探究文本提示於端對端發音訓練系統之應用

Exploiting Text Prompts for the Development of an End-to-End Computer-Assisted Pronunciation Training System

鄭宇森 Yu-Sen Cheng, 羅天宏 Tien-Hong Lo, 陳柏琳 Berlin Chen

國立臺灣師範大學資訊工程學系

Department of Computer Science and Information Engineering
National Taiwan Normal University

sam841009@yahoo.com.tw
teinhonglo@gmail.com
berlin@csie.ntnu.edu.tw

摘要

近年來，電腦輔助發音訓練(Computer assisted pronunciation training, CAPT)系統的需求日益上升。然而，現階段基於端對端(End-to-End)類神經網路架構之系統在錯誤發音檢測(Mispronunciation detection)的效能仍未臻完美，其原因是此類系統的內部模型本質上仍是屬於自動語音辨識(Automatic speech recognition, ASR)模型。ASR 目的是儘量正確地辨識出語者所說內容，縱使其發音是有偏誤的；而 CAPT 目的恰巧相反，是要能儘量正確地偵測出語者的錯誤發音。有鑒於此，本論文基於 CAPT 任務通常會有文本提示的特殊性，嘗試將文本提示資訊融入於端對端模型架構。我們研究使用兩個編碼器(Encoders)分別處理發音特徵以及文本特徵，並以分層式注意力機制(Hierarchical attention mechanism, HAN)來動態地結合不同編碼器產生特徵表示。本論文在一套華語學習者語料庫進行一系列實驗；透過不同評估準則所獲得結果顯示，我們所提出的方法較現有方法有較佳的錯誤發音檢測效能。

Abstract

More recently, there is a growing demand for the development of computer assisted pronunciation training (CAPT) systems, which can be capitalized to automatically assess the pronunciation quality of L2 learners. However, current CAPT systems that build on end-to-end (E2E) neural network architectures still fall short of expectation for the detection of

mispronunciations. This is partly because most of their model components are simply designed and optimized for automatic speech recognition (ASR), but are not specifically tailored for CAPT. Unlike ASR that aims to recognize the utterance of a given speaker (even when poorly pronounced) as correctly as possible, CAPT manages to detect pronunciation errors as subtlety as possible. In view of this, we seek to develop an E2E neural CAPT method that makes use of two disparate encoders to generate embedding of an L2 speaker's test utterance and the corresponding canonical pronunciations in the given text prompt, respectively. The outputs of the two encoders are fed into a decoder through a hierarchical attention mechanism (HAM), with the purpose to enable the decoder to focus more on detecting mispronunciations. A series of experiments conducted on an L2 Mandarin Chinese speech corpus have demonstrated the effectiveness of our method in terms of different evaluation metrics, when compared with some state-of-the-art E2E neural CAPT methods.

關鍵詞：端對端語音辨識、電腦輔助發音訓練、分層式注意力機制、發音檢測、發音診斷

Keywords: end-to-end speech recognition, Computer assisted pronunciation training, hierarchical attention mechanism, mispronunciation detection, mispronunciation diagnosis.

一、緒論

近年來，不少語言學習者透過智慧型裝置在網際網路(Internet)上學習，主要的原因在於資訊科技的日漸普及讓許多不容易得到學習資源的人，譬如偏遠地區的學子、生活貧困的人與不易抽出時間上課的人，都可以透過網際網路簡單的取得學習資源。線上學習存在著諸多優點，譬如學習者可以自行選擇合適的教材編排自己的進度、不管通勤或者在家都可以隨時隨地的進行學習，達到 24 小時沉浸式學習(Immersive learning)的效果，並且可以避免因為群聚而得到傳染病的風險。然而目前線上學習技術在語言學習上，仍舊存在著不足。語言學習可以分為聽、說、讀和寫四大部分，而其中口說尤為困難，由於學習者難以自行察覺發音錯誤，一般需要藉由具備專業知識的專家進行評估，才能判斷學習的成果。因此在口說的部份，本論文希望可以藉由電腦輔助人類專家，檢測出學習者的錯誤發音，以利於學習者進行修正。而這樣的技術就被稱為電腦輔助發音訓練(Computer assisted pronunciation training, CAPT) [1][2]。

在 CAPT 過去的研究中[3-5]，多採用語音辨識(Automatic speech recognition, ASR)和似然機率比例(Likelihood ratio)。兩者皆可使用高斯混合模型結合隱藏式馬可夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM) [3][4]或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) [5][6]。在語音辨識的方法中，透過計算最短編輯距離(Edit distance)將模型預測結果與標準答案強制對齊(Forced alignment)，使用對齊後的結果判斷使用者是否存在發音錯誤[2]。另一方面，似然機率比例則是以與人類專家成高度相關的 GOP (Goodness of pronunciation) [7]最為知名。近年來，由於 DNN-HMM 模型架構過於繁雜訓練不易，研究者[2][8]多採用端對端自動化語音辨識(End-to-end ASR) [9]簡化傳統模型繁複的訓練流程。

然而，上述做法都忽視了 CAPT 與 ASR 的目標相異性，CAPT 的目標是要能儘量正確地偵測出語者的錯誤發音，但 ASR 在使用上希望儘量正確地辨識出語者所說內容，縱使其發音是有偏誤的，因此 ASR 系統會傾向將錯誤發音辨識為正確發音。因為上述原因 ASR 在錯誤發音檢測(Mispronunciation detection)以及錯誤發音診斷(Mispronunciation diagnosis)上無法達到最佳化的效果。另一方面，當採用傳統的 ASR 方法執行 CAPT 任務時，忽視了 CAPT 任務相較於傳統 ASR 擁有發音詞彙相應的文本提示。因此也有相關學者[10][11]陸續開始研究文本提示在端對端 CAPT 任務的重要性。在[10]的研究中，將文本提示的資訊加入注意力模型(Attention model)的權重計算，並將具有文字加權影響的聲學隱藏向量(Audio hidden vector)與原始聲學隱藏向量串接，然後作為解碼器的輸入進行預測。而[11]採用多視角(Multi-view)架構，將文本提示視為額外的提示資訊，擁有獨立的編碼器(Encoder)且共享解碼器(Decoder)的參數，結合輸出端的聲學損失函數(包含 CTC 和 Attention)，用以輔助模型判斷聲音的正確與否。

本論文主要探討如何在端對端架構中使用 CAPT 的文本提示增強錯誤發音檢測。模型採用多編碼器(Encoder)架構平行擷取文本提示的音素(Phone)文本特徵與發音的聲學特徵，並透過分層式注意力機制(Hierarchical attention mechanism, HAN)[12][13]結合兩種特徵。過去 HAN 在[13]中被應用於多麥克風陣列的語音辨識，透過注意力機制(Attention)[14][15][16]動態分配權重給來自不同麥克風的資訊並且合併為單一陣列。本

論文希望藉由 HAN 的合併資訊的特性，動態結合來自不同編碼器的文本特徵或聲學特徵中的資訊。解碼器則針對聲學資訊參考[9]所提出的架構，使用連接時序分類(Connectionist temporal classification, CTC)[17]模型對注意力機制的結果進行限制來取得更高的效能。實驗顯示採用文本特徵後的模型在各個評估標準(F-measure, accuracy, precision, and recall)下，我們所提出的方法較現有方法有較佳的錯誤發音檢測效能。

二、對於錯誤發音檢測的端對端語音辨識技術

本論文實驗模型主要採用與[2][8][9][11]相同的端對端架構，並在本章節中介紹與此架構相關的主要技術。

2.1 連結時序分類(CTC) 模型

連結時序分類最早於 2006 年由[17]提出，類似 DNN-HMM 架構，基於條件獨立假設使用貝氏決策法則找出最大事後機率。其過程尋求輸出符號(字母、單字或音素)序列$C = c_1, c_2, \dots, c_L$，在輸入的聲學特徵序列$\mathbf{X} = \mathbf{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_T}$，發生時出現的最大機率。並且公式可以分解下方的定義:

$$\mathrm{P}_{\mathrm{ctc}}(C|\boldsymbol{X}) \approx \sum_{s} P(C|S)P(S|\boldsymbol{X}) \tag{1}$$

其中S代表每個音框的標籤序列，而 CTC 為了避免重複出現的字造成辨識錯誤，在訓練時引入了額外的區塊(blank)標籤< b >作為標籤間的區隔，因此S可表示為$S = \{s_t \epsilon U \cup \{< b >\} | t = 1, \dots T\}$。

2.2 注意力模型(Attention model)

注意力機制過去在機器翻譯(Machine translation)上取得了卓越的成果，近年來也在語音領域上得到優秀的成果[14][15]。其特點是可以在不需要條件獨立假設的情況下直接計算輸出符號對應輸入聲學向量序列的事後機率，注意力模型目標函式可定義為：

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

上式 2 中的$\mathrm{P}(\mathrm{c_l}|\mathbf{X}, \mathrm{c_{1:l-1}})$藉由編碼器與解碼器的交互作用取得。可以由下列式子推導：

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

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

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

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

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

其中$\mathbf{h}_{\mathrm{t}}$為編碼器的隱藏向量，a_{lt}是由e_{lt}經由 Softmax 函數轉換為機率分佈後得到的注意力機制權重，而 γ為 Sharpen Factor，用於在強調權重的分佈，$\mathbf{q}_l$是 Decoder 每一層的隱藏向量。

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

注意力模型允許非序列化的對齊，這在機器翻譯或者其他不強調順序的任務中沒有問題。然而，語音辨識是種序列化的任務，因此非單調的對齊會讓其訓練時收斂較慢，但注意力模型因為不需要條件獨立假設，更加貼近真實環境因此在辨識時可以取得優越的表現。與之相對的是，CTC 具有由左至右嚴格單調的對齊。然而，傳統上 CTC 必須搭配其他額外的語言模型才能達到最佳效果。其原因是 CTC 的條件獨立假設會使它與真實環境偏離對效能造成負面影響。因此就有學者[9]提出了結合兩者的優點彌補彼此缺點的 CTC-Attention 混合模型。藉由注意力模型可以得到非條件獨立的前後資訊，並且藉由 CTC 的嚴格單調特性限制注意力模型的計算範圍，在[9]的實驗中指出，這樣的模型能夠比注意力模型更快收斂得到更高的效能。模型訓練時以λ作為兩種損失函數的線性相加參數，新的損失函數定義如下：

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

需要注意的是因為文字資訊不需要 CTC 的對齊，所以本論文的雙編碼器架構中的文字資訊注意力模型，並沒有使用 CTC 輔助。

三、文本提示在端對端發音檢測與診斷的使用

本論文將在本章節介紹受到[13]的多編碼器架構與 HAN 啟發的採用文本提示的多編碼器端對端架構。模型採用兩個平行獨立的編碼器，分別為發音編碼器與文本提示編碼器，透過這兩個編碼器分別抽取聲學和文本的特徵。接著透過 HAN 技術動態整合兩種不同維度的特徵。以下部分將針對本架構細節部分進行說明。

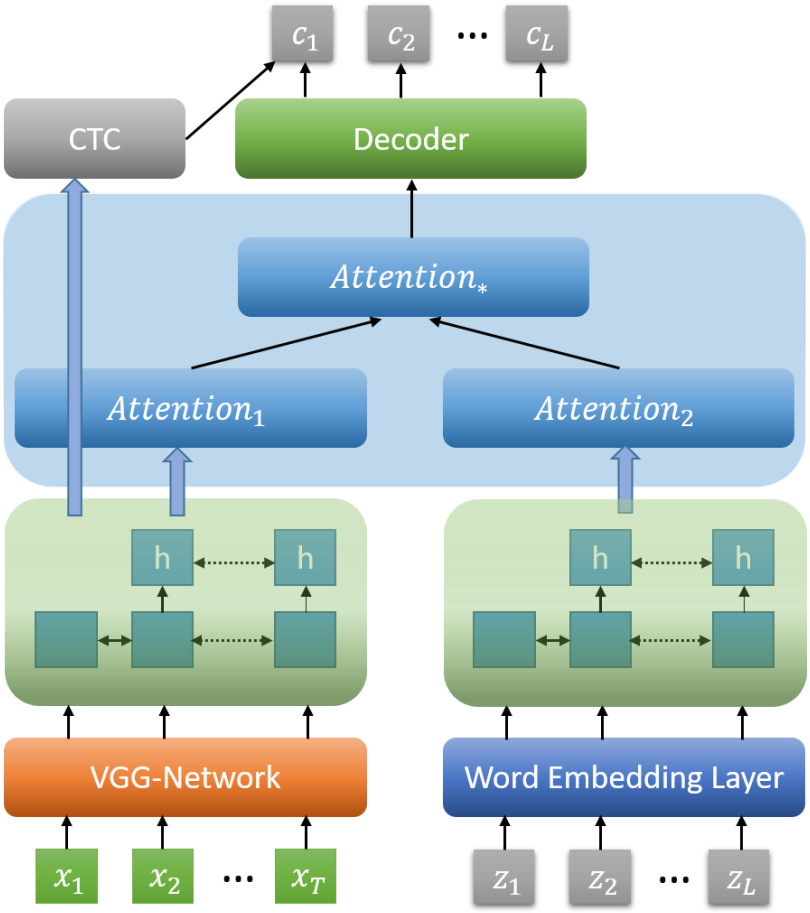

圖 1、多編碼器端對端錯誤發音檢測系統

3.1 文本提示的添加

如圖 1 所示，文本提示以音素層級的符號序列$\mathrm{Z} = \mathrm{z}_1, z_2, \ldots, z_L$被輸入到音素向量層轉換為對應的向量，並輸入到長短期記憶類神經網路(Long short-term memory, LSTM)[18][19]中抽取文本特徵，然後將文本特徵輸入到 HAN 層中動態分配與聲學特徵合併的權重。

3.2 聲學特徵處理

針對聲學特徵的處理，本論文參考[8][9]的 CTC-Attention 混合模型架構，採用 VGG-LSTM 的架構，透過 VGG 層提取聲學特徵，然後用 LSTM 學習聲學特徵中的時序資訊，之後將編碼器輸出的具聲學資訊的隱藏向量輸入到 HAN 層與文本特徵合併。

3.3 分層式注意力機制(Hierarchical attention mechanism, HAN)

由於聲學與文本兩種資訊的空間維度不同，無法直接合併，本論文採用 HAN 技術整合兩種來自不同維度的資訊。公式定義如下:

$$\mathrm{h}_{\mathrm{t}}^{1} = Encoder_1(X), \qquad h_t^2 = Encoder_2(Z) \tag{9}$$

$$\mathrm{r}_l^{\mathrm{i}} = \sum_{t=1}^{T^1,\ L^2} \alpha_{lt}^i h_t^i, i \in \{1,2\} \tag{10}$$

$$\mathrm{r}_{\mathrm{l}}^{*} = \beta_{l1} r_l^1 + \beta_{l2} r_l^2 \tag{11}$$

$$\beta_{\mathrm{li}} = Attention\left(q_{l-1}, r_l^i\right), i \in \{1,2\} \tag{12}$$

公式中(9)的兩個$\mathrm{Encoder}_{\{1,2\}}$分別發音與文本提示的編碼器，這裡定義$i \in \{1,2\}$作為編碼器與對應隱藏向量的索引。第(10)式中的α是兩個編碼器各自的隱藏向量輸入注意力模型所得到的輸出，具體計算過程請參照公式(4)與(5)。透過公式(12)對公式(10)的兩個加權向量$\mathrm{r}_{\mathrm{l}}^{i}$計算權重，並於公式(11)相加得出最後的加權向量$\mathrm{r}_{\mathrm{l}}^{*}$作為公式(7)解碼器的輸入。

四、實驗設定

在這個章節會介紹本實驗所使用的語料集，以及實驗相關參數設定還有開發架構，以利於其他研究者覆現本實驗之結果。

4.1 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[20]，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，我們將 L1 語料均視為正確發音，而 L2 的部分則標記有文本提示以及學習者的真正發音。為了貼近訓練與測

試的條件，我們將 L1 訓練集與 L2 訓練集合併作為我們的訓練集，而測試集本實驗採用的是 L2 測試集，詳統計資訊如表 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

4.2 模型設定

本論文的實驗皆採用開源端對端語音辨識工具“Espnet”[21]完成，在參數上聲學部分參考[2]的模型設定，聲學部分採用 VGG 連接雙向 LSTM，並且設置了投影層，透過大於[2]的節點數量得到更好的效果，文本部分考量文字與聲音的相異性將 VGG 層換成音素向量層對文本部分編碼，具體設定如表 2 所示，名詞表示參考“Espnet”設定所需使用之名詞。

表2、實驗參數設定

	聲學模型($Encoder_1$)	文本模型($Encoder_2$)
特徵	80-dim fbank + 3-dim pitch	pytorch word2vec
編碼器種類	VGGBLSTMP	BLSTMP
編碼器層數	3	3
編碼器節點數	1024 (BLSTMP)	1024 (BLSTMP)
解碼器種類	LSTM	LSTM
解碼器層數	2	
解碼器節點數	1024	
CTC/Attention 混合比	0.5/0.5	0/1

五、實驗結果與分析

在本章節會將展示使用華語學習者語料庫進行的一系列實驗的數據結果，並且對數據進行 nbest 以及華語在語言學上的分析。

5.1 辨識結果

為了驗證採用文本提示後，系統在 CAPT 任務的有效性，本實驗對前述(4.1)測試集進行多種評估標準(F-measure, accuracy, precision, and recall)研究，比較基線為未採用文本提示的傳統端對端 CTC-Attention 混合模型[8][9]，其參數設定與多編碼器模型中聲學部分相同。

從評估結果可以發現本論文提出的模型在各個評估標準下皆優於原始沒有採用文本提示的基線。此外本論文比較了沒有使用 CTC 輔助聲學部分對齊的結果，發現在 precision 的部分得到了提升，然而其他部分是下降的，可以判斷當沒有 CTC 參與時，聲學部分的辨識結果會往文本提示過度擬合。具體數據參見表 3。

另一方面，本論文將相近研究[11]納入比較後，發現本實驗的自動分配權重的方法能夠與[11]在解碼器人工調適文本提示與聲學損失函數結合權重的結果旗鼓相當，且因為少了人工調適的過程，在開發上成本上相對較少。另一方面，從[11]的數據可以表較使用 DNN-HMM 的模型(GOP+MFC)的 F1 65.2%相較低於端對端基線的 F1 69.2%，所以目前端對端方法在錯誤發音檢測上，明顯優於過去傳統的 DNN-HMM 方法。另外一份相近研究[10]研究英語學習者，但本論文主要在探討華語學習者的 CAPT，因此不列入比較。

表3、L2測試集的音素錯誤率與音節錯誤率

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
GOP [11]	-	-	-	51.8%	63.5%	57.0%
GOP+MFC [11]	-	-	-	69.5%	61.3%	65.2%
CTC-ATT(SR)	-	-	-	70.8%	67.9%	69.2%

CTC-ATT(SR) +PS [11]	-	-	-	**71.8%**	68.4%	70.2%
Baseline	87.7%	89.1%	88.4%	70.7%	67.7%	69.1%
Multi-encoder (without CTC)	86.9%	89.2%	88.0%	66.4%	**71.2%**	68.7%
Propose (with CTC)	**88.3%**	**89.4%**	**88.9%**	71.3%	69%	**70.2%**

5.2 N-best 結果

參考[2][11]的實驗，本論文將表3中具有 CTC 的本實驗模型，採用了 N-best 作為比較，可以發現隨著條件的放寬對於 CAPT 的影響是負面的，對於錯誤發音檢測而言，雖然 Precision 提高了，但 Recall 會嚴重下降，具體數據見於表5。

表5、N-best 對於 CAPT 任務的效能影響

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
1-best	**88.3%**	89.4%	**88.9%**	**71.3%**	69%	**70.2%**
2-best	61.3%	85.8%	71.5%	40.5%	72.3%	51.9%
3-best	55.0%	87.3%	67.5%	38.8%	78.1%	51.8%
4-best	53.7%	88.6%	66.9%	39.0%	81.1%	52.7%
5-best	53.3%	**90.0%**	66.9%	39.5%	**83.7%**	53.7%

5.3 語言學分析

本論文以語言學的角度分析系統的辨識能力。在表 7 中探討聲母(Initial)與韻母(Final)在四種量測象限的分布情況，量測象限定義如下:

1. 正確接受(True accept, TA): 系統辨識此發音為正確發音，且確實為正確發音。
2. 錯誤接受(False accept, FA): 系統辨識此發音為正確發音，但實際為錯誤發音。
3. 正確拒絕(True rejection, TR): 系統辨識此發音為錯誤發音，且確實為正確發音。
4. 錯誤拒絕(False rejection, FR): 系統辨識此發音為錯誤發音，但實際為正確發音。

量測象限對應關係可以參考表 6，其中 CP 代表正確發音，MP 代表錯誤發音，Ground Truth 為文本提示。

表6、錯誤發音量測象限

		Ground Truth	
		CP	MP
Model Prediction	CP	True accept (TA)	False accept (FA)
	MP	True rejection (TR)	False rejection (FR)

表7、聲母韻母量測象限分布

	TA	FA	TR	FR
Initial	19595 (76%)	1318(05%)	3347(13%)	1615(06%)
Final	14364 (54%)	2701(10%)	6654(25%)	2874(11%)
Total	33959(65%)	4019(08%)	10001(19%)	4489(09%)

數據上可以發現不管是系統的辨識或者是使用者本身的發音，在聲母上的正確率較高。本論文推測這是因為在韻母的發音上具有聲調(Tone)，所以不利於使用者發音，且也不利於系統辨識。因此在表 8 中本論文嘗試觀察中文的 5 種聲調對於辨識的影響，可以發現在一聲與四聲的 TA 比例較高，而三聲與輕聲的比例較低，其中輕聲因為資料所佔比例較少，所以容易判斷錯誤。本論文進一步在表 9 調查當韻母聲調錯誤時，L2 學習者在面對各個時，會發成何種聲調，藉此了解學習者在聲調上發音錯誤的狀況，其中 None 代表這種 L2 學習者的聲調無法被歸類在五種聲調中，可能是介於兩種聲調之間，或者為受到其 L1 語言影響的聲調。從表 9 可以觀察到，三聲最常與二聲混淆，本論文判斷主要的原因是因為相較於其他聲調，三聲的 F0 輪廓類似於 V 字型[22]，後段上揚部分容易被視為二聲。其中四聲發音混淆的比例最低，本論文判斷是因為相較於其他聲調，四聲相對低頻，因此可以讓人耳以及模型獲得更多資訊以利判斷。

表8、發音聲調量測象限分布

	TA	FA	TR	FR

Tone 1	4413(66%)	536(08%)	1142(17%)	592(09%)
Tone 2	2544(47%)	603(11%)	1498(28%)	726(14%)
Tone 3	2386(35%)	885(13%)	2677(40%)	789(12%)
Tone 4	4780(67%)	604(08%)	1047(15%)	700(10%)
Tone 5	241(36%)	73(11%)	290(43%)	67(10%)
Total	14364(54%)	2701(10%)	6654(25%)	2874(11%)

表9、辨識結果聲調混淆矩陣

		Non-native Pronunciation					
		Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	None
Canonical	Tone 1	**6071(91%)**	183(03%)	95(01%)	234(04%)	9(00%)	91(01%)
	Tone 2	335(06%)	**3840(71%)**	1010(19%)	27(01%)	8(00%)	151(03%)
	Tone 3	86(01%)	2239(33%)	**4088(61%)**	122(02%)	7(00%)	195(03%)
	Tone 4	101(01%)	16(00%)	110(01%)	**6757(95%)**	82(01%)	65(01%)
	Tone 5	178(27%)	20(03%)	24(04%)	101(15%)	**344(51%)**	4(01%)

六、結論

本論文實驗了在端對端錯誤發音檢測系統上使用多編碼器結構處理文本提示的特徵，並且以 HAN 動態的合併不同來源的資訊。實驗部分相較於端對端基線在多種評估標準下都取得了良好的進步，證明了這個方法的有效性。另一方面，使用了混淆矩陣進行分析，具體地顯示出此方法是如何影響評估結果。在未來的部分，希望對於文本特徵與發音特徵可以找到更有效的結合法。此外考量模型可能會過於依賴文本提示而造成偏差，希望尋找一個有效的方法，針對文本特徵與發音特徵嘗試計算其是否映射到相同結果，如果是則加強文本特徵的影響，否則降低文本特徵的影響或反向拉遠希望預測結果偏離文本特徵的映射目標。最後，本論文目前只有使用一個資料集進行實驗，未來希望可以測試在更多大型資料集上是否會有不一樣的表現。

參考文獻

[1] Eskenazi, Maxine, "*An overview of spoken language technology for education*," Speech

Communication, vol. 51, no. 10, pp. 832–844, 2009.

[2] Chang, Hsiu-Jui et al., "*Investigating on Computer-Assisted Pronunciation Training Leveraging End-to-End Speech Recognition Techniques*," ROCLING, 2019.

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

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

[5] 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.

[6] Metallinou, Angeliki, and Jian Cheng, "*Using deep neural networks to improve proficiency assessment for children English language learners*," Interspeech, 2014.

[7] Witt, Silke M., and Steve J. Young, "*Phone-level pronunciation scoring and assessment for interactive language learning*," Speech communication vol. 30.2-3, pp. 95-108, 2000.

[8] Leung, Wai-Kim et al., "*CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis*," ICASSP, 2019.

[9] Watanabe, Shinji et al., "*Hybrid CTC/attention architecture for end-to-end speech recognition*," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240-1253, 2017.

[10] Feng, Yiqing et al., "*SED-MDD: Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis*," ICASSP, 2020.

[11] Lo, Tien-Hong et al., "*An Effective End-to-End Modeling Approach for Mispronunciation Detection*," arXiv, 2020.

[12] Yang, Zichao et al., "*Hierarchical attention networks for document classification*," Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016.

[13] Wang, Xiaofei et al., "*Stream attention-based multi-array end-to-end speech recognition*," ICASSP, 2019.

[14] Chorowski, Jan et al., "*End-to-end continuous speech recognition using attention-based recurrent NN: First results*," arXiv, 2014.

[15] Chorowski, Jan, et al., "*Attention-based models for speech recognition*," Advances in neural information processing systems, 2015.

[16] Vaswani, Ashish et al., "*Attention is all you need*," Advances in neural information processing systems, 2017.

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

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

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

[20] Hsiung, Y. et al., "*Development of Mandarin annotated spoken corpus (MAS Corpus) and the learner corpus analysis*," WoALF, 2014.

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

[22] Lin, Ju et al., "*Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks*," Journal of Signal Processing Systems 90.7, 2018.