後認為是不對的，因為我們的方式是將位置資訊以音框的形式加入，而不是位置的特徵，因此我們修改了程式碼，讓位置資訊能以特徵的方式加入，而觀察實驗一、實驗二、實驗三、實驗四的結果，此方法在 3 維位置資訊中確實能提升 PESQ。

因為受到疫情的關係，皆由謝維負責音檔的錄製與混音，而我負責進行音檔的剪輯，一開始在利用 Audacity 進行剪輯時相當不熟悉，常常會把不同通道錄製到的音檔剪輯成不等長，後來再熟悉了功能之後就能夠比較有效率的剪完 320 句語音。模型的調整與訓練是由我負責，我透過遠端程式連上元智的 IP，使用學校的計算力平台搭配 WinSCP 進行操作，再把降噪後的音檔傳給謝維請他評分，最後再由我將數據製成圖表，以便進行觀察與分析。

我們一開始在實驗三時，遇到很大的困難，因為設備的問題，因此我們還原出來的音檔效果相當糟糕，我們本來以為是程式的問題，因此有將音檔轉時頻資訊的程式改成另一種由王博提供的方法，但是效果還是不好，後來王博建議我們可以利用模擬的方式產生音檔，於是我們上網找了一個 matlab 的模擬程式，讓我們的實驗可以在空間中產生 320 個不同聲源的音檔,最後的模擬的結果也還算不錯。最後的實驗四我們一開始在毫米波雷達板的使用上有很大的問題，這邊想要特別感謝實驗室徐學長的協助,有了他的幫忙我們才有辦法將點雲的位置資訊取出，也才能順利完成最後的實驗。模型調整方面我也有嘗試使用過 batch normalization，不過還原出來的音檔聽起來有些許非噪聲的雜音，loss 圖也沒有確實收斂，因此後來就沒有使用 BN。專題報告撰寫我負責第四章、第五章、文獻回顧以及前面章節的圖表製作。最後，我想要在此特別感謝我的專題組員—謝維，有了他的協助我們才有辦法完成這份完整的專題報告，特別感謝他幫忙錄製音檔的部分，一個人要在半徑一公尺的圓內不斷移動位置，並且一邊講 320 句話真的很辛苦。

第八章、經費規劃及使用

本專題實驗設備主要由四支手機麥克風、毫米波雷達所構成，這些設備由皆為實驗室原先就有的，因此花費僅有用來布置實驗場域時需要用來量測長度的捲尺與標記用的有色膠帶。

表 8.1 經費使用規劃表

項目	單價	數量	總價	應用說明
手機(麥克風)	0	4	0	實驗場域布置
毫米波雷達	0	1	0	實驗場域布置
捲尺	132	1	132	實驗場域布置
有色膠帶	16	1	16	實驗場域布置
總價			148	

元智大學電機工程學系甲組
成果報告授權書

題目：基於位置資訊在分散式麥克風中之語音強化。

共同作者：謝雄、林伯翰、_____、_____。

指導老師：方士豪、_____。

本授權書所授權之報告為授權人在電機工程學系甲組「畢業專題製作/畢業專題」所著作之成果報告內容。
授權人於自由意志下，同意將上列報告所擁有著作權部份之全文(含摘要、圖表、數據和內容之全部)，非專屬、無償授權予指導教授與元智大學電機工程學系甲組，不限地域、時間與次數，以印刷或數位之方式將作品之部分或全部內容重製及公開發表不另致酬，指導教授得在其著作引用或複製本報告之文字或內容。且該作品未曾參與其他競賽獲獎，保證未涉及抄襲，如有抄襲、過度參考之情事發生，授權人無任何異議承擔所有法律責任。

此致 元智大學電機工程學系甲組

授權人簽章：
謝雄、
林伯翰、
_____、
_____。
(須全體成員親筆簽名或蓋章)

指導教授簽章：方士豪 12/15、_____。

中華民國 110 年 12月 15日