The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 103-120

融合多種深層類神經網路聲學模型與分類技術於華語錯誤發音檢測之研究

Exploring Combinations of Various Deep Neural Network based Acoustic Models and Classification Techniques for Mandarin Mispronunciation Detection

許曜麒 Yao-Chi Hsu [a], 楊明翰 Ming-Han Yang [a], 洪孝宗 Hsiao-Tsung Hung [a],
熊玉雯 Yuwen Hsiung [b], 宋曜廷 Yao-Ting Sung [c], 陳柏琳 Berlin Chen [a]
國立臺灣師範大學資訊工程學系 [a]
中原大學應用華語學系 [b]
國立臺灣師範大學教育心理與輔導系 [c]
{ychsu, mh_yang, alexhung, sungtc, berlin}@ntnu.edu.tw
ywhsiung@cycu.edu.tw

摘要

錯誤發音檢測(mispronunciation detection)為電腦輔助發音訓練(computer assisted pronunciation training, CAPT)研究中十分重要的一個環節，其目的是回饋給語言學習者是否在其讀誦一段話中的出現錯誤發音。一般而言，錯誤發音檢測流程可分為兩部分：1)前端特徵擷取模組，基於學習者所念誦的音素或語句段落和聲學模型(acoustic model)的比對以擷取對應的具有鑑別性之發音檢測特徵；2)後端分類模組，基於所求得發音檢測特徵，判斷音素或語句段落所歸屬類別(正確發音或錯誤發音)。在本篇論文延續錯誤發音檢測研究而主要有三項貢獻：1)比較並結合當前基於深層類神經網路(deep neural networks, DNN)與摺積類神經網路(convolutional neuron networks, CNN)之先進的聲學模型以產生更具鑑別性發音檢測特徵；2)我們比較並結合不同分類方法，以期能達到更佳的發音檢測表現；3)針對錯誤發音檢測所包括的模組，進行一系列廣泛且深入的實驗分析與討論。從一套以華語做為第二語學習目標語言的大量語料庫之實驗結果顯示，我們所提出融合多種深層類神經網路聲學模型與分類技術的方法的確能較基礎方法有顯著的效能提升。

關鍵字：錯誤發音檢測、自動語音辨識、深層類神經網路、摺積類神經網路

Abstract

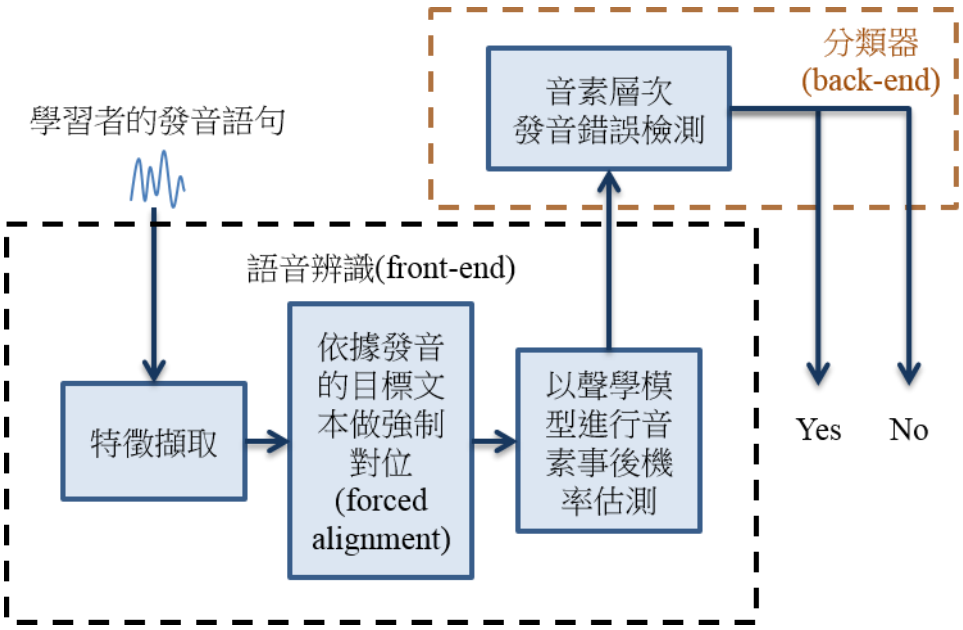
Automatic mispronunciation detection plays a crucial role in a computer assisted pronunciation training (CAPT) system. The main purpose of mispronunciation detection is to judge whether the pronunciations of a non-native speaker are correct or not. In general, the process of mispronunciation detection can be divided into two parts: 1) a front-end feature extraction module that generates pronunciation detection features based on an input speech segment and its associated reference acoustic models; and 2) a back-end classification module that determines the correctness of the pronunciation of the speech segment according to the output of a classifier that takes the pronunciation detection features of the segment as the input. The main contributions of this work are three-fold. First, we investigate the use of two state-of-the-art acoustic models, respectively based on deep neural networks (DNN) and convolutional neural networks

(CNN), and compare their effectiveness for the extraction of discriminative pronunciation detection features. Second, we experiment with different types of classification methods and propose a novel integration of DNN- and CNN-based decision scores at the back-end. Third, we provide an extensive set of empirical evaluations on the aforementioned two modules and associated methods based on a recently compiled corpus for learning Mandarin Chinese as the second language. The experimental results reveal the performance utility of our approach in relation to several existing baselines.

Keywords：Mispronunciation detection, Automatic Speech Recognition, Deep Neural Networks, Convolutional Neural Networks

一、緒論

現今全球化的時代裡，精通兩種或兩種以上的語言不僅是優勢更是必要的能力。在十幾年以前，英語還是國際通用的語言；但近年來，由於中國市場的快速發展，全球華語學習熱潮席捲而來，學習華語的人數預估已經超過一億，在許多非華語語系的亞洲、歐洲以及美洲國家，華話已經逐漸成為一種必須學習的語言[1][2]。語言學習又分為聽(listening)、說(speaking)、讀(reading)和寫(writing)等四類學習面向。隨著第二外語學習者(second language learner)的人數與日俱增，華語師資的需求也越來越大；尤其在語言學習中，說與寫的對錯往往需要透過專業的語言教師來評斷，但語言教師的人數遠遠不及華語學習者數量。因此，電腦輔助語言學習(computer assisted language learning, CALL)的研究領域越來越重要，本篇論文將專注此研究領域有關於電腦輔助發音訓練(computer assisted pronunciation training, CAPT)－「說」的技術發展與探討。



圖一、自動發音檢測之流程

一般而言，電腦輔助發音訓練(CAPT)包括兩個部分：分別是錯誤發音檢測(mispronunciation detection)與錯誤發音診斷(mispronunciation diagnosis)。錯誤發音檢測

系統是請學習者讀誦口說教材，針對學習者念誦的錄音，標記學習者的發音是正確發音(correct pronunciation)或錯誤發音(mispronunciation)，標記的目標可以是音素(phone)層次或詞(word)層次；錯誤發音診斷是當系統偵測到使用者的發音出現錯誤時給予有幫助的回饋，假設教材題目為「師範(shi1 fan4)」，但學習者念成「吃範(chi1 fan4)」，系統除了判斷出學習者有錯誤發音之外，還可以回饋學習者念的「師(shi1)」可能念成「吃(chi1)」。而本篇論文將聚焦在如何改善錯誤發音檢測之效能。目前，在錯誤發音檢測的評估方式中，召回率(recall)和精準度(precision)的曲線與接收者操作特徵曲線(receiver operating characteristic curve, ROC)是最常被採用來評估效能之優劣。我們認為相較於正確發音檢測(correct pronunciation detection)，錯誤發音檢測對於學習者而言是較為重要；所以，本篇論文後續在召回率和精準度曲線的評估實驗中，我們將集中討論錯誤發音檢測的效能表現。

自動錯誤發音檢測的研究大部分是基於現有的語音辨識技術而發展，希望能達到像專業語言教師一樣地給予語言學習者所念誦語句適當的發音評估。在本論文中，我們將語音辨識模組視為錯誤發音檢測系統的前端(front-end)，而錯誤發音檢測(分類)模組視為系統的後端(back-end)。前端的語音辨識模組如果能藉由聲學模型的使用，產生音框(frame)或者段落(segment)層次的事後機率來做為具鑑別性的發音檢測特徵，則後端偵測錯誤發音時就能基於這些發音檢測特徵來精準地判斷學習者的發音正確與否。因此，語音辨識模組中聲學模型所產生的回饋將是我們評斷發音好壞與否的重要依據。在語音辨識研究上，有別於傳統使用梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)之語音特徵的高斯混合模型-隱藏式馬可夫模型(gaussian mixture model-hidden markov model, GMM-HMM)的聲學模型，近年來由於機器學習演算法[3][4][5]與電腦硬體的進步，訓練多隱藏層(hidden layers)及大量輸出神經元(neurons)類神經網路的方法也更有效率在學術界與實務界激起了深層學習(deep learning)的浪潮，顛覆了幾十年來的研究生態。許多學者與實務家研究將深層類神經網路(deep neural networks, DNN)當作語音辨識的聲學模型的重要組成，取代傳統 GMM 的角色來計算每個音框所對應 HMM 狀態的觀測機率(observation probability)或相似度值(likelihood)。雖然 DNN 在語音辨識領域已經有相當優異的效果，但也有許多研究指出摺積類神經網路(convolutional neuron networks, CNN)在音素辨識[6]以及大詞彙連續語音辨識[7]的任務上的表現更優於 DNN；這可歸功於 CNN 能從語音特徵中擷取出發音中細微的位移不變(shift invariance)的特性。透過 CNN 來做為發音檢測特徵的擷取模組，期望能夠從不同國家的華語學習者之發音訊號中求取出對發音檢測有幫助、具鑑別性的發音檢測特徵(能提供更具鑑別力的事後機率來幫助錯誤發音檢測)，提升自動檢測錯誤發音的能力。本篇論文對於錯誤發音檢測研究有三項主要貢獻：首先，我們比較並結合當前基於深層類神經網路(DNN)與摺積類神經網路(CNN)之先進的聲學模型以產生更具鑑別性發音檢測特徵；再者，我們比較並結合不同分類方法，以期能達到更佳的發音檢測表現；最後，針對錯誤發音檢測之構成模組，進行一系列廣泛且深入的實驗分析與討論。

本篇論文的安排如下：第二小節將介紹錯誤發音檢測相關研究的發展近況；第三小節則是介紹錯誤發音檢測前端模組的聲學模型，分別有 GMM、DNN 與 CNN 三種模型與 HMM 的結合；第四小節介紹三種錯誤發音檢測的方法，分別是發音優劣程度(goodness of pronunciation, GOP)、支持向量機(support vector machine, SVM)與邏輯迴歸(logistic regression, LR)；第五小節則是分析不同聲學模型(DNN-HMM 和 CNN-HMM)在不同分類器(GOP、SVM 和 LR)中的表現，與將兩種聲學模型經過分類器 LR 所產生的發音檢測分數值作線性組合後的結果，以及基於 CNN 聲學模型在不同分類器所產生的

輸出發音檢測分數對應之排序取調和平均做為結合後的分類結果；最後，在第六小節，我們提出結論與一些未來可能的研究方向。

二、相關研究

在大多數的錯誤發音檢測研究中幾乎都是以自動語音辨識為前端，而將後端視為分類問題[8]。例如，Franco 等人[9]使用母語者的 HMM 之對數相似度值(log-likelihood)與非母語者的 HMM 之對數相似度值計算比值，稱為對數相似度比值(log-likelihood ratio, LLR)，該論文的實驗顯示使用對數相似度比值(LLR)對於錯誤發音檢測之表現勝過直接使用對數相似度值。Witt 等人[10]提出 GOP 作為錯誤發音檢測之評估方式，該方法基於聲學模型所產生的事後機率(posterior probability)對音素層次的發音計算評估分數，並訂定門檻值(threshold)來區分正確發音與錯誤發音；陸續也有其它研究是基於 GOP 的方法進行改良[11][12]。另一方面，Huang 等人[8]將鑑別式訓練應用在 GOP 估測，以最小大化 F 度量(F-measure)為目標作鑑別式訓練。Ito 等人[13]使用決策樹(decision tree)的方法並針對不同錯誤發音的情況定義各自的門檻值來進行錯誤發音檢測；該論文的實驗證明其效果勝過所有發音共用相同的門檻值。Truong 等人[14]比較決策樹與線性鑑別分析(linear discriminant analysis, LDA)用於荷蘭語學習者的錯誤發音檢測任務。廣義上來看，GOP 也屬於一種二元分類的方法，但 GOP 只有考慮到目標(正確)音素與它的混淆音素的對數相似度值。有鑒於此，Wei 等人[15]使用目標音素與其它所有音素的對數相似度值做為輸入分類器的發音檢測特徵，並將 SVM 做為分類器來辨認音素特徵對應的輸出為正確發音或錯誤發音標記。但除了每一個音素的對數相似度值來作為發音檢測特徵，Hu 等人[16]不只使用[15]提出的發音檢測特徵，還額外地將目標音素與其它音素的對數相似度比值加入成為額外輸入的發音檢測特徵，並使用特殊結構的邏輯迴歸來進行錯誤發音檢測，該結構透過共享隱藏層來解決部分音素資料稀疏(data sparse)的問題。不同於[16]的貢獻，我們認為藉由良好的聲學模型產生之事後機率而得的具鑑別性發音檢測特徵，應有助於錯誤發音檢測的效果；因此，本論文將聚焦於前端聲學模型的比較與融合。

上述的方法皆是運用聲學模型所擷取的發音檢測特徵進行錯誤發音的檢測，除了將音素或語句分類為正確發音與錯誤發音外，也有研究著重在評斷語句的發音品質。Neumeyer 等人[17]使用 HMM 計算出對數相似度值與強制對位(forced alignment)後的音素發音持續時間(duration)資訊，並據此對非母語學習者語句層次的發音品質進行評估。Chen 等人[18][19][20]提出詞層次的發音品質評估，共分成 5 個等級來區分發音的品質，並使用資訊檢索的排序學習法(learning to rank)來結合不同發音檢測特徵用於發音品質評估；其中，在[20]比較各類發音檢測特徵的影響力與 4 種音素層次轉換到詞層次的發音檢測特徵轉換方法。

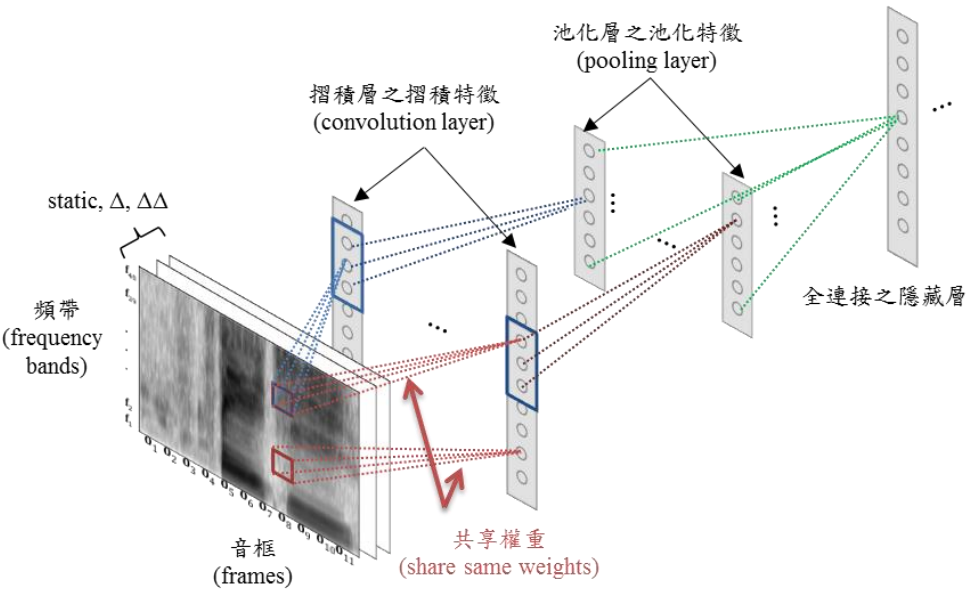
而在聲學模型方面，與傳統 GMM-HMM 相比，DNN-HMM 在語音辨識準確率上已被證實能有顯著的效能提升[21][22]，這主要可能歸功於 DNN 能夠模擬任意的函數，能替語音訊號所內含的複雜對應關係建立模型，表達能力比 GMM 更強。優良的事後機率蘊藏豐富的發音鑑別性資訊，使得錯誤發音檢測的效果更好，有許多 DNN-HMM 應用在 CAPT 的效果已被驗證勝過傳統的 GMM-HMM[2][16][23]，因此聲學模型在計算事後機率的任務中扮演著非常關鍵的角色[24]，而基於深層類神經網路的聲學模型計算而得對發音檢測有幫助的事後機率可使 GOP 與其它分類器達到最佳的檢測效果。相較於 DNN，CNN 被視為是另一種更有效率的深層類神經網路，可用於擷取語音訊號中的頻

譜變化的位移不變性並且能針對頻譜的相關性建立模型[6][7]。CNN 與 DNN 不同在於：神經元間的連接不是全連接的(fully-connected)以及同一層的某些神經元間會共享連接的權重(weight sharing)。Sainath 等人[7]提出 CNN 作為聲學模型更勝於 DNN 的原因是因為他們認為 DNN 有兩項缺點。首先，DNN 的架構中沒有明確地處理語音訊號中的不變特徵的功能，例如不同語者說話方式不同，在頻譜上會有細微的位移。DNN 需要運用各種語者調適(speaker adaptation)技術來降低特徵的變化，DNN 同時需要巨大的網路規模及大量的訓練樣本(training sample)來達到這件事；但 CNN 能透過摺積核(conventional filter)沿著頻譜的時間與頻率掃描，以較少的參數數量捕捉到頻譜平移的不變性。其次，DNN 忽略了輸入的拓撲(topological)結構，它的輸入特徵可以以任何順序輸入網路，而不影響最後的效能[21]；然而語音訊號所對應的頻譜內容著實含有豐富的關聯性，而能夠善用頻譜的局部相關性而建立模型的 CNN 在許多任務上的效果都明顯優於 DNN[25][26][27][28]。因此，本論文將融合兩者的優點，並探討兩種類神經網路所訓練的聲學模型(DNN-HMM 與 CNN-HMM)對於錯誤發音檢測的效果。

三、聲學模型

3.1 深層類神經網路

傳統語音辨識系統透過 HMM 來處理語音訊號在時間上的變異，並使用生成模型 GMM 來建立聲學模型，但是使用高斯混合模型的問題在於如何選出最佳的混合高斯函數的數量，反而導致 GMM 受到侷限。而近年來，在語音辨識的領域中，取代以往的生成模型(generative model)，透過可視為鑑別式模型(discriminant model)的類神經網路[29]來估測音素層次的 HMM 狀態之事後機率的研究越來越多。



圖二、摺積類神經網路之示意圖

DNN 是一種前饋式(feed-forward)的類神經網路，它的輸入層與輸出層之間包含一層以上的隱藏層[30]，每一個隱藏層的神經元通常使用邏輯函數(logistic function)將輸入映射到上一層，邏輯函數通常使用 sigmoid 函數。假設輸入層表示為第0層，輸出層表

示為第L層，表示有$L+1$層的深層類神經網路，此前饋運算可以表示為：

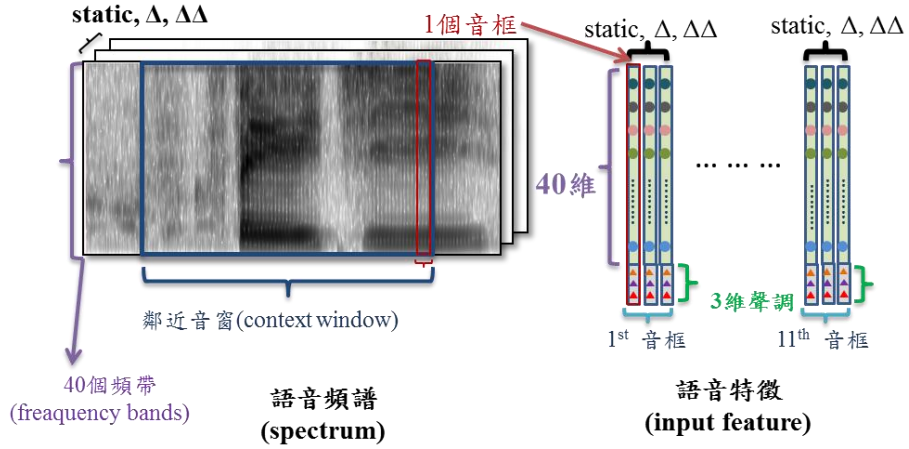
$$\boldsymbol{v}^{\ell} = \sigma\left(\boldsymbol{z}^{\ell}\right) = \sigma\left(\mathrm{W}^{\ell}\mathbf{v}^{\ell-1} + \mathbf{b}^{\ell}\right),\ \ (\ \ell = 0, 1, 2, 3, \ldots, L\ ) \tag{1}$$

$$\sigma(z) = \frac{1}{1+\exp(-z)},\ \ 0 < \sigma(z) < 1 \tag{2}$$

式(1)中，$N_{\ell} \in \mathbb{N}$，為第ℓ層的神經元數量。$\mathbf{v}^{\ell} \in \mathbb{R}^{N_{\ell}\times 1}$，$\mathrm{W}^{\ell} \in \mathbb{R}^{N_{\ell}\times N_{\ell-1}}$，$\mathbf{b}^{\ell} \in \mathbb{R}^{N_{\ell}\times 1}$，$\mathbf{v}^{\ell}$為第$\ell$層的輸出向量，$\mathrm{W}^{\ell}$為第$\ell$層的權重矩陣，通常採取隨機初始化(random initial)來當作網路初始的權重W。近年來有學者提出透過限制性波茲曼機(restricted boltzmann machine, RBM)的非監督式預訓練(unsupervised pre-training)[31][32][31][33]，逐層往上預訓練(pre-training)DNN 的參數。待預訓練完畢後，基於預訓練參數再進行監督式訓練，取代傳統隨機初始化參數的方法來改善語音辨識的正確率[34][35][36]，我們每層 DNN 參數皆使用 RBM 來預訓練權重的初始值，$\mathbf{v}^{\ell-1}$為第$\ell-1$層的輸出向量，$\mathbf{b}^{\ell}$為第ℓ層的偏移量向量。$\mathbf{v}^{0} = \mathbf{o}_t \in \mathbb{R}^{N_0\times 1}$表示為輸入語音音框對應之語音特徵或與相鄰音框對應之語音特徵所串接而成的特徵，N_0為特徵的維度。式(2)中，$\sigma(z)$為 sigmoid 函數，其值域範圍在 0 到 1 之間。

DNN 運用於類別(如音素狀態或更小單位)事後機率預測問題上時，每一個輸出神經元都表示一種類別，總共可分為$\mathcal{H}$類，表示為$i \in \{1, \ldots, \mathcal{H}\}$，則第$i$個輸出神經元的值$\mathbf{v}_i^L$表示輸入語音音框對應語音特徵$\mathbf{o}_t$對應到類別$i$的機率$P(i|\mathbf{o}_t)$，假設輸出向量$\mathbf{v}^L$滿足多項式分佈(multinomial distribution)，那麼$\mathbf{v}^L$需要滿足$\mathbf{v}_i^L \geq 0$及$\sum_{i=1}^{\mathcal{H}} \mathbf{v}_i^L = 1$，可以透過軟式最大化(softmax)做到，如：

$$\mathbf{v}_i^L = \mathrm{softmax}(\mathbf{z}^L, i) = \frac{\exp(\mathbf{z}_i^L)}{\sum_{j=1}^{\mathcal{H}} \exp(\mathbf{z}_j^L)} \tag{3}$$



圖三、摺積類神經網路之特徵架構

在訓練階段，首先對每個輸入語音音框對應的特徵做強制對齊，產生狀態標籤(state label)的序列，這些標籤用於監督式訓練來最小化交叉熵(cross entropy)目標函數$-\sum_i \mathbf{d}_i \log \mathbf{v}_i^L$，意義是要最小化 DNN 預測的 softmax 輸出與其對應的參考標籤$\mathbf{d}_i$的差異。假設反向傳播演算法(back-propagation)[24]使用隨機梯度下降(stochastic gradient descent algorithm)來最小化目標函數，則每個權重矩陣W的更新可透過式(4)：

$$\Delta \mathrm{W}^{\ell} = \epsilon \cdot (\mathbf{v}^{\ell-1})' \boldsymbol{\varepsilon}^{\ell} \tag{4}$$

其中ϵ為學習率(learning rate)，$\boldsymbol{\varepsilon}^{\ell}$為第$\ell$層的錯誤訊號(error signal)。

3.2 摺積神經網路

CNN 由數組的摺積層(convolution layers)和池化層(pooling layers)所組成，摺積層和池化層的運算分別稱為摺積(convolution)及池化(pooling)。摺積層透過摺積核掃描輸入的特徵圖，摺積核就像是生物視覺神經的感受區[37]，每一個摺積核能夠獲取輸入特徵的局部特徵；而池化目標是將摺積層的特徵做降維。已知輸入的語音特徵序列，當計算音框t時，需左右各取C個音框，所組成的特徵圖(feature maps)矩陣表示為$\mathcal{O}_t$，摺積運算後的類別特徵圖表示為$\mathrm{Q}_j(j = 1, 2, \dots, J)$，由$J$個摺積特徵圖所組成，則摺積運算可以視為透過權重矩陣$\mathrm{W}_{t,j}(t = 1, \dots, T;\ j = 1, \dots, J)$，將輸入特徵$\mathcal{O}_t$映射到摺積特徵$\mathrm{Q}_j$的矩陣乘法，如式(5)表示：

$$\begin{aligned} \mathcal{O}_t &= [\ \mathbf{o}_{t-\mathrm{C}}, \dots, \mathbf{o}_{t-1}, \mathbf{o}_t, \dots, \mathbf{o}_{t+1}, \mathbf{o}_{t+\mathrm{C}}] \\ \mathrm{Q}_j &= \sigma\big(\mathcal{O}_t * \mathrm{W}_{t,j} + b_j\big), \qquad (j = 1,2, \dots, J) \end{aligned} \tag{5}$$

其中$*$表示為摺積運算，$\mathrm{W}_{t,j}$為將第t個輸入特徵映射到第j個摺積特徵的區域權重矩陣，b_j為偏移量。更多的細節請參考[26]。摺積層中的權重同樣能透過反向傳播來學習[38]。摺積層與全連接隱藏層的差別有兩點：1)摺積層只從局部感受野接收區域的輸入特徵，換句話說，摺積層的每個元素都表示輸入的區域特徵。2)摺積層中的每個摺積特徵可以視為特徵圖，圖中的每個元素都共享相同的權重，但它們各自是濃縮自前一層之不同區域的特徵而來。接下來是池化的部分，池化層是從摺積層產生對應的池化層，每一個池化特徵圖都是由前一層摺積層的摺積特徵圖做池化運算而來，因此池化特徵圖的數量也會與摺積特徵圖的數量相同，也具備摺積特徵所包含的的區域不變性(local invariance)的特性，池化運算分成最大池化(max-pooling)及平均池化(average-pooling)兩種，以最大池化最多人使用[39]。影像處理中所使用的 CNN，其池化窗(pooling window)不會互相重疊，池化窗之間彼此並排沒有空隙；在本篇論文中，我們的池化運算也採取這樣的做法。

四、 錯誤發音檢測

4.1 發音優劣程度(goodness of pronunciation, GOP)

GOP 是替每一詞彙所包含的每一個音素建立一個評估分數，並制定一個門檻值來區分該音素是否發音正確。而我們基於語音辨識聲學模型所給予的對數相似度值來計算 GOP，若已知語音段落的語音特徵序列O在其目標(正確)發音為音素a之對數事後機率$\log p(a|\mathrm{O})$在(本論文中語音特徵序列O是為基於 MFCC 或 mel-filter bank 輸出的語音特徵所構成)，則 GOP 的公式可以定義成：

$$\mathrm{GOP}(a, \mathrm{O}) = \log p(a|\mathrm{O}) \tag{8}$$

$$= \log \frac{p(\mathrm{O}|a)P(a)}{p(\mathrm{O})} \tag{9}$$

$$= \log \frac{p(\mathrm{O}|a)P(a)}{\sum_{i=1}^{M} p(\mathrm{O}|b_i)P(b_i)} \tag{10}$$

$$\cong \log \frac{p(\mathrm{O}|a)}{\max\limits_{i=1,2,\ldots,M,\ b_i \neq a} p(\mathrm{O}|b_i)} \tag{11}$$

由於無法窮舉語句對應的所有語音訊號，我們無法對語音段落對應特徵序列O建立機率模型，因此式(8)可藉由貝式定理將事後機率轉換成相似度值$p(\mathrm{O}|a)$乘上事前機率$P(a)$除以特徵序列O的機率，如式(9)所示。而式(9)的事前機率$p(\mathrm{O})$可以轉換成將所有音素的對數相似度值加總。如式(10)的分母項，常數M表示目標語言中音素的總數量，在錯誤發音檢測的任務中不應受到音素本身在訓練資料中的數量影響，所以我們假設所有音素的事前機率皆相等$(P(a) = P(b_i))$，且式(10)的分母項約等於音素b_i的相似度值取最大值，因此式(10)可以被簡化成式(11)。接著在定義門檻值τ來預測發音是否正確：

$$\mathrm{GOP}(a, \mathrm{O}) > \tau \begin{cases} Yes & correct\ pronunciation \\ No & mispronunciation \end{cases} \tag{12}$$

其中式(11)的相似度值$p(\mathrm{O}|a)$在語句中都會橫跨數個音框，因此我們將音素a的起始時間t_s到結束時間t_e取平均，因此音素a的相似度值可以寫成：

$$\log p(\mathrm{O}|a) = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log p(\mathbf{o}_t|a) \tag{13}$$

在 GOP 發音的評估方法中，可以基於聲學模型的事後機率(可視為一種發音檢測特徵)來進行計算，並透過門檻值τ來分辨發音正確與否。因此，我們可直覺地將 GOP 看成是一種分類器，但因為 GOP 只有觀測目標發音(正確)的音素a的事後機率，下一小節將透過觀測其它非目標音素的事後機率並使用不同的分類技術來改善 GOP 的不足。

4.2 分類器(Classifier)

此小節將討論兩種分類器(SVM 與 LR)被實際運用於錯誤發音檢測的作法。無論是 SVM 或是 LR 分類器，都需要輸入發音檢測特徵$\boldsymbol{f}_{a_{ui}}$與對應的 2 種輸出結果$\{\mathcal{C}, \mathcal{M}\}$作為訓練的樣本，其中$\mathcal{C}$代表正確發音，$\mathcal{M}$代表錯誤發音，$\boldsymbol{f}_{a_{ui}}$表示第$u$個語句的第$i$個音素的發音檢測特徵，$a_{ui}$表示該特徵對應的目標發音的(正確)音素。輸入發音檢測特徵$\boldsymbol{f}_{a_{ui}}$由對數音素事後機率(log phone posterior, LPP)[11][17]與對數事後機率比值(log posterior ratio, LPR)[16]所組合而成，我們接續 4.1 小節所提及的事後機率計算式(13)，對於任意音素a_{ui}我們將 LPP 定義成：

$$\mathrm{LPP}(a_{ui}, \mathrm{O}_{ui}) = \log p(a_{ui}|\mathrm{O}_{ui}) \tag{14}$$

除此之外我們還需要知道目標發音(正確)的音素a_{ui}與其它任意音素b_j的比值，也就是 LPR，其公式可以定義成：

$$\mathrm{LPR}(b_j, a_{ui}, \mathrm{O}_{ui}) = \mathrm{LPP}(b_j, \mathrm{O}_{ui}) - \mathrm{LPP}(a_{ui}, \mathrm{O}_{ui}) \tag{15}$$

接著就可以建立音素層次的音素發音檢測特徵，我們定義目標發音的音素a_{ui}所對應的發音檢測特徵$\boldsymbol{f}_{a_{ui}}$可以表示為式(16)：

$$\begin{aligned}\boldsymbol{f}_{a_{ui}} = [&\mathrm{LPP}(b_1, \mathrm{O}_{ui}), \mathrm{LPP}(b_2, \mathrm{O}_{ui}), \dots, \mathrm{LPP}(b_M, \mathrm{O}_{ui}),\\ &\mathrm{LPR}(b_1, a_{ui}, \mathrm{O}_{ui}), \mathrm{LPR}(b_2, a_{ui}, \mathrm{O}_{ui}), \dots, \mathrm{LPR}(b_M, a_{ui}, \mathrm{O}_{ui})]\end{aligned} \tag{16}$$

$\mathrm{LPP}(a_{ui}, \mathrm{O}_{ui})$會等於$\mathrm{LPP}(b_1, \mathrm{O}_{ui}), \mathrm{LPP}(b_2, \mathrm{O}_{ui}), \dots, \mathrm{LPP}(b_M, \mathrm{O}_{ui})$的其中一項，且音素$b_j(j = 1, 2, 3, \dots, M)$的其中一項等於音素$a$時，$\mathrm{LPR}(b_i, a_{ui}, \mathrm{O}_{ui})$會為 0。而 4.1 節提到的 GOP 評估值等同於發音檢測特徵$\boldsymbol{f}_{a_{ui}}$後半部的其中一個維度之倒數；因此。利用特徵$\boldsymbol{f}_{a_{ui}}$訓練出的分類器將會比 GOP 擁有更多關於發音的訊息。接著將介紹本論文嘗試比較的兩種分類器。

表一、單音節語料庫與雙音節語料庫之內容

-	單音節				雙音節			
母語(L1) 第二外語(L2)	L1		L2		L1		L2	
人數(人)	62		63		115		40	
音素層次 正確發音(T) 音素層次 正確發音(F)	T	F	T	F	T	F	T	F
時間(小時)	9.32	1.04	13.79	9.03	3.97	0	0.89	0.94
語句數(句)	37,976	4,827	50,856	32,726	10,384	0	1,994	2,003
音素數量(個)	76,638	4,976	119,512	36,862	38,939	0	12,449	2,539

LR 被廣泛利用在二類分類問題的任務中[16][18]，利用 sigmoid 的特性來表示資料的分佈，但在錯誤發音檢測的任務中，不同音素應該使用不同的 LR 分類器，若將所有音素混在一起進行迴歸分析可能導致過度混淆。以下先介紹分類器 LR 對正確發音、錯誤發音樣本的機率表示如式(17)：

$$\begin{aligned}p(\mathcal{C}|\boldsymbol{f}_{a_{ui}}) &= \sigma(\mathbf{w}_{a_{ui}}^T \boldsymbol{f}_{a_{ui}})\\ p(\mathcal{M}|\boldsymbol{f}_{a_{ui}}) &= 1 - p(\mathcal{C}|\boldsymbol{f}_{a_{ui}})\end{aligned} \tag{17}$$

$\sigma(.)$為 sigmoid 函數，$p(\mathcal{C}|\boldsymbol{f}_{a_{ui}})$為已知有發音檢測特徵$\boldsymbol{f}_{a_{ui}}$下發生$\mathcal{C}$的機率，$p(\mathcal{M}|\boldsymbol{f}_{a_{ui}})$為已知發音檢測特徵$\boldsymbol{f}_{a_{ui}}$發生$\mathcal{M}$的機率，$\mathbf{w}_{a_{ui}}$則是透過學習來更新的權重(weight)，不同語句中相同的音素也會使用相同的權重，接著定義相似度值函數L：

$$L = \prod_{u=1}^{U} \prod_{i=1}^{N_u} p(\mathcal{C}|\boldsymbol{f}_{a_{ui}})^{t_{ui}} p(\mathcal{M}|\boldsymbol{f}_{a_{ui}})^{1-t_{ui}} \tag{18}$$

$$E = -ln(L) \tag{19}$$

其中式(18)的$t_{ui} = \{0, 1\}$，0 表示錯誤發音，1 表示發音正確，t_{ui}使得發音檢測特徵$\boldsymbol{f}_{a_{ui}}$

對應的輸出之機率不會為 1，並定義函數E為最小化交叉熵目標函數如式(19)，接著使用隨機梯度下降法來最小化目標函數E，如式(20)：

$$\frac{\partial E}{\partial \mathbf{w}_{a_{ui}}} = \sum_{u=1}^{U}\sum_{i=1}^{N_u}\left(p\left(\mathcal{C}|\boldsymbol{f}_{a_{ui}}\right) - t_{ui}\right)\cdot \boldsymbol{f}_{a_{ui}} \tag{20}$$

$$\Delta \mathbf{w}_{a_{ui}} = \gamma \cdot \frac{\partial E}{\partial \mathbf{w}_{a_{ui}}} \tag{21}$$

式(21)中的參數γ為權重$\mathbf{w}_{a_{ui}}$更新時的學習率，學習率將隨著更新的次數進行調整，經過數次更新後直到權重$\mathbf{w}_{a_{ui}}$的改變過小則收斂，接著當輸入發音檢測特徵為$\boldsymbol{f}_{a_{ui}}$時，該段發音為正確發音的機率則為$p\left(\mathcal{C}|\boldsymbol{f}_{a_{ui}}\right) = \sigma\left(\mathbf{w}_{a_{ui}}^{T}\boldsymbol{f}_{a_{ui}}\right)$。

SVM[15]是一種效能表現良好的分類器，他可以透過將特徵轉換到更高維度的空間來解決資料線性不可分的問題，我們定義函數s(.)用來表示 SVM 給予特徵的$\boldsymbol{f}_{a_{ui}}$決策值，並將$\mathrm{s}(\boldsymbol{f}_{a_{ui}})$代入 sigmoid 函數$\sigma(.)$用以表示正確發音的機率$p\left(\mathcal{C}|\boldsymbol{f}_{a_{ui}}\right) = \sigma\left(\mathrm{s}(\boldsymbol{f}_{a_{ui}})\right)$。本篇論文使用 python 的現有模組“scikit-learn[40]”所提供的 SVM 與 LR 工具，核心函數為徑向基函數核(radial basis function kernel)。

表二、單音節與雙音節在不同 HMM 的字錯誤率(character error rate, CER)與音素錯誤率(phone error rate, PER)

ASR performance	單音節 (%)				雙音節 (%)			
	CER		PER		CER		PER	
	L1	L2	L1	L2	L1	L2	L1	L2
GMM	66.53	80.16	46.00	58.70	55.83	57.29	39.66	39.45
DNN	22.25	37.11	13.34	24.71	15.62	24.37	10.20	16.46
CNN(a)	21.17	36.32	12.76	24.23	16.06	22.61	10.37	14.95
CNN(b)	20.15	36.05	12.01	24.32	17.16	24.37	11.89	16.08

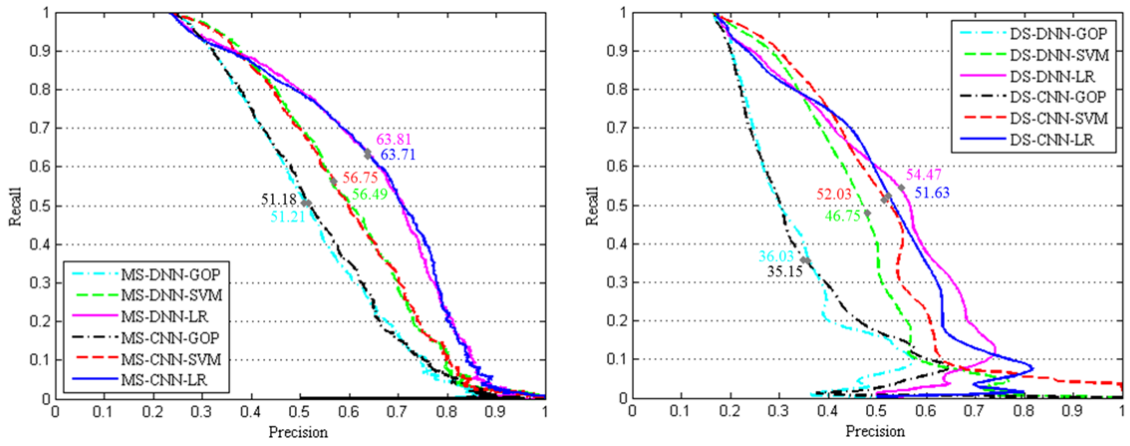
五、 實驗

每種語言的錯誤可分成三種：替換(substitution)、插入(insertion)、刪除(deletion)[41]。對華語來說，每個字(character)都屬於一個音節，而每個音節又可拆成三個部分：聲母(initial)、韻母(final)、聲調(tone)。對於有華語基礎知識的學習者而言並不易發生插入及刪除的錯誤，但華語是一種聲調語言(tonal language)，聲調的發音相較於聲母、韻母則更容易念錯。[8][42][43] 的研究不探究聲調的影響，而本論文將聲調依附在韻母之後，也就是一個音節可拆成聲母及聲調韻母(tonal final)兩個音素。

5.1 語料庫

我們的語料庫使用臺灣師範大學邁向頂尖大學計畫的華語學習者口語語料庫，分成雙音節語料庫及單音節語料庫兩部分，如表一所示。雙音節語料庫中，男女語料的比例為 2:3，母語為華語(L1)的語料全是台灣語者所錄製，只收錄正確的發音，沒有錯誤的發音，而非母語的華語學習者(L2)的語料包含日本及韓國兩種外國口音，收錄了正確發

音與錯誤發音，雙音節每個語句由 2 個中文字組成，意即每個語句可拆解成 4 個音素，但是不代表每個語句的音素都是念錯，因此語句層次的錯誤樣本應該要參考音素層次那欄，同樣的道理也套用在單音節語料庫。單音節語料庫中，男女語料的比例為 21:34，母語為華語(L1)的語料皆為台灣人口音所錄製，非母語的華語學習者(L2)收錄的口音包括美國、瓜地馬拉、越南、韓國、日本、西班牙、阿根廷等 23 國的學習者口音，單音節 L1 及 L2 皆收錄了正確與錯誤的發音，單音節中每個語句都是一個中文字，每個中文字可拆解成 2 個音素。兩種語料庫在訓練聲學模型時，都只使用語句完全正確的樣本來訓練聲學模型，而在訓練錯誤發音檢測模型時則會使用錯誤發音與正確的語句，以音素層次的發音來訓練錯誤發音檢測模型。



圖四、比較單音節(左側)與雙音節(右側)從不同 HMM 萃取出的特徵使用不同分類器之 Recall-Precision 曲線

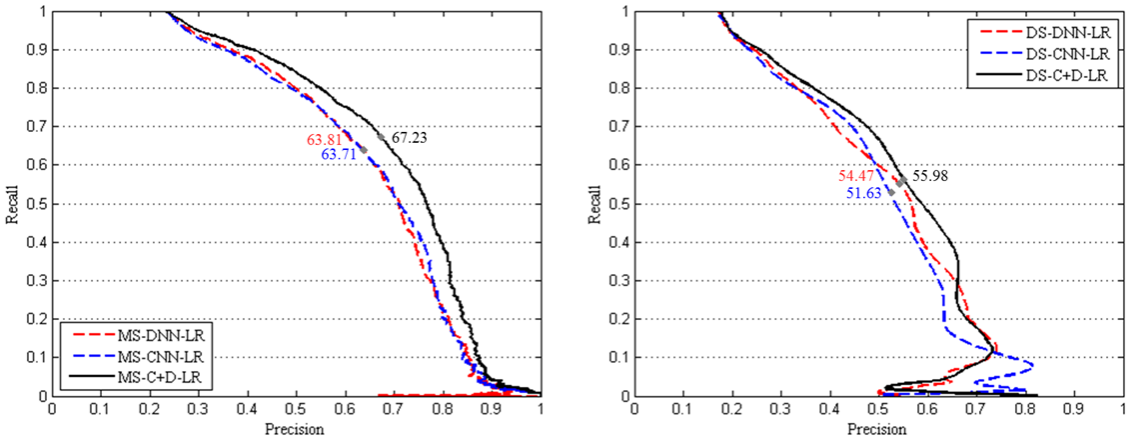
5.2 實驗設定

錯誤發音檢測系統的優劣與語音辨識系統的表現息息相關，因此我們先分析語音辨識系統的表現。我們將語料庫中的正確發音分成訓練集(training set)、發展集(development set)與測試集(test set)，使用高斯混合模型針對訓練集來學習語音訊號的分佈，以及基於 GMM-HMM 所校準的文字與音框之對應關係來藉由 DNN-HMM 與 CNN-DNN-HMM 學習語音訊號的分佈，為了描述方便，將聲學模型 GMM-HMM、DNN-HMM 與 CNN-DNN-HMM 簡稱為 GMM、DNN 與 CNN。發展集目的在於類神經網路等相關的模型在訓練時容易發生過度擬合(over fitting)，因此我們切出一塊發展集來引導模型在訓練時不要過度傾向訓練集。接著再使用 GMM、DNN 與 CNN 所訓練出的語音辨識器對測試集進行辨識，辨識結果如表二。

我們基於 Kaldi 語音辨識工具[44]，將華語學習者發音的語音訊號切成音框後，鄰近的數個音框整合成鄰近音窗(context window)，通常採用前後各 5 個音框，總共 11 個音框來當作一個鄰近音窗的大小，從音窗的語音訊號抽取出 13 維的 MFCC 特徵加上 3 維度的音調(pitch)，並對 16 維語音特徵取相對的一皆差量係數(delta coefficient)和二皆差量係數(acceleration coefficient)當作 DNN 的輸入特徵。先透過 GMM 對語音特徵訓練單連音素(monophone)的聲學模型，單音節與雙音節語料庫中皆有 183 個單連音素(聲母有 24 個，聲調韻母有 159 個)，接著保留 GMM 計算出來的初始機率、轉移機率與強制

對齊的資訊，取代 GMM，訓練產生每個音框所對應 HMM 狀態的機率，再根據強

制對齊的資訊取得每個音素所對應到音框數量，來計算每個音素的事後機率，當作 HMM 的觀測機率。在 CNN 的輸入特徵設定方面，我們使用從梅爾頻譜係數(mel-scale



圖五、比較雙音節(左側)與單音節(右側)從不同 HMM 擷取出的發音檢測特徵使用 LR 分類器與不同 HMM 發音檢測之 LR 分類器輸出分數的線性組合(M/DS-C+D-LR)之 Recall-Precision 曲線

frequency spectral coefficients, MFSC)取得的對數能量特徵並透過濾波器組(filter banks)所產生的 40 維輸出作為 CNN 的輸入語音特徵，鄰近音窗我們採用前後各 5 個音框，共含 11 個音框，每個音框皆為 40 維的 filter banks 輸出加上 3 維度音調特徵，並對 43 維語音特徵取相對的一皆差量係數(delta coefficient)和二皆差量係數(acceleration coefficient)，則輸入的語音特徵就會得到 11 個 129 維的特徵向量。我們讓 CNN 沿著特徵頻率軸做摺積，並使用 2 層的 CNN，取代 DNN 作為特徵抽取的工具，使經過網路得到的事後機率富含發音鑑別力的資訊。

CNN(a)和(b)使用 40 維度 filter banks 特徵加上 3 維度音調特徵。在 DNN 與 CNN 的隱藏層數量與各層神經元數量的選擇中，DNN 使用基本的 4 層隱藏層，各層有 1024 個神經元；CNN(a)使用 2 組 CNN 層加上 2 層各有 512 個神經元的 DNN 隱藏層；CNN(b)使用 2 組 CNN 層加上 2 層各有 1024 個神經元的 DNN 隱藏層。由於本論文的目的為音素層次的錯誤發音檢測，因此我們將選擇對於華語學習者(L2)且音素錯誤率較低的聲學模型做為產生錯誤發音檢測所需的特徵。

5.3 實驗結果

圖四我們比較了聲學模型 DNN、CNN 分別使用 GOP、SVM、LR 等分類器所產生的 6 種結果，每種結果都是由不同分類器所產生的輸出分數並透過調整門檻值來繪製圖四、五與六的 Recall-Precision 曲線，我們將曲線中召回率與精準度相同的點作為評估標準。其中召回率與精準度所顯示的數值是對於錯誤發音的樣本所做的計算，由於發音正確的樣本數多過於錯誤的發音，因此在本論文的實驗中將不額外探討正確發音的 Recall-Precision 曲線。首先分析單音節的部分(圖四左)，在分類器 GOP、SVM、LR 使用不同聲學模型(CNN 與 DNN)所產生的發音檢測特徵上的表現十分接近。若比較在 DNN 聲學模型中不同分類器的改善，LR 則是勝過 GOP 約 12.60%的大幅度改進。在雙音節中整體表現不如單音節來的優秀，GOP 的曲線在 DNN 與 CNN 中只得到 36.03%與 35.15%，是非常不可靠的分類器，但是雙音節(圖四右)的 DNN 聲學模型在分類器 LR 中的表現相較於 GOP 提升約 18.44%，進步的幅度比單音節更為劇烈，因此我們可以從圖

四的實驗中觀察到，若能給予分類器更多的事後機率做為特徵，將可以得到更好的錯誤發音檢測結果。

整體而言，雙音節的表現皆不如單音節，原因有兩點：首先，進行錯誤發音檢測前，必須先透過聲學模型來擷取事後機率做為檢測用的特徵；而聲學模型皆是用發音正確的語句訓練而成，但是強制對位的音素邊界(boundary)是根據正確語句所訓練的聲學模型而得，因此錯誤發音的強制對位結果將無法預期；這樣的情況在單音節中也會發生，且在雙音節或多音節的語句中將會更嚴重。第二個可能的原因則是雙音節的資料量相較於單音節還要少許多，因此一些較特別的錯誤發音並未在訓練資料中出現。

無論是單音節或雙音節中，可以觀察到分類器 LR 的表現皆優於 SVM；我們使用聲學模型 DNN 產生的檢測特徵所訓練的錯誤發音檢測模型在單音節訓練資料中進行測試(test on train set)，會發現分類器 SVM 的模型對於錯誤發音檢測的 Recall-Precision 相同時可以達到 99.49%，而分類器 LR 則是 86.73%，但是換到測試資料時則是 LR 表現勝過 SVM。我們對於 SVM 效果不如 LR 分類器的現象有兩種解釋：1)由上述現象可觀察到 SVM 發生過度擬合的現象，使得轉換到測試資料進行錯誤發音檢測時的表現不如預期；2)我們使用的 SVM 核心函數會將特徵轉換到較高的維度以便進行線性迴歸分析，可能轉換的方法並不完全適用於測試資料。因此，我們在接下來的實驗將探討分類器 LR 在不同聲學模型以及不同檢測模型之輸出分數在線性組合上的表現。

聲學模型 DNN 與 CNN 在分類器 LR 下各自的表現與結合後的錯誤發音檢測之表現如圖五，我們將 CNN-LR 的分類機率函數定義成$\varphi_{LR}^{CNN}(.)$，DNN-LR 的分類機率函數定義成$\varphi_{LR}^{DNN}(.)$，延續 4.2 小節的特徵$\boldsymbol{f}_{a_{ui}}$，其中$\varphi_{LR}^{CNN}(\boldsymbol{f}_{a_{ui}})$、$\varphi_{LR}^{DNN}(\boldsymbol{f}_{a_{ui}})$可以表示成：

$$\varphi_{LR}^{DNN}(\boldsymbol{f}_{a_{ui}}) = \sigma\left(\left(\mathbf{w}_{a_{ui}}^{DNN}\right)^{T}\boldsymbol{f}_{a_{ui}}\right) \tag{22}$$

$$\varphi_{LR}^{CNN}(\boldsymbol{f}_{a_{ui}}) = \sigma\left(\left(\mathbf{w}_{a_{ui}}^{CNN}\right)^{T}\boldsymbol{f}_{a_{ui}}\right) \tag{23}$$

權重$\mathbf{w}$會因為聲學模型的不同而使用不同的權重($\mathbf{w}^{DNN}$與$\mathbf{w}^{CNN}$)，$\mathbf{w}_{a_{ui}}$表示對應音素a_{ui}的權重，在 4.2 小節有說明$\mathbf{w}_{a_{ui}}$的訓練方式以及提到每個音素應分開訓練，因為各音素的對錯情況各有不同，應避免在同一分類器中產生不必要的混淆。接著我們在定義一個參數λ，其值域為$0 \le \lambda \le 1$，該參數用來線性結合φ^{CNN}與φ^{DNN}的結果：

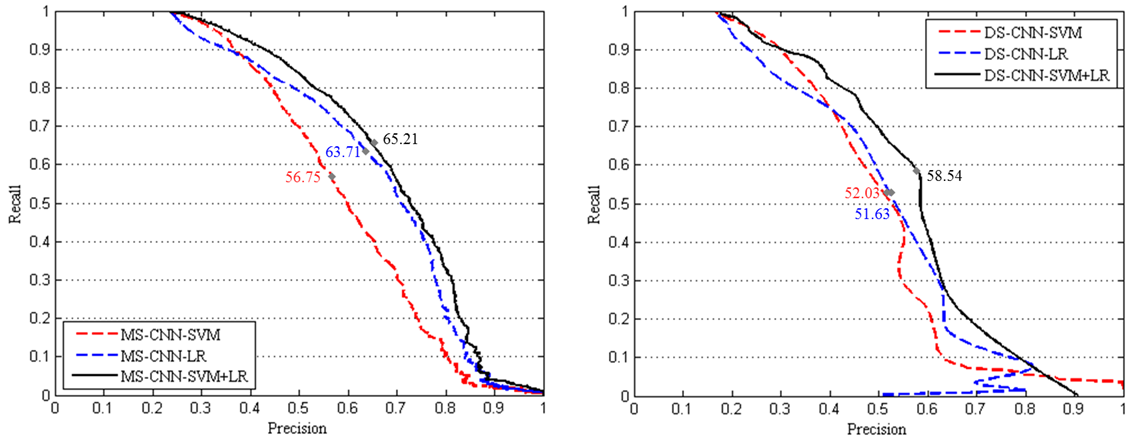
$$g(\boldsymbol{f}_{a_{ui}}) = \lambda \cdot \varphi_{LR}^{DNN}(\boldsymbol{f}_{a_{ui}}) + (1-\lambda)\cdot\varphi_{LR}^{CNN}(\boldsymbol{f}_{a_{ui}}) \tag{24}$$

$g(\boldsymbol{f}_{a_{ui}})$則為錯誤發音檢測模型 DNN-LR 與 CNN-LR 輸出分數的結合，如同 4.1 小節的式(8)定義門檻值τ來決定發音為正確或錯誤，在圖五的實驗中我們將λ設定為 0.5，並調整門檻值τ畫出圖五的曲線，從圖五(左)單音節實驗中可以觀察到線性結合兩種特徵所產生的機率值將可以得到不錯的成效，由 DNN-LR 的 63.81%進步至線性結合後的 67.23%約有 3.42%的進步，而雙音節(圖五右)經過線性結合後從 54.47%到 55.98%得到 1.51%的進步。

接著在圖六中我們將探討使用 CNN 聲學模型所擷取的特徵在不同分類器(SVM、LR)結合的效果，由於不同分類器的輸出值域並不一致，所以我們不使用式(24)的結合方式，在此我們基於每個音素在不同分類器之結果的排名，並對排名結果計算調和平均數(harmonic mean)，我們定義iRank$\left(\varphi_{SVM}^{CNN}(\boldsymbol{f}_{a_{ui}})\right)$表示成特徵$\boldsymbol{f}_{a_{ui}}$在測試集的分類器 SVM 輸出分數由低到高排名，也就是從錯誤發音排到發音正確，因此定義調和平均函

數$h(.)$可表示為：

$$h(\boldsymbol{f}_{a_{ui}}) = \frac{2 \cdot \text{iRank}\left(\varphi_{SVM}^{CNN}(\boldsymbol{f}_{a_{ui}})\right) \cdot \text{iRank}\left(\varphi_{LR}^{CNN}(\boldsymbol{f}_{a_{ui}})\right)}{\text{iRank}\left(\varphi_{SVM}^{CNN}(\boldsymbol{f}_{a_{ui}})\right) + \text{iRank}\left(\varphi_{LR}^{CNN}(\boldsymbol{f}_{a_{ui}})\right)} \tag{25}$$



圖六、比較雙音節(左側)與單音節(右側)從 CNN-DNN 萃取出的特徵使用不同分類器(SVM 與 LR)輸出分數的線性組合(M/DS-CNN-SVM+LR)之 Recall-Precision 曲線

因此函數$h(.)$的輸出分數如同函數$\varphi(.)$和$g(.)$，越高表示正確發音、越低表示錯誤發音，函數$h(.)$與iRank(.)的值域為$1\sim L$，常數L表示測試集的音素層次樣本數。從圖六(左)可以發現聲學模型 CNN 之特徵用於分類器 SVM 與 LR 之結合在單音節的表現(65.21%)不如圖五(左)的不同特徵用於分類器 LR 之結合(67.23%)。雙音節的表現則是相反，在分類器 SVM 與 LR 結合的成效勝過模型 DNN 與 CNN 的結合，結果分別為 58.54%(如圖六右)與 55.98%(如圖五右)。

除了探討在 Recall-Precision 曲線的表現外，我們也在圖七與圖八個別列出單音節與雙音節在 ROC 曲線上的表現，而 ROC 空間值個數可分為四種：真陽性(true positive, TP)：系統推測為正確發音，實際上也是正確發音；真陰性(true negative, TN)：系統推測為錯誤發音，實際上也是錯誤發音；偽陽性(false positive, FP)：系統推測為正確發音，實際上為錯誤發音；偽陰性(false negative, FN)：系統推測為錯誤發音，實際上為正確發音，而圖七與圖八皆是藉由調整門檻值得到不同的真陽性率(true positive rate, TPR)與偽陽性率(false positive rate, FPR)所繪製而成的曲線，TPR 與 FPR 的計算方式如式(26)與式(27)：

$$\text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{26}$$

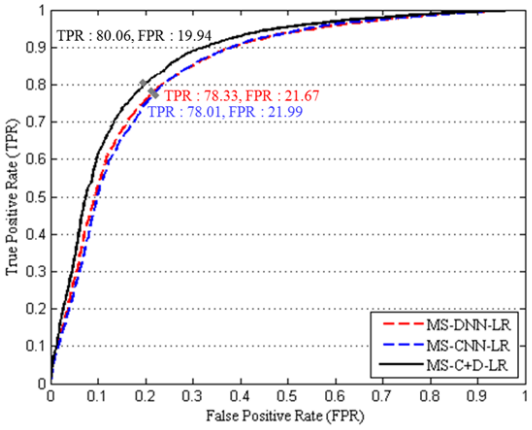
$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})} \tag{27}$$

單音節曲線下面積(area under the curve of ROC, AUC)顯示於表三，實驗中的 AUC 是使用梯形法(trapezoid method)求得，從 DNN 結合 CNN 的模型(如表三的 C+D)之表現來看，TP 與 TN 都有所提升，而 FP 與 FN 也有明顯的下降，AUC 的部分相較 DNN 的 84.42%則進步了 2.28%達到約 86.70%，從圖七中可以清楚看到 DNN 與 CNN 結合之聲學模型的表現優於 CNN 與 DNN 各自使用；我們將圖七左上與右下之對角線相連求出

TPR 與假 FPR 相加趨近於 1 的點，該點所表示的 FPR 稱作相等錯誤率(equal error rate, EER)，在表三的 ROC 空間值個數(TP、TN、FP、FN)是利用該點求得；圖七可以發現 DNN 的 EER 為 21.67%，而經過結合的模型(如圖七的 C+D)可降低至 19.94%。在表四與圖八則是顯示雙音節的 ROC 空間值個數、ROC 曲線與 EER，圖八可以發現 DNN 的 EER 為 24.30%，而經過結合的模型(如圖八的 C+D)可降低至 23.08%；表四在 AUC 的部分可以觀察到模型結合後從 CNN 的 80.90%進步到 82.58%。

表三、單音節在分類器 LR 在不同聲學模型的 ROC 空間值(TP、TN、FP 與 FN)與曲線下面積(AUC)之比較

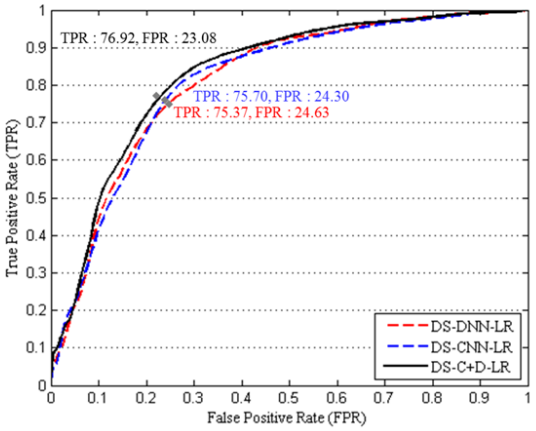
-	TP	TN	FP	FN	AUC (%)
DNN	9,609	2,896	805	2,658	84.42
CNN	9,569	2,880	821	2,698	84.02
C+D	9,821	2,967	734	2,446	86.70



圖七、單音節在分類器 LR 在不同聲學模型的 ROC 曲線

表四、雙音節在分類器 LR 在不同聲學模型的 ROC 空間值(TP、TN、FP 與 FN)與曲線下面積(AUC)之比較

-	TP	TN	FP	FN	AUC (%)
DNN	921	185	61	301	80.90
CNN	925	186	60	297	81.18
C+D	940	189	57	282	82.58



圖八、雙音節在分類器 LR 在不同聲學模型的 ROC 曲線

六、 結論與未來研究展望

本論文探討兩種聲學模型(DNN 和 CNN)以及它們的結合對於發音檢測效能的影響。另一方面，從實驗結果可以發現，本論文所使用的三種分類方法(GOP、SVM 和 LR)中無論是單音節或雙音節皆以 LR 表現最佳。雖然 DNN-LR 與 CNN-LR 兩種錯誤發音檢測模型之表現十分相近，但經過簡單的線性組合後依然可以在單音節錯誤發音檢測的召回率與精準度相同時得到 3.42%的進步並達到 67.23%的表現；同時，在雙音節錯誤發音檢測上，經過線性組合後也得到 1.51%的進步並提升至 55.98%。而 ROC 曲線在單音節

跟雙音節皆因為模型的結合使得 EER 與 AUC 的表現都有所提升。雖然 DNN-LR 與 CNN-LR 各自使用的結果並無明顯的差異，但結合時的效果卻出乎意料，這表示不同的聲學模型產生的發音檢測特徵可能具有互補性。希望在未來的研究中可以使用更好的聲學模型特徵(如鑑別式訓練後的聲學模型所產生的特徵)，除了聲學模型所提供的相似度值特徵外，未來嘗試加入韻律(prosodic)特徵並探討錯誤發音檢測結果的影響；另一方面希望探究不同結合方式與各式分類技術在錯誤發音檢測的表現，並且更詳細與廣泛地探討各種聲學模型所擷取的發音特徵之優缺點。

致謝

本論文之研究承蒙教育部－國立臺灣師範大學邁向頂尖大學計畫(104-2911-I-003-301)與行政院科技部研究計畫(MOST 104-2221-E-003-018-MY3, MOST 103-2221-E-003-016-MY2, NSC 103-2911-I-003-301)之經費支持，謹此致謝。

七、 參考文獻

[1] "40 million people worldwide study Chinese," http://english.people.com.cn/90001/90782/90872/7112508.html .

[2] W. Hu, Y. Qian, and F. Soong, "A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training," in *Proc. ICASSP*, pp. 3230–3234, 2013.

[3] G. Hinton, L. Deng, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[4] Y. Bengio, "Practical recommendations for gradient-based ttraining of deep architectures," *Neural Networks: Tricks of the Trade*, K.R. Muller, G. Montavon, and G.B. Orr, eds., Springer 2013.

[5] D. E. Rumelhart, G.E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, 1986, vol. 323, pp. 533–536.

[6] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. ICASSP*, pp. 4277–4280, 2012.

[7] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. ICASSP*, pp. 8614–8618, 2013.

[8] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-score discriminative training criterion for automatic mispronunciation detection," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 5 pp. 787–797, April. 2015.

[9] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. Eurospeech*, pp. 851–854, 1999.

[10] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[11] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. H. Wang, "Automatic mispronunciation detection for Mandarin," in *Proc. ICASSP*, pp. 5077–5080, 2008.

[12] Y.B. Wang and L.S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. ICASSP*, pp. 5049–5052, 2012.

[13] Ito, A., Lim, Y., Suzuki, M., Makino, S., "Pronunciation error detection method based on

error rule clustering using a decision tree", in *Proc. EuroSpeech*, pp. 173–176, 2005.
[14] K. Truong, A. Neri, C. Cuchiarini, and H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," in *Proc. of the InSTIL/ICALL Symposium*, pp. 135–138, 2004.
[15] S. Wei, G. Hu, Y. Hu, and R. H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication.*, vol. 51, no. 10, pp. 896–905, 2009.
[16] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.
[17] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. Int. Conf. Spoken Lang. Process.*, pp. 1457–1460, 1996.
[18] L. Y. Chen and J. S. R. Jang, "Automatic pronunciation scoring using learning to rank and DP-based score segmentation," in *Proc. Interspeech*, pp. 761–764, 2010.
[19] L. Y. Chen and J. S. R. Jang, "Improvement in automatic pronunciation scoring using additional basic scores and learning to rank," in *Proc. Interspeech*, 2012.
[20] L. Y. Chen and J. S. R. Jang, "Automatic pronunciation scoring with score combination by learning to rank and class-normalized DP-based quantization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 11 pp. 787–797, November. 2015.
[21] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "An application of pretrained deep neural networks to large vocabulary speech recognition," submitted for publication.
[22] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. ICASSP*, 2013.
[23] X. Qian, H. Meng, and F. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in *Proc. Interspeech*, pp. 775–778, 2012.
[24] Y. Ke. "Acoustic model optimization for automatic pronunciation quality assessment," in *Proc. ICMFI*, 2014.
[25] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Proc. Neural Information and Processing Systems*, 2012.
[26] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.
[27] Y. Le Cun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA: MIT Press, pp. 255–258, 1995.
[28] D. Hau and K. Chen, "Exploring hierarchical speech representations using a deep convolutional neural network," in *Proc. UKCI*, 2011.
[29] D. Yu and L. Deng, "Automatic speech recognition - a deep learning approach", Springer, 2014.
[30] G. Hinton, L. Deng, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[31] G. E. Dahl, M. Ranzato, A. Mohamed, and G. E. Hinton, "Phone recognition with the mean-covariance restricted boltzmann machine," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. ShaweTaylor, R.S. Zemel, and A. Culotta, Eds. Cambridge, MA: MIT Press, pp. 469–477, 2010.
[32] R. Salakhutdinov and G.E. Hinton, "Deep boltzmann machines," in *Proc. AISTATS*, pp.

448-455, 2009.
[33] H. Lee, P. Pham, Y. Largman, and A. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Cambridge, MA: MIT Press, pp. 1096–1104, 2009.
[34] A. Mohamed, G. Hinton, and G. Penn, “Understanding how deep belief networks perform acoustic modelling,” in *Proc. ICASSP*, pp. 4273–4276, 2012.
[35] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pretrained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. Audio Speech Lang. Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[36] A. Mohamed, T. N. Sainath, G. E. Dahl, B. Ramabhadran, G. E. Hinton, and M. Picheny, “Deep belief networks using discriminative features for phone recognition,” in *Proc. ICASSP*, pp. 5060–5063, 2011.
[37] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular in teraction, and functional architecture in the cat’s visual cortex,” *J. Physiology (London)*, vol. 160, pp. 106–154, 1962.
[38] J. Bouvrie, “Notes on convolutional neural networks,” 2006.
[39] D. Scherer, A. Muller, and S. Behnke, “Evaluation of pooling operations in convolutional architectures for object recognition,” in *Proc. ICANN*, pp. 92–101, 2010.
[40] Python – scikit-learn. http://scikit-learn.org/dev/index.html
[41] X. Qian, H. M. Meng, and F. K. Soong, “Discriminatively trained acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT),” in *Proc. Interspeech*, 2010.
[42] S. Wei, H. Wang, Q. Liu, and R. Wang, “CDF-matching for automatic tone error detection in Mandarin CALL system,” in *Proc. ICASSP*, pp. 205–208, 2007.
[43] J. Cheng, “Automatic tone assessment of non-native Mandarin speakers,” in *Proc. Interspeech*, pp. 1299–1302, 2013.
[44] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motl´ıcek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *Proc. IEEE ASRU*, 2011.