深度學習語音增強技術於聲電混合刺激電子耳之應用

亞洲大學聽力暨語言治療學系 王玉賢助理教授
中央研究院資訊科技創新研究中心 曹昱研究員

前言

從1960年代至今人工電子耳技術已有長足地進步。然而，如何能夠進一步提升人工電子耳使用者在現實環境中的語音聽辨（speech perception）和語音清晰度（speech intelligibility）仍是研究者努力的目標。針對低頻聽力相對好，但高頻聽損嚴重的個案，聲電混合刺激電子耳（Electro-acoustic stimulation，EAS）被視為更好的輔具選擇。與傳統電子耳相同，EAS包含電子耳植入體和外部語音處理器。接受EAS手術的個案大多在250-750 Hz，約20-60分貝聽閾有殘存聽力，仍可藉由擴音方式聆聽。因此只會植入一部分的電子耳植入體在個案耳蝸。在語音訊號處理上，EAS系統（圖1）將低頻訊號透過擴大器放大後，傳給接收器進行電能及聲能的轉換；放大後的聲音經由塞入式耳塞或耳蝸傳遞。而高頻訊號則被傳送至傳輸線圈再利用無線射頻方式將訊號傳遞至內部植入體。部份研究顯示，相較於僅是放大聲學訊號的助聽輔具及傳統電子耳，EAS在詞彙及句子階級皆能達到更好的語音辨識效果（Dorman & Gifford, 2010; Gantz & Turner, 2003）。僅管有許多優點，現階段EAS在噪音環境中的語音辨識表現仍有不少進步空間。然而，在語音訊號增強（speech enhancement, SE）領域中，關於EAS系

統語音增強研究非常稀少，而以中文語料進行的研究更是罕見。有鑑於此，本研究探討以深度學習為基礎的語音增強模型是否能適用於EAS聲學訊號處理。此外，本研究檢測過去被用於傳統電子耳的語音增強方法是否也能有效地優化EAS語音辨識。

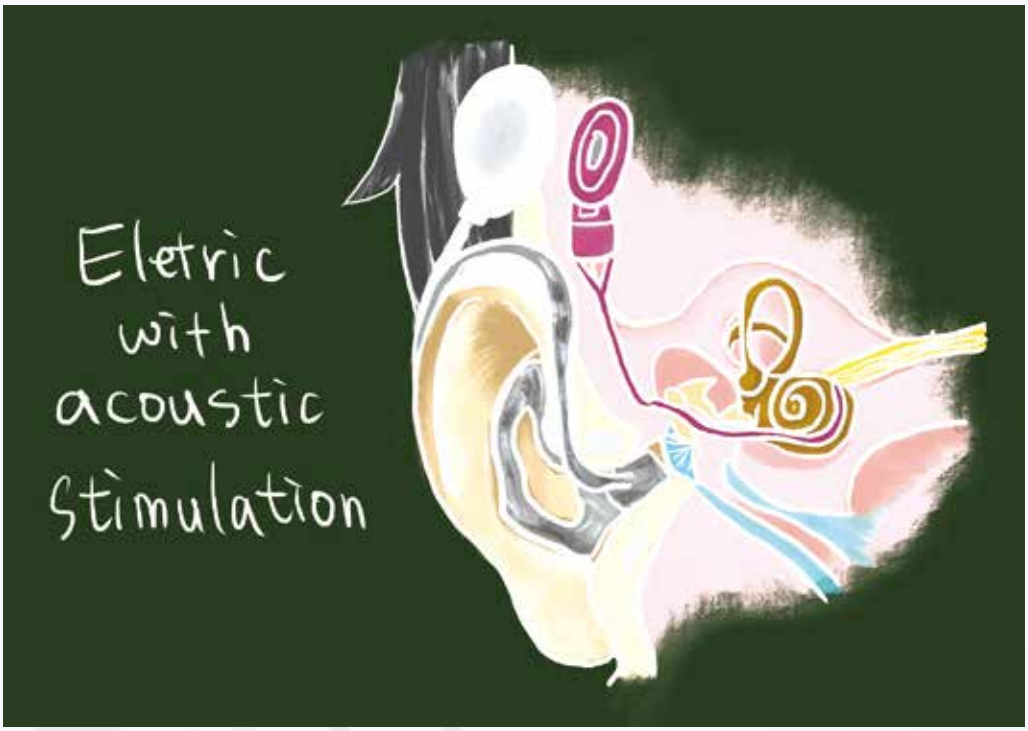

圖 1、EAS系統主要包含內部的傳輸線圈和電子耳植體（紫色部份）以及外部的擴大器（白色部份）。

目前已有許多針對單一麥克風和多麥克風裝置的噪音情況提出不同的語音增強方法。一般而言，針對多麥克風的語音增強方法更能有效地區分語音和雜訊在空間上的差異，但在混響環境中這類方法的有效性會顯著下降。此外，由多個麥克風所接收到的訊號在傳至EAS或傳統電子耳時，終究會被融為單通道語音訊號。因此，單一麥克風語音增強方法對於提升EAS和傳統電子耳系統的語音區辨表現非常重要。現有的單一麥克風語音增強方法大致可分為監督式學習（supervised learning）和非監督式學習（unsupervised learning）兩種。非監督式學習方法的其中一類是由訊號統計分析而來，包含：最小均方誤差短時譜幅度估計（minimum-mean square-error, MMSE）和維納濾波等。另一類非監督式學習方法則是基於訊號子空間觀念及頻譜消去法的語音增強技術，包含Karhunen-Loeve 轉換和主成份分析（principle component analysis）。

此類技術主要利用噪音能量及語音訊號能量的特質差異來分離兩者並消除噪音。在一段語音及噪音混雜的聲學訊號中，語音訊號能量分佈於某一子空間，而噪音能量則較為均勻地分佈於訊號所在的向量空間。將語音及噪音訊號分開後再利用語音子空間中的資訊將訊號還原為乾淨的語音。在電子耳訊號處理方面，這些針對單一麥克風訊號所開發的非監督式學習方法能夠有效地消除穩態雜訊（如：引擎聲），但在非穩態雜訊（如：嘈雜聲）抑制上則表現欠佳。然而，現實環境中存在著各式難以預測的非穩態雜訊，因此近年許多研究者嘗試將深度學習（deep learning）演算法用於單通道語音增強領域。

深度學習是機器學習（machine learning）的一種類別。深度學習語音增強模型通常對於訊號中的語音及噪音訊號不具有強烈的統計假設，而是以資料驅動模式為核心。深度學習模型架構中包含多個層次，其中的噪音抑制模型可用於截取語音以進一步進行語音訊號增強 （Xu et al., 2015; Wang & Chen, 2018）。這樣的層次結構也能更明確地定義語音及噪音訊號間的複雜關聯。相較於傳統的單通道語音增強模型，基於深度學習的模型更能夠大幅提升語音清晰度。較為知名的深度學習模型包含：深度除噪自編碼（deep denoising autoencoder, DDAE）和卷積神經網絡（convolutional neural networks, CNNs）。近年，以深度學習為基礎的語音增強方法主要發展重點分為兩個層面。一是開發適用於語音增強系統的訊號輸入及輸出模式；二是發展任務導向的目標函數（objective function）以訓練語音增強模型。傳統單麥克語音增強模型（如：深度除噪自編碼模型）的輸入模式，多採用聲學特徵能量頻譜或其對數型態做為輸入訊號。此類模型藉由轉換包含噪音訊號的聲學特徵能量頻譜以增強聲學特徵，並且盡可能地輸出與參照目標（無噪音）相近的乾淨語音。由於噪音語音中並沒有清晰的相位譜結構，較難經由相位譜推估無噪音狀態下的相位資訊。因此在語音增強過程中，利用噪音語音中包含的相位資訊來重建語音波形是較為普遍的做法。然而，直接使用噪音語音中相位資訊的做法並不理想，可能導致經增強後語音品質降低。

目前已有不少研究者針對相位估計可能產生的問題提出解決方法，這些方法依其特性可分為兩大類。其中一類採用複雜頻譜做為聲學特徵。此類的深度學習模型運用遮蓋函數（masking function）來截取噪音訊號中的語音部份，並同時估計語音訊號的相位和振幅資訊。相對於將聲波轉換為頻譜特徵的做法，另一類方法則是直接針對原始語音波形進行增強。近年發展出來的全卷積神經網絡（fully convolutional neural networks, FCN）語音增強模型即

是保留語音波形中的鄰項訊息並用於重建語音中高頻及低頻部份（Fu等人, 2017, 2018）。相較於CNN和深度神經網絡（deep neural network），FCN能夠處理較多不同類型的目標函數，且僅需要較少的參數就能達到較佳的語音增強效果。另一方面，依據人類聽覺系統特性找出適用參數也是近年深度學習語音增強領域的研究重點。傳統的深度學習語音增強系統利用均方誤差（mean-square error, MSE）量測被增強的聲學訊號與乾淨語音參照訊號（reference clean speech signal）間的差異。然而，近年提出的深度學習模型陸續檢視了採用其它評估矩陣（evaluation metrics）建構目標函數的訓練成效。例如，Fu等人提出以短時客觀理解度（short-time objective intelligibility, STOI）當作損失函數來訓練FCN，此模型稱為FCN（S）。傳統的以深度學習為基礎的語音增強模型大多是作在時頻域（time-frequency domain）上，並且以幅（frame）為單位作處理，因而很難直接最佳化跨越幅計算的STOI。而Fu等人提出的FCN（S）則是直接作用在時域的波型（waveform）上，並且以整個句子為處理單位。主觀語音辨識任務顯示，相較於以MSE作為目標函數的FCN模型，FCN（S）模型有較佳的語音增強效果。

近年，在電子耳語音訊號處理領域已陸續有研究探討以深度學習為基礎的語音增強模型的效果。例如，Lai等人（2017）利用由聲碼器（vocoder）所模擬噪音語音來檢視DDAE（deep denoising autoencoder）模型用於電子耳語音增強的成效。此研究比較了DDAE和三種常用的單一麥克風語音增強方法（包含對數振幅MMSE、Karhunen-Loeve轉換、及維也納濾波），並採用主觀及客觀的評估方式檢測各模型的語音增強表現。客觀評估是以語音增強前後STOI參數的變化做為依據；而主觀評估則是以聽力測驗的方式進行。聽力測驗以聽力正常之成人為受測者進行語音辨識任務，直接比較四種語音增強模型所輸出的語音清晰度。研究結果顯示，深度學習模型能有效地改善電子耳系統的語音清晰度。然而，目前仍缺乏深度學習語音增強用於EAS系統的研究，因此這類模型用於EAS語音增強的成效仍有待驗證。有鑑於此，本研究以聲碼器（vocoder）模擬EAS環境下的噪音語音，並採用三個深度學習模型（FCN（S）、DDAE、及MMSE）分別進行語音增強，以直接比較不同的深度學習模型用於EAS系統的表現。其中，FCN（S）為Fu等人近年所開發的深度學習模型，此模型在傳統電子耳系統中的語音增強表現也尚屬未知。而DDAE及MMSE則已被用於不少電子耳系統語音增強研究中。因此，本研究主要目的可分為兩個層面：一是檢視FCN（S）是否適用於EAS系統；二是比較以句子為

單位，且直接針對波形進行增強（FCN（S））在EAS系統中的表現，是否優於另外兩種以幅為單位，且針對時頻域進行增強的模型。

基於STOI目標函數的FCN模型

本研究所使用的語音增強模型為FCN，其架構如圖2所示。FCN 模型可以接受任意長度的輸入訊號，因此是一種特別適合使用於語音訊號處理的模型架構。圖2中的 FCN語音增強模型是一個端到端（帶噪語音輸入和乾淨語音輸出）的架構，不需要前處理（特徵提取）和後處理（語音還原），FCN由多層架構組成，每一層是由多個一維卷積式濾波器 （1-D Convolutional Filters）組成，每個濾波器都與前一層生成出的波形訊號進行卷積，並輸出濾波後的訊號。此FCN的最後一層由單一濾波器組成，目標是產生單通道的語音增強訊號。

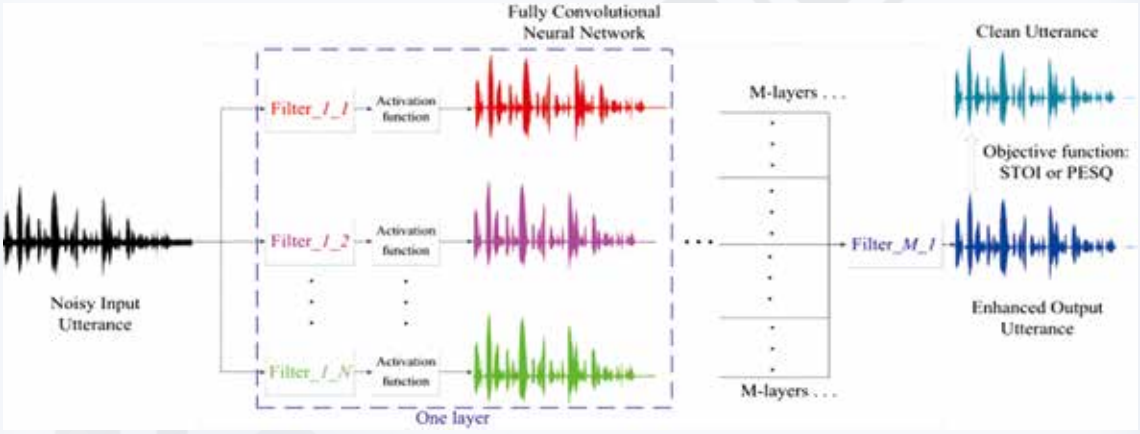

圖2、基於全卷積神經網路（FCN）的語音增強系統架構

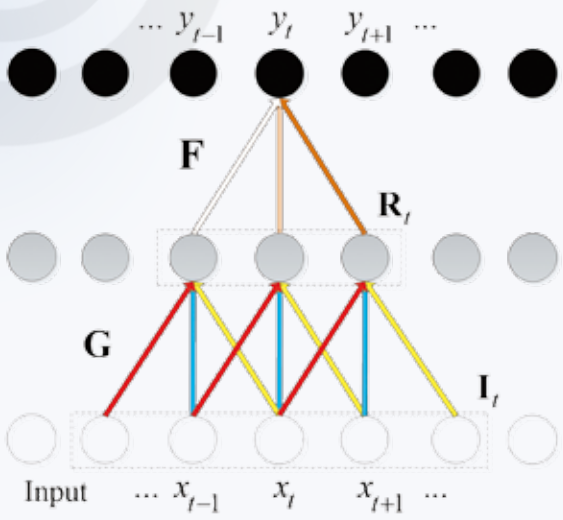

圖3、全卷積網路架構中輸出層與最後一個隱藏間的關係

許多研究指出，全連接的隱藏層無法有效地對語音訊號的高頻和低頻分量同時建立模型，而經由移除全連接層，FCN可以更有效地對語音訊號建立模型，同時也有大幅減少模型參數量的優點。如圖3所示，FCN模型的輸出樣本 y_t 與連接的隱藏節點 R_t 之間的關係為

$$y_t = \mathbf{F}^{\mathrm{T}}\mathbf{R}_t \quad (1)$$

其中 $\mathbf{F} \in \mathbb{R}^{f\times 1}$ 為一個一維的濾波器，f 是過濾器的大小。

為了估測FCN中的參數，我們需要定義一個目標函數。因為希望能夠提升語音理解度，我們開發一套新穎的目標函數，此目標函數是以提升語音的短時客觀理解度（Short-Time Objective Intelligibility, STOI）為目的，其數學式如下：

$$\mathcal{L}(\theta) = -\frac{1}{U}\sum_u stoi(\boldsymbol{w_y}(u), \boldsymbol{w_q}(u)) \quad (3)$$

其中 θ 為FCN的模型參數，w_y（u）及 w_q（u）是增強後的語音以及參考語音訊號，U 則是所有訓練的語音數量，$stoi(.)$ 指的是STOI計算函數，此函數的計算包括以下五個步驟：

（1）對增強語音以及參考語音訊號移除非語音片段

（2）對留下來的增強語音及參考語音訊號執行短時傅立葉轉換

（3）對增強語音以及參考語音訊號執行三分之一倍頻分析

（4）計算增強語音及參考語音訊號之間的相關性

（5）基於不同頻帶計算相關性的平均值，依此推估STOI分數

在先前的研究中，我們發現使用結合均方差（Mean-Square Error，MSE）及STOI的目標函數，可以更加有效地提升語音理解度以及語音品質。因此本研究中我們採用以下的目標函數：

$$\mathcal{L}(\theta) = \frac{1}{U}\sum_u \left(\frac{\alpha}{L_u}\left\|\boldsymbol{w_y}(u) - \boldsymbol{w_q}(u)\right\|_2^2 - stoi(\boldsymbol{w_y}(u), \boldsymbol{w_q}(u))\right) \quad (4)$$

其中 $\left(\frac{1}{L_u}\left\|\boldsymbol{w_y}(u) - \boldsymbol{w_q}(u)\right\|_2^2\right)$ 是增強語音以及參考語音訊號間的MSE，L_u為第 u句語音的長度，α 則是MSE與STOI間的權重值。

EAS聲碼器

本研究以聲碼器模擬EAS使用者所聽到的語音做為語料。聲碼器用於分析和重新合成人類語音訊號的語音處理系統，此系統廣泛用於音頻數據壓縮、語音加密、語音轉換及合成。在電子耳研究領域，聲碼器被廣泛地用於模擬電子耳使用者所聽到的語音。使用聲碼器的模擬語音，研究者可藉由觀察正常聽力受試者在電子耳語音環境下的表現，藉以預測電子耳使用者的聽覺體驗。這樣的作法除了避免電子耳使用者反覆接

受測試，也可以排除個案的個別差異所造成的測試偏差。因此，在許多電子耳研究中，聲碼器是一項被廣泛使用的工具（Zeng, 2004; Lai et al., 2017）。近年以聲碼器模擬的EAS語料也被用於各種電子耳相關研究，包括評估電子耳和EAS系統在語音清晰度表現上的差異，EAS系統在噪音情境下的語音清晰度（Seldran et al., 2014），以及操弄訊號處理參數對於EAS系統在語音辨識和清晰度上的影響（Chen & Loizou, 2010a, 2010b, 2011）。

本研究中所採用的EAS聲碼器結構如圖4。圖中語音X輸入聲碼器後即進入語音增強，增強後的語音訊號以Y表示。增強後的語音訊號經過聲學模擬（語音的低頻部分）和電刺激模擬（語音的高頻部分）兩條路徑處理。對於聲學模擬路徑，語音訊號直接經過低通濾波器處理，電刺激模擬路徑採用了四個帶通濾波器，將語音訊號分成四個頻帶，截止頻率分別為 500 Hz、1017 Hz、1901 Hz、3414 Hz和 6000 Hz，分頻帶之後的訊號使用全波整流器提取包絡訊號，再以白噪聲訊號來激發每個頻帶的包絡訊號，之後再經過同一組帶通濾波器進一步濾波。最後，將四個頻帶的電訊號與濾波後的聲學訊號相加，並對能量進行正規化，得到最後EAS 語音編碼器的語音訊號（圖4中的 Z）。

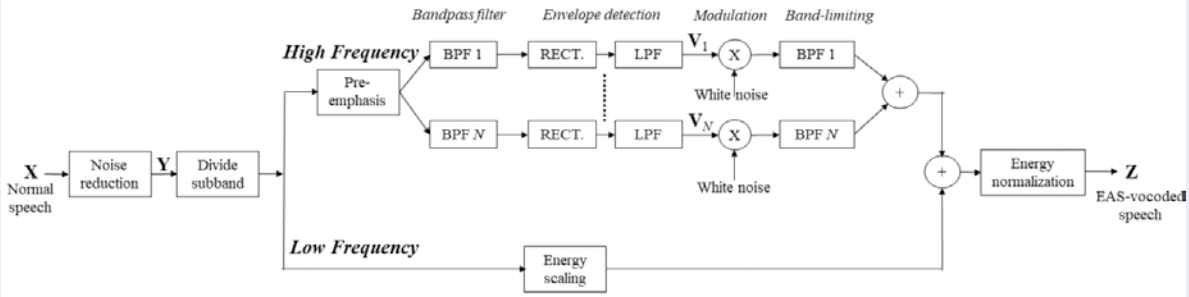

圖4、EAS語音模擬聲碼器架構：此系統中包含兩個路徑，分別處理電能和聲能。電能處理路徑是由一個包含N通道的噪音聲碼器組成；聲能處理路徑則是一個語音訊號的低通濾波器。

實驗配置及結果

本研究著重於檢視FCN（S）模型在訓練及測試語料不匹配（mismatch training and testing）情況下的語音增強表現。也就是說，訓練語料和測試語料是不相同的、由不同的語者錄製、且混和的噪音種類也不同。訓練所採用的語料庫由八位語者（4男、4女）錄製，每位分別錄了320句語料，共2,560句華語語句。語音資料庫是根據台灣地區漢語語音噪音下聽辨測試（Taiwan MHINT）腳本所建立。每個句子皆包含10個字且句長皆在3~4秒之間。所有語句皆於安靜的隔音空間錄製，音訊取樣頻率為16 kHz。FCN（S）模型訓

練過程中，我們首先以其中六位語者（男、女各三位）的前200句語句做為訓練語料；另兩位語者的最後120句語句做為測試語料。為了創造噪音語音，我們將100種不同的環境音混入訓練語料中並調整其信噪比（speech-to-noise ratio, SNR），並製作出五個介於-10 dB和10 dB之間的訊噪比情境，包含SNR -10 dB、-5 dB、0 dB、5 dB、10 dB。也就是說，本實驗共採用600,000個語句（200個句子x 6位語者x100種環境音）做為訓練語料。

在FCN（S）模型中，端對端訓練（end-to-end training）僅用於訓練語音增強模型，而針對推論（Inference）部份則以幅為單位進行處理。也就是說，FCN（S）在訓練階段以整句為單位的方式進行語音增強，而在測試階段則是以幅為單位進行語音增強。最後訓練階段所用的語料為原訓練語料中隨機挑選的3,000個語句，並以不同於原訓練用的兩種環境音（引擎和街道噪音）創造出-3 dB和1 dB的測試。同樣的訓練語料也用於DDAE和MMSE模型，以直接比較三者在一般語音和聲碼器語音增強上的表現。

本研究以兩組種評估方式分別檢視FCN（S）、DDAE、和MMSE模型在一般語音和EAS語音增強的成效。針對一般語音，我們以寬頻語音在經由各模型進行語音增強後的STOI數質做為評估言語清晰度的標準化評估指標（standardized evaluation metric）。STOI分數介於0至1之間；分數越高表示言語清晰度越高。另外，我們以聽測評估這三個模型在EAS系統中的語音增強表現。在聽測中，由聽能正常的受試者聆聽經語音增強及EAS聲碼器處理過的語句，並以受試者的語音辨識表現做為評估依據。

一、一般語音評估

本研究分別以穩態（引擎）和非穩態噪音（街道雜音）進行混噪；圖5呈現這兩類噪音的頻譜特性。經由觀察聲學訊號在不同時間點上的特徵變化，可看出三種語音增強模型的不同特性。圖6（c）至（e）分別呈現了信噪比-3 dB的非穩態語音在經過MMSE、DDAE、和FCN（S）處理後的波形及頻譜特徵。由圖6（c）可看出利用MMSE增加語音能夠去除部份高頻雜訊，但此模型也造成中低頻的語音訊號扭曲。更明確地說，在中低頻部份，MMSE去除中低頻雜訊的能力有限，且容易發生語音訊號與雜訊一併移除的情況（如虛線框（1）、（2）所示）。另一方面，DDAE雖然有效地移除高頻雜訊，卻也同時將中低頻語音訊號（圖6（d）虛線框部份）一併移除。也就是說，此模型難以區分中低頻區段中的語音和噪音訊號，因此將兩者重疊的部份全都移除。相較於DDAE，FCN（S）模型對於高頻雜訊的處理能力較弱（如圖6（e）虛線框部份所示）。然而FCN（S）是三個模型中保留了最多且最完整中低聲學訊

號的語音增強模型。由於語音訊號大多集中於中低頻區段，因此這些頻段的訊號完整性對於語音清晰度較高頻訊號來的重要。

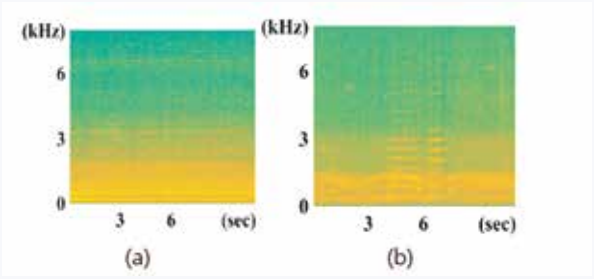

圖5、用於合成吵雜語音（noisy speech）的兩種噪音訊號頻譜:（a）引擎聲和（b）街道雜音。

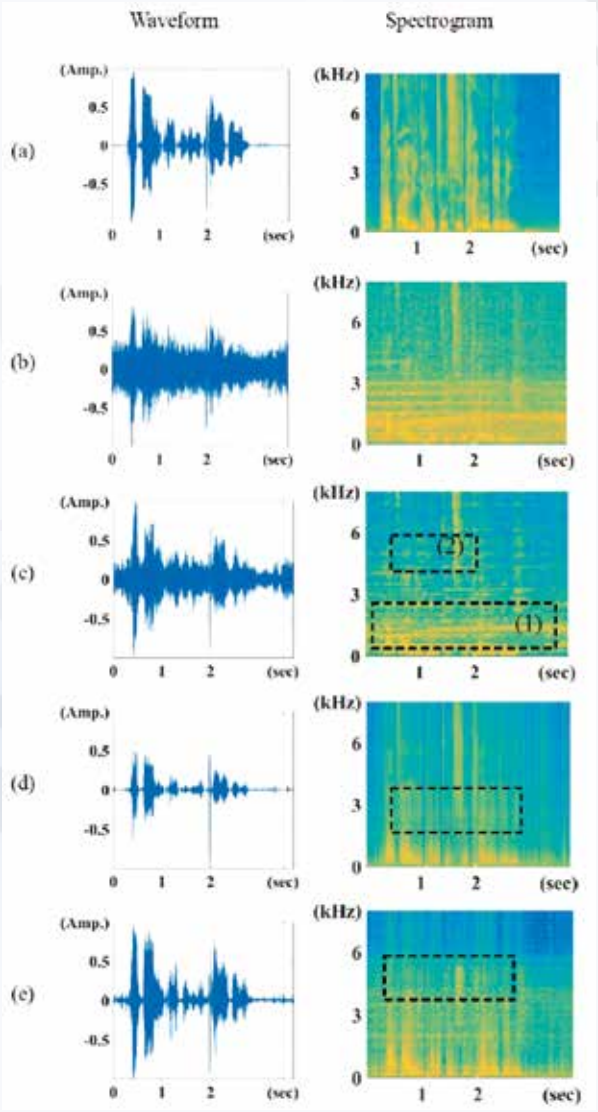

圖6、單一語段在各實驗情境下的波形及頻譜:（a）乾淨語音、（b）噪音語音（街道雜音; -3 dB SNR）、以及經（c）MMSE、（d）DDAE、和（e）FCN（S）增強後的語音（Wang et al., 2020）。

此外，我們以STOI分數做為客觀比較各模型語音增強表現的依據。藉由比較噪音語音的STOI分數和經語音增強模型處理的STOI分數，可得知各模型在不同信噪比情況下的表現。圖7（a）和（b）分別呈現各模型在不同信噪比等級的穩態和非穩態噪音情況下的STOI分數。圖中，未經語音增強的噪音語音以noisy表示，做為各模型STOI比較的基準。在穩態噪音情境下，當信噪比為-11 dB時，{Noisy, MMSE, DDAE, FCN（S）}的平均STOI分數為{0.15, 0.17, 0.30, 0.31};在-7 dB時，平均STOI分數為{0.25, 0.28, 0.44, 0.49}；在 3 dB時為{0.38, 0.41, 0.59, 0.64}；在1 dB時為{0.51, 0.54, 0.71, 0.73}，在5 dB時為 {0.64, 0.67, 0.78, 0.79}；在9 dB時為{0.73, 0.80, 0.81, 0.83}。而在非穩庇噪音情境下，當信噪比為-11 dB時，{Noisy, MMSE, DDAE, FCN（S）}的平均STOI分數為{0.20, 0.22, 0.37, 0.42}；在-7 dB時，平均STOI分數為{0.32, 0.33, 0.54, 0.60}；在 3 dB時為{0.44, 0.45, 0.66, 0.71}；在-1 dB時為{0.55, 0.55, 0.75, 0.77}；在5 dB時為{0.65, 0.67, 0.80, 0.81}；在9 dB時為{0.74, 0.77, 0.82, 0.85}。

由此可知FCN（S）模型無論用於穩態或非穩態噪音語音增強都能有效提升語段的平均STOI分數；且在不同信噪比情境下的效果都顯著優於MMSE和DDAE。此外，由平均STOI分數可知，深度學習語音增強模型[DDAE和FCN（S）]的表現均優於基於訊號統計分析的語音增強方法（MMSE）。而DDAE和FCN（S）的主要差異在於FCN（S）在低信噪比情境下（-7 dB、-3 dB、及1 dB）的語音增強表現優於DDAE。

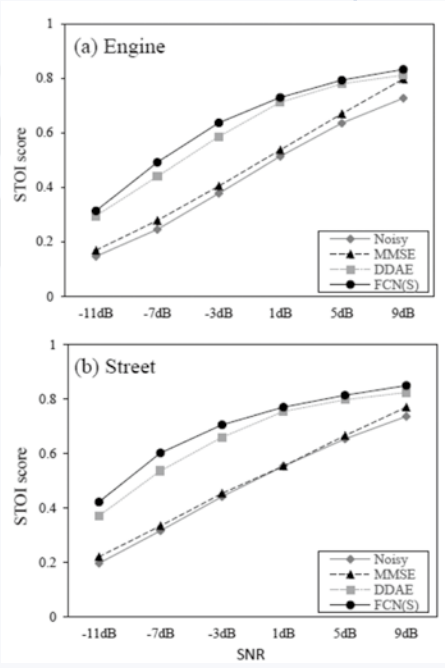

圖7、（a）穩態及（b）非穩態噪音語音的六個信噪比等級和其相對應的語段平均STOI分數（Wang et al., 2020）。

二、EAS語音評估

針對各模型在EAS語音強化上的表現，我們藉由分析增強後的語音振幅輪廓（amplitude envelope）和頻譜（圖8）進行質化的比較，並以聽測做為客觀評估依據。振幅輪廓、波形、和頻譜圖都是分析特定時域內訊號變化情形的常用指標。且振幅輪廓的調制深度（modulation depth）與言語清晰度呈正相關。因此，比較格模型處理後的EAS語音的振幅輪廓可初步預測言語清晰度。圖8（a）和（b）分別呈現乾淨EAS語音和EAS噪音語音的振幅輪廓和頻譜圖。圖8（c）至（e）則分別是經由MMSE、DDAE、FCN（S）模型處理後的EAS語音。由圖8（c）可觀察到，MMSE造成振幅資訊喪失，可能導致語音清晰度下降。相對地，DDAE保留較多振幅資訊但調制深度明顯減弱。而FCN（S）不但保留了振幅資訊且調制深度與乾淨EAS語音相近。這個結果顯示FCN（S）整體表現顯著優於MMSE和DDAE；也就是說，以FCN（S）進行EAS語音增強能有最佳的語音清晰度。另一方面，藉由觀察EAS語音的頻譜資訊我們發現，三個模型在處理EAS語音上的表現與處理一般語音相似。如頻譜圖8（c）所示，MMSE無法有效地去除噪音訊號，且無法移除中低頻與語音訊號重疊（虛線框部份）的雜訊。而DDAE（圖8（e））雖大致保留了語音訊號資訊，但在各頻帶皆有過度移除訊號的情況。相對而言，FCN（S）雖然無法完全移除高頻雜訊，但有效地移除了大部份的中低頻雜訊，且保留大部份的語音資訊。因此，我們預期經FCN（S）增強的EAS語音能有較佳的語音清晰度。

為了更客觀地評估這些語音強化模型在是否適用於EAS系統以及是否能有效地提升語音清晰度，本研究招募了兩組18~39歲的聽常受試者（mean = 24.9歲；SD=4.2歲；男女比例為1:1）進行聽力測驗，每組各30人。為了避免測驗過程中出現聽覺疲勞的情況，各組受試者只會聽到其中一個信噪比等級（-3 dB或1 dB）下的EAS語音。而為了進一步瞭解語音增強模型在不同噪音情境下的表現，聽測中包含兩個穩態及非穩態兩種情境。也就是說，一位受試者會聽到八種語音情境：1種信噪比等級（3 dB或1 dB）x 2種噪音情境（引擎和街道雜音）x 4種語音增強情境（Noisy、MMSE、DDAE、FCN（S））。每個實驗情境包含10個語段，每個語段皆包含10個字。聽測所用的語料皆由男性語者所錄製。本研究執行前已通過中央研究院研究倫理委員會核可。

所有受試者都在安靜的環境下配耳罩式耳機（型號：Sennheiser HD565）接受測驗。所有聽覺刺激播放音量皆以60 dB為基準，並可依受試者要求調升或調降5 dB。測驗過程中，受試者被告知其任務為聆聽句子並在聽完後說出聽

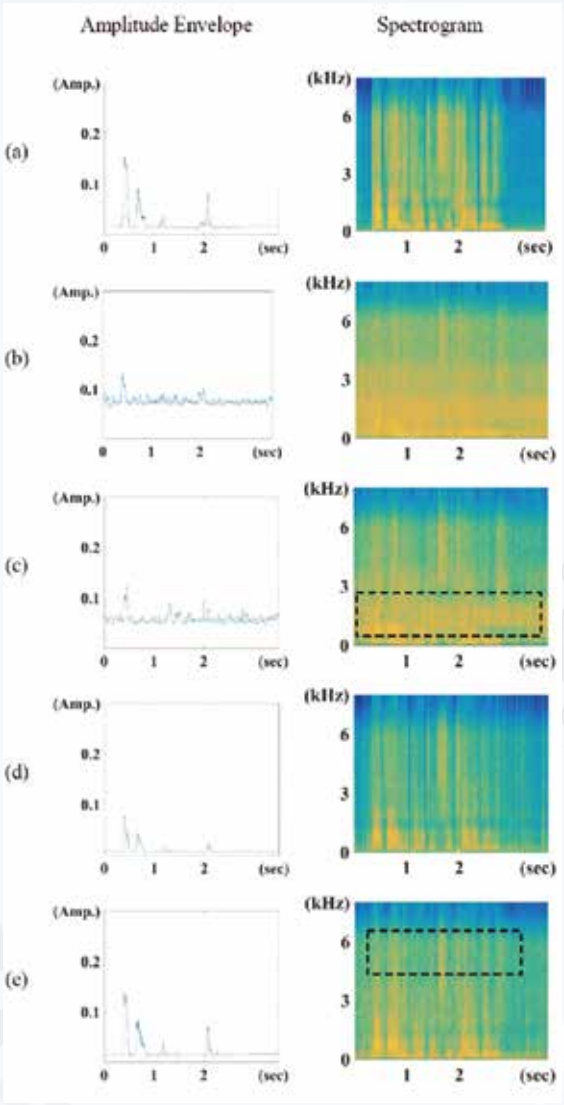

圖8、單一語段在各實驗情境下的振幅輪廓及頻譜:（a）乾淨語音、（b）噪音語音（街道雜音; -3 dB SNR）、以及經（c）MMSE、（d）DDAE、和（e）FCN（S）增強後的語音（Wang et al., 2020）。

到了哪些語音，部份語音可能包含雜訊，請務必專心且仔細聆聽，每個句子只能重複聽一次。此外，受試者也被告知實驗過程中總共會聽到80個句子。但受試者對於實驗情境並不知情。聽測為Matlab軟體所支援的電腦化測驗；由施測者經由實驗介面播放音檔，並紀錄各題答對字數，測驗結束時程式自動計算正確率（character correct rate, CCR）。正式測驗開始前所有受試者皆做過練習並熟悉實驗流程；練習階段包含五個句子，這五個句子並不會出現在正式聽測中。每個受試者在正式聽測中聽到的語句與實驗情境的搭配皆為隨機。聽測總時長約為45分鐘，測驗過程中受試者可隨時暫停並稍作休息。

各實驗情境皆包含10個句子，因此每個情境的CCR與原史分數相等，最高為100。圖9（a）為穩態噪音語音情境下{Noisy, MMSE, DDAE, FCN（S）}的聽測CCR和平均值標準誤差。信噪比為1 dB時，各模型聽測分數為{69.8 ± 1.7, 65.2 ± 1.3, 61.6 ± 1.3, 76.9 ± 1.7}；信噪比為-3 dB時，各模型聽測分數為{56.7 ± 2.5, 36.8 ± 2.7, 51.8 ± 2.1, 58.4 ± 2.5}。圖9（b）則為穩態噪音語音情境下{Noisy, MMSE, DDAE, FCN（S）}的聽測CCR和平均值標準誤差。信噪比為1 dB時，各模型聽測分數為{72.5 ± 1.7, 68.9 ± 2.1, 72.0 ± 2.2, 80.6 ± 1.4}；信噪比為-3 dB時，各模型聽測分數為{57.2 ± 2.3, 49.4 ± 2.3, 56.5 ± 2.1, 70.3 ± 2.5}。由圖9可觀察到FCN（S）在不同噪音類型和不同信噪比情境下都能有效地增強EAS語音；其CCR高於MMSE和DDAE。另一方面，DDAE的表現與MMSE相近，CCR分數皆與未增強的實驗情境相近，甚至在部份實驗情境中呈現低於基準線（noisy）的分數。而DDAE與MMSE的主要差異在於DDAE較能有效地移除非穩態噪音。此外，MMSE在信噪比低（-3 dB）的情境中，表現不如以深度學習為基礎的語音增強模型。

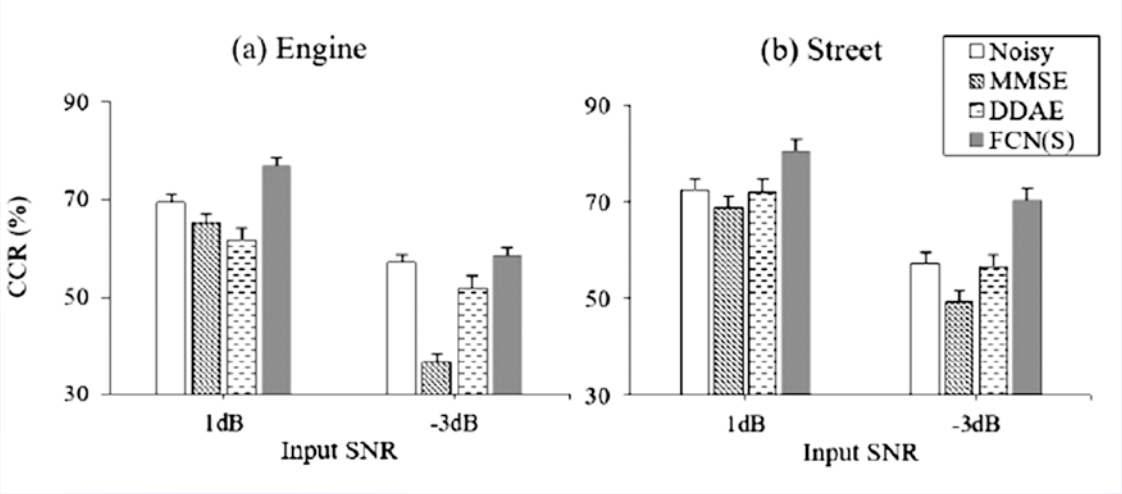

圖9、（a）穩態和（b）非穩態噪音語音的各實驗情境和其相對應的語音辨識分數，以CCR（%）為單位。誤差棒（error bar）為平均值標準誤差（Wang et al., 2020）。

為了進一步驗證FCN（S）模型是否優於MMSE和DDAE，我們將兩種噪音類型的CCR分數分別進行單因子變異數分析（One-way ANOVA），以直接比較三個模型在不同信噪比情境下的表現。單因子異數分析以各實驗情境的CCR做為依變項，自變項為語音增強模型類別。表1統整了分析結果：在信噪比低的非穩態噪音情境中，三種不同的語音增強方法有顯著差異，而FCN

（S）增強的語音清晰度顯著地高於DDAE和MMSE。相對的，在1 dB信噪比情境下，FCN（S）的表現雖然優於其它兩個模型，但其差異並未達到統計顯著。EAS聲碼器聽測的結果大致與一般語音增強的STOI分數走勢相同。

表1、單因子變異數分析及Boferroni事後檢定結果（Wang et al., 2020）。

test conditions	models	Mean	*df*	*F*-value	*p*-value	η^2	post-hoc comparisons
Engine (1dB)	noisy MMSE DDAE FCN(S)	69.4 65.11 61.6 76.6	(3, 104)	11.380	< 0.001	0.247	(FCN(S), DDAE); (FCN(S), MMSE); (FCN(S), Noisy); (DDAE, Noisy)
Engine (-3dB)	noisy MMSE DDAE FCN(S)	56.9 36.3 51.8 58.3	(3, 104)	19.121	<0.001	0.355	(FCN(S), DDAE); (FCN(S), MMSE) (DDAE, Noisy); (DDAE, MMSE) (MMSE, Noisy)
Street (1dB)	noisy MMSE DDAE FCN(S)	72.5 68.9 72.0 81.0	(3, 104)	2.782	0.045	0.074	(FCN(S), MMSE)
Street (-3dB)	noisy MMSE DDAE FCN(S)	57.2 49.4 56.5 70.3	(3, 104)	15.093	<0.001	0.303	(FCN(S), Noisy) (FCN(S), MMSE) (FCN(S), DDAE)

討論

本研究是第一個探討深度學習模型用於EAS聲碼器語音增強的研究。本研究專注於比較近期開發的FCN（S）語音增強模型與DDAE和MMSE的差異。其中，DDAE與FCN（S）同為以深度學習為基礎的語音強化模型，但FCN（S）打破過去以幅為單位的處理模式改以整段為單位，針對語音波形直接增強。因此，我們預期FCN（S）無論在一般語音增強或EAS模擬語音增強上都會有較佳的表現。而本研究中的客觀評估和聽測結果均顯示FCN（S）在各實驗情境下的表現皆優於MMSE和DDAE。而FCN（S）的優點可分為幾個面向討論：首先，FCN（S）在穩態及非穩態噪音語音增強上皆有優於其它兩者的表現，可見其適用的噪音情境可能較廣泛。此外，FCN（S）的增強效果在信噪比低的非穩態噪音情境下顯著優於DDAE和MMSE；可見FCN（S）更適用於處理極具挑戰性的噪音情境，而這類噪音正是日常生活中最常出現的噪音形態。最後，FCN（S）模型訓練過程中所需的參數較DDAE和MMSE來得少，但卻能達到更高的語音清晰度。這顯示了無論是要提升一般語音或EAS模擬語音的清晰度，FCN（S）都是一個效率和效果更好的選擇。

本研究討論了將深度學習語音增強模型運用於EAS系統的可行性。但仍有

不少需要進一步釐清的問題，而這些問題也反映出目前此研究領域還有許多不足之處。首先，本研究結果是基於EAS聲碼器進行模擬。將深度學習模型導入EAS系統中是否能有同樣的效果仍屬未知。再者，本研究中的語音處理速度有20毫秒的短暫延遲；然而，這樣的處理速度可能無法滿足助聽輔具需在10毫秒內即使處理的需求。因此，深度學習模型的處理速度也是有待解決的問題。最後，FCN（S）模型目前處理高頻雜訊的能力仍有待提升。以上幾點皆是未來應進一步研究的方向。

參考文獻

1. F. Chen and P. C. Loizou （2010a）. Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing. Ear Hearing, 31（2）, 259-267.
2. F. Chen and P. C. Loizou （2010b）. Speech enhancement using a frequency specific composite Wiener function. in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 4726-4729.
3. F. Chen and P. C. Loizou （2011）. Predicting the intelligibility of vocoded and wideband mandarin chinese. J. Acoust. Soc. Amer., 129（5）, 3281-3290.
4. F.-G. Zeng （2004）. Trends in cochlear implants. Trends Amplification, 8（1）, 1-34.
5. F. Seldran, S. Gallego, H. Thai-Van, and C. Berger-Vachon （2014）Influence of coding strategies in electric-acoustic hearing: A simulation dedicated to EAS cochlear implant, in the presence of noise. Appl. Acoust., 76, 300-309.
6. S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai （2018）. Raw waveform-based speech enhancement by fully convolutional networks. in Proc. APSIPA, 6-12.
7. S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao （2017）. Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery. IEEE Trans. Biomed. Eng., 64（11）, 2584-2594.
8. Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee （2017）. A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation. IEEE Trans. Biomed. Eng., 64（7）, 1568-1578.
9. Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee （2015）. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio, Speech, Language Process., 23（1）, 7-19.
10. N. Y.-H. Wang, H.-L. S. Wang, T.-W. Wang, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao （2020）. Improving the Intelligibility of Speech for Simulated Electric and Acoustic Stimulation Using Fully Convolutional Neural Networks. IEEE Transactions on Neural Systems & Rehabilitation Engineering, 29, 184-195