Computational Linguistics and Chinese Language Processing
Vol. 25, No. 2, December 2020, pp. 85-98

基於深度聲學模型其狀態精確度最大化之強健語音特徵擷取的初步研究

The Preliminary Study of Robust Speech Feature Extraction based on Maximizing the Accuracy of States in Deep Acoustic Models

張立家*、洪志偉*

Li-Chia Chang and Jeih-weih Hung

摘要

在本研究中，我們提出一種新穎的強健性語音特徵擷取技術，以增進雜訊干擾環境下的語音辨識效能。此新技術利用語音辨識系統中後端的原聲學模型所提供的資訊，在不重新訓練聲學模型的前提下，藉由深度類神經網路架構，學習得到最大化聲學模型狀態之精確度對應的語音特徵，進而使此語音特徵擁有對雜訊的強健性，相較於其他改善聲學模型以達到雜訊強健性的技術，本研究所提出的新技術具有計算量小且訓練快的優點。

在初步實驗中，我們使用了 TIMIT 此中型語料庫來評估，實驗結果顯示所提之新語音特徵擷取法，相對於基礎實驗，能有效地降低各種雜訊種類與雜訊程度之環境下語音的音素錯誤率，凸顯此方法的效能及發展價值。

Abstract

In this study, we focus on developing a novel speech feature extraction technique to achieve noise-robust speech recognition, which employs the information from the backend acoustic models. Without further retraining and adapting the backend acoustic models, we use deep neural networks to learn the front-end acoustic speech feature representation that can achieve the maximum state accuracy

*國立暨南國際大學電機工程學系
Department of Electrical Engineering, National Chi Nan University
E-mail: s108323518@mail1.ncnu.edu.tw; jwhung@ncnu.edu.tw

obtained from the original acoustic models. Compared with the robustness methods that retrain or adapt acoustic models, the presented method exhibits the advantages of lower computational complexity and faster training.

In the preliminary evaluation experiments conducted with the median-vocabulary TIMIT database and task, we show that the newly presented method achieves lower word error rates in recognition under various noise types and levels compared with the baseline results. Therefore, this method is quite promising and worth developing further.

Keywords: Noise-robust Speech Feature, Speech Recognition, Deep Learning

關鍵詞：雜訊強健性之語音特徵、語音辨識、深度學習

1. 緒論 (Introduction)

在現今的時代，電子產品和相關服務，例如手機、助聽器、耳機、電話會議系統等，在我們的生活中逐漸變成日常且不可或缺的需求，在這些設備和服務中，語音的功能和應用（語音互動，語音通話，語音辨識等）是一個相當重要的環節。然而，現實通訊環境中中存在各種干擾源，而這些干擾源會干擾語音訊號，因此減損了上述語音功能和應用的性能。這些干擾源包括加成性雜訊、通道干擾和混響等，近幾十年來，各方已經研究開發出多種技術來降低這些干擾效應、以改進語音相關的功能。沿此方向，在本研究中，我們著眼於語音訊號中加成性雜訊的干擾的問題，提出了一種專門用於提高語音辨識準確度的新型降噪方法。

語音通話和辨識中，環境雜訊的存在很可能反映在語音訊號的品質和能辨度以及辨識的準確度上。為了降低雜訊的影響，研究者從語音處理系統的不同角度提出了多種方法，例如前端訊號處理(front-end signal processing)，聲學特徵擷取(acoustic feature representation)和後端聲學模型(back-end acoustic model)等。

在聲學特徵擷取方法中，例如，相對頻譜法(relative spectral analysis, RASTA) (Hermansky & Morgan, 1994)設計一個基於聲學知識的無限脈衝響應濾波器，其應用在語音訊號的對數頻譜中可以有效的抑制訊號中非語音的成分，著名的 RASTA-PLP (Hermansky, Morgan, Bayya & Kohn, 1991)語音特徵表示就是將感知線性估計(perceptual linear prediction, PLP) (Hermansky, 1990)的語音特徵經由 RASTA 處理。此外，對語音特徵序列進行不同階級的正規化可有效地減輕訓練與測試資料的統計不匹配，而研究中也指出藉此可以同時降低雜訊的影響，相關的方法包括平均正規化法(mean normalization, MN) (Liu, Stern, Huang & Acero, 1993)、正規化法(mean and variance normalization, MVN) (Viikki & Laurila, 1998)和統計圖等化法(histogram equalization, HEQ) (Torre *et al.*, 2005)，上述方法分別對語音特徵正規化了平均、平均與變異數、機率密度函數。

在後端聲學模型中，聲學模型之調適法旨在調整聲學模型去適應嘈雜環境下的輸入語音特徵，一些知名的方法例如最大後驗機率自適應模型(maximum a posteriori

adaptation, MAP) (Su, Tsao, Wu & Jean, 2013)、最大似然線性回歸(maximum likelihood linear regression, MLLR) (Stolcke, Ferrer, Kajarekar, Shriberg & Venkataraman, 2005)與最大似然線性轉換(maximum likelihood linear transformation, MLLT) (Gales, 1998)應用於聲學模型的參數（例如高斯混和模型的平均與共變異數）並進行映射轉換。此外，鑑別式聲學模型同樣有非常好的效果，其透過設計訓練時的目標函式，以達到直接改善辨識時的準確率，例如，在單語句辨識中，最小化分類錯誤(minimum classification error, MCE) (Juang, Hou & Lee, 1997)的目標函式是最佳化其辨識的分類結果，而非只是模擬輸入的語句；在大辭彙連續語音辨識中，最小化音素錯誤(minimum phone error, MPE) (Povey, 2003)和最小化詞錯誤(minimum word error, MWE) (Kuo & Chen, 2005)所得的聲學模型訓練目標是對訓練資料能最小化對辨識錯誤的平滑估計。

由於近年來深層神經網絡(deep neural network, DNN)技術的蓬勃發展，語音處理的前後端方法都得到了顯著改善和進步，從而獲得更好的效能，例如，在語音強化領域中，可以使用 DNN 透過大量成對的雜訊語音與乾淨語音，來學習二者之間的映照(mapping)關係；在聲學特徵擷取與聲學模型方面，DNN 同樣也部份甚至全面地取代了傳統的方法，例如，ANN-HMM(artificial neural network-hidden Markov model) (Bourlard & Morgan, 1994)的雙向系統藉由 ANN 更精確地估計出語音特徵的似然分數，此外，TANDEM 系統(Hermansky, Ellis & Sharma, 2000)訓練 DNN 產生語音特徵的後驗機率，並將其作為額外的資訊去訓練傳統的聲學模型，研究中也指出此方法可使模型具有更好的強健性，同樣地，瓶頸特徵(bottleneck feature)技術(Grezl, Karafiat, Kontar & Cernocky, 2007)擷取了 ANN 的中間特徵作為傳統聲學模型的輸入，也可以有效地提高其辨識準確度。

特別一提的是，由於語音強化或是嘈雜語音特徵的映射(Han, He, Bagchi, Fosler-Lussier & Wang, 2015)旨在將雜訊語音訊號或是其特徵轉換回受雜訊干擾前的原始值，這是機器/深度學習中典型的回歸(regression)問題，因此，在其方法中 DNN 的訓練經常使用均方誤差(mean squared error, MSE)作為損失函數、藉由其最小化來學習 DNN 模型參數。然而，在評估方法的性能時，通常會使用其他一些客觀的指標，例如語音品質的感知評估(perceptual evaluation of speech quality, PESQ) (Rix, Beerends, Hollier & Hekstra, 2001)、短時客觀能辨度(short-time objective intelligibility, STOI) (Taal, Hendriks, Heusdens & Jensen, 2010)或詞錯誤率(word error rate, WER)。這些評估分數不一定與還原後的語音和原始語音之間的均方誤差(MSE)有直接的相關，亦即 DNN 訓練目標與評估指標並不一致，因此降低 MSE 未必可直接提升這些評估分數。有鑑於此，在一些近年開發的基於深度學習的語音強化法中(Zhang, Zhang & Gao, 2018)，直接將 PESQ 和 STOI 作為 DNN 模型訓練的目標函數、加以最佳化，而獲得更好的效能。

受上述觀察和其他文獻的啟發(Fu, Liao, Tsao & Lin, 2019; Xia & Bao, 2014)，本研究提出一種基於深度學習模型之強健性特徵擷取的新方法，其利用了與 MSE 無關的目標函數來訓練其中的深度網路。在此新方法中，我們使用的目標函數是給定語音特徵序列下，對應的聲學模型其中狀態序列(state sequence)與真實狀態序列相較之下的精確度，與語音識別的精確度有直接的相關。簡而言之，我們訓練一個深度神經網絡來進行語音特徵映

射，用以最大化地提高語音辨識系統中後端聲學模型狀態的後驗機率(posterior probability)。我們預期得到的新語音特徵在辨識準確度方面將優於原始特徵，並且具有對雜訊的強健性。

在以下章節中，我們介紹新提出的語音特徵擷取方法，並探討它的特點與可能的優勢。然後進行實驗與分析結果。而後以結論作終。

2. 基於最大化狀態精確度的語音特徵擷取法 (Proposed Method: Feature Extraction based on Maximizing State Accuracy)

在本研究中，我們提出了一種使用深度學習之強健語音特徵擷取法，這種方法目的在於增強原始語音辨識系統中，聲學特徵的抗噪能力，但毋需更改其後所使用的聲學模型。該方法在訓練其中的深度模型時，所使用的目標函數與辨識準確度有直接的相關，簡而言之，我們提出的方法的主要想法是找到在雜訊干擾環境中的聲學語音特徵序列$\bar{O}$，從而使後端的隱藏式馬可夫聲學模型 (hidden Markov model, HMM)對應的狀態序列 $\bar{q}'$ 相較於真實序列 $\bar{q}$（其由乾淨語音特徵序列對應之 HMM 的狀態序列）的相似度（精確度）最大化。該方法的流程圖如圖 1 所示，它包括以下步驟：

步驟 1:

對乾淨狀態(clean-condition)語音與訓練集與多條件狀態(multi-condition)的訓練集中的每個句子計算其 MFCC 與 FBANK 特徵序列，之後對於這些特徵序列使用平均值與變異數正規化法(Viikki & Laurila, 1998)加以處理，這些特徵以$\{\bar{o}_t\}$表示，其中t是音框的索引。

步驟 2:

利用訓練集中的 MFCC 特徵，透過 Kaldi (Povey *et al*., 2011)所提供的標準程序訓練單音素(monophone)與三連音素(triphone)的高斯混合(Gaussian mixture model) —隱藏式馬可夫的聲學模型(GMM-HMM)。值得一提的是，在訓練過程中，線性鑑別分析(linear discriminant analysis, LDA) (Haeb-Umbach & Ney, 1992)、最大相似度線性變換(Gales, 1998)和語者自適應訓練(speaker adaptive training, SAT) (Anastasakos, McDonough, Schwartz & Makhoul, 1996)幾種演算法在聲學模型訓練期間也應用於語音特徵上。在訓練後，我們得到訓練集中每個特徵序列根據 GMM-HMM 所對齊(aligned)之狀態序列 $\bar{q}$，此狀態序列包括單音標籤和三音標籤。

步驟 3:

藉由步驟 1 所得之訓練集的 FBANK 特徵以及步驟 2 所得之特徵對應的聲學模型狀態序列標籤，我們訓練相應的 DNN-HMM 聲學模型(Hinton *et al*., 2012)，其目標是找到語音特徵$\bar{o}_t$及其對應的狀態標籤q_t之間的轉換，因此這是一個多元分類的問題，在 DNN 的最後一層可以產生每個特徵$\bar{o}_t$的狀態觀察機率。值得注意的是，我們是使用多條件狀態(multi-condition)的訓練集來訓練聲學模型，其語音訊號摻雜了不同種類與訊雜比(signal-to-noise ratio, SNR)的雜訊，因此，預期產生的 DNN-HMM 會比使用乾淨狀態的訓練集對應的聲學模型具有更好的抗噪能力。

此外，我們使用乾淨狀態的訓練集特徵（與多狀態訓練集有相同的原始乾淨語句內容），另外訓練一套 DNN-HMM，藉此 DNN-HMM，我們為每句語音特徵求其對應的狀態序列$\bar{q}$，我們把它視為真實狀態序列(ground-truth state sequence)，因為它是由乾淨語音求得，沒有雜訊干擾。

步驟 4:

此步驟是我們方法的核心。我們訓練一個去噪深度神經網路(denoising network)，用來將原始語音特徵$\bar{o}_t$ 轉換為$\bar{o}_t'$，如下所示：

$$\bar{o}_t' = f_{DN}(\bar{o}_t) \tag{1}$$

其中$f_{DN}(.)$表示欲訓練之去噪網路函數，其訓練目標是使新的特徵$\bar{o}_t'$在步驟 3 中創建的 DNN-HMM 裡，可以預測出更接近真實狀態的聲學模型狀態序列，其數學式如下：

$$f_{DN} = \text{argmax}_f(\text{Acc}(\bar{q}, \bar{q}' = g(f(\bar{o}_t) \mid \lambda)) \tag{2}$$

其中λ是 DNN-HMM 聲學模型、g 是一個給定模型λ的一個函數，用來產生新特徵$f(\bar{o}_t)$對應的最高相似度(maximum likelihood)狀態序列 $\bar{q}'$ 、$\bar{q}$ 是由前一步驟所得之真實狀態序列(ground-truth state sequence)、Acc是對數相似度(log-likelihood)函數，用於評估 $\bar{q}'$ 相對於$\bar{q}$ 的精確度。

當訓練好去噪深度神經網路f_{DN}後，在辨識過程中，我們將其用於雜訊干擾的語音特徵$\bar{o}_t$映射到新特徵$\bar{o}_t'$，然後將$\bar{o}_t'$輸入至原本 (無須重新訓練) DNN-HMM 聲學模型與語言模型、分別生成最高相似度狀態序列與詞序列。與原始特徵 $\bar{o}_t$ 相比，新特徵$\bar{o}_t'$預期有更強的抗噪能力，因為它是在多條件訓練的 DNN-HMM 幫助下創建的，並有乾淨訓練的 DNN-HMM 所提供的真實狀態，相當於整合了雜訊環境對映至乾淨狀態的資訊；同時，由於新特徵 $\bar{o}_t'$所對應的狀態序列，相較於原始 $\bar{o}_t$ 而言應會具有較高的狀態精確率，因此它們在辨識中理應產生較低的詞錯誤率。

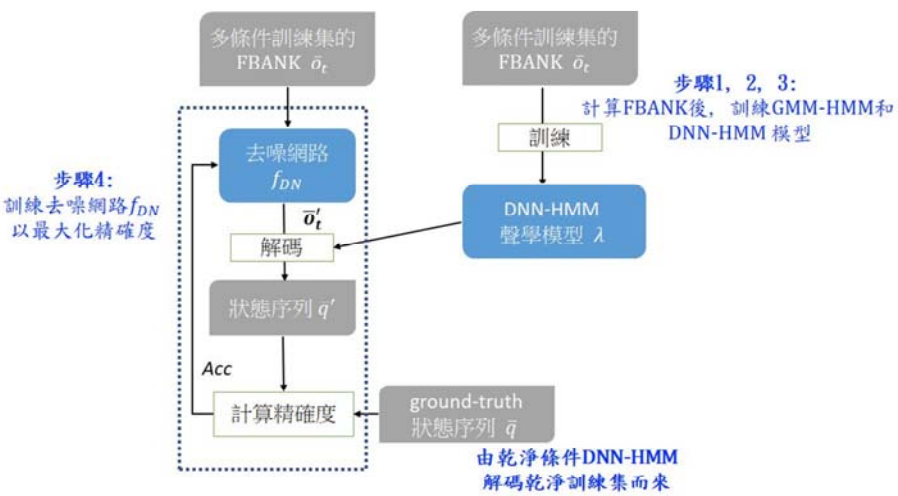

圖 1. 所提方法之流程圖

[Figure 1. The flowchart of the presented method]

與使用平均平方誤差(mean squared error, MSE)作為損失函數的之 DNN 求取抗噪之語音特徵方法相比(Garofolo, Lamel, Fisher, Fiscus & Pallett, 1993)，我們提出的方法具有以下潛在優勢：

1. 我們的方法其中的抗噪網路其訓練目標是在雜訊環境下，最大化聲學模型的狀態精確率，進而直接降低語音辨識的詞錯誤率(word error rate, WER)。相對而言，直接最小化訓練集內雜訊語音特徵與原始乾淨語音特徵之間的平方誤差(MSE)的 DNN 抗噪網路，可能存在如前一章節討論的目標不匹配問題，並不能保證在測試的雜訊環境下降低辨識錯誤率。
2. 我們的方法採用多條件訓練集依序獲得 GMM-HMM 和 DNN-HMM 聲學模型，藉此訓練我們的降噪神經網路。但是，我們預期當使用*乾淨無雜訊的訓練集所訓練的聲學模型*時，此降噪神經網路仍然可以使輸入雜訊語音特徵之輸出，對應到較高辨識率。根本的原因是整個降噪架構要改善原始語音特徵來擬合後端的聲學模型。在後面的章節中，我們將會在評估實驗中觀察並討論這方面的結果。

3. 實驗設置 (Experimental Setup)

我們使用著名的 TIMIT 語料庫(Garofolo *et al*., 1993)進行實驗，TIMIT 包含來自不同性別和方言的美國英語使用者的語句，其標註語句對應的音素和詞彙序列，TIMIT 中的語句在我們的評估實驗中用作訓練和測試集。在訓練語句上，除非特別提及之例外，我們是使用多條件狀態(multi-condition)的訓練集，先任意挑選 1000 句乾淨語音、再個別摻入不同種類及不同訊雜比(signal-to-noise ratio, SNR)的雜訊干擾，雜訊有三種，分別為：Babble、Car、Street，而訊噪比有五個等級，分別為 -5 dB、0 dB、5 dB、10 dB 和 15 dB，因此訓練集共有 15,000 句語音。在測試語句上，選擇了與訓練集不同的 400 句乾淨語音，再分別摻入三類別(White, Engine 和 Jackhammer)及六種訊雜比(-6 dB, -3 dB, 0 dB, 3 dB, 6 dB 和 12 dB)的雜訊干擾，共有 7,200 句語音。

訓練集中的語音特徵用於訓練上下文相關(context-dependent)的三連音素(tri-phone)聲學模型，它們被訓練成兩種不同的結構，分別是 GMM-HMM 和 DNN-HMM，具體來說，GMM-HMM 和 DNN-HMM 分別是使用 GMM 和 DNN 表示 HMM 的各種狀態。對於 GMM-HMM，每個單音素(monophone)的語音訊號和靜音分別由具有 3 個狀態的 HMM(總共 1000 個 Gaussian)來表示，而每個三連音素由具有 3 個狀態的 HMM 來表示，總共 2500 個 leaves。總共有 15000 個 Gaussian。此外，在三連音素模型訓練期間，將 LDA、MLLT 和 SAT 應用於語音特徵。另一方面，對於 DNN 的結構，使用了 5 層隱藏層，每個隱藏層包含 1024 個節點，並且分別連接到 DNN-HMM 中用於三連音和單音的兩個獨立輸出層。此 DNN 的訓練使用 Dropout 法，比例為 15%，進行 24 個 epochs 和使用 SGD 優化器，並且使用對數相似度(log-likelihood)作為目標函數。在模型訓練中，會將三連音素和單音素的誤差相加，並將其最小化。我們使用 Kaldi 工具包(Povey *et al*., 2011)來創建 GMM-HMM，而 Pytorch-Kaldi (Ravanelli, Parcollet & Bengio, 2019)工具包則用於創建 DNN-HMM。

此外，透過 Kaldi 的標準程序，我們建構了一組用於訓練語音的三元語法的語言模型(tri-gram)。

對於訓練和測試集中的每個語句，我們使用 69 維的 FBANK 特徵（每個音框 23 維的 FBANK 以及其 delta 和 delta-delta，音框長度為 20 毫秒，每次位移 10 毫秒）來作為基礎特徵(baseline feature)。我們提出的降噪 DNN 框架將 FBANK 作為輸入，按照上個章節中的步驟產生新的特徵，以進行後續辨識。去噪 DNN 模型是一個卷積神經網絡(convolutional network)，具有 4 個相同尺寸的一維卷積層，每層 kernel 數為 30，kernel 大小為 5，padding 數為 2。此外，這四個卷積層後面接著兩個相同的全連接層，各層具有 759 個節點。每層輸出的激活函數是線性整流函數(ReLU)。該降噪框架的訓練過程使用 Adam 優化器進行了 30 個 epochs，並使用了對數相似度(log-likelihood)作為目標函數。

4. 實驗結果與討論 (Experimental Results and Discussions)

在本章節中，我們呈現實驗結果並加以討論，為了方便討論，我們將所提出的方法命名為"最大化聲學模型狀態精確率法"，英文為"maximum state accuracy"，以縮寫"MSA"表示，同時，我們使用了兩種語音強化法進行比較，分別為最小均方誤差短時頻譜強度估測(minimum mean-square error short-time spectral amplitude estimation, 縮寫為 MMSE-STSA) (Ephraim & Malah, 1984)，及理想比例遮罩法(ideal ratio masking, 縮寫為 IRM) (Wang, 2005)。此外，我們採用一種基於 DNN 求取聲學特徵轉換的方法(Han *et al*., 2015)進行比較，該方法主要使用深度神經網絡(DNN)來轉換輸入的 FBANK 特徵，透過直接最小化多條件訓練集中雜訊—乾淨配對(noisy-clean pair)之語音的 FBANK 特徵之間的均方誤差(mean squared error, MSE)來學習其 DNN，這種方法稱為 feature-based MSE，縮寫為 FMSE。

在這裡，我們的實驗結果分為兩部分，分別為多條件訓練模式(multi-condition training mode)和乾淨狀態訓練模式(clean-condition training mode)。值得注意的是，我們所提出的方法 MSA 中的降噪模型在兩個模式下皆是藉由多條件訓練集所學習而得，但我們想測試由此產生的強化後語音特徵在兩種模式下可否都能表現良好。

- 多條件訓練模式(multi-condition training mode)之結果與討論

當利用多條件訓練集所訓練而得的聲學模型時，表 1、表 2 與表 3 列出了我們的方法 MSA 及三種比較法 FMSE、MMSE-STSA 與 IRM 在三種雜訊測試集所得之詞錯誤率(word error rate, WER)，值得注意的是，MMSE-STSA 與 IRM 二種語音強化法只使用於測試集的語句，並未作用於訓練集之語句，原因是我們之前實驗發現，若它們同時作用於訓練集，將使測試集之詞錯誤率明顯增加。從這三個表中，我們有以下觀察結果：

1. 平均而言，各種方法在 Jackhammer 雜訊環境中得到的詞錯誤率明顯低於 White 和 Engine 雜訊環境中的 WER，這表明與 White 和 Engine 雜訊相比，Jackhammer 雜訊對語音訊號的失真較小。但是，我們發現所有的方法均無法改善 Jackhammer 雜訊下的基礎實驗結果(baseline)，這表明語音強化或強健性特徵方法可能會對干擾較少的語句引

入更多可觀察到的失真。

2. 對於 MMSE-STSA 與 IRM 兩種語音強化法而言，MMSE-STSA 效果明顯比 IRM 差，且比基礎實驗得到較高的詞錯誤率，而 IRM 法相較於基礎實驗結果而言，只能在部分 SNR 環境小幅較低詞錯誤率，此結果部分驗證了語音強化方法雖可改善語音品質，但未必能有效提升語音辨識精確度。
3. 在 White 和 Engine 雜訊環境中，新提出的 MSA 法在大多數 SNR 環境下可以得到較低的詞錯誤率，並且更勝過其他方法，這驗證了 MSA 藉由提高語音特徵之狀態精確度，可改善特徵對雜訊的強健性並增加辨識準確率。特別的是，新提出的 MSA 中的降噪網路，其訓練所使用之語音包含的雜訊種類並非是測試集之 White 雜訊與 Engine 雜訊。因此，MSA 在某種程度上顯示出其一般化(generalization)的能力，在未知雜訊(unseen noise)環境下，仍可提升語音特徵的強健性。
4. FMSE 方法是為了最小化雜訊語音和乾淨語音其 FBANK 特徵之間的均方誤差(MSE)，然而在幾乎所有雜訊情況下，其效果都比基礎實驗結果差，如同之前討論，其可能原因是其評估與優化指標的不匹配，造成 FMSE 方法轉換後的語音特徵反而產生較差的辨識準確率，另一個原因是，在 FMSE 中學習到的 DNN 過度擬合訓練資料，因此無法很好地改善測試資料的失真問題。

表 1. 多條件訓練模式在“White”雜訊環境下測試集的基礎實驗、MSA、FMSE、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 1. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the White noise-corrupted test set under the multi-condition-training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	66.1	62.1	57.0	49.8	44.8	34.4
MSA	65.5	61.0	54.9	48.9	43.2	34.7*
FMSE	69.2*	63.3*	57.2*	50.4*	44.8	35.3*
MMSE-STSA	70.3*	66.8*	60.4*	55.0*	49.7*	40.1*
IRM	65.8	61.9	56.5	50.4*	44.4	34.3

表2. 多條件訓練模式在 "Engine" 雜訊環境下測試集的基礎實驗、MSA、FMSE 、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 2. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Engine noise-corrupted test set under the multi-condition-training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	65.3	61.7	55.1	48.2	41.5	31.1
MSA	65.5*	60.2	54.6	47.9	41.7*	32.3*
FMSE	70.4*	64.5*	56.2*	48.2	41.0	31.2*
MMSE-STSA	68.3*	63.6*	57.3*	51.4*	44.9*	34.2*
IRM	65.9*	61.4	54.8	48.2	41.2	31.1

表 3. 多條件訓練模式在"Jackhammer"雜訊環境下測試集的基礎實驗、MSA 、FMSE 、 MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 3. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Jackhammer noise-corrupted test set under the multi-condition-training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	31.4	27.8	25.9	23.8	23.0	21.9
MSA	32.6*	29.4*	27.5*	25.7*	25.1*	23.9*
FMSE	34.4*	31.3*	29.1*	27.6*	27.0*	26.1*
MMSE-STSA	32.9*	29.0*	27.2*	25.2*	24.4*	23.2*
IRM	32.4*	29.8*	26.6*	25.2*	23.8*	22.8*

- 乾淨訓練模式(clean-condition training mode)之結果與討論

利用乾淨訓練集所訓練而得的聲學模型，表 4、表 5 與表 6 列出了我們的方法 MSA 及三種比較法 FMSE、MMSE-STSA 與 IRM 在三種雜訊測試集所得之詞錯誤率(word error rate, WER)，值得注意的是，由於訓練集是乾淨語音，我們並不使用任何方法對其特徵作進一步強化，意即我們使用乾淨訓練集中的原始 FBANK 特徵來訓練聲學模型，而各個方法只用在測試集上。從這三個表，我們有以下觀察：

1. 若與前三個表(表 1、2、3)相比，乾淨訓練模式得到的基礎實驗(baseline)結果比多條件訓練模式所得到的基礎實驗結果較差(前者得到較高的詞錯誤率)，這很可能是因為與多條件訓練語句相比，乾淨訓練語句與測試集的雜訊語句之間之不匹配更為明顯。
2. 類似於之前的觀察，相較於基礎實驗結果，兩種語音強化法 (MMSE-STSA 和 IRM) 只能得到相近或更高的詞錯誤率，這再次顯示了直接改善語音的品質並不一定能提高其

辨識準確率。

3. 對於大多數雜訊之狀態(除了 SNR 高於-3 dB 的 Jackhammer 雜訊環境)，我們所提出的 MSA 法相較於基礎實驗結果，獲得明顯較低的詞錯誤率，這些結果表明，即使訓練集特徵未經 MSA 處理，若測試語句特徵經過 MSA 法處理後，仍可改善其語音辨識精確度。我們認為，這再次證實了我們先前的陳述，即 MSA 具有一般化的能力，可克服未見雜訊之問題。
4. 作用在特徵上的 FMSE 法，在部分雜訊環境下能比基礎實驗結果表現較佳(即得到較低的詞錯誤率)，但其效果仍不及我們所新提出的 MSA 法。

表 4. 乾淨訓練模式在"White"雜訊環境下測試集的基礎實驗、MSA 、FMSE 、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 4. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the White noise-corrupted test set under the clean-condition training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	67.6	64.6	61.0	55.6	50.9	41.2
MSA	64.3	60.6	56.3	50.8	45.4	36.6
FMSE	68.6*	64.3	58.9	53.0	47.7	38.9
MMSE-STSA	69.9*	67.1*	63.7*	59.1*	54.4*	45.6*
IRM	67.4	64.3	60.6	55.0	50.5	40.7

表 5. 乾淨訓練模式在"Engine"雜訊環境下測試集的基礎實驗、MSA 、FMSE 、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 5. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Engine noise-corrupted test set under the clean-condition training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	67.3	64.7	61.1	56.3	50.4	39.4
MSA	64.4	60.9	55.2	49.8	43.8	34.7
FMSE	69.9*	65.5*	59.7	52.3	46.8	36.6
MMSE-STSA	69.1*	65.9*	62.5*	57.3*	52.1*	40.7*
IRM	66.8	65.1*	60.6	55.7	49.7	39.4

表6. 乾淨訓練模式在"Jackhammer"雜訊環境下測試集的基礎實驗、MSA、FMSE、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 6. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Jackhammer noise-corrupted test set under the clean-condition training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	35.3	31.5	28.5	26.7	24.9	23.1
MSA	33.8	30.6	28.7*	27.0*	26.3*	24.9*
FMSE	35.0	32.2*	29.7*	28.1*	27.0*	25.4*
MMSE-STSA	36.5*	32.8*	29.6*	27.7*	25.3*	23.7*
IRM	34.8	31.8*	28.8*	26.6	24.9	23.3*

5. 結論與未來展望 (Conclusion and Future Work)

在本研究中，我們主要關注在自動語音辨識中的雜訊問題，提出一種基於深度學習的新方法來建立抗噪語音特徵，該方法利用深度神經網路來最大化語音特徵所對應的聲學模型的狀態精確度。初步實驗表明，新提出的方法可以提高 FBANK 特徵的辨識準確率，特別是在中度和重度雜訊干擾狀態中；且無論聲學模型的訓練集是在多條件環境下或是乾淨環境，它都能表現良好。關於未來的改良方向，我們將透過採用更多種類的訓練數據或增加訓練資料量來進一步增強此降噪神經網路，然後將其與其他基於特徵或基於模型的雜訊強健性演算法組合，以實現更好的性能。

參考文獻 (References)

Anastasakos, T., McDonough, J., Schwartz, R., & Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proceedings of Fourth International Conference on Spoken Language Processing (ICSLP) 1996*. doi : 10.1109/ICSLP.1996.607807

Bourlard, H. & Morgan, N. (1994). *Connectionist Speech Recognition: A Hybrid Approach*. New York, NY: Springer.

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans on Acoustics, Speech, and Signal Processing*, *32*(6), 1109-1121. doi: 10.1109/TASSP.1984.1164453

Fu, S.-W., Liao, C.-F., Tsao, Y., & Lin, S.-D. (2019). MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *Proceedings of the 36th International Conference on Machine Learning 2019*, 2031-2041.

Gales, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, *12*(2), 75-98. doi: 10.1006/csla.1998.0043

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., & Pallett, D. (1993). *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1* (NASA STI/Recon Technical Report N, vol. 93, p. 27403).Retrieved from https://ui.adsabs.harvard.edu/abs/1993STIN...9327403G

Grezl, F., Karafiat, M., Kontar, S., & Cernocky, J. (2007). Probabilistic and bottleneck features for lvcsr of meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2007*. doi: 10.1109/ICASSP.2007.367023

Haeb-Umbach, R. & Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 1992*. doi : 10.1109/ICASSP.1992.225984

Han, K., He, Y., Bagchi, D., Fosler-Lussier, E., & Wang, D. (2015). Deep neural network based spectral feature mapping for robust speech recognition. In *Proceedings of INTERSPEECH 2015*, 2484-2488

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738-1752. doi: 10.1121/1.399423

Hermansky, H., Ellis, D. P. W., & Sharma, S. (2000). TANDEM connectionist feature extraction for conventional hmm systems. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2000*. doi: 10.1109/ICASSP.2000.862024

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 578-589. doi: 10.1109/89.326616

Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1991). Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proceedings of EUROSPEECH 1991*, 1367-1370.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., … Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97. doi: 10.1109/MSP.2012.2205597

Juang, B.-H., Hou, W. & Lee, C.H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, *5*(3), 257-265. doi : 10.1109/89.568732

Kuo, J.-W. & Chen, B. (2005). Minimum word error based discriminative training of language models. In P*roceedings of Interspeech'2005 - Eurospeech*, 1277-1280.

Liu, F. H., Stern, R. M., Huang, X., & Acero, A. (1993). Efficient cepstral normalization for robust speech recognition. In *Proceedings of the workshop on Human Language Technology (HLT '93)*, 69-74. doi: 10.3115/1075671.1075688

Povey, D. (2003). *Discriminative training for large vocabulary speech recognition* (Doctoral dissertation). University of Cambridge, UK.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. … Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding 2011*.

Ravanelli, M., Parcollet, T., & Bengio, Y. (2019). The PyTorch-Kaldi speech recognition toolkit. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. doi: 10.1109/ICASSP.2019.8683713

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*. doi: 10.1109/ICASSP.2001.941023

Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., & Venkataraman, A. (2005). MLLR transforms as features in speaker recognition. In *Proceedings of Eurospeech 2005*, 2425-2428.

Su, Y.-C., Tsao, Y., Wu, J.-E., & Jean, F.-R. (2013). Speech enhancement using generalized maximum a posteriori spectral amplitude estimator. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2013*. doi: 10.1109/ICASSP.2013.6639114

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2010*. doi: 10.1109/ICASSP.2010.5495701

Torre, A., Peinado, A., Segura, J., Pérez-Córdoba, J., Benitez, C., & Rubio, A. (2005). Histogram equalization of the speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *13*(3), 355-366. doi: 10.1109/TSA.2005.845805

Viikki, O. & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, *25*(1-3), 133-147. doi: 10.1016/S0167-6393(98)00033-8

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi P. (eds) *Speech Separation by Humans and Machines* (pp. 181-197), Springer, Boston, MA. https://doi.org/10.1007/0-387-22794-6_12

Xia, B. & Bao, C.-c. (2014). Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Communication*, *60*, 13-29. doi: 10.1016/j.specom.2014.02.001

Zhang, H., Zhang, X., & Gao, G. (2018). Training supervised speech separation system to improve STOI and PESQ directly. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2018*. doi: 10.1109/ICASSP.2018.8461965