結合端對端語音辨識與發音模型於英文錯誤發音偵測之研究

Exploring the Integration of E2E ASR and Pronunciation Modeling for English Mispronunciation Detection

Hsin-Wei Wang[1], Bi-Cheng Yan[1], Yung-Chang Hsu[2], Berlin Chen[1]
[1] Department of Computer Science and Information Engineering National Taiwan Normal University
[2] EZ-AI Inc.
Taipei, Taiwan
{ hsinweiwang, bicheng, berlin }@ntnu.edu.tw
mic@ez-ai.com.tw

摘要

電腦輔助發音訓練需求日益漸增，它能讓第二外語學習者根據電腦所產生的回饋來改進發音並重複練習。然而現階段基於語音辨識的發音檢測系統在該任務的效能上仍未臻完美，在缺少非母語學習者的語料下，電腦輔助發音訓練系統的表現常受到自動語音辨識性能不佳而影響。有鑑於此，本論文發展一個兩階段的英文錯誤發音偵測方法：第一階段針對學習者的語音輸入進行端對端自動語音辨識，而第二階段將自動語音辨識產生的前 N 個最佳音素序列假設輸入到發音模型以預測出最接近學習者實際所發出音素序列的假設來與提示文字的音素序列進行比對，藉此提升錯誤發音偵測的效能。本論文經由在一套英語標竿資料集所進行的一系列實驗確認了我們所提出方法的可行性。

Abstract

There has been increasing demand to develop effective computer-assisted language training (CAPT) systems, which can provide feedback on mispronunciations and facilitate second-language (L2) learners to improve their speaking proficiency through repeated practice. Due to the shortage of non-native speech for training the automatic speech recognition (ASR) module of a CAPT system, the corresponding mispronunciation detection performance is often affected by imperfect ASR. Recognizing this importance, we in this paper put forward a two-stage mispronunciation detection method. In the first stage, the speech uttered by an L2 learner is processed by an end-to-end ASR module to produce N-best phone sequence hypotheses. In the second stage, these hypotheses are fed into a pronunciation model which seeks to faithfully predict the phone sequence hypothesis that is most likely pronounced by the learner, so as to improve the performance of mispronunciation detection. Empirical experiments conducted a English benchmark dataset seem to confirm the utility of our method.

關鍵字：端對端語音辨識、發音檢測與診斷、N-best 重新排序

Keywords : End-to-End Speech Recognition , Mispronunciation Detection and Diagnosis , N-best Rescoring

1 緒論 (Introduction)

在全球化趨勢下，外語變成現今國際化人才最需具備的能力之一。近年教育方式也順應科技日新月異不斷改變，許多研究早已開始探討如何利用資訊科技及網際網路的優勢來加速語言的學習 (Mark Warschauer, 1995; Mark Warschauer et al., 2000)。語言學習的熱潮讓電腦輔助發音訓練 (Computer Assisted Pronunciation Training, CAPT)的研究逐漸受到重視，讓電腦具備與英語教師相當的專業能力。

學習者會根據 CAPT 系統提供的文本提示(prompt)進行朗讀，系統即時針對錄音結果進行偵測並診斷發音，最後提供回饋以便學習者可以改進並重複練習。

開發電腦輔助發音訓練系統與語音辨識技術息息相關，常見的實踐方法是透過識別學生的發音音素序列，將其與作為規範音素(canonical phone)的母語人士發音音素序列進行比對。最近端對端混合模型架構已逐漸取代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)(Geoffrey Hinton et al., 2012)作為主流語音辨識模型架構；使用單一的深度網絡架構取代複雜的模組組合，大大簡化了傳統語音識別系統的建立過程。常見 CAPT 系統基於語音辨識能以高準確度自動識別音素的假設下，直接從語音辨識結果診斷錯誤發音。然而在非英語母語者(non-native speaker)發音標記資料相對少量的訓練情況下，系統診斷準確率常受語音辨識性能下降影響。因此本論文設計發音模型(pronunciation model)結合端對端語音辨識，用來提升英文錯誤發音偵測性能與錯誤發音診斷準確率。論文還分別採用兩種不同的編碼/解碼器的端對端語音辨識模型進行實驗與效能評估。

2 端對端自動語音辨識技術 (E2E ASR)

2.1 CTC (Connectionist Temporal Classification)

連結時序分類最早於 2006 年提出(Alex Graves et al., 2006)，概念為給定一段長度為 T 的聲學特徵序列 X，$X = \{x_t \in \mathbb{R}^D | t = 1, \dots, T\}$ (x_t表示為第 t 音框的 D 維語音特徵向量) 及一段長度 L 的標籤序列 C，$C = \{c_l \in U | l = 1, \dots, L\}$ (U 為存在的標籤集合)，目標估計聲學特徵對應字符的後驗概率 $P(C|X)$。CTC 在訓練時引入了額外的空白標籤(blank symbol)，作為標籤間的分界，每個音框的標籤序列可表示為 $Z = \{z_t \in U \cup \{< b >\} | t = 1, \dots, T\}$ 。CTC 的目標函數表示如下：

$$P_{ctc}(C|X) \approx \sum_Z \prod_{t=1}^{T} P(z_t|z_{t-1}, C) P(z_t|X) \quad (1)$$

其中 $P(z_t|z_{t-1}, C)$ 表示為狀態轉移機率。$P(z_t|X)$ 為 CTC 聲學模型，可由透過鏈式法則展開後利用條件獨立的假設求得。

2.2 注意力機制 (Attention Mechanism)

與 CTC 方法不同，基於注意力機制的方法(Jan Chorowski et al., 2015)不做任何條件獨立假設，而是直接估計後驗機率 $P(C|X)$。注意力模型的目標函數表示如下：

$$P_{att}(C|X) \approx \prod_{l=1}^{L} P(c_l|c_1, \dots, c_{l-1}, X) \quad (2)$$

上式中的 $P(c_l|c_1, \dots, c_{l-1}, X)$ 可由下列式子求得：

$$h_t = Encoder(X) \quad (3)$$

$$a_{lt} \begin{cases} ContentAttention(q_{l-1}, h_t) \\ LocationAttention(\{a_{l-1}\}_{t=1}^{T}, q_{l-1}, h_t) \end{cases} \quad (4)$$

$$r_l = \sum_{t=1}^{T} a_{lt} h_t \quad (5)$$

$$P(c_l|c_1, \dots, c_{l-1}, X) = Decoder(r_l, q_{l-1}, c_{l-1}) \quad (6)$$

第(3)與(6)式分別為編碼器(encoder)網路與解碼器(decoder)網路。上述 4 式的符號定義分別如下：h_t為 encoder 的隱藏向量、 a_{lt}為注意力權重、 q 表示為前一個解碼隱藏向量。r_l表示字母級別的隱藏向量。注意力機制與 CTC 的差異在於注意力機制計算時會考慮過去輸出的字元。

2.3 CTC-Attention 混和模型 (Hybrid CTC-Attention Model)

由 Shinji Watanabe (2017)等人於提出的 CTC-Attention 混和模型架構如圖 1 所示，以 CTC 目標函數作為輔助任務，運用多任務學習框架訓練注意力模型的編碼器。CTC-Attention 混和模

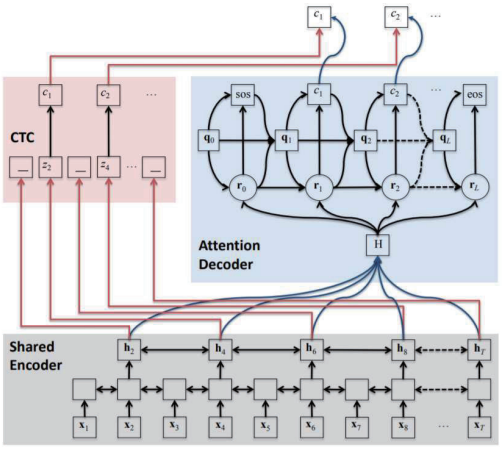

圖 1. CTC-Attention 混和模型端對端架構 [1].

[1] 圖片取自論文 Hybrid CTC/Attention Architecture for End-to-End Speech Recognition.

型共享編碼器的網路，藉由注意力模型的前後資訊改善了CTC對每個音框的對應字符輸出的獨立假設所面臨與真實情況偏離的問題，並透過 CTC 的嚴格單調特性減少對齊計算範圍加快注意力模型對齊過程。CTC-Attention 混合模型訓練的損失函數為兩種模型目標函數的線性組合，表示如下：

$$\mathcal{L}_{MOL} = \lambda log P_{ctc}(C|X) + (1-\lambda) log P^{*}_{att}(C|X) \quad (7)$$

本論文提出的英文錯誤發音偵測實驗中第一階段採用 CTC-Attention 架構進行訓練的語音辨識模型，並實驗兩種編碼與解碼器的模型(VGG-BiLSTM & Transformer)在兩階段錯誤發音偵測上的性能。

3 發音模型 (Pronunciation Model)

語言學家將第二外語學習者常見的發音錯誤情形分為以下三類(以 North, ***/n ao r th/***為例)：①語言轉移(Language transfer)：學習者使用母語發音去近似目標發音。例如：以中文發音諾斯進行朗誦 ***/ no^{4} ssu/***。②不正確字母的聲音轉換：某些不常見的word，學習者使用拼音知識去猜測。例如：***/n ow r th/***。雖然發音聽起來可能很相似，但以音素級別角度進行診斷時，其音素序列(phone sequence)還是與標準發音的音素序列不同。③誤讀文本提示：學習者朗讀與文本內容不相關的語句。

基於端對端語音辨識模型的錯誤發音偵測與診斷方法是透過計算編集距離將辨識結果與文本提示進行對齊，並直接給予回饋。開放的語料庫中英語母語者的標記訓練語料資源相對豐富，針對英語非母語學習者的標記語料反而很少或是不易取得。由於訓練受缺少非母語者發音標記的語料限制，導致自動語音辨識系統(ASR)針對錯誤發音的辨識率下降，進而影響僅一階段的電腦輔助發音訓練系統整體性能。其實錯誤發音診斷與自動語音辨識目標任務本質不同，如圖 2 所示，比起紅色為學習者可能的候選錯誤發音路徑，語音辨識系統更有可能優先輸出藍色的正確發音路徑。因此本論文提出兩階段的電腦輔助發音訓練系統，於第二階段加入發音模型(pronunciation model)輔助提升錯誤發音偵測的性能與錯誤發音診斷任務的準確率。

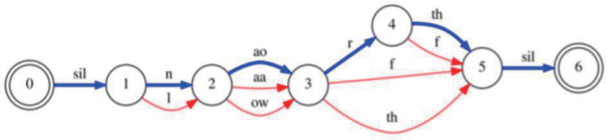

圖 2. 單字 North 的可能發音路徑.

具體來說給定一組候選 N-best 語音辨識結果，利用共享 Bi-LSTM 參數萃取出對各個候選發音路徑有用的資訊，讓發音模型學習重新選擇辨識結果。模型架構如圖 3 所示。首先將候選 N-best 結果逐一透過嵌入層(embedding layer)轉成指定大小的音素嵌入(phone embedding)；接者輸入進Bi-LSTM，Bi-LSTM 會自動編碼成序列表示法 ***h****(如圖 3 深綠色長方形所示)；最後由於這是一個分類問題，因此需要再將最後一個時間點的輸出 ***h**** 經過一層線性層(linear layer)轉換，方可進行多類別預測任務。發音模型的訓練首先計算第一階段語音辨識產生的候選N-best 序列在錯誤發音偵測任務的 F1 分數，將分數高低排名作為優異與否的分類依據，成為第二階段發音模型多類別分類(Multi-class Classification)的預測目標。交叉熵被廣泛應用於許多多類別分類任務中，本篇論文訓練也採用交叉熵(Cross Entropy)作為損失函數。測試時取分類類別為最優的發音序列作為發音模型的輸出，以進行後續發音偵測任務與診斷任務。

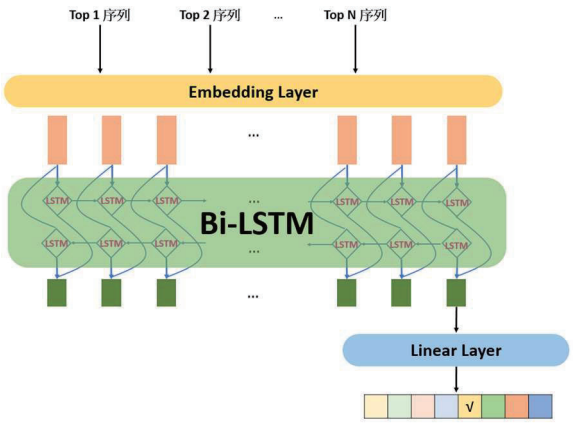

圖 3. N-best 辨識結果與聲學分數分類模型架構.

4 實驗 (Experiments)

4.1 資料集

實驗使用兩個資料集，分別為美式英語母語者(L1 speaker) 的 ***TIMIT*** (J. S. Garofolo et al., 1993)以及英語非母語者(L2 speaker)的 ***L2-ARCTIC*** (Guanlong Zhao et al., 2018)。

TIMIT 由來自美國八個主要方言地區的 630 位美式英語母語人士錄製指定的 10 句文本提示 (prompt)，共計 5.4 小時，所有的錄音都提供經時間對齊的音素級別轉錄。該資料集語句分為 3 種類型，詳細資料如表 1 所示，SA 為方言語句，SX 與 SI 則為一般類型語句。將該資料集切割為訓練集 3.15 小時與驗證集 0.34 小時作為實驗使用，統計如表 2 所示。

L2-ARCTIC 包括二十四位英語非母語人士的錄音，每位語者錄製大約一小時取自 CMU ARCTIC 的文本提示。錄音者的母語分別為印度語、韓語、華語、西班牙語、阿拉伯語和越南語，共計 6 種不同語言，詳細資料如表 3 所示。該資料集除提供經強制對齊的音素級別轉錄外，也提供每位語者 150 句的專家標記(annotation)語句，其中 150 句分別選自相同的 100 句文本提示，以及 50 句針對各母語特性挑選的易犯發音錯誤的文本提示，增加發音錯誤情形。將該資料集切割為訓練集 2.66 小時，驗證集 0.12 小時及測試集 0.88 小時作為實驗使用，統計如表 4 所示。

表 1. TIMIT 資料集.

類型	SA	SX	SI	統計
總文本數	2	450	1890	2342
各文本所含語者數	630	7	1	-
各語者所唸文本數	2	5	3	10
總句數	1260	3150	1890	6300

表 2. 實驗中 TIMIT 資料統計.

	訓練	驗證
總語者數	462	50
各語者所唸句數	8	8
總句數	3696	400
總音素個數	139,940	15,342
時長統計(hrs)	3.15	0.34

表 3. L2-ARCTIC 資料集.

語者	母語
ABA / SKA / ZHAA / YBAA	阿拉伯語
BWC / LXC / NCC / TXHC	華語
ASI / RRBI / SVBI / TNI	印度語
HJK / HKK / YDCK / YKWK	韓語
EBVS / ERMS / MBMPS / NJS	西班牙語
HQTV / PNV / THV / TLV	越南語

表 4. 實驗中 L2-ARCTIC 資料統計.

	訓練	驗證	測試
語者數	2549	150	900
正確發音音素個數	71,935	4,054	25,690
錯誤發音音個數	13,236	903	4,314
時長統計(hrs)	2.66	0.12	0.88

4.2 發音偵測與診斷任務評估指標

在發音偵測任務我們會分別關注正確發音以及錯誤發音的判定成效，兩者的評估指標 Recall(RE)、Precision(PR)與 F 度量(F1)計算方式如下：

- 正確發音：

$$\mathrm{RE}_{cor} = \frac{\text{正確接受數量}}{\text{實際正確發音數量}} = \frac{TA}{TA+FR} * 100\% \quad (8)$$

$$\mathrm{PR}_{cor} = \frac{\text{正確接受數量}}{\text{系統判定正確發音數量}} = \frac{TA}{TA+FA} * 100\% \quad (9)$$

$$\mathrm{F1}_{cor} = \frac{2\times \mathrm{RE}_{cor}\times \mathrm{PR}_{cor}}{\mathrm{RE}_{cor}+\mathrm{PR}_{cor}} * 100\% \quad (10)$$

- 錯誤發音：

$$\mathrm{RE}_{mis} = \frac{\text{正確拒絕數量}}{\text{實際錯誤發音數量}} = \frac{TR}{TR+FA} * 100\% \quad (11)$$

$$\mathrm{PR}_{mis} = \frac{\text{正確拒絕數量}}{\text{系統判定錯誤發音數量}} = \frac{TR}{TR+FR} * 100\% \quad (12)$$

$$\mathrm{F1}_{mis} = \frac{2\times \mathrm{RE}_{mis}\times \mathrm{PR}_{mis}}{\mathrm{RE}_{mis}+\mathrm{PR}_{mis}} * 100\% \quad (13)$$

錯誤發音偵測評估指標中的正確拒絕 TR，其實又可以分成診斷正確 CD (不僅檢測出學習者唸錯，又正確診斷唸錯成哪個音素)與診斷錯誤 DE (雖然判定學習者唸錯，但無法正確診斷唸錯成哪個音素)。因此錯誤發音診斷正確率(DAR_{mis})的計算方式如下：

$$\mathrm{DAR}_{mis} = \frac{\text{正確診斷數}}{\text{正確診斷數+診斷錯誤數}} = \frac{\mathrm{CD}}{\mathrm{CD+}} * 100\% \quad (14)$$

4.3 實驗設定

本論文的第一階段實驗使用了開源端對端語音辨識工具 **Espnet** (Shinji Watanabe et al., 2018) 完成，端對端語音辨識模型使用 CTC-Attention 混合模型架構，並進行兩個不同的實驗設定。分別採用 VGG-BiLSTM 與 Transformer 的編碼器與解碼器進行實驗，具體設定如表 5 所示。

表 5. 端對端語音辨識實驗設定.

VGG-BiLSTM			
feature		80-dim fbank + 3-dim pitch	
encoder / decoder		BiLSTM	
encoder		decoder	
layers	2	layers	3
hidden size	1024	hidden size	1024
CTC/Attention 混和比		0.6/0.4	
Transformer			
feature		80-dim fbank + 3-dim pitch	
encoder / decoder		Transformer	
encoder		decoder	
attention heads	8	attention heads	8
linear units	2048	linear units	2048
blocks	12	blocks	6
dropout rate	0.1	dropout rate	0.1
CTC/Attention 混和比		0.3/0.7	

4.4 實驗結果與討論

實驗分別使用兩個語音辨識模型進行解碼，將產生的 5-best 候選結果用於第二階段發音模型的訓練。訓練完成後，將發音模型分別測試在測試集解碼的 5-best 候選結果。

在第一階段的語音辨識結果與專家標記(annotation)進行比對的音素錯誤率(phone error rate)表現如表 6 所示。可以看到 VGG-BiLSTM 的語音辨識結果表現均較 Transformer 表現差，這將會影響採用 VGG-BiLSTM 的電腦輔助發音訓練系統在後續錯誤發音診斷任務上的準確率不如採用 Transformer 的發音訓練系統表現。

表 6. L2-ARCTIC 測試集中各語者音素錯誤率.

語者	VGG-BiLSTM	Transformer
NJS	23.3	15.7
TLV	25	18.2
TNI	32.3	19.7
TXHC	28	18.3
YKWK	24.9	15.7
ZHAA	26.1	15.3
平均	26.6	17.1

測試集經過發音模型後最後輸出的發音序列與文本提示(prompt)進行比對的發音檢測與診斷表現如表 7 與表 8 所示。Baseline 為一階段基於語音辨識的發音訓練系統結果，N=5 為本論文提出的兩階段電腦輔助發音訓練系統結果，另外針對 Transformer 產生的候選結果還簡單連接(concatenated)了聲學模型分數一起作為發音模型的輸入進行實驗與測試，實驗結果如表 7 中的 N=5#所示。

實驗結果可以看到，發音模型重新選擇候選結果均能改進了發音檢測任務的各項指標。雖

然採用 VGG-BiLSTM 的電腦輔助發音訓練系統在偵測任務指標表現進步幅度略小，但可以看到發音模型讓診斷錯誤的音素個數減少(diagnose error，DE)，正確診斷的音素個數增加，進而讓診斷任務的準確率有所提升。

表 7. 採用 VGG-BiLSTM 編/解碼器表現

指標	baseline	N=5
Correct Pronunciation		
PR(%)	94.57	**94.59**
RE(%)	79.53	**79.54**
F1(%)	86.4	**86.41**
Mispronunciation		
PR(%)	35.72	**35.77**
RE(%)	71.36	**71.46**
F1(%)	47.61	**47.67**
CD(音素個數)	1803	**1809**
DE(音素個數)	1120	**1118**
DIA(%)	61.68	**61.8**

表 8. 採用 Transformer 編/解碼器表現

指標	baseline	N=5	N=5 #
Correct Pronunciation			
PR	92.14	**92.16**	**92.21**
RE	91.13	**91.48**	90.98
F1	91.63	**91.65**	91.59
Mispronunciation			
PR(%)	47.95	**48.05**	47.79
RE(%)	51.24	**51.34**	**51.78**
F1(%)	49.54	**49.64**	**49.71**
CD(音素個數)	1526	**1528**	**1553**
DE(音素個數)	573	575	**568**
DIA(%)	72.7	72.66	**73.22**

在採用 Transformer 電腦輔助發音訓練系統的表現上，發音偵測的各項指標都明顯提升，尤其正確發音偵測的 recall 表現。在實務上，對正確發音的偵測性能提升，可以讓學生更願意相信並使用系統進行練習。雖然受錯誤診斷的音素個數影響，診斷錯誤率略為下降，但在更進一步的實驗中串聯聲學模型表現分數作為輸入，可以有效提升診斷準確率。正確診斷發音的音素個數明顯增加，錯誤診斷的音素個數下降，讓診斷準確率提升至 73.22 是所有實驗中表現最好的結果。實驗結果也表明在發音模型中加入更多元的資訊可以幫助電腦輔助發音訓練系統中診斷任務性能進一步有效提升。

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

本論文實踐了二階段電腦輔助發音訓練系統，將基於端對端語音辨識的系統結合發音模型進行實驗。發音模型對候選序列重新排序，可以有助於校正第一階段辨識學習者的發音結果並提升發音檢測的指標。實驗結果顯示透過簡單模型架構，就能改進電腦輔助發音訓練系統檢測性能。未來將會加大第一階段候選者數作為發音模型輸入的實驗，也會考慮更多元的特徵與候選結果進行串聯，例如：多元的語音特徵、音素時長(duration)，甚至是學習者的發音特徵...等，來提升發音診斷的準確率。本論文僅是對發音模型進行初步的實驗，未來會持續改進發音模型，相信音素級別候選結果的重新選擇及更嚴謹的發音模型設計架構，能再讓發音檢測與診斷系統更趨完善。

參考文獻 (References)

Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton. 2013. *Speech recognition with deep recurrent neural networks*. in ICASSP 2013.

Alex Graves, Santiago Fern´andez, Faustino Gomez and J¨urgen Schmidhuber. 2016. *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks*. in ICML 2006.

Eskenazi, Maxine. 2019. An overview of spoken language technology for education. in *Speech Communication*. Vol. 51, No. 10, 832–844.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. in *IEEE Signal Processing Magazine*, Vol. 29, No. 6, 82-97. DOI: 10.1109/MSP.2012.2205597.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai and Ivana Lucic. 2018. *L2-ARCTIC: A Non-native English Speech Corpus*. in INTERSPEECH 2018.

Hsiu-Jui Chang, Tien-Hong Lo, Tzu-En Liu and Berlin Chen. 2019. *Investigating on Computer-Assisted Pronunciation Training Leveraging End-to-End Speech Recognition Techniques*. in ROCLING 2019.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus and D. S. Pallett. 1993. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. *NIST speech disc 1*-1.1. NASA STI/Recon technical report, No. 93.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho and Yoshua Bengio. 2015. *Attention-Based Models for Speech Recognition*. in NIPS 2015.

L. R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. in *Proceedings of the IEEE*. Vol. 77, No. 2, 257-286. DOI: 10.1109/5.18626.

Mark Gales and Steve Yang. 2018. The Application of Hidden Markov Models in Speech Recognition. in *Signal Processing*. Vol. 1, No. 3, 195–304. DOI: 10.1561/2000000004.

Mark Warschauer, Heidi Shetzer, and Christine Meloni. 2000. *Internet for English teaching.Alexandria*. VA: Teachers of English to Speakers of Other Language, Inc.

Mark Warschauer. 1995. *Virtual connections: Online activities and projects for networking language learners*. Universit y of Hawaii: SecondLanguageTeaching & Curriculum Center.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala and Tsubasa Ochiai. 2018. *ESPnet: End-to-End Speech Processing Toolkit*. in INTERSPEECH 2018.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey and Tomoki Hayashi. 2017. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition. in *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, No. 8, 1240-1253. DOI: 10.1109/JSTSP.2017.2763455.

Suyoun Kim, Takaaki Hori, Shinji Watanabe. 2017. *Joint CTC-Attention based end-to-end speech recognition using multi-task learning*. in ICASSP 2017.

Wai-Kim Leung, Xunying Liu and Helen Meng. 2019. *CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis*. in ICASSP 2019.

Ying Qin, Yao Qian, Anastassia Loukina, Patrick Lange, Abhinav Misra, Keelan Evanini and Tan Lee. 2021. *Automatic Detection of Word-Level Reading Errors in Non-native English Speech Based on ASR Output*. in ISCSLP 2021.