Computational Linguistics and Chinese Language Processing
Vol. 24, No. 1, June 2019, pp. 39-50

探究端對端混合模型架構於華語語音辨識

An Investigation of Hybrid CTC-Attention Modeling in Mandarin Speech Recognition

張修瑞*、趙偉成*、羅天宏*、陳柏琳*

Hsiu-Jui Chang, Wei-Cheng Chao, Tien-Hong Lo, Berlin Chen

摘要

近年來端對端(End-to-End)語音辨識的出現，簡化了許多傳統語音辨識的繁複流程。端對端語音辨識中，最主要的模型架構分別為連結時序分類(Connectionist Temporal Classification, CTC)與注意力模型(Attention Model)。本論文嘗試結合上述兩種模型架構(即 CTC-Attention 混合模型)於華語會議語音辨識之使用，以期能進一步提升語音辨識的效能。為此，我們分析模型結合時混合權重調整的影響，並進一步探究 CTC-Attention 混合模型對於短句的辨識效果。在中文會議語料的實驗結果顯示，相較於傳統語音辨識的 TDNN-LFMMI 模型，CTC-Attention 混合模型在語句較短時，可具有較好的一般化能力(Generalization)。

Abstract

The recent emergence of end-to-end automatic speech recognition (ASR) frameworks has streamlined the complicated modeling procedures of ASR systems in contrast to the conventional deep neural network-hidden Markov (DNN-HMM) ASR systems. Among the most popular end-to-end ASR approaches are the connectionist temporal classification (CTC) and the attention-based encoder-decoder model (Attention Model). In this paper, we explore the utility of combining CTC and the attention model in an attempt to yield better ASR performance. we also analyze the impact of the combination weight and the

* 國立台灣師範大學資訊工程研究所
Department of Computer Science and Information Engineering, National Taiwan Normal University
E-mail: {60647061S , 60647028S , teinhonglo , berlin}@ntnu.edu.tw

performance of the resulting CTC-Attention hybrid system on recognizing short utterances. Experiments on a Mandarin Chinese meeting corpus demonstrate that the CTC-Attention hybrid system delivers better performance on short utterance recognition in comparison to one of the state-of-the-art DNN-HMM settings, namely, the so-called TDNN-LFMMI system.

關鍵詞：CTC、Attention、端對端中文語音辨識、短句辨識
Keywords: CTC, Attention-based Encoder-Decoder, End-to-End Mandarin Chinese Speech Recognition, Short Utterance Recognition

1. 緒論 (Introduction)

隨著近幾年來深度學習技術的長足發展，在語音辨識任務上，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) (Hinton *et al*., 2012)與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) (Rabiner, 1989) (Gales & Yang, 2008)相比，在字錯誤率(Character Error Rate, CER)和詞錯誤率(Word Error Rate, WER)有了大幅度的下降。然而，儘管 DNN-HMM 已取得不錯的成果，但 DNN 聲學模型仍無法充分利用語音信號之時間依賴性的缺點，為了更好地捕捉該性質，過往學者們引入了遞歸類神經網路(Recurrent Neural Network, RNN) (Hochreiter & Schmidhuber, 1997) (Gers, Schmidhuber & Cummins, 1999)及長短期記憶模型(Long Short-Term Memory, LSTM) (Graves, Mohamed & Hinton, 2013) (Graves, Jaitly & Mohamed, 2013) (Sak, Senior & Beaufays, 2014) (Sak, Vinyals & Heigold, 2014) (Li & Wu, 2015)組成聲學模型。這類的聲學模型與 DNN 相同，在訓練時仍是使用最小交互熵(Cross Entropy, CE)的準則，並且也能夠再進一步結合序列式鑑別式訓練(Kingsbury, Sainath & Soltau, 2012) (Veselý, Ghoshal, Burget & Povey, 2013)得到更好的辨識效果。

語音辨識可以視為一種序列對序列的任務，將輸入的語音訊號對應輸出的文字序列。在傳統語音辨識器的訓練中，分別由聲學模型、語言模型及發音詞典構成，並且在訓練 DNN 前，還得透過預先訓練的 GMM-HMM 將聲音與文字強制對齊，因此需要額外的冗餘步驟。有別於傳統的語音辨識訓練，CTC 訓練準則使得聲學模型可直接將聲學特徵透過類神經網路輸出對應到的字符(Character)或音素(Phone) (Graves *et al*., 2013) (Graves, Fernández, Gomez & Schmidhuber, 2006)，甚至在資料量夠大(通常大於 3000 小時)時能夠直接對應到單詞(Soltau, Liao & Sak, 2016) (Li, Ye, Das, Zhao & Gong, 2018)，並且在解碼時可以不需要語言模型，這樣的做法稱之為端對端的訓練方式。另一方面，有鑑於 CTC 端對端模型的成功，且基於 Attention 的遞歸類神經網路已被廣泛應用於各個研究領域(Bahdanau, Cho & Bengio, 2015) (Xu *et al*., 2015)，(Chorowski, Bahdanau, Serdyuk, Cho & Bengio, 2015)也將此模型應用於語音辨識的任務上，得到接近 CTC 的 WER。在後續其他學者研究中，在大量語料的情況下，Attention 模型的 WER 甚至能逼近辨識效果很好的 CLDNN-HMM 模型(Convolutional Long Short-Term Memory, Fully Connected Deep

Neural Networks, CLDNN) (Chan, Jaitly, Le & Vinyals, 2016)。

雖然端對端的訓練方式相較於傳統的 DNN-HMM 訓練更加簡單，但在少量語料下，其效能仍與傳統的 DNN-HMM 模型有一段差距。為此，(Kim, Hori & Watanabe, 2017) (Watanabe, Hori, Kim, Hershey & Hayash, 2017)，使用 CTC-Attention 模型 (Hybrid CTC-Attention Model)。該方法為結合 CTC 與 Attention 模型的多任務學習架構，目的是希望利用 CTC 彌補 Attention 模型對齊錯誤(Misalignment)及收斂慢的問題。在(Kim *et al*., 2017) (Watanabe *et al*., 2017)的實驗結果顯示，CTC-Attention 模型可在少量語料下，能夠更接近甚至低於 DNN-HMM 模型的辨識率。因此，本篇論文希望基於此模型對於中文會議語料的辨識做研究探討，我們的貢獻可分為：

1. 不同 Attention 機制的辨識結果：在長句測驗集實驗結果中發現使用 Coverage Location 效果比 Location 機制好，而在短句實驗則反之。
2. CTC 的權重對於辨識結果之影響：一般來說情況下，多任務架構訓練之聲學模型可優於傳統 CTC 或 Attention 模型。
3. CTC-Attention 混合模型於短語句測試之影響：短句辨識任務上，當使用較大的 CTC 權重作為解碼參數，可以得到最好的效果。

2. 方法 (Method)

2.1 CTC (Connectionist Temporal Classification)

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$\mathrm{C} = \{c_l \in \mathrm{U} | \mathrm{l} = 1, \dots, L\}$，U 為存在的標籤集合。並且 CTC 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$S = \{s_t \epsilon \mathrm{U} \cup \{< blank >\} | t = 1, \dots T\}$。 X 對應 C 的後驗機率可表示為：

$$P(C|X) = \sum_{S} P(C|S,X)P(S|X)$$

$$\approx \sum_{S} P(C|S)P(S|X) \quad (1)$$

由於 CTC 假設每一時間下的聲音輸入對應字符為條件獨立，因此$P(C|S,X) \approx P(C|S)$，其中$P(C|S)$可以視為 CTC 標籤模型，可以分別由貝氏定理(Bayes' Rule)、鏈式法則(Chain Rule)展開。最後帶入條件獨立的假設可推導為：

$$P(C|S) = \frac{P(S|C)P(C)}{P(S)}$$
$$= \prod_{t=1}^{T} P(s_t|C, s_{1:t-1}) \frac{P(C)}{P(S)}$$
$$\approx \prod_{t=1}^{T} P(s_t|s_{t-1,}C) \frac{P(C)}{P(S)} \quad (2)$$

其中，$P(C)$為字符級別的語言模型，$P(S)$為每一狀態的先驗機率，$P(s_t|s_{t-1,}C)$為狀態轉移機率，為了使輸出有空白標籤，CTC 將上述長度 L 的標籤序列 C 調整為：

$$c' = \{< blank >, c_1, < blank >, c_2, < blank >, \dots c_L\}$$
$$= \{c'_l \in \mathrm{U} \cup \{< blank >\} | l = 1, \dots 2L + 1\} \quad (3)$$

狀態轉移機率$p(s_t|s_{t-1,}C)$可以表示為：

$$P(s_t|s_{t-1,}C) \begin{cases} 1 & s_t = c'_l \; and \; s_{t-1} = c'_l \; for \; all \; possible \; l \\ 1 & s_t = c'_l \; and \; s_{t-1} = c'_{l-1} \; for \; all \; possible \; l \\ 1 & s_t = c'_l \; and \; s_{t-1} = c'_{l-2} \; for \; all \; possible \; even \; l \\ 0 & otherwise \end{cases} \quad (4)$$

其依序分別為相似於 HMM 的自我轉移(Self-loop)，轉移至下一狀態，而第三個則是在l為偶數時且c'_l及c'_{l-2}皆屬於標籤序列 S 時跳過 blank 狀態，如同下圖的拓樸結構：

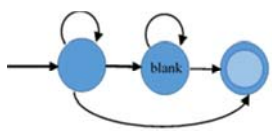

圖 1. CTC 拓撲結構
[Figure 1. CTC's topology]

另一方面，P(S|X)為 CTC 聲學模型，由鏈式法則展開後，再帶入條件獨立的假設可以表示為：

$$P(S|X) = \prod_{t=1}^{T} P(s_t|s_1, \dots, s_{t-1}, X)$$
$$\approx \prod_{t=1}^{T} P(s_t|X) \quad (5)$$

其中$P(s_t|X)$為 softmax 輸出的結果，綜合上述式(2)、式(5)，可以得到：

$$P(C|X) \approx \sum_s \prod_{t=1}^{T} P(s_t|s_{t-1}, C) P(s_t|X) \frac{P(C)}{P(S)} \tag{6}$$

而 CTC 的目標函數通常不包含$\frac{P(C)}{P(S)}$ ，因此可定義為：

$$P_{ctc}(C|X) \approx \sum_s \prod_{t=1}^{T} P(s_t|s_{t-1}, C) P(s_t|X) \tag{7}$$

上式為CTC目標函數，而訓練時希望最小化損失函數(Loss Function)便是$-\ln P_{ctc}(C^*|X)$，C^*為訓練語料的正確字符序列的標籤，損失函數越小等同於輸出正確標籤的機率越大。

2.2 Attention 模型 (Attention-based Encoder-Decoder Network)

有別於 CTC 對於聲音對應字符的條件獨立假設，Attention 模型直接估測聲學特徵對應到字符的後驗機率，其目標函式可定義為：

$$P_{att}(C|X) = \prod_{l=1}^{L} P(c_l|X, c_{1:l-1}) \tag{8}$$

$P(c_l|X, c_{1:l-1})$可以由下列式子推得：

$$h_t = Encoder(X) \tag{9}$$

$$e_{l<t} = \begin{cases} LocationAttention: & \\ \boldsymbol{F}_l = \boldsymbol{K} * \boldsymbol{a_{l-1}} & (10) \\ \mathbf{g}^T \tanh\left(W_q \boldsymbol{q_{l-1}} + W_h \boldsymbol{h_t} + W_f \boldsymbol{f_{lt}}\right) & (11) \\ CoverageLocationAttention: & \\ \boldsymbol{F_l} = \boldsymbol{K} * \boldsymbol{a_{l-1}} & (12) \\ \boldsymbol{v_l} = \sum_{l'=1}^{l-1} \boldsymbol{a l'} & (13) \\ \mathbf{g}^T \tanh\left(W_q \boldsymbol{q_{l-1}} + W_h \boldsymbol{h_t} + W_f \boldsymbol{f_{lt}} + W_v \boldsymbol{v_{lt}}\right) & (14) \end{cases}$$

$$a_{lt} = \frac{\exp(\gamma e_{lt})}{\sum_l \exp(\gamma e_{lt})} \tag{15}$$

$$\boldsymbol{r_l} = \sum_{t=1}^{T} a_{lt} \boldsymbol{h_t} \tag{16}$$

$$p(c_l|X, c_{1:l-1}) = \text{Decoder}(\boldsymbol{r_l}, \boldsymbol{q_l}, c_{l-1}) \tag{17}$$

其中$\mathrm{h_t}$為 Encoder 的隱藏狀態向量，$\mathrm{a_{lt}}$為 Attention 的權重由$\mathrm{e_{lt}}$作 Softmax 得到，而 γ 為強調權重的 Sharpen Factor，而我們可藉由 Decoder 的隱藏狀態向量$\mathrm{q_{l-1}}$ 為 Query 去查找做為 Key-Value 的$\mathrm{h_t}$得到$\mathrm{e_{lt}}$，g、$\mathrm{W_q}$、$\mathrm{W_h}$、$\mathrm{W_f}$、$\mathrm{W_v}$為可訓練的矩陣參數。F_l為 Location Attention 機制(Chorowski *et al*., 2015)中由一維摺積層 K 對於過去的 Attention 向

量 $\{a_1, a_2, \ldots a_{l-1}\}$抽取的向量集合，Fl={fl1, fl2, …, flT}。v_l為 Coverage Attention 機制(Watanabe *et al*., 2017)中負責紀錄所有 Decoder 過去的 Attention 權重分佈，加入該機制的目的是希望能夠減少插入錯誤(Insertion)與刪除錯誤(Deletion)的出現，以達到更低的 WER 或 CER。Attention 模型訓練時損失函數也同樣希望最小化$-\ln P_{att}(C^*|X)$。Attention 模型與 CTC 損失函數差異在於前者計算時必須考慮過去輸出的字符。

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

由於語音的每個音框間彼此相關，所以 CTC 中對於每個音框對應文字輸出的獨立性假設是飽受批評。另一方面，Attention 模型有著非單調的左到右對齊和收斂較慢的缺點。(Kim et al., 2017) (Watanabe et al., 2017)通過使用 CTC 目標函數作為輔助函數，將 Attention 模型與 CTC 結合作多任務學習。這種訓練方式可保留 Attention 模型的優勢，並能有效改善 Attention 模型的收斂速度與對齊錯誤的問題。綜合式(7)及式(8)，CTC-Attention 混合模型透過線性組合兩種模型的目標函數，其訓練的損失函數可以表示成：

$$\mathcal{L}_{MOL} = -\big(\lambda ln P_{ctc}(C|X) + (1-\lambda) ln P_{att}(C|X)\big) \tag{18}$$

其中 λ 的範圍為 0 ≤ λ ≤ 1，而在解碼時，，我們可同時使用 CTC 及 Attention 模型的輸出，可表示為：

$$\begin{aligned} &logp(c_n|c_{1:n-1}, h_{1:T'}) \\ &= \alpha logp_{ctc}(c_n|c_{1:n-1}, h_{1:T'}) + (1-\alpha) logp_{att}(c_n|c_{1:n-1}, h_{1:T'}) \end{aligned} \tag{19}$$

2.4 聲學模型 (Acoustic model)

本篇論文在聲學模型的 Encoder 部分使用的是兩層的 VGG 層加上八層 Long Short-Term Memory Projection(LSTMP)，LSTMP (Sak *et al*., 2014)是 LSTM 的變形，通過添加投影層來進一步優化 LSTM 的速度和效能。而 VGG 與(Chan *et al*., 2016)的金字塔型的 LSTM 結構作為 Encoder 相比，使用 VGG 的效果在(Watanabe *et al*., 2018)說明了在大多數情況會優於金字塔型的 LSTM，因此我們採用 VGG-LSTMP 作為 Encoder，完整模型架構如圖 2，其中 X 代表輸入特徵，C 代表輸出的字符序列。解碼算法採用光束搜尋，搜尋時的分數結合可參考 2.3 節式 19。

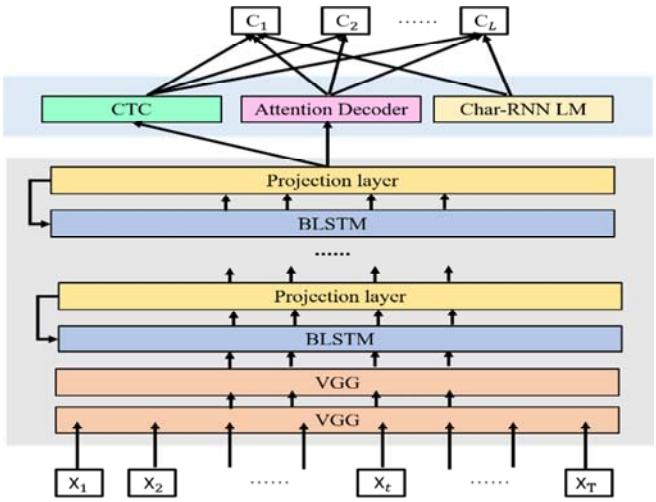

圖 2. CTC-Attention 混合模型架構
[Figure 2. Hybrid CTC-Attention model architecture]

3. 實驗結果與分析 (Experiments and Results)

3.1 實驗語料與設定 (Corpus and Setup)

本論文實驗使用的語料為華語會議語料，該語料為國內企業所收集整理的語料庫。其中談話內容沒有經過設計，而是一般公司在實際開會中討論面臨的問題與技術，而說話方式屬於正常交談，所以會有不少停頓、口吃、中英文轉換等情形，相較於新聞語料，較具有挑戰性。其訓練集為 230 小時，而測試集則為 2.6 小時兩場會議的內容，另外還有一額外 3 小時短句測試集，其內容為多為在訓練語料中未曾出現的專有名詞，在辨識上更有難度。

表 1.語料庫訓練集、測試集小時數與句數
[Table 1. hours of training set and test set]

	總小時數	句數
訓練集	230	367434
測試集	2.6	2306
短句測試集	3	2809

特徵部份，我們使用 80 維的 Filterbank 加 Pitch 特徵；聲學模型部分，我們使用兩層 VGG 層及八層 LSTMP 作為 Encoder，每層 LSTMP 各有 320 個單元，Decoder 部分則使用單層 300 個單元的 LSTM，如圖 2 所示。Attention 機制分別為 Location 及 Coverage Location。語言模型部分我們用訓練集的轉寫作為語料訓練字符級別的 RNN 語言模型，訓練時 CTC 權重設為 0.5，在解碼時使用(Watanabe *et al.*, 2017)的解碼算法並利用

Shallow Fusion (Gulcehre *et al*., 2015)的方式，插入額外的語言模型分數以提升整體辨識效能，實作上使用 Espnet (Watanabe *et al*., 2018)工具，另外為我們也使用了 Kaldi (Povey *et al*., 2011)工具實作時延式類神經網路(Time-delay Neural Network, TDNN)結合 Lattice-free Maximum Mutual Information (LF-MMI) (Povey *et al*., 2016)訓練的聲學模型與端對端混和模型做比較。

3.2 實驗結果 (Experiment result)

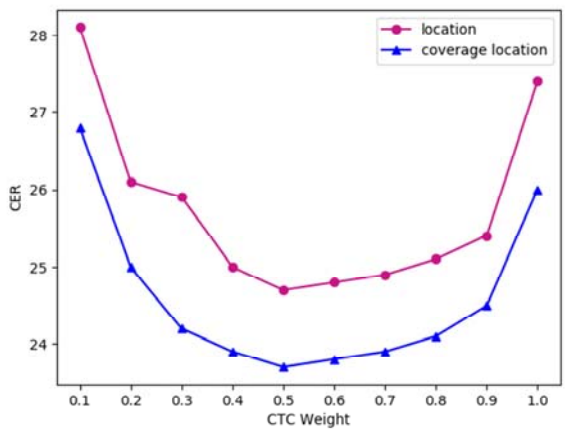

圖 3.不同的 CTC 權重對於測試集 CER 的影響
[Figure 3. Character error rate when using different CTC weight in test set]

圖 3 橫軸代表 CTC 的權重，而縱軸代表 CER。由於 CTC 的權重在解碼時是可以變動的，我們利用窮舉的方式嘗試不同的權重組合。由實驗結果得知，我們發現 Location 及 Coverage Location 皆發現權重設為 0.5 在測試集上表現最好，而權重偏向 CTC 或是 Attention 都使 CER 有上升趨勢。當 CTC 權重為 1.0 時可視為傳統 CTC 模型，反之當權重為 0.0 時為傳統 Attention 模型。另一方面，Coverage Location 在任一權重下其 CER 皆比 Location Attention 模型低，因此我們進一步去分析其解碼結果。

表 2.不同 Attention 機制的表現
[Table 2. Different attention mechanism performance in test set]

Attention	CER	#Deletion	#Insertion
location	24.7	3637	1474
Coverage location	**23.7**	3378	1467

由圖 3 已知道 CTC 的權重設為 0.5 時其 CER 為最低，因此表 1 為該權重下的辨識率，CER 分別為 24.7 及 23.7。在實驗的結果中，我們發現由 Coverage 機制的模型解碼後，插入錯誤與刪除錯誤數有些微但一致的進步，其結果也反映在 CER 上。其中可能的原因是 Coverage 機制，該機制避免了模型的注意力過度集中在同個音框的語音特徵上。

另外 TDNN-LFMMI 於此測試集的 CER 為 17%，相較之下我們的方法仍有進步空間。

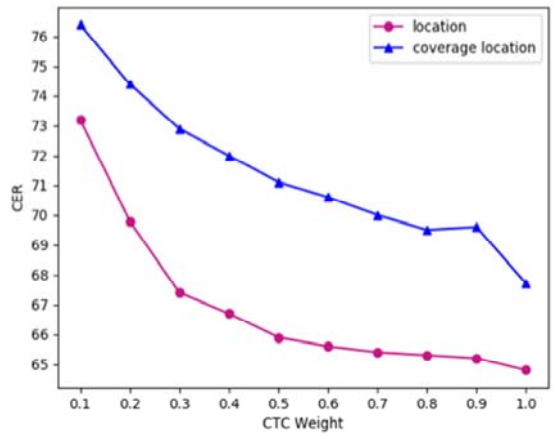

圖 4. 不同的 CTC 權重對於短句測試集 CER 的影響
[Figure 4. Character error rate when using different CTC weight in external short utterance test set]

在這次的實驗中，我們額外比較 CTC-Attention 混合模型於短句辨識任務上的表現，由圖 4 可以得知在任一權重下的 CER，與前一個測試集的實驗相反，Location 機制的模型反而較 Coverage Location 好。推測其原因可能在於語句過短，使得 Coverage Location 模型無法發揮 Coverage 機制的作用，因而表現較差。而 CTC 權重為 1.0 時，即僅使用 CTC 解碼，兩種模型皆為最佳表現，其原因可能在於 CTC 模型是為了解決輸出的文字序列長度小於輸入的聲音長度的情況而設計，而 Attention 模型，也出現了如同(Chan *et al.*, 2016) 的實驗結果，當測試語句與訓練語句長度差異太大時，解碼出來的 CER 變差許多，然而因為 CTC 權重的可變動性，可以看到 CTC-Attention 混合模型具有因應不同語句長度的彈性。

表 3.不同 Attention 機制於短句測試集表現
[Table 3. Different attention mechanism performance in in external short utterance test set]

Model	CER
location	**64.8**
Coverage location	**67.7**
TDNN-LFMMI	85.5

4. 結論與未來展望 (Conclusion and Future works)

本篇論文探討了兩種端對端語音辨識的主流方法，以及 CTC-Attention 模型權重對於語句長短的辨識效果，我們發現在短語句辨識上 CTC-Attention 模型相不僅相較於 TDNN-LFMMI 的表現更加出色，同時具有能夠依據語句長短改變權重解碼的彈性。另一方面，並且由於使用字符級別的預測目標及語言模型，更能有效處理未知詞的問題。

近年來在序列對序列模型上有學者提出許多優化訓練的方法如(Pereyra, Tucker, Chorowski, Kaiser & Hinton, 2017)，能夠避免 Overconfidence，以及 Cold Fusion (Sriram, Jun, Satheesh & Coates, 2018) 在訓練聲學模型時加入預先訓練語言模型，以上方法都能夠有更好的泛化效果與收斂速度，我們在未來也將在訓練中嘗試加入該方法。其次，在語言模型則將加入目前訓練集外以外的語料，並希望能針對語種切換做額外研究；最後，聲學模型方面也希望能夠再多嘗試不同的 Attention 機制，以及不同的類神經網路架構對於華語語音辨識的效果，以期待未來能夠得到更低的字錯誤率。

參考文獻 (References)

Bahdanau, D., Cho, K.H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of ICASSP 2016*. doi: 10.1109/ICASSP.2016.7472621

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In *Proceedings of NIPS 2015*, 577-585.

Gales, M. & Yang, S. (2008). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends® in Signal Processing*, *1*(3), 195-304. doi: 10.1561/2000000004

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. In *Proceedings of ICANN 1999*. doi: 10.1049/cp:19991218

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML 2006*, 369-376. doi: 10.1145/1143844.1143891

Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *Proceedings of ASRU 2013*. doi: 10.1109/ASRU.2013.6707742

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of ICASSP 2013*.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., …Bengio, Y. (2015). On Using Monolingual Corpora in Neural Machine Translation. In arXiv preprint arXiv: 1503.03535

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., …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. doi: 10.1109/MSP.2012.2205597

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Kim, S., Hori, T., & Watanabe, S. (2017). Joint CTC-Attention based end-to-end speech recognition using multi-task learning. In *Proceedings of ICASSP 2017*. doi: 10.1109/ICASSP.2017.7953075

Kingsbury, B., Sainath, T. N., & Soltau, H. (2012). Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization. In *Proceedings of Interspeech 2012*, 10-13.

Li, X. & Wu, X. (2015). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *Proceedings of ICASSP 2015*. doi: 10.1109/ICASSP.2015.7178826

Li, J., Ye, G., Das, A., Zhao, R., & Gong, Y. (2018). Advancing Acoustic-to-word CTC model. In *Proceedings of ICASSP 2018*. doi: 10.1109/ICASSP.2018.8462017

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., & Hinton, G. (2017). Regularizing neural networks by penalizing confident output distributions. In *Proceedings of ICLR 2017*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., …Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU 2011*.

Povey, D., Peddinti, V., Galvez, D., Ghahrmani, P., Manohar, V., Na, X., …Khudanpur, S. (2016). Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proceedings of Interspeech 2016*. doi: 10.21437/Interspeech.2016-595

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257 - 286. doi: 10.1109/5.18626

Sak, H., Senior, A. & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proceedings of INTERSPEECH-2014*, 338-342.

Sak, H., Vinyals, O., & Heigold, G. (2014). Sequence discriminative distributed training of long short-term memory recurrent neural networks. In *Proceedings of Interspeech 2014*, 1209-1213.

Soltau, H., Liao, H., & Sak, H. (2016). Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. In arXiv preprint arXiv: 1610.09975

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2018). Cold Fusion: Training Seq2Seq Models Together with Language Models. In *Proceedings of ICLR 2018*.

Veselý, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence discriminative training of deep neural networks. In *Proceedings of Interspeech 2013*, 2345-2349..

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., …Ochiai , T. (2018). ESPnet: End-to-End Speech Processing Toolkit. In *Proceedings of Interspeech 2018*, 2207-2211. doi: 10.21437/Interspeech.2018-1456

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayash, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1240-1253. doi: 10.1109/JSTSP.2017.2763455

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., …Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML 2015*, 2048-2057.