The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 212-225

使用生成對抗網路於強健式自動語音辨識的應用

Exploiting Generative Adversarial Network for Robustness Automatic Speech Recognition

楊明璋 Ming-Jhang Yang, 趙福安 Fu-An Chao, 羅天宏 Tien-Hong Lo,
陳柏琳 Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{60547076S, 60747002S, teinhonglo, berlin}@ntnu.edu.tw

摘要

在過去幾年中，深度學習技術的發展在許多領域中大放異彩，應用在語音辨識中也一樣表現優異。儘管語音辨識有了大幅度的改進，然而「雜訊」仍然一定程度的干擾語音辨識之準確度。諸如:背景人聲、火車、公車站牌、汽車噪音、餐館背景雜音…以上皆為易影響語音辨識結果的環境噪音。因此語音辨識的強健性技術研究仍扮演著重要角色。過往於強健性技術的研究主要可區分為以特徵為基礎，以及以模型為基礎兩大面向。以特徵為基礎的強健性技術又可分為特徵正規化以及語音訊號增益。本研究主要採用生成對抗網路(Generative Adversarial Network , GAN)以語音訊號增益方式使用在調變頻譜特徵上。我們的目的在於把受到吵雜環境干擾，或被通道效應破壞的語音特徵轉換成接近乾淨環境下錄製之語音特徵，此方法比起原始梅爾倒頻譜係數特徵可以有效的提升辨識率。

Abstract

In the recent past, deep learning techniques have reached record-breaking performance in a wild variety of applications like automatic speech recognition (ASR). Even though cutting-edge ASR systems evaluated on a few benchmark tasks have already reached human-like

performance, they, in reality, are not robust, in the manner that humans are, to disparate types of environmental noise such as babble, train, bus station, car driving, restaurant, and among others. In view of this, this paper embarks on an effort to develop effective enhancement methods, stemming from the so-called generative adversarial networks (GAN), for use in the modulation domain of speech feature vector sequences. A series of experiments conducted on the Aurora-4 database and task seem to demonstrate the practical merits of our methods.

關鍵詞：生成對抗網路、語音訊號增益、語音強健性技術、強健性語音辨識

Keywords: Generative Adversarial Network, Speech Enhancement, Robustness Techniques, Robust Speech Recognition

一、緒論

在自動語音辨識技術（Automatic Speech Recognition, ASR）的發展中，我們發現環境噪音會大幅度的影響辨識率。因此，為了降低噪聲的影響以及提升辨識率，強健性語音辨識的發展便應運而生。目前強健性語音辨識技術大致可以分為特徵為基礎(Feature-Based)以及模型為基礎之方法(Model-Based)兩種。前者著重於特徵正規化(Feature Normalization)以及特徵增益(Feature Enhancement)兩方面。後者則主要專注在改良聲學模型上，將其「加深」、「加廣」以及其他特殊訓練方法用於提升語音辨識的強健性效果。

時至今日，已有多種新穎之強健性技術可以為語音辨識帶來更好表現，其中在特徵處理方法中有多項採用調變頻譜分析的研究指出在頻率較低之 4Hz 附近存在諸多語意資訊 [1]，而這將有助於提升語音辨識的效果。因此諸多調變頻譜正規化的研究便由此而生 [2] [3] [4]。從以上研究得到啟發，學者發現探索語音特徵的子空間結構可以得到更佳語音辨識效果 [5]，故萌生了子空間學習的概念。依循著這一個脈絡，目前子空間學習已發展出: 字典學習法結合稀疏編碼(Sparse Coding) 、低序表示法(Low Rank Representation, LRR) 等主要方法。

除了語音特徵正規化之外，語音訊號增益法(Speech Enhancement)亦是一種強健性方法。部分採用此作法的研究直接針對語句之波形圖濾波。另一部份則採用深度學習技術生成接近無干擾的語音特徵，例如導入自動編碼器 [6] 用來抑制噪聲干擾。做為新起之秀的生成對抗網路(Generative Adversarial Networks, GAN) [7]也可以作為語音訊號增益的手段，其自動生成與鑑別是否正確的功能在增益特徵強健性上被認為有助益。雖然最初設計是用於影像處理，但是目前已有多種變形應用於語音強健性研究上，著名研究有 SEGAN[8]、Whispered-to-voiced GAN[9]、RSRGAN[10]以及 FSEGAN [11]。這類方法除了可以處理波形圖外亦可處理時域(Time Domain)特徵和頻率域(Frequency Domain)特徵。本研究即是從上述生成對抗網路方法中得到啟發，導入風格轉換概念，並結合調變頻譜相關研究想法，以生成對抗網路處理頻率域特徵，關於詳細方法將於後續章節依序介紹。我們發現使用生成對抗網路處理調變頻譜特徵，可以使語句的特徵分布更接近未受干擾語句，從而提升語音辨識效果。

二、文獻回顧

以處理語音特徵為基礎的強健性技術，目的在於不需要重新設計聲學模型，透過語音訊號增益、特徵向量補償、頻譜補償或正規化等方式還原出乾淨完整的語音特徵。語音訊號增益(Speech Enhancement)技術之目的在於增強語音訊號的可讀性和品質，以便在包含雜訊的情況下可以順利被聽懂亦可用於進行語音辨識 [12]。遮罩與濾波技術是一種很直觀的訊號增益方法，維納濾波就是著名的方法之一。維納濾波器之概念於 1949 正式出版於數學家諾伯特·維納（Norbert Wiener）的著作中 [13]，是一種採用最小化平均誤差(Mean Square Error, MSE)當作最佳化函數的線性濾波器(Linear Filter)，也就是說:在給定約束條件下計算濾波器輸出與期望的輸出之間的平方誤差之最小值，便是維納濾波器的核心概念。簡而言之:目的在於使得經過濾波後的訊號能盡量的接近未受干擾的真實訊號，主要運算可以由方程式(1)表示，其中 ***s(t)***是我們要估計的原始訊號，***n(t)***代表疊加的雜訊，輸入訊號由 ***s(t)***和 ***n(t)***組成和濾波器 ***g(t)***進行摺積運算後得到濾波後的訊號 ***x(t)***:

$$\boldsymbol{x}(\mathrm{t}) = \boldsymbol{g}(\mathrm{t}) * \big(\boldsymbol{s}(\mathrm{t}) + \boldsymbol{n}(\mathrm{t})\big), \tag{1}$$

隨著時光荏苒，在諾伯特·維納之後又有諸多學者以濾波器為題，提出多種可以有效增益語音訊號之遮罩方法 [14] [15]。直接遮罩(Direct Masking) [16]，很直觀的採用$\lambda_{t,f}^{(S)}$作為一個遮罩用來增益嘈雜語音訊號$Y_{t,f}$，經過 [34]改寫，增益後的訊號可以由下列方程式表示:

$$\hat{S}_{t,f} = f_{\theta DM}^{(DM)}(Y,t,f) = \lambda_{t,f}^{(S)} \ dot Y_{t,f}, \tag{2}$$

然而直覺遮罩往往會帶來大量失真，因此便有學者採用由維納濾波器改良成的帶參數維納濾波(Parametric Wiener Filter, PW)來解決這個問題，PW 在抑制噪訊與控制失真這兩者的權衡之間擁有更多靈活性 [17] [18] [19]，PW 具體運算可表示如下:

$$\hat{S}_{t,f} = f_{\theta pw}^{pw}(Y,t,f) = \left|\frac{|Y_{t,f}|^p - l \ .|\widehat{N}_{t,f}|^p}{|Y_{t,f}|^p}\right|^{1/q}.Y_{t,f}, \tag{3}$$

另一方面，除了濾波與遮罩法之外，近年導入深度學習技術，也為語音訊號增益帶來新的想法。我們知道神經網路可以學習與調整一群資料的分布，因此深度學習技術在語音訊號增益中的用途大多以映射為主，將嘈雜訊號映射成乾淨語句的分布便可以達到我們的目的。以下讓我們回顧經幾個經典作法，雖然他們的目的在於消除殘響(Reverberation)，但是仍然對我們想要抑制疊加雜訊干擾以及抑制通道摺積效應有莫大啟發。

以深度遞歸神經網路(Regression Neural Network)做為主要結構 [20]，首先採用乾淨語句提取出的對數功率譜(Log-Power Spectrum, LPS)當作特徵，用以訓練深度學習模型。如此一來模型將學習到乾淨語句的特徵分布，接著輸入嘈雜語句的特徵便可以輸出增益

過後接近原始乾淨語句的語音特徵，再由此重構出波形圖及原始聲音，就可以得到還原後的語句。 [21]之研究發現將語音特徵由 LPS 映射到 MFCC，可以更進一步改善語音辨識結果，由此現象可以得知深度學習技術也可以學習不同特徵之間轉換的對應關係。

上述研究多關注在以深度神經網路生成特徵，但是其估算出的特徵是否夠接近我們的預期呢?這點除了在整體計算完畢後進一步評估語音訊號品質如(Perceptual Evaluation of Speech Quality, PESQ)或是語音辨識結果之詞錯誤率(Word Error Rate, WER%)之外，我們在訓練的同時並無從得知，也無法同時最佳化模型參數，因此強健性效果有可能會大打折扣。此時，導入生成對抗網路中鑑別器(Discriminator)的概念便是一個新的契機，將原本用來生成特徵的模型當作生成器，同時另外訓練一個網路當作鑑別器，由鑑別的結果自動更新生成器的參數，就可以更周全的訓練模型。

三、生成對抗網路應用於語音強健性技術

本研究主要採用生成對抗網路(Generative Adversarial Network, GAN)結合調變頻譜特徵進行語音訊號增益處理，並結合自動語音辨識(Automatic Speech Recognition, ASR)用來達成增進強健性表現之目的。本節將針對本研究使用的生成對抗網路模型與方法進行討論。

生成對抗網路是一種可以減少人類知識介入，而得到更佳學習效果的一種深度學系技術，這項關鍵就在於「生成」與「鑑別」，也有人稱之「新手畫家」 與「鑑賞家」。在訓練過程中，新手畫家不斷臨摹名畫，而鑑賞家持續鑑定畫作，在兩造交手若干次之後，新手畫家有了弄假成真的本事，而鑑賞家漸漸分不出贋品與真品，這便是我們的目的。這樣子對抗學習的過程可以看成是生成器(G)與鑑別器(D)在進行最大與最小值的對局(Minimax-Game)。本研究採用 LSGAN 中的方均根誤差(Mean Square Error, MSE)當作損失函數，在訓練過程中我們需要將 G 與 D 串接起來，因此整體目標函數可以改寫成:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[\log D(\boldsymbol{x})] + \mathbb{E}_{\hat{x}\sim P_{\hat{x}}(\hat{x}),z\sim P_z(z)}\left[log\left(1-D\left(G(\hat{\boldsymbol{x}}+\boldsymbol{z})\right)\right)\right], \quad (4)$$

我們用 $\boldsymbol{x}$ 表示乾淨情境樣本(Clean Condition)，用$\hat{\boldsymbol{x}}$表示噪訊情境樣本(Noisy/Multi Condition)，再這裡 $\boldsymbol{z}$ 為一組 Latent Vector，將其設為介於 0 到 1 之間的隨機雜訊。於訓練時將噪訊樣本加上隨機雜訊輸入 G 使其盡可能有能力把雜亂資料轉換成我們期望的乾淨樣本。

一個完整的生成對抗網路之訓練步驟主要可以分為三個階段: (1)訓練鑑別器認識真實樣本(2)訓練鑑別器認識生成器生成之假樣本(3)固定鑑別器的參數，同時更新生成器參數以達成訓練目標。

圖 1: 生成對抗網路-鑑別器訓練-1

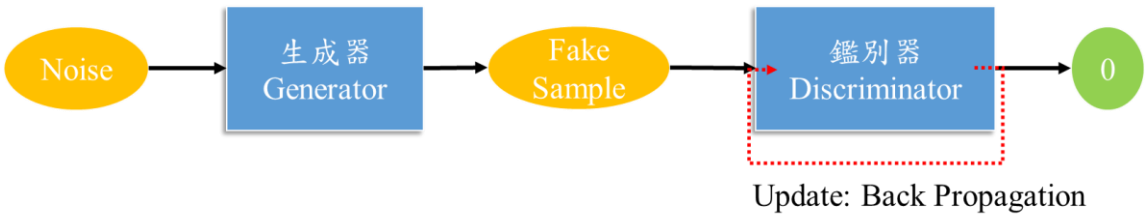

圖 2: 生成對抗網路-鑑別器訓練-2

圖 3: 生成對抗網路-生成器訓練

以上為 GAN 之訓練方法及其示意圖，接著本段將介紹本研究使用之神經網路架構。我們採用兩種不同結構，分別命名為 CAGAN 以及 DNN-LSGAN，前者採用類似於摺積自動編碼器(Convolution Auto Encoder, CAE)作為生成器的主要結構，後者則採用全連接 DNN 作為生成器的主結構，並以此為原則命名。啟發於多項類似於自動編碼器(Auto Encoder, AE)與 GAN 結合的研究[29][22]，加上目前普遍認為 CNN 在學習時間-頻率特徵或圖像的能力比起 DNN 還有更好效果。而我們以調變頻譜特徵作為輸入，在頻率域上進行訊號增益，概念類似電腦視覺領域中於影像降噪的研究，也就是說我們以摺積運算結合自動編碼器當作一項取得強健性特徵的方法。

在消除噪訊干擾效應的深度學習技術中，降噪自動編碼器(Denoise Autoencoder, DAE)與摺積自動編碼器(Convolutional Autoencoder, CAE)是很有效方法。可以輸入被噪訊破壞的原始資料，並還原出未受干擾的資料。而本研究生成對抗網路方法之一就是受到他們啟發，CAGAN 之生成器就是類似於摺積自動編碼器的結構。

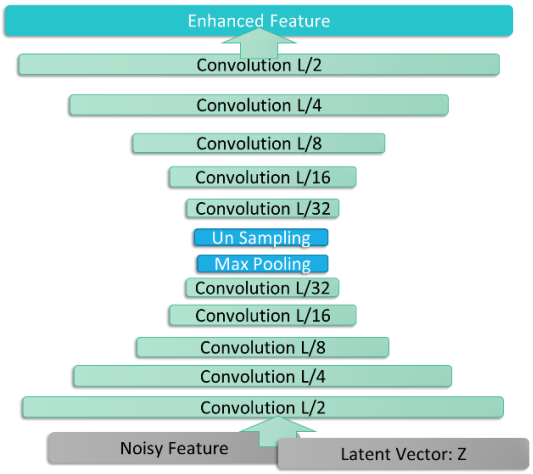

圖 4: CAGAN-生成器結構

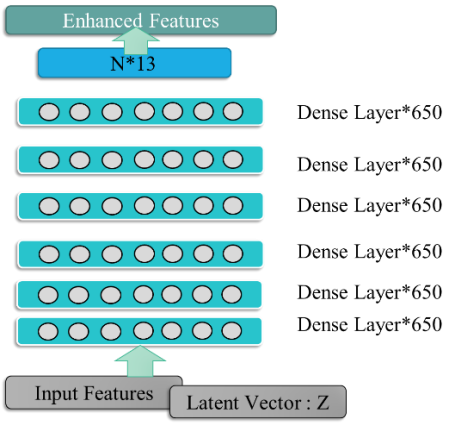

圖 5: DNN-LSGAN-生成器結構

我們知道自動編碼器可以分成編碼階段與解碼階段兩大部分。在編碼階段，隨著深度增加，我們將摺積層的特徵圖(Feature Map)大小減半，並在每一次摺積運算後進行池化(Max Pooling)，目的在於將有效的特徵往下傳遞並且減少不必要的網路參數，使之更方便訓練。在解碼階段，其結構可以視為將編碼階段水平鏡射的對稱關係，唯一不同處在於相對於最大池化法(Max Pooling)，我們在每一次摺積運算之後採用反取樣法(Up-Sampling)，將維度還原成原始大以利進行後續語音辨識步驟。在訓練時我們採用嘈雜環境語料結合隨機雜訊作為輸入資料，其結構圖 4 所示。

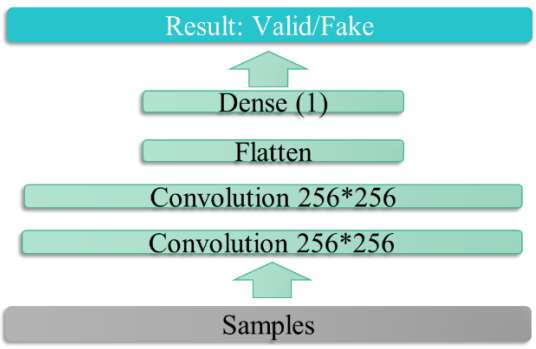

圖 6: CAGAN 之鑑別器

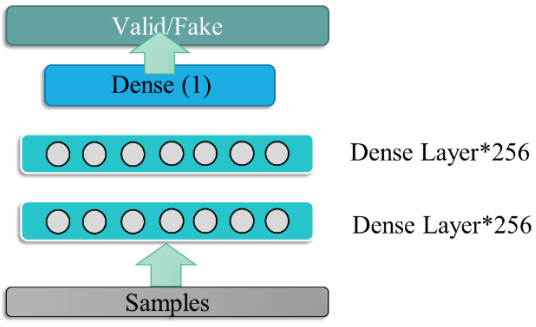

圖 7: DNN-LSGAN 之鑑別器

此外，顧慮到摺積運算比起全連接 DNN 需要更多運算資源，因此也採用以 DNN 結構和 LSGAN 為基礎的 DNN-LSGAN 作為強健性方法之一。為了減輕運算量，我們

捨棄摺積運算，總共使用六層全連接 DNN 網路作為生成對抗網路的生成器(G)，如圖 5。在上述兩個生成對抗網路中，我們分別搭配不同的鑑別器使用，如圖 6 及圖 7 所示。

四、實驗與討論

本研究採用 Aurora-4 作為實驗語料庫，收錄了華爾街日報(Wall Street Journal, WSJ)之朗讀發音，並包含-5dB 至+15dB 的雜訊。簡而言之 Aurora-4 是以華爾街日報為基礎錄音並加上 6 種不同情境下的噪音來組成的，是一個專門設計用來從事語音強健性技術研究的語料庫。其中包含 8KHz 與 16KHz 兩種音頻取樣率並且採用兩種麥克風錄音(Sennheiser, Secondary-Mic)。其訓練資料集可分為無雜訊干擾(Clean-Condition)，和多情境雜訊混合(Multi-Condition)兩種，測試集則包含的 6 種雜訊，分別包含 330 個發音，其種類如下:人聲(Babble)、汽車(Car)、機場(Airport)、火車(Train)、街道(Street)、餐廳(Restaurant)。另一方面，將測試集分為 A、B、C、D 四種子集合，其詳細介紹如表: 1 所示，此外本研究均採用 16Khz 作為取樣率。 語音特徵部分我們採用 13 維 MFCC，聲學模型部分我們採用 5 層 TDNN-F 網路，每一層 650 維，並將瓶頸層設為 128 維，並訓練 8 個 epoch。

表: 1　Aurora4 簡介

取樣率	8kHz/16kHz
語音內容	WSJ 5000 詞
長度	約 15 小時，每一句約 5~12 秒鐘
訓練資料	Clean:7138 個語句 Multi:7137 個語句
測試資料	A 組:330 個無雜訊語句
	B 組:1980 個語句，包含六種環境噪音
	C 組: 受通道效應干擾的 330 個無雜訊的語句
	D 組: 受通道效應干擾的 1980 個包含雜訊的語句

表: 2 實驗結果

Clean Condition Training (WER%)					
	A	B	C	D	AVG
MFCC+TDNN-F	3.61	41.96	33.18	60.01	34.69
WAV+ SEGAN+TDNN-F	4.15	35.20	40.84	55.28	33.86
Modulation Spectrum +LSGAN+TDNN-F	10.29	34.11	23.03	47.48	28.73
Modulation Spectrum +CAGAN+TDNN-F	6.91	30.17	20.08	42.46	**24.95**

本研究比較經典的生成對抗網路方法應用於強健式語音辨識的效果，以及本研究設計的其餘結合生成對抗網路與語音訊號增益法應用於強健式語音辨識的方法。此外，由於 TDNN-F 為新穎的聲學模型，其除了擁有考慮時間資訊的功能之外，也能夠透過因子分解捨去不必要資訊，使整體模型更容易訓練，因此我們以 TDNN-F 做為其餘實驗的 ASR 系統。

從表: 2 中，我們可以看出使用生成對抗網路出直接作用於波形圖上的增益方法，雖然有些微效果，但是對整體語音辨識率的幫助有限。這是由於波形圖中包含太多資訊，並不是所有都有助於我們了解一句話的語意。但是此方法確實可以幫助改善人類聽覺的效果，不過對 ASR 系統幫助有限。所以在 ASR 中，我們如果提取出特徵，再行進一步處理，將會有更好效果。於是本研究之 CAGAN 與 DNN-LSGAN，即採用 MFCC 特徵轉換成調變頻譜特徵，並取出其中之強度頻譜，利用深度學習技術中的映射功能進行訊號增益還原出乾淨語句的特徵，同時也由其他研究得到啟發重新設計 GAN 內部的網路

結構，使其針對我們的任務有更好表現。未來希望能夠再結合傳統強健性技術，使本研究之方法在語音辨識任務下能有更佳效果。

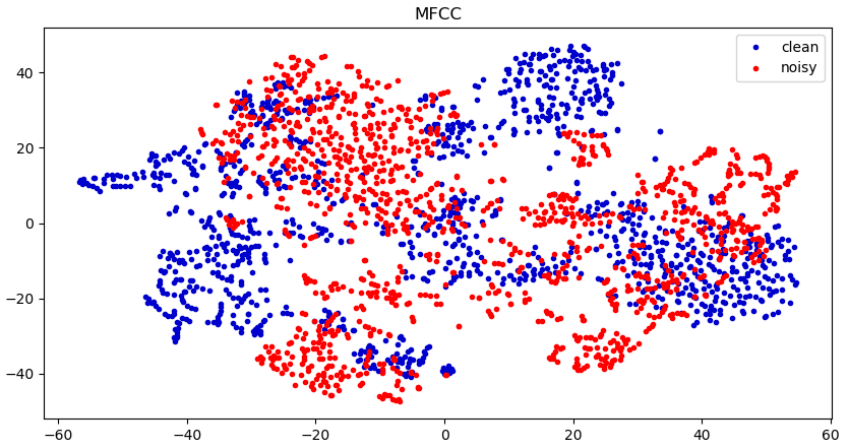

圖 8: MFCC 特徵分布圖

經過 T-SNE 降維後，可將多維度資料投影至某一平面上，方便我們觀察。我們以視覺化方式來討論 MFCC 之特徵分布以及經過強健性技術處理後之差異。乾淨語句與受噪聲干擾語句之 MFCC 分布圖如圖 8 所示，由此可以看出噪聲干擾的確扭曲了語音特徵的分布結構。

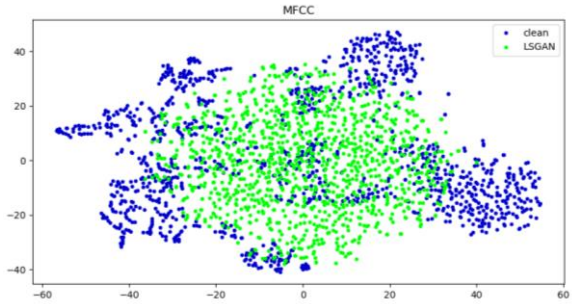

圖 9: DNN-LSGAN 處理後之 MFCC 分布

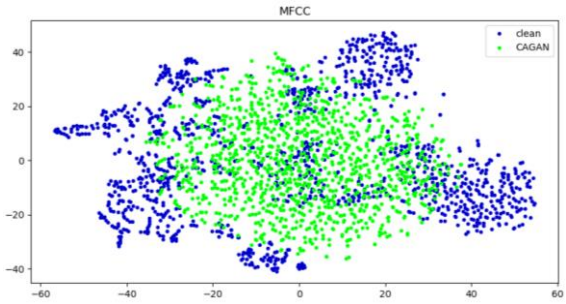

圖 10: CAGAN 處理後之 MFCC 分布

由圖:9 與圖:10 我們可以觀察出一些現象，經過生成對抗網路增益後的語句特徵大幅度的調整了受噪聲干擾的分布，使其與乾淨語句的特徵分布有較多靠近與重疊處。雖

然不尚明顯，但是經過 DNN-LSGAN 與 CAGAN 處理後之特徵分布有較接近乾淨語句分布的趨勢，以上現象也將反映在語音辨識效果上。

五、結論

本研究探討強健式語音辨識的新穎技術，並採用 Aurora-4 語料庫做為實驗基礎，用以比較語音強健技術基於生成對抗網路方法之下對於自動語音辨識的影響。由於調變頻譜可以呈現語音特徵更大尺度變化，所以我們就順著這個脈絡，由調變頻譜特徵著手研究並比較時域特徵與頻率域特徵運用於在語音訊號增益方法上對於提升語音辨識效果的幫助。此外，生成對抗網路的一大特色就是可以自動鑑別生成特徵準確與否，以往多運用在影像處理與電腦視覺領域研究中，用來轉換不同風格圖片，或是用來將含有噪訊之影像映射成特定類型之乾淨影像。本研究由此得到啟發，採用生成對抗網路作為一種語音強健性技術。本研究主要運用生成對抗網路來實現訊號增益方法，從 CAGAN 與 DNN-LSGAN 的實驗中，我們發現在調變頻譜上應用訊號增益方法比起其他媒介更能夠有效提升語音辨識率的效果。與原始 MFCC 比較本研究之方法可分別降低 5.96(WER%)與 9.74(WER%)。

未來希望除了強度頻譜之外，採用相位頻譜也能亦或是結合傳統強健性方法也能成為研究方向之一。此外，由於本研究主要探討以特徵為基礎的強健性方法較少關注以模型為基礎的強健性技術。未來也可採用資料增強法(Data Augmentation)用來增加訓練資料的變異度以及加入更多種模擬雜訊，使用多情境訓練方式來訓練聲學模型，使聲學模型可以學習到更多種情境的資訊，也可以大幅降低訓練與測試的環境不匹配問題，從而大幅提升語音辨識效果。因此，未來希望不只是專注在特徵上，雖然模型方法的缺點在於需要更多計算量，但是隨著硬體運算技術進步，我們可以將精神轉移到資料增強法和模型調適方法上，或許能有更多突破，並且能使研究更具實用價值。

致謝

本論文之研究承蒙行政院科技部研究計畫 (MOST 105-2221-E-003-018-MY3 和 MOST 107-2221-E-003-013-MY2、MOST 108-2221-E-003-005-MY3 和 MOST 108-2634-F-008 -004 -) 之經費支持，謹此致謝。

參考文獻

[1] 汪逸婷, “運用調變頻譜分解技術於強健語音特徵擷取之研究,” *國立臺灣師範大學碩士論文*, 2014.

[2] 朱紋儀, “調變頻譜正規化用於強健式語音辨識之研究,” *國立臺灣師範大學 碩士論文*, 2011.

[3] 張庭豪, “調變頻譜分解之改良於強健性語音辨識,” *國立臺灣師範大學 碩士論文*, 2015.

[4] 顏必成 石敬弘 劉士弘 陳柏琳, “使用字典學習法於強健性語音辨識 The Use of Dictionary Learning Approach for Robustness Speech Recognition,” 於 *ROCLING, ACLCLP*, 2016.

[5] Bi Cheng Yan, Chin Hong Shih, Shih Hung Liu, Berlin Chen, "Exploring Low-Dimensional Structures of Modulation Spectra for Robust Speech Recognition," in *INTERSPEECH*, 2017.

[6] Pierre Baldi, "Autoencoders, Unsupervised Learning, and Deep Architectures," in *JMLR: Workshop and Conference Proceedings*, 2012.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Networks," in *NIPS*, 2014.

[8] Santiago Pascual, Antonio Bonafonte, Joan Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *INTERSPEECH*, 2017.

[9] Santiago Pascual, Antonio Bonafonte, Joan Serrà, Jose A. Gonzalez, "Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks," in *arXiv*, 2018.

[10] Wang, Ke and Zhang, Junbo and Sun, Sining and Wang, Yujun and Xiang, Fei and Xie, Lei, "Investigating Generative Adversarial Networks based Speech Dereverberation for Robust Speech Recognition," in *INTERSPEECH*, 2018.

[11] Chris Donahue, Bo Li, Rohit Prabhavalkar, "Exploring Speech Enhancement With Generative Adversarial Networks," in *ICASSP*, 2018.

[12] P.C.Loizou, Speech Enhancement: theory and Practice, Boca Raton, FL, USA: CRC Press, 2013.

[13] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series, NewYork: WILEY, 1949.

[14] Jahn Heymann, Lukas Drude, Aleksej Chinaev, Reinhold Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *ASRU*, 2015.

[15] Tobias Menne, Ralf Schlüter, Hermann Ney, "Speaker adapted beamforming for multi-channel automatic speech recognition," in *SLT*, 2018.

[16] Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, Shinji Watanabe, "Building state of the art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *INTERSPEECH*, 2018.

[17] Tobias Menne, Ralf Schluter, Hermann Ney, "INVESTIGATION INTO JOINT OPTIMIZATION OF SINGLE CHANNEL SPEECH ENHANCEMENT AND ACOUSTIC MODELING FOR ROBUST ASR," in *ICASSP*, 2019.

[18] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang, Springer Handbook of Speech Processing, Berlin Heidelberg: Springer-Verlag, 2008.

[19] JAE.S. Lim, ALAN.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE,* no. 67,no12, p. 1586–1604.

[20] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters,* vol. 1, no. 21, p. 65–68, 2014.

[21] Kun Han, Yanzhang He, Deblin Bagchi, Eric Fosler-Lussier, DeLiang Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *in Sixteenth Annual Conference of the International Speech Communication Association*, 2015.