Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 35-46

結合鑑別式訓練聲學模型之類神經網路架構及優化方法的改進

Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method

趙偉成*、張修瑞*、羅天宏*、陳柏琳*

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

摘要

本論文探討聲學模型上的改進對於大詞彙連續中文語音辨識的影響。在基礎聲學模型的訓練上，有別於以往語音辨識通常使用交互熵(Cross Entropy)作為深度類神經網路目標函數，我們使用 Lattice-free Maximum Mutual Information (LF-MMI)做為序列式鑑別訓練的目標函數。LF-MMI 使得能夠藉由圖形處理器(Graphical Processing Unit, GPU)上快速地進行前向後向運算，並且找出所有可能路徑的後驗機率，省去傳統鑑別式訓練前需要提前生成詞圖(Word Lattices)的步驟。針對這樣的訓練方式，類神經網路的部分通常使用所謂的時間延遲類神經網路(Time-Delay Neural Network, TDNN)做為聲學模型可達到不錯的辨識效果。因此，本篇論文將基於 TDNN 模型加深類神經網路層數，並藉由半正交低秩矩陣分解使得深層類神經網路訓練過程更加穩定。另一方面，為了增加模型的一般化能力(Generalization Ability)，我們使用來回針法(Backstitch)的優化算法。在中文廣播新聞的辨識任務顯示，上述兩種改進方法的結合能讓 TDNN-LF-MMI 的模型在字錯誤率(Character Error Rate, CER)有相當顯著的降低。

關鍵詞：中文大詞彙連續語音辨識、聲學模型、鑑別式訓練、矩陣分解、來回針法

* 國立臺灣師範大學資訊工程研究所
Institute of Linguistics, National Taiwan Normal University
E-mail: {60647028S, 60647061S, teinhonglo, berlin}@ntnu.edu.tw

Abstract

This paper sets out to investigate the effect of acoustic modeling on Mandarin large vocabulary continuous speech recognition (LVCSR). In order to obtain more discriminative baseline acoustic models, we adopt the recently proposed lattice-free maximum mutual information (LF-MMI) criterion as the objective for sequential training of component neural networks in replace of the conventional cross entropy criterion. LF-MMI brings the benefit of efficient forward-backward statistics accumulation on top of the graphical processing unit (GPU) for all hypothesized word sequences without the need of an explicit word lattice generation process. Paired with LF-MMI, the component neural networks of acoustic models implemented with the so-called time-delay neural network (TDNN) often lead to impressive performance. In view of the above, we explore an integration of two novel extensions of acoustic modeling. One is to conduct semi-orthogonal low-rank matrix factorization on the TDNN-based acoustic models with deeper network layers to increase their robustness. The other is to integrate the backstitch mechanism into the update process of acoustic models for promoting the level of generalization. Extensive experiments carried out on a Mandarin broadcast news transcription task reveal that the integration of these two novel extensions of acoustic modeling can yield considerably improvements over the baseline LF-MMI in terms of character error rate (CER) reduction.

Keywords: Mandarin Large Vocabulary Continuous Speech Recognition, Acoustic Model, Discriminative Training, Matrix Factorization, Backstitch

1. 緒論 (INTRODUCTION)

近幾年來，語音辨識技術已有了長足的進步。其中，隨著深度學習技術以及電腦運算能力的突破性發展，聲學模型化技術已從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)，轉變成以使用交互熵(Cross Entropy)作為損失函數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) (Hinton *et al.*, 2012)。DNN-HMM 將以往用 GMM 計算的生成機率透過 DNN 的輸出層所代表的事後機率來近似，輸入特徵使用當前幀還有相鄰的幀，輸出則和 GMM-HMM 常用的 Triphone 共享狀態相同，以得到更低的詞錯誤率(Word Error Rate, WER)或字錯誤率(Character Error Rate, CER)。另一方面，進一步透過鑑別式訓練估測的聲學模型在語音辨識的表現上往往比僅以交互熵做為深度類神經網路損失函數的訓練方式來的好。但由於傳統上進行鑑別式訓練需要使用先進行交互熵訓練的聲學模型來產生詞圖(Word Lattices)，才能再進行下一步聲學模型鑑別式訓練(Bahl, Brown, de Souza & Mercer, 1986) (Veselý, Ghoshal, Burget & Poveyet, 2013)。近年來為了減少時間及空間複雜度，有學者對於 Maximum

Mutual Information (MMI)訓練，提出了所謂的 Lattice-free 的方式，使產生詞圖的步驟能夠在 GPU 上完成(Povey *et al.*, 2016)，因而讓鑑別式訓練得以做到端對端的訓練方式(Hadian, Sameti, Povey & Khudanpur, 2018)，因而大幅縮減了聲學模型訓練所需時間。

傳統 DNN-HMM 模型用於語音辨識的缺點在於無法充分利用語音信號之時間依賴性；而如同在(Graves, Mohamed & Hinton, 2013)所提到，基於遞迴類神經(Recurrent Neural Network, RNN)能對於序列性資料能夠有好的建模效果的想法所發展的 RNN-HMM 聲學模型其辨識效果卻是不如 DNN-HMM 模型來的好，因此以長短期記憶(Long Short-Term Memory, LSTM)取代簡單 RNN 所形成的聲學模型(LSTM-HMM) (Sak, Senior & Beaufays, 2014)，解決了 RNN-HMM 梯度消失的問題，在語音辨識上能夠達到比 DNN-HMM 好的效果。但在實務上，這樣的聲學模型很難像 DNN-HMM 一樣平行化訓練(Pascanu, Mikolov & Bengio, 2013)，以致於模型訓練時間的增加。另一方面，也由於其模型架構較為複雜使得運算量較大，較不適合需即時反應的語音辨識任務；相對來說時間延遲類神經網路(Time-Delay Neural Network, TDNN) (Waibel, Hanazawa, Hinton, Shikano & Lang, 1989)可以包含歷史和未來輸出、對長時間依賴性的語音訊號建模，使 TDNN-HMM 與傳統 DNN-HMM 訓練效率也相仿，因此在使用 LF-MMI 進行鑑別式訓練時，聲學模型的類神經網路部分通常是使用 TDNN。

從經驗上來看，類神經網路的深度對模型的性能非常重要(Ba & Rich, 2014)，增加層數之後能有更加複雜的特徵擷取能力。對於 TDNN 而言，增加層數可以說是提取更長時間的特徵；我們希望加深 TDNN 的網路層數來達到更好的結果，但以往的實驗發現深度的網路常有退化問題，類神經網路的深度之增加準確率反而會下降。因此本篇論文將比較並結合當前先進的聲學模型訓練方法，例如(Povey *et al.*, 2018)對網路的矩陣分解訓練可以使網路訓練更穩定，以期達到更佳的語音辨識表現。另一方面，梯度下降是執行優化的最流行的算法之一，也是迄今為止優化類神經網絡的最常用方法。而常見的優化算法有隨機梯度下降法(Stochastic Gradient Descent, SGD)、RMSprop、Adam、Adagrad、Adadelta (Ruder, 2016)等演算法；其中，SGD 算法在語音辨識任務上最被廣為使用。而本論文則採用來回針法(Backstitch) (Wang *et al.*, 2017)做為模型優化的演算法；它是一種基於 SGD 上的改進，希望能夠藉由兩步驟的更新 Minibatch，以達到更好的效果。

總合以上所述，我們認為加入對網路的矩陣分解來可順利訓練更深層的類神經網路模型；同時，使用 Backstitch 亦可提升模型泛化性，最終能使辨識結果更加進步。因此，本論文將分別比較使用 TDNN-LF-MMI，TDNN-LF-MMI 加入半正交低秩矩陣分解，TDNN-LF-MMI 加入半正交低秩矩陣分解及來回針法優化算法的辨識效果，最終在 TDNN-LF-MMI 加入半正交低秩矩陣分解及來回針法優化算法達到較佳的中文廣播新聞語音辨識的 CER 表現。

2. 聲學模型 (ACOUSTIC-MOLEL)

2.1 基本聲學模型-時間延類神經網路 (Time-Delay Neural Network, TDNN)

TDNN 在 1989 年被提出(Waibel *et al.*, 1989)，最初用於音素辨識；基於 TDNN 所產生的模型架構，能適用於處理語音所擁有特徵向量序列之時間長度不一致的特性。TDNN 對每一個隱藏層的輸出在時間上進行擴展，即每個隱藏層收到的輸入會有前一層在不同時刻的輸出。語音在考慮上下文長時間相關性很重要，TDNN 的優點在可以比傳統 DNN 看更長的時間，而且速度不會比 DNN 在訓練和辨識(解碼)時來的慢。

2.2 半正交低秩矩陣分解 (Semi-Orthogonal Low-Rank Matrix Factorization)

減少類神經網路參數的方法之一是透過 SVD 分解已經估測好的權重矩陣；近期有學者(Povey *et al.*, 2018)提出基於一個隨機的初始參數，用同樣的分解架構開始訓練類神經網路聲學模型，但是要讓其中一個分解的矩陣保持正交，避免有不穩定的問題。

在實作上，我們可以在進行若干次 SGD 後強迫參數矩陣變成半正交的更新，假設M是參數矩陣，定義$\mathrm{P} \equiv \mathrm{MM}^{\mathrm{T}}$目標是要讓 P 變成單位矩陣，學習率(Learning Rate)決定了權值更新速率的快慢，愈大的學習率會更快達到半正交的結果，但是設置太大會變得很不穩定，在接近半正交矩陣的時候 0.125 的設置是最好的，數學上可以達到平方收斂。令X是每次M更新的值，所以我們做一次更新$\mathrm{M} \leftarrow \mathrm{M} + \mathrm{X}$，我們希望$\mathrm{tr}(\mathrm{MX}^{\mathrm{T}}) = 0$以達到正交效果，下式為更新公式:

$$\mathbf{M} \leftarrow \mathbf{M} - \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\alpha}^{\mathbf{2}}}(\mathbf{MM}^{\mathbf{T}} - \boldsymbol{\alpha}^{\mathbf{2}}\mathbf{I})\mathbf{M} \tag{1}$$

α是一個縮放的參數，I是單位矩陣不考慮常數項，我們要使$\mathrm{tr}\left(\mathrm{MM}^{\mathrm{T}}(\mathrm{P} - \alpha^2\mathrm{I})\right) = 0$，因為$\mathrm{MM}^{\mathrm{T}} = \mathrm{P}$，所以$\mathrm{tr}(\mathrm{P}^2 - \alpha^2\mathrm{P}) = 0$移項之後$\alpha = \sqrt{\frac{tr(\mathrm{P}^2)}{tr(P)}}$，因為 P 是對稱矩陣所以$\mathrm{P}^2 = \mathrm{PP}^{\mathrm{T}}$，為了計算上較快會使用$\alpha = \sqrt{\frac{tr(\mathrm{PP}^{\mathrm{T}})}{tr(\mathrm{P})}}$。圖 1 是 TDNN+NF(Networks Factorized)內部架構，1536 維的隱藏層經矩陣分解後變成 1536*160*1536，SMAT 是要做正交限制的矩陣，後面再接上線性整流函數(ReLU)和批次標準化(Batch Normalization)。

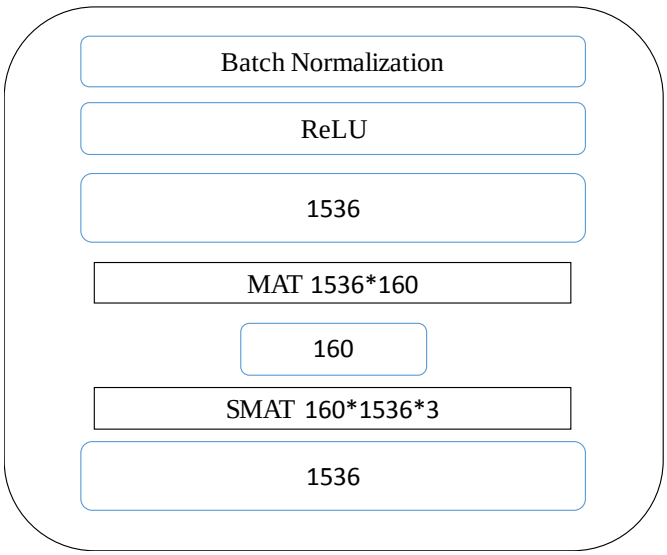

圖 1. 矩陣分解
[Figure 1. Matrix Factorization]

2.3 來回針法 (Backstitch)

Backstitch 是修改 SGD 增進對沒有看過資料的效能(Wang *et al*., 2017)；這個方法分成兩個步驟，先用一個較小的負學習率，再跟著一個較大的學習率，和對抗訓練的概念很像，可以消除有限數據集的偏見，用同樣的 Minibatch 但重新計算梯度，傳統的 SGD 單一的迭代如下式:

$$\boldsymbol{\theta_{n+1} \leftarrow \theta_n - \nu g(x_n, \theta_n)} \tag{2}$$

θ 是更新的參數，x_n 是第 n 個迭代的樣本 g(x_n, θ _n)是 f(x, θ)關於 θ 的導函數

$$\boldsymbol{\theta'_{n+1} \leftarrow \theta_n + \alpha \nu g(x_n, \theta_n)} \tag{3}$$

$$\boldsymbol{\theta_{n+1} \leftarrow \theta'_n - (1+\alpha) \nu g(x_n, \theta'_{n+1})} \tag{4}$$

其中，式(3)為 Backstitch 第一步驟退回更新，而式(4)為 Backstitch 第二步驟前進更新；α 這個常數決定要做多大步伐的更新，我們可以調整要幾個 Minibatch 做一次這種更新，根據原始論文比較有效率的設定是 α =1 和 n=4。

3. 模型架構和訓練方法 (METHODS)

3.1 聲學模型之類神經網路架構(Structure)

每層 TDNN 使用 1,536 維，加上兩層分解過後的矩陣 160 維，加上 ReLU 和 Batch

Normalization 合起來稱為 TDNNF 層，層數可比以前的 TDNN 網路(9 層以內)都還深。如圖 2 所示，最下層為隨著時間輸入之特徵，最上層為多任務的輸出，LF-MMI 的目標函數(Povey *et al.*, 2016)對應由決策樹定義的 Senone 函數，Cross Entropy Regularization 為輔助正規化訓練輸出，中間的 TDNNF 層有捷徑連結(Skip Connections) 。

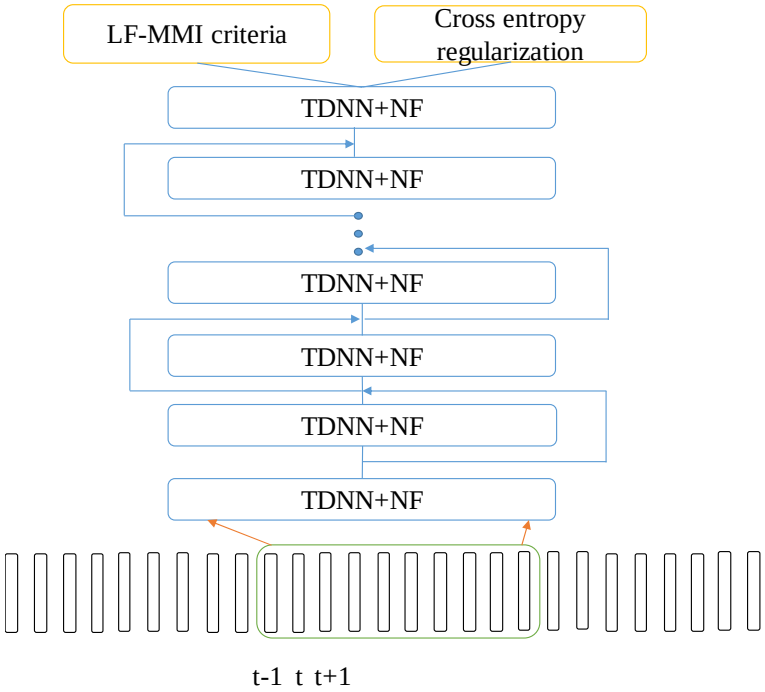

圖 2. 聲學模型中基於 TDNN 之類神經網路架構
[Figure 2. TDNN artitecture]

3.2 LF-MMI

近年來，透過可視為鑑別式模型(Discriminant Model)的類神經網路能夠有效提升系統效能。訓練開始之前先要有所有單詞序列組合，透過交叉熵訓練一個模型，配合語言模型生成一個詞圖，詞圖要有正確結果的路徑和足夠靠近的其他路徑，鑑別式模型用目標是要提高走正確路徑之機率，降低走相似路徑之機率。

MMI（Maximum Mutual Information）準則是要加大正確路徑和其他路徑的機率差。另一方面，LF-MMI 透過在類神經網路輸出計算所有可能的序列，根據這些序列計算 MMI 的梯度來訓練，因為 LF-MMI 在訓練中計算所有路徑的後驗機率(Posterior Probability)，所以不用事先生成訓練語句之詞圖。由於 TDNN 相鄰節點的變化通常不大，而且訊息重複機率很高，所以可以跳過一些幀的計算。LFMMI 的實驗設定在(Povey *et al.*, 2016)是降低幀率和將傳統 3-state 的 HMM 拓撲改為 2-state 的 HMM，端對端的鑑別式訓練也借鑒 CTC 上特殊的空白標籤(Graves, Fernández, Gomez & Schmidhuber, 2006)，在少量語料下

也達到很好的效果，在辨識率和解碼速度都有很大提升。

3.3 結合半正交限制和來回針法訓練 (Combine Training)

我們嘗試兩種訓練參數的方法用在聲學模型的效果，在 Natural Gradient (NG) (Povey, Zhang & Khudanpur, 2014)上面做改變，做 SGD 更新時每若干次在 backstitch 第一步驟後做半正交限制的更新，第二步驟維持原本的做法，在退回和前進步驟間做正交限制。

3.4 Dropout

在機器學習中，模型的參數太多會發生過度擬合現象，在每個批次訓練中忽略一些特徵檢測器，在 TDNN 中是橫跨時間的，(Povey *et al*., 2018)不是用二值的零一丟棄遮罩，而是用一個連續型均勻分布[1-2α, 1+2α]。我們使用一個丟棄排程表，在訓練一開始設定α=0.2，訓練到一半提升到 0.5，最後又下降到 0，這個設定在普通未分解的 TDNN 看起來沒有效果，但在分解後的網路架構中有明顯改善。

3.5 捷徑連結 (Skip Connections)

根據影像辨識裡 VGG (Simonyan & Zisserman, 2014)的發展，經驗上增加層數可以增加準確度，但是會增加訓練上的難度。ResNet (He, Zhang, Ren & Sun, 2016)在其上進行修改在網路上增加捷徑，可以防止類神經網路太深而無法訓練，我們在 TDNNF 裡也做上類似的機制，每層加上輸入更前面一層的三分之二和前一層相加當成新的輸入。

4. 實驗 (EXPERIMENTS)

4.1 實驗設定 (Experiment Setups)

表 1. 中文廣播新聞的實驗語料
[Table 1. MATBN]

	長度(小時)	句數
訓練集	114.7	38,556
發展集	3.7	2,001
測試集一	3.6	1,957
測試集二	1.4	307

本論文實驗語料來自公視新聞(Wang, Chen, Kuo & Cheng, 2005) (Mandarin Across Taiwan-Broadcast News, MATBN)。公視新聞語料是 2001 年至 2003 年間由中研院資訊所口語小組與公共電視台合作錄製，共計 197 個小時，取其中部分用於實驗，表 1 為實際用於實驗的語料長度和句數，包含內場新聞與外場新聞，其中內場新聞為新聞主播語料，外場包含採訪記者語受訪者語料。背景語言模型使用 5-gram 語言模型，訓練語料來自

2001 年至 2002 年中央通訊社(Central News Agency, CNA)的文字新聞語料，內含一億五千萬個中文字，經斷詞後約有八千萬個詞(本論文使詞典約七萬二千詞)。本論文是使用 SRI Language Modeling Toolkit(SRILM) (Stolcke, 2002) 來訓練語言模型。語言模型的訓練集由 2001 及 2002 年的新聞語料所篩選出來的。測試集一包含五場錄音在 2003/01/28，2003/01/29，2003/02/11，2003/03/07 和 2003/04/03，測試集二為只選擇了採訪記者語料並濾掉了含有語助詞之語句。另一方面，聲學模型都是在開源的語音辨識工具 Kaldi (Povey *et al*., 2011)上訓練，首先在語音語料庫上訓練具語者調適性(Speaker-adaptive)高斯混和隱馬可夫模型(GMM-HMM)，並利用該模型來獲得所有訓練語句的詞圖來準備後續聲學模型之類神經網路訓練。然後使用 TDNN-LF-MMI 準則來訓練出一類神經網路；其中，最佳化的部分則使用 NG 和 Backstich，遵循(Povey *et al*., 2016)中描述的方法來創建聲學模型，即對於 5,600 個依賴於上下文的語音中的每一個具有一個狀態的 HMM 拓撲，以原始幀速率的三分之一操作。有用變速擾動的資料擴充，特徵使用 40 維的 MFCC 和 3 維的聲調特徵加上 100 維的 i-vectors 做調適。

4.2 實驗結果 (Experiment Results)

表 2 比較了基本的 TDNN 和其他論文的 Attention 模型(Povey, Hadian, Ghahremani, Li & Khudanpur, 2018)包含 TDNN 和 3 層 LSTM 訓練而成，從實驗可以看出交互熵(CE)會遜於 LF-MMI 訓練結果，所以之後的改進模型皆使用 LF-MMI 訓練。基礎 TDNN 模型有 9 層隱藏層，每一個隱藏層有 625 維，前後文音窗各 15 幀，TDNN+NF 模型使用 1536 維的隱藏層，矩陣分解瓶頸 160 維，前後文音窗各 33 幀。TDNN+NF 在隱藏層維度變大時效果較好(Waibel *et al*., 1989)，而基本的 TDNN 沒有這種變化。

表 2. 基礎語音辨識實驗
[Table 2. Baseline experiment results]

	WER	CER	Parameters	RTF
Attention(LF-MMI)	26.76	18.96	50M	1.66
TDNN(LF-MMI)	**26.22**	**18.34**	**15M**	**0.42**
TDNN(CE)	27.84	19.17	**15M**	0.47

改進模型實驗結果如表 3，分別實做了 TDNN 和 TDNN+NF 及各自加上 ackstitch 的實驗，分解後的網路在參數上只多了一些，解碼速度相當但是字錯誤率降低很多，在不同深度的 TDNN+NF 比較實驗中 15 層表現最好。表 4 可見在不同難度測試集解碼後最終模型的字錯誤率都有顯著改善。

表 3. 改進模型在測試集一的實驗結果
[Table 3. Experiment results for test sets]

	WER	CER	Parameters	RTF
TDNN+Backstitch(9 層)	25.14	17.45	15M	0.42
TDNN+NF(15 層)	23.98	16.27	18M	0.47
TDNN+NF+Backstitch(10 層)	23.30	15.64	**13M**	**0.37**
TDNN+NF+Backstitch(15 層)	**22.56**	**15.15**	18M	0.47
TDNN+NF+Backstitch(20 層)	22.75	15.26	23M	0.58

表 4. 改進模型在測試集二與發展集的實驗結果
[Table 4. Experiment results for other test sets]

	測試集二 (CER)	發展集 (CER)
TDNN	5.05	33.39
TDNN+Backstitch	4.88	33.15
TDNN+NF	3.69	23.56
TDNN+NF+Backstitch	**3.67**	**22.73**

RTF(Real Time Factor)是一個常用於度量自動語音辨識系統解碼速度的值，如果處理一段長度為 a 的音訊信號需要花費時間 b，則 RTF 為 b/a，圖 3 是不同模型解碼速率比較，可以看出使用 LSTM 會提升大量時間，而其他模型隨著參數提昇會稍微增加時間。

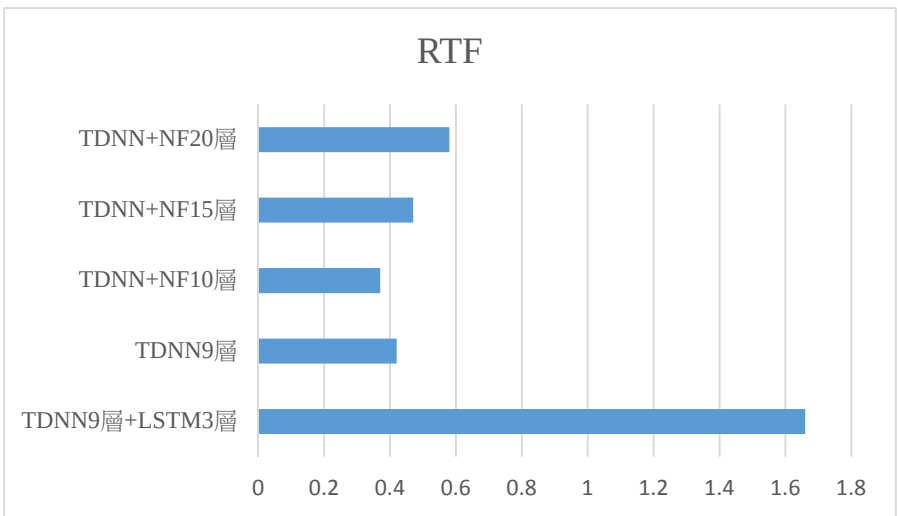

圖 3. 不同模型解碼速率比較
[Figure 3. Decoding speed comparison]

從訓練過程的準確率來看，有沒有使用 Backstitch 看不出太大差異，但是反應在字的錯誤率上有進步。有做矩陣分解的模型因為有正交限制的更新，在迭代 160 次後準確率會超越基本的 TDNN。關於 Backstitch 方面，圖 4 中虛線為訓練集，實線為驗證集，藍色 α =1 相較紅色 α =0.3 需要多兩倍的迭代才會收斂到相同準確率，可能是第二步驟沒有做正交化延遲了收斂的速度。

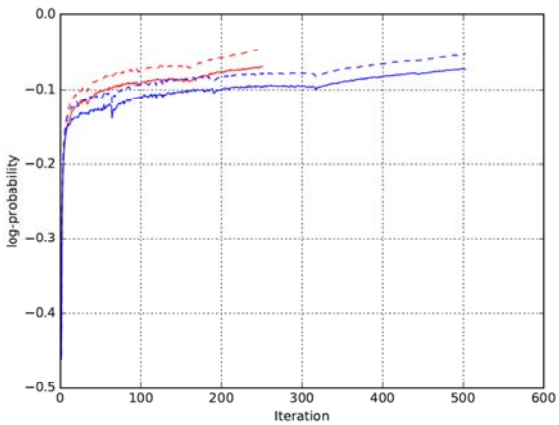

圖 4. 不同 Backstitch 的更新步伐之比較
[Figure 4. Different Backstitch steps comparison]

5. 結論 (CONCLUSION AND FUTURE WORK)

本論文探討聲學模型中的類神經網路權重參數更新並優化對於語音轉成文字錯誤率的影響。且應用了幾個其他改進，像是捷徑連接和隨著時間改變網路節點丟失的方法，從實驗結果發現，加上矩陣分解的時延遲網路用結合半正交限制和交叉針法的訓練效果最佳，在不同的測試集的字錯誤率上都有顯著的進步，訓練時間和解碼速度也不比基本的模型差，未來希望繼續探究不同結合方式在自動語音辨識的表現，並且更詳細與廣泛地探討各種聲學模型之優缺點。

參考文獻 (REFERENCES)

Ba, J., & Rich, C. (2014). Do deep nets really need to be deep? In *Proceedings of NIPS 2014*, *2*, 2654-2662.

Bahl, L., Brown, P., de Souza, P., & Mercer, R. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proceedings of ICASSP 1986*. doi: 10.1109/ICASSP.1986.1169179

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

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML '06*, 369-376. doi: 10.1145/1143844.1143891

Graves, A., Mohamed, A.-r., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of ICASSP 2013*. doi: 10.1109/ICASSP.2013.6638947

Hadian, H., Sameti, H., Povey, D., & Khudanpur, S. (2018). End-to-end speech recognition using lattice-free MMI. In *Proceedings of Interspeech 2018*. doi: 10.21437/Interspeech.2018-1423

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. (2016). In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2016)*. doi: 10.1109/CVPR.2016.90

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

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of ICML 2013*, *28*, 1310-1318.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohamadi, M., …Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proceedings of Interspeech 2018*. doi: 10.21437/Interspeech.2018-1417

Povey, D., Ghoshal, A., Boulianne, G., Goel, N., Hannemann, M., Qian, Y., …Stemmer, G. (2011). The Kaldi speech recognition toolkit. In *Proceedings of ASRU 2011*.

Povey, D., Hadian, H., Ghahremani, P., Li, K., & Khudanpur, S. (2018) A time-restricted self-attention layer for ASR. In *Proceedings of ICASSP 2018*. doi: 10.1109/ICASSP.2018.8462497

Povey, D., Peddinti, V., Galvez, D., Ghahremani, 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

Povey, D., Zhang, X., & Khudanpur, S. (2014). Parallel training of DNNs with natural gradient and parameter averaging. Retrieved from https://arxiv.org/abs/1410.7455

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

Ruder, S. (2016). An overview of gradient descent optimization algorithms. Retrieved from https://arxiv.org/abs/1609.04747

Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. Retrieved from arXiv:1402.1128

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Retrieved from https://arxiv.org/abs/1409.1556

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, 901-904.

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

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(3), 328-339. doi: 10.1109/29.21701

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese broadcast news corpus. *International journal of computational linguistics & Chinese language processing, Special Issue on Annotated Speech Corpora*, *10*(2), 219-236.

Wang, Y., Peddinti, V., Xu, H., Zhang, X., Povey, D., & Khudanpur, S. (2017). Backstitch: counteracting finite-sample bias via negative steps. In *Proceedings of Interspeech 2017*. doi: 10.21437/Interspeech.2017-1323