Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 55-70 


# 評估尺度相關最佳化方法於華語錯誤發音檢測之研究

# Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection

許曜麒*、楊明翰*、洪孝宗*、林奕儒*、陳冠宇+、陳柏琳*

**Yao-Chi Hsu, Ming-Han Yang, Hsiao-Tsung Hung,**

**Yi-Ju Lin, Kuan-Yu Chen and Berlin Chen**

**摘要**

錯誤發音檢測(Mispronunciation Detection)與錯誤發音診斷(Mispronunciation Diagnosis)為電腦輔助發音訓練系統的一部分，它們能輔助第二外語學習者準確地找出語句中錯誤發音的部位以增進學習者的口說熟練度。本論文延續過去學者的研究，大致可將貢獻分為三點：1) 比較不同的發音分數做為錯誤發音檢測的評估依據，並探討對於錯誤發音檢測效能的影響；2) 我們透過最佳化評估尺度相關訓練法則估測深層類神經網路聲學模型的參數以及發音檢測決策函數之參數；3) 使用 $F_1$ 度量作為目標函數時，若將二類的 $F_1$ 度量線性組合並調整權重，可有效處理資料類別不平衡的問題。一系列的實驗將建立在華語錯誤發音檢測與診斷任務，從實驗中可以觀察到我們提出的方法之優點。

**關鍵詞：**電腦輔助發音訓練、錯誤發音檢測、自動語音辨識、鑑別式訓練與深層類神經網路。

*國立台灣師範大學資訊工程學系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {ychsu, mh_yang, alexhung, lin_yj, berlin}@ntnu.edu.tw
+中央研究院資訊科學所
Institute of Information Science, Academia Sinica
E-mail: kychen@iis.sinica.edu.tw

**Abstract**

Mispronunciation detection and diagnosis are part and parcel of a computer assisted pronunciation training (CAPT) system, collectively facilitating second-language (L2) learners to pinpoint erroneous pronunciations in a given utterance so as to improve their spoken proficiency. This thesis presents a continuation of such a general line of research and the major contributions are three-fold. First, we compared the performance of different pronunciation features in mispronunciation detection. Second, we propose an effective training approach that estimates the deep neural network based acoustic models involved in the mispronunciation detection process by optimizing an objective directly linked to the ultimate evaluation metric. Third, we can linearly combine two $F_1$-score when we consider $F_1$-score as final objective function. It can effectively deal with the label imbalance problem. A series of experiments on a Mandarin mispronunciation detection task seem to show the performance merits of the proposed methods.

**Keywords:** Computer Assisted Pronunciation Training, Mispronunciation Detection, Automatic Speech Recognition, Discrimetive Training, Deep Neural Networks.

## 1. 緒論

全球化時代來臨，為提升個人的競爭力，外語能力已列為基本的技能之一。因此電腦輔助語言學習(Computer Assisted Language Learning, CALL)在現今已是非常具有潛力的研究；其目的是透過電腦自動判斷外語學習者的學習狀況並給予有幫助的回饋。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，華語已經逐漸成為一種必須學習的語言(Hu, Qian, Soong & Wang, 2014)。語言學習可分為聽(Listening)、說(Speaking)、讀(Reading)和寫(Writing)等四類學習面向，其中口說與書寫測驗的評量往往需要專業的語言教師來評斷，但語言教師的培養尚無法滿足遍佈全球的需求。本篇論文將專注於電腦輔助發音訓練(Computer Assisted Pronunciation Training, CAPT)，也就是「說」的技術進行討論。

電腦輔助發音訓練最主要目的就是要讓第二外語(Second-Language, L2)學習者有更多的機會練習發音；過去第二外語學習者要進行發音練習都需要配合語言教師的授課時間，若將電腦輔助發音訓練普及到現有的智慧型行動裝置，將會有更多的第二外語學習者因此受惠。電腦輔助發音訓練中的首要任務為自動錯誤發音檢測；檢測過程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(Witt & Young, 2000)、音節(Syllable)層次(Zhang, Huang, Soong, Chu & Wang, 2008)或詞(Word)層次(Chen & Jang, 2015)。當系統指出學習者的錯誤發音時，將可以針對該錯誤發音進行

偏誤回饋，該階段被稱為錯誤發音診斷(Harrison, Lau, Meng & Wang, 2008; Harrison, Lo, Qian & Meng, 2009; Lo, Zhang & Meng, 2010; Wang & Lee, 2012; Wang & Lee, 2015)。錯誤發音檢測為電腦輔助發音訓練中的第一步，當錯誤發音檢測可以精準的預測學習者的發音狀況時，才能有效的進行錯誤發音診斷。本研究主旨在探討如何提升錯誤發音檢測之效能？錯誤發音檢測可視為二類分類問題，對於發音檢測的結果在語言專家與系統的決策之間可以產生四種結局：若學習者的發音正確，系統卻判斷為發音錯誤稱為是錯誤的拒絕(False Rejections, FR)；而學習者發音錯誤，系統認定為發音正確則稱為錯誤的接受(False Acceptances, FA)；學習者發音正確，系統判斷為發音正確稱為正確的接受(True Acceptances, TA)；學習者發音錯誤，系統判定為發音錯誤稱為正確的拒絕(True Rejections, TR)。上述的四種指標可以計算出其它評估的標準，例如召回率(Recall)與精準度(Precision)，有許多發音檢測的研究皆以該評估方式作為評量系統優劣的準則(Hu *et al.*, 2015; Huang, Xu, Wang & Silamu, 2015)。我們可更進一步使用召回率與精準度的調和平均－$F_1$度量($F_1$-Score)做為準則，$F_1$度量在自然語言處理(Natural Language Processing, NLP)與資訊檢索(Information Retrieval, IR)等研究中廣為使用，甚至有許多任務直接將該指標作為模型訓練的目標(Fujino, Isozaki & Suzuki, 2008; Dembczynski, Waegeman, Cheng & Hüllermeíer, 2011; Ye, Chai, Lee & Chieu, 2012)。在錯誤發音檢測任務中也有類似想法的研究已被探討(Huang *et al.*, 2015; Qian, Soong & Meng, 2010; Huang, Wang & Abudureyimu, 2012)。

近年來，在語音辨識系統中的聲學模型已由深層類神經網路(Deep Neural Network, DNN)取代傳統的高斯混合模型(Gaussian Mixture Model, GMM)，並在語音辨識任務上取得巨大的進步(Hinton *et al.*, 2012)。在錯誤發音檢測的相關研究中也因為深層類神經網路聲學模型的使用而在效能上有顯著的提升(Hu *et al.*, 2015; Qian, Meng & Soong, 2012; Hu *et al.*, 2014)。基於上述研究的啟發，我們延續過去學者以最大化錯誤發音檢測任務的效能(Huang *et al.*, 2015; Hsu, Yang, Hung & Chen, 2016)為目標函數對模型進行調整的想法，並實作於深層類神經網路聲學模型的架構上探討對於錯誤發音檢測任務的影響。

本篇論文在第二節將介紹錯誤發音檢測相關研究的發展近況；第三節簡單的回顧錯誤發音檢測任務中較常被使用的方法；第四節則是討論基於第三節的發音檢測方法如何實現最大化錯誤發音檢測$F_1$度量之訓練；第五節則是從實驗中探討最大化錯誤發音檢測$F_1$度量之訓練對於發音檢測任務的影響；最後，在第六節，我們提出結論與一些未來可能的研究方向。

## 2. 文獻探討

錯誤發音檢測大致可分為基於門檻值(Thresholding-Based)與基於分類器(Classification-Based)等兩種做法。兩者差別在於是否使用明確的門檻值來判斷發音為正確或錯誤；基於分類器則是整合多種特徵並訓練二元分類器來決定發音是否合格。基於門檻值等方法早期由(Hsu *et al.*, 2016)提出三種發音檢測特徵：對數相似度值(Log-Likelihood)、對數事後機率(Log Posterior Probability)、段落區間長度(Segment

Duration)對於發音檢測效果的影響。學者 Kim 在實驗中指出對數事後機率為表現較好的發音檢測分數 (Kim, Franco & Neumeyer, 1997)。之後則有學者簡化事後機率的計算方式並將其稱作 GOP (Goodness of Pronunciation) (Witt & Young, 2000)，之後也有研究針對 GOP 等方法進行改良(Zhang *et al*., 2008)。因為基於門檻值之方法展現了簡潔有效的優點，所以學者們提出以最大化錯誤發音檢測之$F_1$度量作為目標對聲學模型進行鑑別式訓練 (Huang *et al*., 2012)。

而基於分類器的發音檢測方法，較早是由(Wei, Hu, Hu & Wang, 2009)所提出的，陸續也有許多不同的發音特徵(Lee & Glass, 2012; Laborde *et al*., 2016)或是不同分類模型 (Hu *et al*., 2015)。事實上，在錯誤發音診斷任務中，早已開始整合多種特徵，例如引入韻律特徵(Strik, Truong, De Wet & Cucchiarini, 2007)。但這類的做法都只有在特定的發音才能使用(例如荷蘭語的/x/或/k/)，且韻律特徵容易因為不同語者而產生無法預期的變化。然而，也有學者基於語音辨識模組來進行發音診斷(Hu *et al*., 2015)，但從實驗數據看來距離理想的準確率還有一段差距。有些學者認為錯誤發音檢測與診斷應該要視為語音辨識的任務(Harrison *et al*., 2008; Harrison *et al*., 2009; Qian *et al*., 2012)，將訓練資料的錯誤型態(Error Pattern)都記錄在模型中；倘若測試資料出現訓練時從未的錯誤型態，辨識結果將會無法預期，且該情況會因為外語學習者的母語不同使得更容易發生。

## 3. 錯誤發音檢測

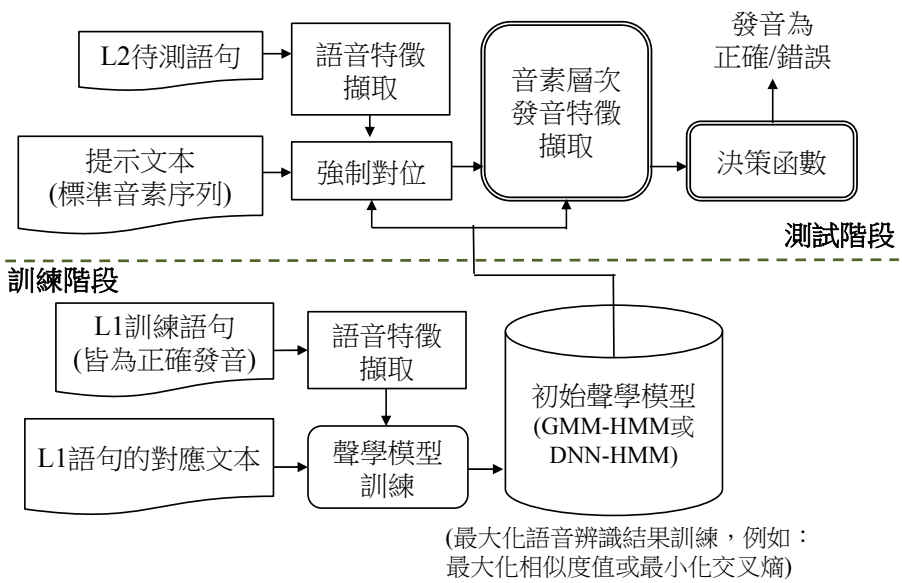


***圖 1. 基礎錯誤發音檢測流程圖***
***[Figure 1. The flowchart of the mispronunciation detection process.]***

錯誤發音檢測的基本流程如圖 1 所示。我們首先使用母語者的語料庫訓練語音辨識所需的聲學模型，在將外語學習者的發音語句與正確的文本做強制對位；接著將聲學模型算出的事後機率作為發音檢測特徵進行錯誤發音檢測。錯誤發音檢測的早期研究中，有學者延續(Kim *et al*., 1997)的觀察並將事後機率改良並稱作 GOP (Witt & Young, 2000)，也是最常被使用的發音檢測方法。GOP 的計算方式如下：

$$\mathrm{GOP}(u,n) \equiv \frac{1}{T_{u,n}} \log P\left(q_{u,n}|\mathbf{O}_{u,n}\right) \tag{1}$$

$$= \frac{1}{T_{u,n}} \log \frac{p(\mathbf{O}_{u,n}|q_{u,n})P(q_{u,n})}{\sum_{\tilde{q}\in Q_{u,n}} p(\mathbf{O}_{u,n}|\tilde{q})P(\tilde{q})} \tag{2}$$

$$\approx \frac{1}{T_{u,n}} \log \frac{p(\mathbf{O}_{u,n}|q_{u,n})}{\max_{\tilde{q}\in Q_{u,n}} p(\mathbf{O}_{u,n}|\tilde{q})} \tag{3}$$

其中 GOP 是音素段落$\mathbf{O}_{u,n}$對應目標音素$q_{u,n}$的事後機率，其中 $u$ 與 $n$ 表示第 $u$ 個語句的第 $n$ 個音素，根據貝氏定理將式(1)轉換成式(2)；$Q_{u,n}$是該段落對應的音素集合，可以是全部音素或部分較混淆的音素，$T_{u,n}$則是音素段落的經歷時間(Duration)。我們假設每個音素的事前機率相同，且只使用最大相似度值的音素，即最混淆音素做為分母項，如式(3)。其中$p\left(\mathbf{O}_{u,n}|q_{u,n}\right)$是已知音素$q_{u,n}$要取得音素段落$\mathbf{O}_{u,n}$的相似度值，計算$p\left(\mathbf{O}_{u,n}|q_{u,n}\right)$可以透過已知的文本內容對語句進行強制對位取得對應音素$q_{u,n}$的狀態序列$\mathbf{s}^* = \{s_{t_s}, s_{t_s+1}, \dots, s_{t_e}\}$，同時也可以得到音素段落區間對應的起始時間$t_s$與結束時間$t_e$。式(3)所計算的 GOP 分數作為決策發音錯誤與否的評估依據，並經過式(3)決定發音程度的分數。我們定義函數 D(·)表示發音的決策函數：

$$\mathrm{D}(u,n) = \frac{1}{1+\exp(\alpha\cdot\mathrm{GOP}(u,n)+\beta)} \tag{4}$$

而 D(·)接近 1 表示發音可能錯誤，接近 0 則表示發音正確，$\beta$表示決策用的門檻值，而參數$\alpha$用來將 GOP 分數放大或縮小。上述兩個參數皆可以設計為音素相依，若為音素相依則用$\boldsymbol{\alpha}$與$\boldsymbol{\beta}$表示。接著我們利用指示函數判定發音是否錯誤：

$$\mathbb{I}(\mathrm{D}(u,n)) = \begin{cases} 1 & \text{if } \mathrm{D}(u,n) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$\tau$為全域的固定門檻值，大部分都是透過發展集調整至一個較合適的值。然而 GOP 是錯誤發音檢測的方法中較普遍的作法，且不需依賴人工標記好的錯誤發音，屬於非監督式學習(Unsupervised Learning)的方法。

此外，已有學者提出利用深層類神經網路聲學模型的輸出為事後機率$P(s_t|\mathbf{o}_t)$的方法作為發音檢測的分數，稱作對數音素事後機率(Log Phone Posterior, LPP) (Hu *et al*., 2015)。其計算方式為音素段落$\mathbf{O}_{u,n}$對應的狀態事後機率之幾何平均。與 GOP 的算法類似，透過已知的文本內容對語句進行強制對位取得對應目標音素$q_{u,n}$的狀態序列$\mathbf{s}^{(q_{u,n})} =$

$\{s_{t_s}, s_{t_s+1}, \ldots, s_{t_e}\}$，而計算 LPP 的公式可以寫成：

$$\mathrm{LPP}(u,n) = \log P(q_{u,n}|\mathbf{O}_{u,n}; t_s, t_e) \tag{6}$$

$$\approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log P\left(s_t^{(q_{u,n})} \middle| \mathbf{o}_t\right) \tag{7}$$

透過式(7)算出目標音素$q_{u,n}$的 LPP，$\mathbf{s}^{(q_{u,n})}$為音素$q_{u,n}$在音素段落$\mathbf{O}_{u,n}$的最佳路徑所對應的狀態序列。從我們實驗中可以發現使用 LPP 產生發音分數在發音檢測任務的效果與 GOP 相近，但 LPP 的計算複雜度遠低於 GOP。如式(3)所見，GOP 在分母項需要將所有音素的相似度值算出；而 LPP 只需要計算目標音素$q_{u,n}$的狀態事後機率之幾何平均，符合深層類神經網路架構的輸出狀態事後機率。當我們取得以 LPP 表示的發音分數後，寫成決策函數的形式則為：

$$\mathrm{D}(u,n) = \frac{1}{1+\exp(\alpha \cdot \mathrm{LPP}(u,n) + \beta)} \tag{8}$$

## 4. 最大化錯誤發音檢測$\mathbf{F_1}$度量之訓練

在發音檢測任務中有許多研究都以改良 GOP 為主軸提升錯誤發音檢測的效能，近期有學者將鑑別式訓練應用在 GOP 估測，以最大化 $F_1$ 度量為目標作鑑別式訓練(Huang *et al.*, 2015)，學者 Huang 使用高斯混合模型-隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)建構聲學模型，並使用 GOP 進行錯誤發音檢測，透過調整聲學模型中的參數來提升錯誤發音檢測的表現。而在本論文，我們將聲學模型改用深層類神經網路-隱藏式馬可夫模型(Deep Neural Networks-Hidden Markov Model, DNN-HMM)，在錯誤發音檢測的部分則用 LPP 做為發音分數，透過決策函數決定發音是否錯誤；並以最大化$\mathrm{F_1}$度量為目標做鑑別式訓練更新深層類神經網路聲學模型的參數以及決策函數的參數。首先，$\mathrm{F_1}$度量的計算方式如下：

$$F_1 = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{9}$$

其中$F_1$為召回率與準確率兩種評估標準的調合平均，而召回率與準確率可以表示為：

$$\mathrm{Precision} = \frac{C_{H \cap D}}{C_D} \tag{10}$$

$$\mathrm{Recall} = \frac{C_{H \cap D}}{C_H^{(\mathcal{M})}} \tag{11}$$

$C_D$表示訓練資料中被系統標記為錯誤發音的音素數量；$C_H^{(\mathcal{M})}$則是訓練資料中被語言專家標記為錯誤發音的音素數量，因此該值是一個固定的常數；而$C_{H \cap D}$就是系統與語言專家同時認為該音素段落為錯誤發音的數量。將式(10)與式(11)代入式(9)並簡化後得到：

$$F_1 = \frac{2C_{H\cap D}}{C_D + C_H^{(\mathcal{M})}} \tag{12}$$

接著將我們在第 3 節定義的錯誤發音決策函數 D(·)透過指示函數𝟙(·)轉成非 1 即 0 的數值，訓練資料的所有音素段落經過決策函數 D(·)與指示函數𝟙(·)的總和為$C_D$；每個音素段落的決策與專家評斷之結果 H(·)相乘的總和則為$C_{H\cap D}$，如式(13)：

$$F_1 = \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathbb{1}(\mathrm{D}(u,n))\cdot \mathrm{H}(u,n)}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathbb{1}(\mathrm{D}(u,n)) + C_H^{(\mathcal{M})}} \tag{13}$$

然而上述定義的$\mathrm{F}_1$度量並不是可微分的函數，因為在計算$C_{H\cap D}$與$C_D$時使用到的指示函數𝟙(·)在基於梯度法(Gradient Based Method)的參數更新方式中較難處理。因此我們定義一個平滑(Smooth)的$\mathrm{F}_1$度量，如式(14)：

$$\Xi(\boldsymbol{\theta}) = \frac{2\cdot C_{H\cap D}^{S}}{C_D^{S} + C_H^{(\mathcal{M})}} \tag{14}$$

$$= \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)\cdot \mathrm{H}(u,n)}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n) + C_H^{(\mathcal{M})}} \approx F_1 \tag{15}$$

由於錯誤發音決策函數 D(·)已將發音檢測分數經過激發函數轉為 0 到 1 之間的值，因此計算$C_{H\cap D}^{S}$與$C_D^{S}$時不使用指示函數 𝟙(·)也可以近似$\mathrm{F}_1$度量的算法，如式(15)。定義了目標函數後，我們使用隨機梯度上升法(Stochastic Gradient Ascent Algorithm)來更新參數。最後對於最大化$\mathrm{F}_1$度量鑑別式訓練的流程列出摘要：

(1) 首先透過華語母語者(L1)訓練資料使用於 DNN-HMM 聲學模型的訓練，且訓練資料皆為正確發音，並以最小化交叉熵為目標函數更新聲學模型。

(2) 基於步驟(1)訓練的聲學模型，透過第 3 節提及的 LPP 算法(式(7))得出每筆訓練資料的發音分數，接著透過決策函數(式(8))將發音分數轉成決策值(值域 0 到 1 之間)。

(3) 接續步驟(2)算出的決策值透過式(15)算出近似的$\mathrm{F}_1$度量作為目標函數並迭代的訓練決策函數的參數$\boldsymbol{\alpha}, \boldsymbol{\beta}$以及 DNN-HMM 聲學模型的參數，而決策函數的參數可為音素相依。

相較於原本的流程在訓練資料中加入了二語(L2)的資料(包含正確與錯誤發音)；且決策函數與聲學模型的參數也以發音檢測任務的目標函數進行調適。

## 5. 實驗結果

### 5.1 華語學習者口語語料庫

本論文使用臺灣師範大學邁向頂尖大學計畫所提供的華語學習者口語語料庫(Hsiung & Sung, 2014)，語者部分包含華語母語者(L1)與華語非母語者(L2)，錄音內容有單音節、雙音節、多音節與短文等情境；其中華語非母語者語料庫為音素層次的發音標記，每筆資

料皆是由 1 至 4 人進行審視，並採用多數決判斷發音為正確或錯誤。我們將語料庫分成訓練集、發展集與測試集，如表 1。

*表 1. 華語學習者口語語料庫*
*[Table 1. Statistics of the Mandarin Annotated Spoken Corpus.]*

| | | 時間(小時) | 語者(個) | 音素數量(個) | 發音錯誤之音素數量(個) |
|---|---|---|---|---|---|
| 訓練集 | L1 | 6.68 | 44 | 72,846 | NA |
| | L2 | 14.04 | 74 | 107,202 | 24,150 |
| 發展集 | L1 | 1.4 | 10 | 14,186 | NA |
| | L2 | 3.39 | 18 | 25,900 | 5,227 |
| 測試集 | L1 | 3.21 | 25 | 32,568 | NA |
| | L2 | 7.49 | 44 | 55,190 | 14,247 |

## 5.2 聲學模型訓練

本研究的語音辨識模組的建立是使用美國約翰霍普金斯大學學者所發展的大詞彙連續語音辨識工具－"Kaldi"(Povey *et al*., 2011)；其它的實驗以 Python 程式語言為主，並參考各種函式庫像是"Scikit-learn"(Pedregosa *et al*., 2011)和"Theano"(Bergstra *et al*., 2010)等，其提供機器學習或深層學習與 GPGPU 運算結合的開發環境。在聲學模型的設定我們採每個音素基於 GMM-HMM 的聲學模型皆由 3 個狀態所組成，而每個狀態至少 16 個高斯混合而成，並以的 L1 訓練集與 L1 發展集作為訓練資料調整聲學模型參數。輸入特徵為梅爾倒頻譜係數，每個音框由 12 維梅爾倒頻譜係數、1 維能量特徵和 3 維度音調(pitch)特徵所組成；並對 16 維語音特徵取相對的一階差量係數(Delta Coefficient)和二階差量係數(Acceleration Coefficient)合併成 48 維的特徵向量；其中取一階和二階差量係數是為了獲得語音特徵在時間的相關資訊。

在 DNN-HMM 聲學模型的部分，每個隱藏層皆使用邏輯函數(sigmoid function)作為激發函數，到輸出層則使用軟式最大化函數(Softmax)轉換成機率。輸入特徵是梅爾頻譜係數(Mel-Scale 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；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 維的濾波器組產生加上 3 維度音調(Pitch)特徵；並對 43 維語音特徵取相對的一皆差量係數和二皆差量係數，則輸入的語音特徵就會得到 11 個 129 維的特徵向量串成一個 1,419 維度的特徵向量。

在自動語音辨識的結果我們以音節錯誤率(Syllable Error Rate, SER)與音素錯誤率(Phone Error Rate, PER)來表示，如表 2；解碼過程為自由音節解碼(Free-Syllable Decoding)且無任何語言模型限制，辨識錯誤率是使用表 1 的 L1 測試集所計算的。從表 2 的辨識結果可以觀察到無論是音節錯誤率或是音素錯誤率皆由 DNN-HMM 聲學模型大幅度勝

過 GMM-HMM 聲學模型。

*表 2. 自動語音辨識實驗結果*
***[Table 2. ASR experimental results.]***

| | 音節錯誤率(%)<br>(syllable error rate, SER) | 音素錯誤率(%)<br>(phone error rate, PER) |
|---|---|---|
| GMM-HMM | 50.87 | 34.30 |
| DNN-HMM | 41.71 | 28.14 |

## 5.3 評估方法

*表 3. ROC 分析的四項指標在發音檢測任務中的定義*
***[Table 3. The definition of the confusion matrix used in the mispronunciation detection task.]***

| | 描述 |
|---|---|
| 錯誤的接受<br>(false acceptances, FA) | 實際上學習者的發音錯誤，系統卻認定為發音正確。 |
| 錯誤的拒絕<br>(false rejections, FR) | 實際上學習者的發音正確，系統卻判斷為發音錯誤。 |
| 正確的接受<br>(true acceptances, TA) | 實際上學習者的發音正確，系統也判斷為發音正確。 |
| 正確的拒絕<br>(true rejections, TR) | 實際上學習者的發音錯誤，系統也認定為發音錯誤。 |

如同本論文在第 1 節提及的，二分類問題會有四種結局(如表 3)，基於這四項指標可以延伸出非常多變的評估方式；例如召回率與精準度是分類問題中經常被使用的評估方式，而召回率與精準度的調和平均，也就是$\mathrm{F}_1$度量更是廣為使用。無論是正確或錯誤發音的檢測結果都是重要的指標，因此我們先定義正確發音檢測的召回率($\mathrm{Recall}_{\mathcal{C}}$)、精準度($\mathrm{Precision}_{\mathcal{C}}$)與$\mathrm{F}_1$度量($\mathrm{F1}_{\mathcal{C}}$)的計算方式：

$$\mathrm{Recall}_{\mathcal{C}} = \frac{\text{正確接受(TA)的個數}}{\text{實際為正確發音的個數}} = \frac{\#\mathrm{TA}}{\#\mathrm{TA}+\#\mathrm{FR}} \tag{16}$$

$$\mathrm{Precision}_{\mathcal{C}} = \frac{\text{正確接受(TA)的個數}}{\text{系統判斷為正確發音的個數}} = \frac{\#\mathrm{TA}}{\#\mathrm{TA}+\#\mathrm{FA}} \tag{17}$$

$$\mathrm{F1}_{\mathcal{C}} = \frac{2\cdot\mathrm{Recall}_{\mathcal{C}}\cdot\mathrm{Precision}_{\mathcal{C}}}{\mathrm{Recall}_{\mathcal{C}}+\mathrm{Precision}_{\mathcal{C}}} \tag{18}$$

而錯誤發音檢測的召回率($\mathrm{Recall}_{\mathcal{M}}$)、精準度($\mathrm{Precision}_{\mathcal{M}}$)與$\mathrm{F}_1$度量($\mathrm{F1}_{\mathcal{M}}$)的計算方式為：

$$\text{Recall}_{\mathcal{M}} = \frac{\text{正確拒絕(TR)的個數}}{\text{實際為錯誤發音的個數}} = \frac{\#\text{TR}}{\#\text{TR}+\#\text{FA}} \tag{19}$$

$$\text{Precision}_{\mathcal{M}} = \frac{\text{正確拒絕(TR)的個數}}{\text{系統判斷為錯誤發音的個數}} = \frac{\#\text{TR}}{\#\text{TR}+\#\text{FR}} \tag{20}$$

$$\text{F1}_{\mathcal{M}} = \frac{2\cdot\text{Recall}_{\mathcal{M}}\cdot\text{Precision}_{\mathcal{M}}}{\text{Recall}_{\mathcal{M}}+\text{Precision}_{\mathcal{M}}} \tag{21}$$

精準度的資訊在其它常見的評估方式(準確率(Accuracy)或ROC曲線等)中不易觀察，但對於錯誤發音檢測任務而言精準度也是非常重要的指標之一，因此同時考慮召回率與精準度的$F_1$度量指標為本論文後續實驗討論最常使用的評估標準。

## 5.4 錯誤發音檢測實驗

延續 5.2 小節設定的初始聲學模型(GMM-HMM 與 DNN-HMM)，並使用第 3 節提到的 GOP 分數(式(3))作為評估發音品質的特徵；並代入決策函數(式(4))與指示函數(式(5))檢測學習者的發音為正確或錯誤，決策函數與指示函數的參數皆使用全域的數值(未調整為音素相依或音素狀態相依)，其結果如表 4 所示。由表 4 可以得知基於 DNN-HMM 作為聲學模型產生 GOP 分數並應用在發音檢測任務效果更勝 GMM-HMM 聲學模型的效果發音檢測的$F_1$度量皆有約 3%的絕對進步(正確發音檢測的$F_1$度量由 0.836 提升至 0.863；錯誤發音檢測的$F_1$度量由 0.546 提升至 0.579)。已有許多學者在實驗中證明了深層學習在發音檢測任務的突破(Hu *et al*., 2015; Qian *et al*., 2012; Hu *et al*., 2014)。

***表 4. 不同聲學模型在發音檢測任務的實驗結果***
***[Table 4. Mispronunciation detection results achieved by using different acoustic models.]***

| GOP | Correct pronunciation detection | | | Mispronunciation Detection | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Recall | Precision | F1 |
| GMM-HMM | 0.828 | 0.844 | 0.836 | 0.562 | 0.532 | 0.546 |
| DNN-HMM | 0.877 | 0.849 | 0.863 | 0.552 | 0.609 | 0.579 |

接著我們探討發音檢測任務使用第 3 節所提到的對數音素事後機率(LPP)作為發音分數，如表 5。GOP 與 LPP 的方法在$F_1$度量的表現相近(正確發音檢測的$F_1$度量由 0.863 降低至 0.854；錯誤發音檢測的$F_1$度量由 0.579 提升至 0.587)，其變化皆在 0.01 之間。如第 3 節提到的 LPP 的計算複雜度遠低於 GOP，因此接下來的實驗將以 LPP 作為主要的發音分數。

***表 5. 基於 DNN-HMM 聲學模型使用不同發音分數的發音檢測基礎實驗***
***[Table 5. Mispronunciation detection results achieved by incorporating the DNN-HMM acoustic model with different decision features.]***

| | Correct pronunciation detection | | | Mispronunciation Detection | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Recall | Precision | F1 |
| GOP | 0.877 | 0.849 | 0.863 | 0.552 | 0.609 | 0.579 |
| LPP | 0.850 | 0.857 | 0.854 | 0.594 | 0.580 | 0.587 |

本論文探討的主題為最大化發音檢測效能之鑑別式訓練。首先我們定義欲進行發音檢測的聲學模型與決策函數，延續 5.2 小節的語音辨識基礎實驗中表現最好的聲學模型 DNN-HMM，以及在發音檢測任務上效果不遜色於 GOP 所提供的發音分數，也就是 LPP 發音分數作為發音分數；並透過非線性的邏輯函數轉換為決策值，供我們計算評估該模型的表現。然而，在第 4 節討論的$\mathrm{F}_1$度量之目標函數為錯誤發音檢測的$\mathrm{F}_1$度量，但是在電腦輔助發音訓練等任務中正確發音檢測也是非常重要的部分，我們不希望系統對於學習者本身正確的發音造成誤判。延續第 4 節的定義的$\mathrm{F}_1$度量目標函數，並再次定義錯誤/正確發音檢測的$\mathrm{F}_1$度量近似算法：

$$\Xi_{\mathcal{M}}(\boldsymbol{\theta}) = \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)\cdot\mathrm{H}(u,n)}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)+C_H^{(\mathcal{M})}} \tag{22}$$

$$\Xi_{\mathcal{C}}(\boldsymbol{\theta}) = \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}(1-\mathrm{D}(u,n))\cdot(1-\mathrm{H}(u,n))}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}(1-\mathrm{D}(u,n))+C_H^{(\mathcal{C})}} \tag{23}$$

最後我們使用參數 $\varphi$ 作為正確發音與錯誤發音的$\mathrm{F}_1$度量之線性組合作為最終的目標函數：

$$\tilde{\Xi}(\boldsymbol{\theta}) = \varphi\cdot\Xi_{\mathcal{M}}(\boldsymbol{\theta}) + (1-\varphi)\cdot\Xi_{\mathcal{C}}(\boldsymbol{\theta}) \tag{24}$$

從我們的實驗中可以觀察到參數 $\varphi$ 對於結果發音檢測發展集的影響，如圖 2，以最大化$\mathrm{F}_1$度量調整決策函數的參數(未更新聲學模型)。從圖中可以發現參數 $\varphi$ 對於錯誤發音檢測結果的影響十分顯著(圖 2 下半部)，當 $\varphi$=0.8 時在錯誤發音檢測有最好的效果($\mathrm{F}_1$度量為 0.566，基礎實驗的$\mathrm{F}_1$度量為 0.527)。但是在正確發音中並非 $\varphi$=0.8 為效果最佳，但參數 $\varphi$ 的調整對於正確發音檢測效能的影響較小，因此我們挑選參數 $\varphi$ 則以錯誤發音檢測之效能為優先考量。有趣的是，在訓練資料中正確發音與錯誤發音的比例正好接近 0.8，這表示透過調整參數 $\varphi$ 的方式巧妙的處理了資料類別不平衡的問題。

基於圖 2 的實驗結果，我們將固定 $\varphi$=0.8 並依續探討最大化發音檢測效能之鑑別式訓練對不同階段的參數進行調整所帶來的影響。在表 6 中我們基於 LPP 所算出的發音分數透過決策函數並算出整體的$\mathrm{F}_1$度量，以最大化$\mathrm{F}_1$度量更新決策函數(+MFC (DF))或聲學模型(+MFC (AM))的參數，甚至是決策函數與聲學模型的參數同時調整(+MFC (Both))。從表 6 可以發現無論是更新任何階段的參數在發音檢測任務上都可以得到顯著的提升。首先討論更新決策函數的參數(+MFC (DF))，與基礎實驗相比(LPP)則有明顯的提升；而

只更新聲學模型參數(+MFC (AM))可以得到更好的效果，若同時更新兩階段的參數(+MFC (Both))效果最佳。從實驗結果中可以發現更新聲學模型參數有著最大幅度的進步，由此可見初始的聲學模型是為語音辨識任務所設計，若經過調適則可以得到更好的效果。

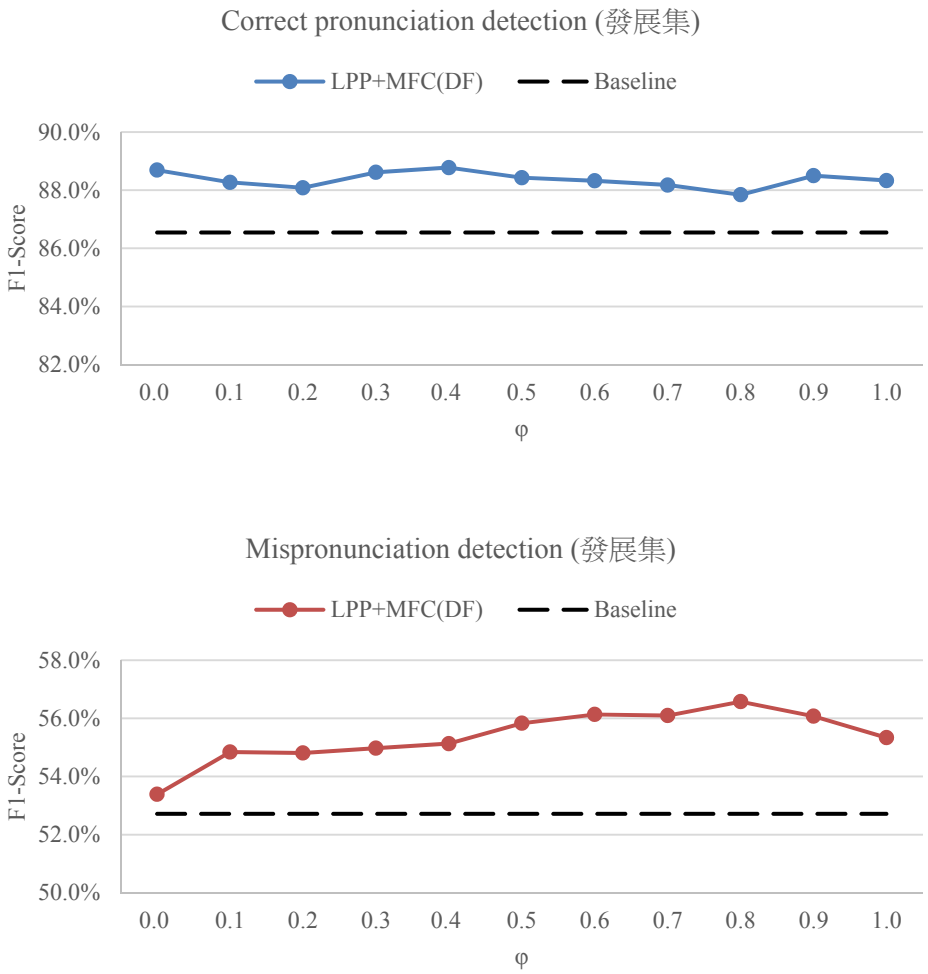


***圖 2. 不同 φ 在發展集的發音檢測效能***

***[Figure 2. Mispronunciation detection results on the development set with different threshold values φ.]***

***表 6. 基於 LPP 最大化 F1 度量鑑別式訓練於不同設定的發音檢測效能***
***[Table 6. Mispronunciation detection results achieved by using LPP features with/without MFC training.]***

| | Correct pronunciation detection | | | Mispronunciation detection | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | F1 | Recall | Precision | F1 |
| LPP | 0.850 | 0.857 | 0.854 | 0.594 | 0.580 | 0.587 |
| +MFC (DF) | 0.863 | 0.866 | 0.865 | 0.617 | 0.611 | 0.614 |
| +MFC (AM) | 0.906 | 0.870 | 0.888 | 0.612 | 0.694 | 0.650 |
| +MFC (Both) | 0.907 | 0.871 | 0.889 | 0.613 | 0.697 | 0.652 |

## 6. 結論與未來展望

本論文著重在電腦輔助發音訓練的錯誤發音檢測任務，並以最佳化錯誤發音檢測效能為主軸進行一系列的實驗。基於過去學者的研究，我們認為以最大化發音檢測之$F_1$度量為目標函數進行模型訓練是非常有潛力的。因此我們延伸該作法至現今語音辨識模組十分熱門的部份－深層類神經網路聲學模型，取代傳統的高斯混合聲學模型。從實驗結果可以發現以最大化$F_1$度量為目標對決策函數或聲學模型的參數進行調整，甚至是同時調整，都可以在效果上得到提升；尤其對於聲學模型參數進行調整的進步幅度令人印象深刻。且以$F_1$度量作為目標進行訓練在不同的評估方式也可以得到進步。在未來我們希望從特徵與模型等兩個面向來探討對於電腦輔助發音訓練任務的影響。在特徵的部分，我們期望從不同角度來獲取跟發音狀況高相關性的特徵，其中韻律特徵非常具有潛力；在模型的部分除了持續探討更新穎的聲學模型外，我們也預期將語音辨識所使用的調適技術移轉到該任務，例如一些非監督式的語者調適或是針對不同語言進行模型調適等。

## 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 和 MOST 105-2221-E-003-018-MY3)之經費支持，謹此致謝。

## 參考文獻 References

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D. & Bengio, Y. (2010). Theano: A CPU and GPU math compiler in Python. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 1-7.

Chen, L. Y. & Jang, J. S. R. (2015). Automatic pronunciation scoring with score combination by learning to rank and class-normalized DP-based quantization. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(11), 1737-1749.

Dembczynski, K. J., Waegeman, W., Cheng, W. & Hüllermeier, E. (2011). An exact algorithm for F-measure maximization. In *Advances in Neural Information Processing Systems*, 1404-1412.

Fujino, A., Isozaki, H. & Suzuki, J. (2008). Multi-label Text Categorization with Model Combination based on F1-score Maximization. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 823-828.

Harrison, A. M., Lau, W. Y., Meng, H. M. & Wang, L. (2008). Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 2787-2790.

Harrison, A. M., Lo, W. K., Qian, X. & Meng, H. (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *Proceedings of the International Symposium on Languages, Applications and Technologies (SLaTE)*, 45-48.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. & 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.

Hsiung, Y. & Sung, Y. (2014). Development of Mandarin Annotated Spoken Corpus (MAS Corpus) and the Learner Corpus Analysis. *1st Workshop on the Analysis of Linguistic Features (WoALF)*.

Hsu, Y. C., Yang, M. H., Hung, H. T. & Chen, B. (2016). Mispronunciation detection leveraging maximum performance criterion training of acoustic models and decision functions. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 2646-2650.

Hu, W., Qian, Y. & Soong, F. K. (2014). A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 3206-3210.

Hu, W., Qian, Y., Soong, F. K. & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, *67*, 154-166.

Huang, H., Wang, J. & Abudureyimu, H. (2012). Maximum F1-score discriminative training for automatic mispronunciation detection in computer-assisted language learning. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 815-818.

Huang, H., Xu, H., Wang, X. & Silamu, W. (2015). Maximum F1-score discriminative training criterion for automatic mispronunciation detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(4), 787-797.

Kim, Y., Franco, H. & Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 649-652.

Laborde, V., Pellegrini, T., Fontan, L., Mauclair, J., Sahraoui, H. & Farinas, J. (2016). Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 2686-2690.

Lee, A. & Glass, J. (2012). A comparison-based approach to mispronunciation detection. In *Proceedings of the International Conference on Spoken Language Technology Workshop (SLT)*, 382-387.

Lo, W. K., Zhang, S. & Meng, H. M. (2010). Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 765-768.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer G. & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Qian, X., Meng, H. M. & Soong, F. K. (2012). The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 775-778.

Qian, X., Soong, F. K. & Meng, H. M. (2010). Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT). In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 757-760.

Strik, H., Truong, K. P., De Wet, F. & Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 1837-1840.

Wang, Y. B. & Lee, L. S. (2012). Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 5049-5052.

Wang, Y. B. & Lee, L. S. (2015). Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(3), 564-579.

Wei, S., Hu, G., Hu, Y. & Wang, R. H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, *51*(10), 896-905.

Witt, S. M. & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2), 95-108.

Ye, N., Chai, K., Lee, W. & Chieu, H. (2012). Optimizing Fmeasures: a tale of two approaches. In *Proceedings of the International Conference on Machine Learning (ICML)*, 289-296.

Zhang, F., Huang, C., Soong, F. K., Chu, M. & Wang, R. (2008). Automatic mispronunciation detection for Mandarin. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 5077-5080.