The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 266-280

探究端對端語音辨識於發音檢測與診斷

Investigating on Computer-Assisted Pronunciation Training Leveraging End-to-End Speech Recognition Techniques

張修瑞 Hsiu-Jui Chang, 羅天宏 Tien-Hong Lo, 劉慈恩 Tzu-En Liu, 陳柏琳 Berlin Chen

國立臺灣師範大學資訊工程學系

{ftes90015, teinhonglo, hane0131}@gmail.com
berlin@csie.ntnu.edu.tw

摘要

電腦輔助發音系統(Computer assisted pronunciation techniques, CAPT)，任務可分為錯誤發音檢測(Mispronunciation detection)以及錯誤發音診斷(Mispronunciation diagnosis)。在過往的研究中，這兩種任務主要依賴於傳統語音辨識系統的強制對齊(Forced alignment)方法，並利用強制對齊產生的音素(Phone)段落與觀測到的全部音素或較混淆的音素計算 GOP (Goodness of pronunciation)分數，並以此作為發音好壞的依據。然而傳統語音辨識系統的訓練流程既冗長且複雜。近年來，端對端語音辨識系統不僅大幅簡化此問題，且效能也有追上傳統語音辨識的趨勢。因此，本論文將基於端對端架構下，分別探討(1) 基於辨識產生的信心分數(Confidence score)；(2) 基於語音辨識結果，兩者對於發音檢測任務的影響。實驗結果顯示，使用端對端架構進行發音檢測與診斷，不僅相較於以往基於傳統語音辨識架構有更少的訓練流程，也大幅提升檢測與診斷的效果。

Abstract

One of the primary tasks of a computer-assisted the pronunciation techniques (CAPT) system is mispronunciation detection and diagnosis. Previous research on CAPT mostly relies on a forced-alignment procedure which is usually conducted with the acoustic models adopted from a traditional speech recognition system, in conjunction with a phoneme paragraph, to calculate the goodness of pronunciation (GOP) scores for the phonemes of spoken words with respect to a text prompt. However, the training process of the traditional speech recognition system is complicated. In recent years, the end-to-end speech recognition system has not only greatly simplified this problem, but also has the trend of catching up with

traditional speech recognition. In view of this, this thesis sets out to conduct mispronunciation detection and diagnosis on the strength of end-to-end speech recognition. To this end, we design and develop two mispronunciation detection methods: 1) method leveraging a recognition confidence measure; 2) method simply based speech recognition results; A series of experiments showed that leveraging end-to-end speech recognition architecture on mispronunciation detection and diagnosis not only reduced the training steps originally required for traditional speech recognition but also improve the performance of detection and diagnosis significantly.

關鍵詞：端對端語音辨識、聲學模型、發音檢測、發音診斷

Keywords: end-to-end speech recognition, acoustic model, mispronunciation detection, mispronunciation diagnosis.

一、緒論

在國際化的時代中，學習外語變成了不可或缺的一部份。當今的人們至少需要學習兩種或兩種以上的語言，而語言學習的過程可分為聽、說、讀和寫四個部分，其中以說和寫最為需要專家知識才得以判斷學習的程度。然而大多數的學習平台較注重於聽力以及閱讀練習，口說的練習則通常藉由教學影片讓使用者複誦，缺乏即時的回饋；此機制使學習者較不易發現發音錯誤，對於學習效果有限，因此電腦輔助發音訓練的研究更顯重要。我們希望電腦具備與專業師資相當的聽力，檢測出學習者的發音錯誤，並且給予回饋，使學習者能夠藉由反覆的練習使口說能力更為進步。

電腦輔助發音的研究與語音辨識器息息相關，過去研究中主要依賴傳統的深度類神經網路結合隱藏式馬可夫模型(Deep neural network-hidden Markov model, DNN-HMM)語音辨識架構。該架構主要由聲學模型(Acoustic model)、語言模型(Language model)、發音詞典(Pronunciation lexicon)所組成，並且在訓練的過程中，必須先由傳統的高斯混合模型結合隱藏式馬可夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM) [1][2]，取得聲音與文字的強制對齊，才得以訓練 DNN-HMM 聲學模型，而目前常用的類神經網路包含多層感知器(Multiple-layer perceptron, MLP)[1][3]、摺積式類神經網路(Convolutional neural networks, CNN)[4]、遞迴式類神經網路(Recurrent neural

network, RNN)、長短期記憶類神經網路(Long short-term memory, LSTM)[5][6]、時延式類神經網路(Time delay neural network, TDNN)[7]以及這些類神經網路的延伸。

傳統語音辨識具有下列幾點問題：(1)訓練流程與多個模組有關連，無法清楚知道哪一部分影響了語音辨識的效果；(2)需要較多的語言及語音知識將詞彙對到相對應的音素序列來產生發音詞典，以及音素的上下文相關決策樹；(3)聲學模型和語言模型各自使用不同準則分開訓練，導致語音辨識或其應用任務的最後評估標準不一致。近年來端對端語音辨識大幅簡化了傳統語音辨識繁複的訓練流程。其主流為連結時序分類(Connectionist temporal classification, CTC)以及注意力模型(Attention model)兩種方法。前者 CTC 訓練準則使得聲學模型可直接將聲學特徵僅透過類神經網路輸出對應到的標籤序列，通常為字符(Character)或音素[5][8]，並且於解碼時可以不需要使用語言模型。而另一方面，鑑於 CTC 對於端對端語音辨識的成功，且注意力模型已在其他領域被廣泛應用[9][10]，[11]將注意力模型應用於語音辨識的任務上，並得到與 CTC 模型可比較的結果，但在少量語料下仍遜於 DNN-HMM 的效能。[12][13]藉由多任務學習的方法結合 CTC 與注意力模型，希望 CTC-Attention 模型利用 CTC 彌補注意力模型對齊錯誤(Misalignment)及收斂慢的問題。實驗結果顯示，CTC-Attention 模型可在缺乏語料的情況下，更接近甚至低於 DNN-HMM 模型的辨識錯誤率。

本篇論文中，主要探討端對端聲學模型如何應用於電腦輔助發音檢測，其主要任務分為錯誤發音檢測(Mispronunciation detection)以及錯誤發音診斷(Mispronunciation diagnosis)。錯誤發音檢測任務希望藉由學習者朗誦第二外語口說教材，在已知朗誦內容的情況下，由電腦評判學習者的發音是否正確。在過往的發音檢測實驗中，都是利用傳統聲學模型的事後機率(Posterior probability)、對數相似度值(Log-likelihood) [14]，或是 GOP [15][16]作為發音檢測特徵，以此判斷發音的對錯。而發音診斷則是當學習者發音出現錯誤時系統所給予的糾正。假設所希望聽到的是「國語(guo2 yu3)」，但當學習者唸成「狗語(gou3 yu3)」，系統除了能得知學習者發音出錯，也能反饋學習者的「國(guo2) 」唸錯成「狗 gou3」了。過往也有學者將發音檢測與診斷視為語音辨識的任務如[17][18][19]。基於端對端架構的發音檢測方法較為稀少，僅[19]提出以 CTC 進行英

文發音檢測與診斷。因此本篇論文將初步探討端對端架構應用於華語 CAPT 任務。實驗中將分別使用信心分數以及語音辨識結果進行發音檢測。在實驗結果顯示利用信心分數進行錯誤發音檢測，能夠在錯誤檢測上達到與 GOP 相同的效果，然而整體來說較 GOP 的判斷更為嚴格。另外直接視為語音辨識的方法超越了[19]，使得檢測與診斷都達到最好的效果。

二、端對端語音辨識技術

隨著近年來端對端語音辨識技術的發展，端對端聲學模型的辨識率已與傳統聲學模型不相上下，以下將針對近年來主要的端對端語音辨識方法進行說明。

2.1 連結時序分類(CTC)

連結時序分類最早於 2006 年提出[20]，作為取代傳統聲學模型訓練使用交互熵的損失函數希望最小化$-\ln \mathrm{P}_{\mathrm{ctc}}(C^*|X)$，即輸出越接近真實標記越好。常見應用於音素辨識以及手寫辨識。其概念為給定一段長度為 T 的聲學特徵序列 X 及一段長度 L 的標籤序列 C，其中$C = \{c_l \in U | l = 1, \dots, L\}$，U 為存在的標籤集合。並且 CTC 在訓練時引入了額外的空白標籤，作為標籤間的分界，每個音框的標籤序列可表示為$S = \{s_t \epsilon U \cup \{< blank >\} | t = 1, \dots T\}$，其損失函數可表示為：

$$\mathrm{P}_{\mathrm{ctc}}(C|X) \approx \sum_s \prod_{t=1}^{T} \mathrm{P}(s_t|s_{t-1}, C)\mathrm{P}(s_t|X) \tag{1}$$

其中$\mathrm{P}(s_t|s_{t-1}, C)$代表狀態轉移機率，$\mathrm{P}(s_t|X)$則為 Softmax 輸出的結果。

2.2 注意力模型(Attention model)

沿用上一小節中符號設定，注意力模型目標函式可定義為：

$$\mathrm{P}_{\mathrm{att}}(C|X) = \prod_{l=1}^{L} \mathrm{P}(c_l|X, c_{1:l-1}) \tag{2}$$

同樣也希望直接估測聲學特徵對應到標籤的事後機率，然而與 CTC 不同在於注意力模型並無條件獨立的假設，如上式 2 所示，每一當前輸出皆考慮過去的輸出。$\mathrm{P}(\mathrm{c_l}|\mathrm{X}, \mathrm{c_{1:l-1}})$可以由下列式子推得：

$$\boldsymbol{h}_t = \mathrm{Encoder}(X) \tag{3}$$

$$e_{lt} = \mathrm{Attention}(\mathbf{q}_{\mathrm{l}-1}, \mathbf{h}_{\mathrm{t}}, a_{l-1}) \tag{4}$$

$$a_{lt} = \frac{\exp(\gamma e_{lt})}{\sum_l \exp(\gamma e_{lt})} \tag{5}$$

$$\mathbf{r}_l = \sum_{\mathrm{t}=1}^{T} a_{lt}\mathbf{h}_{\mathrm{t}} \tag{6}$$

$$\mathrm{p}(c_l|X, c_{1:l-1}) = \mathrm{Decoder}(\mathbf{r}_l, \mathbf{q}_l, c_{l-1}) \tag{7}$$

其中$\mathbf{h}_{\mathrm{t}}$為 Encoder 的隱藏狀態向量，a_{lt}為注意力權重由e_{lt}經由 Softmax 函數得到，而 γ為 Sharpen Factor，目的在強調權重的分佈，**q** 代表的是前一個 Decoder 隱藏狀態向量。可以想成 **q** 是 Query，**h** 是 Key Value 然後我們透過任意注意力的機制計算注意力權重a_{lt}，同樣地，注意力模型訓練的損失函數也希望最小化$-\ln \mathrm{P}_{\mathrm{att}}(C^*|X)$。

2.3 CTC-Attention 混合模型(Hybrid CTC-Attention model)

由於注意力模型有著非單調的左到右對齊和收斂較慢的缺點，CTC 則是必須使用額外的語言模型才能有較好的效果。因此有學者也將兩者結合[12] [13]，以 CTC 給予注意力模型更強的左到右限制，並在進行光束解碼時(Beam search)同時加入兩種模型的輸出，以達到更佳的解碼效果。訓練時以λ作為兩模型混和參數，其損失函數為：

$$\mathcal{L}_{\mathrm{CTC-ATT}} = -\big(\lambda \ln \mathrm{P}_{\mathrm{ctc}}(C|X) + (1-\lambda)\ln \mathrm{P}_{\mathrm{att}}(C|X)\big) \tag{8}$$

三、端對端語音辨識技術於發音檢測與診斷

基於檢測與診斷可被視為語音辨識任務的想法，一個能夠辨識清楚第一語言者與第二語言的學習者所發出音素差異的辨識器，將對於發音檢測與診斷有很大的突破。利用這樣的方法不僅可以同時進行發音診斷與檢測，也省去以往必須再藉由發音分數評估的兩階段步驟。以下將針對端對端語音辨識發音檢測與診斷方法進行說明。

3.1 基於分數之發音檢測

過去進行檢測的首要步驟在於進行強制對齊，無論發出對與錯的音素皆解碼成目標音素，並標記音素出現於聲音之時間段。在端對端的語音辨識採用光束搜尋算法，輸出時經由 Softmax 函數輸出所有標籤之事後機率，保留每一次輸出前 n 高值直到出現句尾符號 <eos>為止。為了達到在搜尋時能產生我們所想要的目標音素，使用了限制解碼的方法，即在每一次 Softmax 函數輸出時只關注我們所想要的音素集合，如下圖 1 所示。在解碼的過程中，找出音素事後機率總和最高之音素組合，作為最終想要的目標音素序列，並記錄每一音素之事後機率$\mathrm{P}(c^*|x)$。得到音素事後機率可帶入式 9 決策函數$\mathrm{D}(\mathrm{P}(c^*|\mathbf{x}))$，使音素事後機率投影到 0-1 的範圍，並根據門檻值τ決定發音好壞。

$$\mathrm{D}(\mathrm{P}(c^*|\mathbf{x})) = \frac{1}{1+\exp(\mathrm{P}(c^*|\mathbf{x}))} \tag{9}$$

$$\mathbb{I}\big(\mathrm{D}(\mathrm{P}(c^*|\mathbf{x}))\big) = \begin{cases} 1 \text{ if } \mathrm{D}(\mathrm{P}(c^*|\mathbf{x})) \geq \tau \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

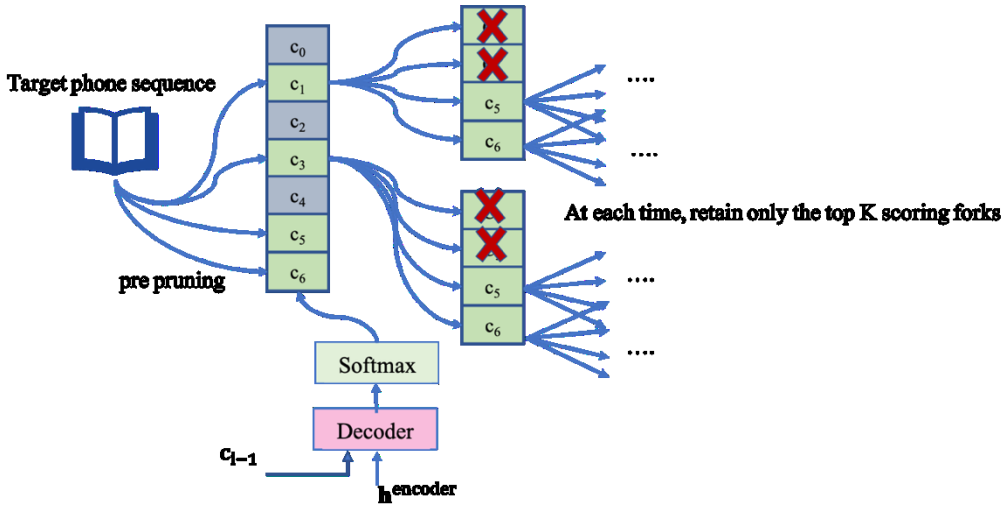

圖1、限制解碼流程

3.2 基於辨識結果之發音檢測與診斷

電腦輔助發音檢測可被視為語音辨識的問題，當語音模型的辨識率為百分之百，發音檢測的問題便可以被解決。儘管當前語音辨識技術仍達不到完美，但已十分進步。在本節中，我們將語音辨識結果與目標語句文字以最短編緝距離演算法(Edit distance)進行對齊

[21]。如下圖 2 所示，紅色箭號代表替換錯誤，藍色箭號為刪除錯誤，綠色箭號為插入錯誤，黑色箭號則為與目標相符。發生替換錯誤與刪除錯誤時則代表發生了發音錯誤，而插入錯誤的發生情況較為特殊，由於中文的一字一音節特性，發生插入錯誤的可能性更低。會發生的情況通常是在學習者發出聲音時意識到自己唸得不夠標準，想要再次發出正確的音所導致，或是環境的噪音被當作語者所說的話。因此，對於插入錯誤的部分我們能夠忽略，僅專注於插入錯誤以外的替換錯誤與刪除錯誤檢測。

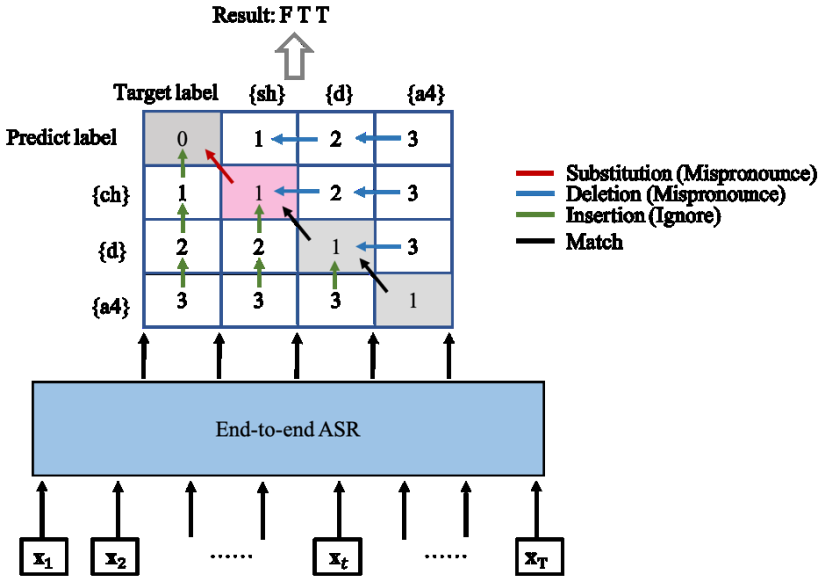

圖2、最短編輯距離發音檢測

四、實驗設定

5.1 語料

本論文使用臺灣師範大學邁向頂尖大學計畫之華語學習者口語語料庫，其中可以分為華語母語者(L1 speaker)以及華語非母語者(L2 speaker)兩部份。語料中的語句包含三種類型分別為單音節(Mono syllable, MS)、雙音節(Double syllable, DS)以及短文(Essay, ES)；詳細統計資訊如表 1 所示。

表1、華語學習者口語語料庫之訓練集、發展集與測試集

		時間(小時)	語者數	音素數量	錯誤發音音素數量
訓練集	L1	6.7	44	72,486	-
	L2	17.4	82	133,102	29,377
發展集	L1	1.4	10	14,186	-
	L2	-	-	-	-
測試集	L1	3.2	25	32,568	-
	L2	7.5	44	55,190	14,247

5.2 聲學模型

本研究分為兩階段，第一階段將比較以 L1 語料訓練傳統聲學模型進行發音檢測與端對端聲學模型進行發音檢測之效果，第二階段則是以端對端聲學模型加入 L2 語料進行發音檢測之效果。傳統聲學模型與端對端聲學模型皆使用美國約翰霍普金斯大學學者發展之大詞彙連續語音辨識工具，分別為“Kaldi”[22]以及“Espnet”[23]。

在傳統聲學模型設定中，初始階段訓練 GMM-HMM 模型輸入特徵為 MFCC 特徵，包含 3 維音調(Pitch)組成共 16 維，並對特徵取一階差量係數(Delta coefficient)，與二階差量係數(Acceleration coefficient)合併為 48 維特徵向量。DNN-HMM 聲學模型輸入特徵則每一音框為 40 維的 Filterbank 特徵加上 3 維音調(Pitch)並且取一階差量係數，與二階差量係數共 129 維。DNN-HMM 模型分別使用了不同層數與神經元數，也嘗試了最新的 DNN-HMM 架構 Factorized TDNN (TDNN-F)以及訓練準則 LF-MMI (Lattice-free maximum mutual information) [24]，也利用了速度擾動進行資料增添分別加快 1.1 倍速與放慢 0.9 倍速，詳細架構如下表 2 所示。

表1、傳統聲學模型架構

	類神經網路層數	每一神經元數
DNN-HMM	6	2048
TDNN-F LFMMI	13	768

端對端聲學模型於第一階段實驗使用的是 CTC-Attention 混合模型，架構主要參考[25][26]，混合參數為 0.5。Encoder 的架構為兩層的 VGG 層加上六層 Long short-term memory projection(LSTMP)一層含 320 個神經元，Decoder 的架構則為一層 LSTM 含 300 個神經元，使用的注意力機制為 Location attention[12]，計算方式如下式 11 所示，為了強化左到右的對齊，除了考慮前一個 Decoder state 以及當前 Encoder state 外，更加入一維摺積層 K 對於過去的 Attention 向量 $\mathbf{a}_{l-1}$抽取的向量。除此之外，也加入了標籤平滑方法(Label smoothing) 參數設為 0.05，目的在於不讓模型過度自信使部分較少出現的標籤也能有點機率分佈使模型更加一般化。由於端對端聲學模型通常需要較大量資料，我們同樣地也使用了速度擾動方法，因此與 TDNN-F LFMMI 之結果較具可比性，詳細架構如下圖 3 所示。

$$e_{lt} = \begin{cases} \mathbf{F_l} = \mathbf{K} * \mathbf{a_{l-1}} \\ \mathbf{g}^{\mathrm{T}} \tanh\left(\mathrm{W_q}\mathbf{q_{l-1}} + \mathrm{W_h}\mathbf{h_t} + \mathrm{W_f}\mathbf{f_{lt}}\right) \end{cases} \tag{11}$$

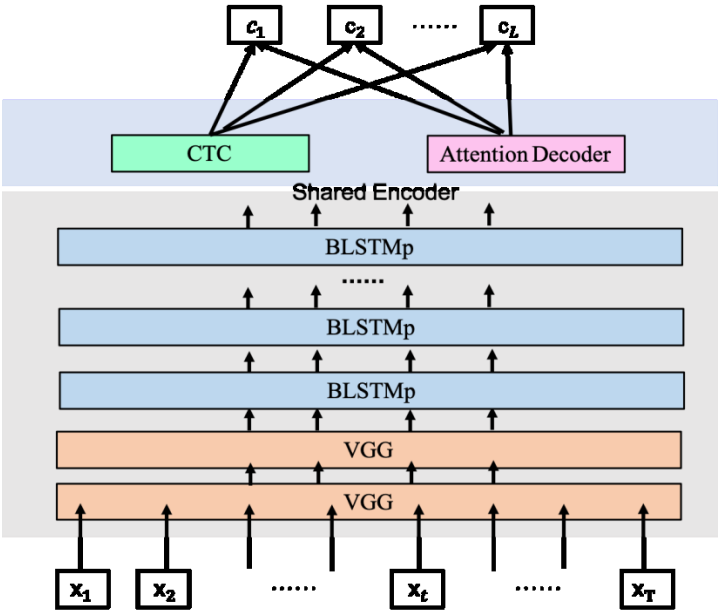

圖3、端對端語音辨識模型架構

第二階段主要探討端對端聲學模型加入 L2 語料對於發音檢測之影響，將基於圖 3 之架構，分別比較只使用 CTC、Attention 模型與 CTC-Attention 混合模型。並且比較加入已知 L2 錯誤模式進行解碼。

五、實驗結果與分析

5.1 L1 辨識結果

為了證實端對端語音辨識於發音檢測的可行性，我們首先不同聲學模型架構於 L1 語料中的表現，下表 3 為不同聲學模型架構於同一語料測試集的音素錯誤率與音節錯誤率，而使用資料量相同的為 TDNN-F LFMMI 與 CTC-Attention 模型。由結果可以得知使用 CTC-Attention 的辨識效果優於任意其他模型，可能原因是端對端聲學模型並不受制於發音詞典，因此相較於傳統 DNN-HMM 模型不會受到未知音素組合的影響，並且 Attention 模型架構設計帶有語言模型的概念，可提升對於已知音素組合的辨識效果。

表3、L1測試集的音素錯誤率與音節錯誤率

	Mono syllable(MS)		Double syllable(DS)	
Model	SER	PER	SER	PER
DNN-HMM	41.8	28.4	28.7	18.0
TDNN-F LFMMI	34.2	26.7	25.3	22.5
CTC-Attention	**32.2**	**18.9**	**6.8**	**5.1**

5.2 基於分數之發音檢測結果

基於門檻值方法，首先我們利用 ROC 曲線觀察門檻值對於發音檢測的評估標準變化，如下圖 4 所示，我們從中發現當門檻值設越小錯誤拒絕率越低，而錯誤接受率會隨之上升，但是增加幅度較小，進而發現當當全域門檻值設為 0.1 時兩者之錯誤接受率與錯誤拒絕率相等。儘管使用門檻值方法在錯誤檢測上與使用 GOP 方法可比較。總體來看分類效果仍然是劣於 GOP 方法。錯誤的拒絕過多導致在判斷錯誤時的精準度偏低。

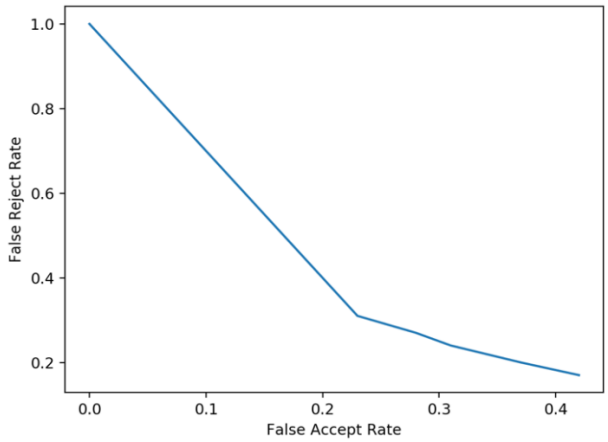

圖4、基於門檻值之ROC曲線圖

表4、第二階段分類效果

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
DNN-HMM (GOP)	0.88	0.85	0.86	0.55	0.61	0.58
End-to-end (Threshold)	0.76	**0.87**	0.81	**0.69**	0.51	**0.59**

5.3 基於辨識之發音檢測與診斷結果

由 5.1 節的實驗結果顯示，CTC-Attention 模型在辨識率上超越其他傳統聲學模型。我們假設此聲學模型所辨識的結果為正確答案，並套用最短編輯距離去對齊目標音素，得到的結果如下表 5 所示：

表5、L1聲學模型發音檢測結果

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
1-best	0.73	**0.87**	0.79	**0.71**	0.49	**0.58**
2-best	0.80	0.84	0.82	0.59	0.52	0.55
3-best	0.84	0.82	0.83	0.52	0.54	0.53
5-best	0.87	0.81	0.84	0.43	0.55	0.49

由表中得知，僅使用 L1 訓練聲學模型判斷 L2 語者發音好壞過於嚴格，也可能因為口音的不同造成模型的誤判。鑑於判斷過於嚴格，我們也逐漸放寬標準，改採以 N-best 的結

果做判斷，然而隨著標準放寬，錯誤接受也隨之上升使得結果缺乏鑑別力。為了使模型能夠判斷發音好壞，接下來的實驗我們將 L2 語料加入訓練加入聲學模型訓練，希望能夠使聲學模型直接學習到 L2 的錯誤模式。

加入含有錯誤模式的L2訓練集後，發音檢測的任務可以直接視為語音辨識的任務，即當對測試集解碼的錯誤率較低，後續的發音檢測與診斷也將有很好的效果。因此我們首先比較音素錯誤率如表 6 所示：

表6、端對端聲學模型音素錯誤率

Model	Weight	Phone error rate
CTC	-	28.3
Attention	-	**24.5**
CTC-ATT	0.2	24.8
CTC-ATT	0.5	**24.6**
CTC-ATT	0.8	24.9

由此實驗得知使用 CTC-Attention 與僅使用 Attention 模型差異不大，然而音素錯誤率的計算有考量到插入錯誤，我們在進行最小編輯距離對齊目標音素時則不考慮插入錯誤因此兩者的效果有待進一步評估，另外也比較[19]基於 CTC 的結果。

發音檢測的效果如下表 7 所示，整體來看使用 CTC-Attention 模型的效果與僅使用 Attention 效果差異不大，但是都比只使用 CTC 做發音檢測效果更好。

表7、端對端聲學模型發音檢測效果

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
CTC	0.831	0.893	0.861	0.706	0.656	0.680
CTC-Att	0.873	0.893	0.883	**0.714**	0.672	**0.693**
Attention	0.875	0.892	0.884	0.710	0.674	0.691

儘管發音檢測效果已經得到良好的結果，對於學習者來說仍然無法得知自己的發音發錯成什麼了，因此繼續探討端對端聲學模型的診斷效果。如下表 8 所示：

表8、發音診斷效果

	Initial	Final	Tone
DNN-HMM	0. 548	0.441	0.752
CTC	0.611	0.582	0.768
CTC-Attention	**0.661**	**0.612**	**0.801**
Attention	0.645	0.609	0.797

由上表診斷結果顯示，儘管在發音檢測 CTC-Attention 與 Attention 的效果差異不大，但是 CTC 與 Attention 的聯合解碼幫助了診斷結果，使診斷更加準確。

六、結論

本論文在端對端語音辨識架構上提出兩種發音檢測方式，分別為使用信心分數以及語音辨識結果，並且比較傳統使用語音辨識器進行發音檢測的方法。在使用信心分數的結果中，仍然遜於 GOP 的方法，希望在未來能夠加入更多發音的特徵，如[27]為了改善單用發音事後機率容易有誤判的情況，額外加入了許多發音特徵，例如發音方式、發音類型、吸氣吐氣等特徵。另一方面，基於語音辨識結果的發音檢測，我們發現加入 L2 語料訓練對於整體的檢測效果影響很大。當僅有母語者語料時，訓練的模型對於 L2 語者來說過於嚴格。而加入 L2 語料不僅能夠使模型進行發音檢測也能夠診斷，並且診斷正確率也超越以往方法。將發音檢測視為語音辨識的問題將使得研究更加簡單，僅需要思考如何讓模型辨識率提升。而未來方向除了改進聲學模型外，也希望能夠處理未知的錯誤。例如在 L2 測試集中有許多不存在於訓練集的錯誤標記，往往是兩個音素或是聲調的組合，而我們的模型診斷結果由於缺少這樣的標記通常只能回饋兩個音素中的其中一種，對於此情況仍然難以解決。期許在未來能夠找到對於未知錯誤的正確回饋方法。另外端對端聲學模型的解碼速度不夠即時，對於實際應用來說還有一段距離，在未來也希望能夠做到即時解碼，如[28][29]，使得我們的架構能被實際應用。

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] Lawrence R. Rabiner et al., "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*," Proceedings of the IEEE, 1989.

[2] Mark Gales and Steve Yang, "*The Application of Hidden Markov Models in Speech Recognition*," Foundations and Trends® in Signal Processing, 2008.

[3] Geoffrey Hinton et al., "*Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*," IEEE Signal processing magazine, 2012.

[4] Ossama Abdel-Hamid et al., "*Convolutional neural networks for speech recognition*," IEEE/ACM Transactions on audio, speech, and language processing, 2014.

[5] Alex Graves et al., "*Speech recognition with deep recurrent neural networks*," ICASSP, 2013.

[6] Haşim Sak et al., "*Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition*," arXiv, 2014.

[7] Vijayaditya Peddinti et al., "*A time delay neural network architecture for efficient modeling of long temporal contexts*," Interspeech,2015.

[8] Alex Graves et al., "*Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*," ICML, 2006.

[9] Dzmitry Bahdanau et al., "Neural machine translation by jointly learning to align and translate," ICLR, 2015.

[10] Kelvin Xu et al., "*Show, attend and tell: Neural image caption generation with visual attention*," ICML, 2015.

[11] Jan Chorowski et al., "*Attention-Based Models for Speech Recognition*," NIPS, 2015.

[12] Suyoun Kim et al., "*Joint CTC-Attention based end-to-end speech recognition using multi-task learning*," ICASSP, 2017.

[13] Shinji Watanabe et al., "*Hybrid CTC/attention architecture for end-to-end speech recognition*," IEEE Journal of Selected Topics in Signal Processing 11, 2017.

[14] Yoon Kim et al., "*Automatic pronunciation scoring of specific phone segments for language instruction*," in Proc. Eurospeech-1997. ISCA, pp. 645–648, 1997.

[15] Silke Witt and Steve Young, "*Language Learning Based On Non-Native Speech Recognition*," European Conference on Speech Communication and Technology, 1997.

[16] Silke Witt and Steve Young, "*Phone-level pronunciation scoring and assessment for interactive language learning*," Speech Communiciation, Vol. 30, No. 2-3, pp. 95–108, 2000.

[17] Kun Li et al., "*Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks*," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016.

[18] Shaoguang Mao et al., "*Applying Multitask Learning to Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech*," ICASSP, 2018.

[19] Wai-Kim Leung et al., "*CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis*," ICASSP, 2019.

[20] Alex Graves et al., "*Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*," ICML, 2006.

[21] Robert A. Wagner and Michael J. Fischer. "*The string-to-string correction problem*," Journal of the ACM (JACM) , Vol. 21.1, pp. 168-173, 1974.

[22] Daniel Povey et.al, "*The Kaldi Speech Recognition Toolkit*," ASRU, 2011.

[23] Shinji Watanabe et al., "*ESPnet: End-to-End Speech Processing Toolkit*," Interspeech, 2018.

[24] Povey, Daniel et.al, "*Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks*," Interspeech, 2018.

[25] Takaaki Hori et al., "*Advances in Joint CTC-Attention based End-to-End speech recognition with a Deep CNN Encoder and RNN-LM*," Interspeeh, 2017.

[26] Haşim Sak et al., "*Long short-term memory recurrent neural network architectures for large scale acoustic modeling*," Interspeech, 2014.

[27] Wei Li et al., "*Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models*," Interspeech, 2017

[28] Chung-Cheng Chiu and Colin Raffel. "*Monotonic chunkwise attention*," ICLR, 2018.

[29] Ruchao Fan et al., "*An Online Attention-based Model for Speech Recognition*," arXiv, 2018.