Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 1-18

使用長短期記憶類神經網路建構中文語音辨識器之研究

A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network

賴建宏*、王逸如+

Chien-hung Lai and Yih-Ru Wang

摘要

近年來類神經網路(Neural network)被廣泛運用於語音辨識領域中，本論文使用遞迴式類神經網路(Recurrent Neural Network)訓練聲學模型，並且建立中文大辭彙語音辨識系統。由於遞迴式類神經網路為循環式連接(Cyclic connections)，應用於時間序列訊號的模型化(Modeling)，較於傳統全連接(Full connection)的深層類神經網路而言更有益處。

然而一般單純遞迴式類神經網路在訓練上隨著時間的遞迴在反向傳播(Backpropagation)更新權重時有著梯度消失(Gradient vanishing)以及梯度爆炸(Gradient exploding)的問題，導致訓練被迫中止，以及無法有效的捕捉到長期的記憶關聯，因此長短期記憶(Long Short-Term Memory, LSTM)為被提出用來解決此問題之模型，本研究基於此模型架構結合了卷積神經網路(Convolutional Neural Network)及深層類神經網路(Deep Neural Network)建構出 CLDNN 模型。

訓練語料部分，本研究使用了 TCC300(24 小時)、AIShell(162 小時)、NER(111 小時)，並加入語言模型建立大辭彙語音辨識系統，為了檢測系統強健度(Robustness)，使用三種不同環境之測試語料，分別為 TCC300(2.4 小時，朗讀

* 國立交通大學電信工程研究所
Institute of Communications Engineering, National Chiao Tung University
E-mail: lsr950082@speech.cm.nctu.edu.tw
+ 國立交通大學電機工程學系
Department of Electronic Engineering, National Chiao Tung University
E-mail: yrwang@cc.nctu.edu.tw

語速)、NER-clean(1.9 小時，快語速，無雜訊)、NER-other(9 小時，快語速，有雜訊)。

關鍵詞：遞迴式類神經網路、長短期記憶、梯度消失(爆炸)、聲學模型、中文、大辭彙語音辨識、卷積類神經網路、深層類神經網路

Abstract

In recent years, neural networks have been widely used in the field of speech recognition. This paper uses the Recurrent Neural Network to train acoustic models and establish a Mandarin speech recognition system. Since the recursive neural networks are cyclic connections, the modeling of temporal signals is more beneficial than the full connected deep neural networks.

However, the recursive neural networks have the problem of gradient vanishing and gradient exploding in the backpropagation, which leads to the training being suspended. And the inability to effectively capture long-term memory associations, so Long Short-Term Memory (LSTM) is a model proposed to solve this problem. This study is based on this model architecture and combines convolutional neural networks and deep neural networks to construct the CLDNN models.

Keywords: RNNs, LSTMs, Gradient Vanishing (Exploding), Acoustic Model, Mandarin, LVCSR, CNNs, DNNs

1. 緒論 (Introduction)

近年來，人工智慧(Artificial intelligence, AI)儼然已成為隨處可聽見的關鍵詞，綜觀歷史，AI 浪潮共出現過三次，而每一次浪潮的興起，都和語音辨識技術的發展脫離不了關係。早期的語音辨識技術是由語言學學者透過研究聲學以及語言學之間的關聯，統整歸納出一套規則法(Ruled-based)的語音辨識系統；但是由於聲學和語言學二者間的變化，無法單單使用規則法完成描述，而後發展出像機器學習(Machine Learning)這樣透過資料驅動(Data-driven)的方法，讓機器從輸入帶有標籤(Label)的資料中，自動剖析並從中獲取規則，並對於未知的資料進行預測。

在近期的大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)系統中，聲學模型部分有別於傳統的高斯混合模型(Gaussian Mixture Model, GMM) (Reynolds, 2009)，使用了深層類神經網路(Deep Neural Network, DNN) (Zhang, Trmal, Povey & Khudanpur, 2014) (Mohamed, 2014) 取代之。而以 DNN 建構的聲學模型，在訓練的過程中，必須使用大量的語料，對於不同的發聲才能有較佳的辨識結果。由於語音為時序相關訊號，因此本研究加入了遞迴式類神經網路(Recurrent neural network)訓練聲學模型，並探討其辨識結果

在語音辨識系統中，語言模型(Language model)亦扮演相當重要的角色，本研究基於

本實驗室擁有的文字語料，選擇八萬詞、十萬詞、十二萬詞的詞典(Lexicon)分別建構出三種 Tri-gram 語言模型，並且對於 TCC300、NER-clean、NER-other 三種不同環境的測試語料進行分析及探討，其中 TCC300 屬於朗讀語速且無雜訊之一般語料，NER-clean 為快語速且無雜訊之自發性語料、NER-other 則為快語速且有背景雜訊之自發性語料。

2. 實驗流程與實驗環境介紹 (Experimental Process and Experimental Environment)

由於訓練深層類神經網路非常耗時，本研究使用繪圖處理器(Graphics Processing Unit, GPU)來訓練含有 DNN、CNN (Abdel-Hamid *et al*., 2014) (Abdel-Hamid, Mohamed, Jiang & Penn, 2012) 及 LSTM (Sak, Senior & Beaufays, 2014a) (Sak, Senior & Beaufays, 2014b) 之聲學模型，並使用 Kaldi speech recognition toolkit (Povey *et al*., 2011) 中 nnet3 所提供的深層類神經網路訓練流程，進行聲學模型訓練。表 1 為實驗所使用之硬體規格；表 2 則為 GPU 規格表。另外為了使矩陣運算效能加速，本研究使用的 Kaldi 經由 Intel 開發之數學運算核心函式庫(Math Kernel Library, MKL)進行編譯。

表 1. 硬體規格描述
[Table 1. Hardware specification:]

CPU	Intel® Core™ i7-8700K @ 3.70GHz
RAM	64 GB DDR4-3000
HDD	4 TB SATA-III 7200RPM
GPU	NVIDIA GeForce GTX 1080TI
OS	Arch Linux 4.17.5-1 64bit

表 2. GPU 規格描述
[Table 2. GPU specification]

型號	NVIDIA GeForce GTX 1080TI
CUDA 核心數	3584
基礎時脈	1480 MHz
加速時脈	1582 MHz
記憶體時脈	11 Gbps
記憶體容量配置	11264 MB
記憶體介面型號	GDDR5X
記憶體介面頻寬	484 GB/s

3. 語料庫介紹 (Databases)

本節將分別介紹用於本實驗中之所有語料庫，其中用來當作訓練語料的有 TCC300、NER 及 AIShell 語料庫，而為了測試本實驗之辨識系統對於不同環境的辨識能力，因此在測試語料的選擇上，使用 TCC300 及 NER 語料庫，其中 NER 為廣播語料，又可細分為背景乾淨無雜訊之 NER-clean 以及背景有人為雜訊或音樂參雜其中之 NER-other。

3.1 TCC300語料庫 (TCC300 Corpus)

實驗中所使用的 TCC300 麥克風語音資料庫[1]是由國立交通大學(National Chiao Tung University, NCTU)、國立成功大學(National Cheng Kung University, NCKU)、國立台灣大學(National Taiwan University, NTU)共同錄製而成，並且由中華民國計算語言學學會(The Association for Computational Linguistics and Chinese Language Processing, ACLCLP)發行，此語料庫屬於麥克風朗讀語音，主要目的為提供台灣腔之中文語音辨認研究使用。

詳細資訊如表 3 所示，台灣大學部分主要包含詞以及短句，文本經過設計，考慮音節與其相連出現之機率，共由 100 人錄製而成；交通大學與成功大學部分則為長文語料，其語句內容透過中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百字，再切割分成 3 至 4 段，每段至多包含 231 字，兩校分別各錄製 100 人而成，且每人朗讀的文章皆不相同。每個學校之錄音取樣頻率皆為 16000 Hz，取樣位元數為 16 位元。

本實驗進一步將整個 TCC300 語料庫分為訓練語料與測試語料，訓練與測試比例約為 9 : 1，分別資訊如下：

- 訓練語料：約為 24.4 小時，共 284 位語者，8633 句發音，304780 個音節數。
- 測試語料：約為 2.4 小時，共 19 位語者，225 句長句發音，26357 個音節數。

表 3. TCC300 語料庫資訊
[Table 3. TCC300 corpus information]

學校名稱	文章屬性	語者總數		音節總數		檔案總數	
台灣大學	短句	男	50	男	27541	男	3425
		女	50	女	24677	女	3084
		總數	100	總數	52218	總數	6590
交通大學	長文	男	50	男	75059	男	622
		女	50	女	73555	女	616
		總數	100	總數	148614	總數	1238
成功大學	長文	男	50	男	63127	男	588
		女	50	女	68749	女	582
		總數	100	總數	131876	總數	1170

[1] Mandarin Microphone Speech Corpus-TCC300. http://www.aclclp.org.tw/use_mat_c.php#tcc300edu

3.2 NER語料庫 (NER Corpus)

NER 語料庫，全名為 NER Manual Transcription Vol1，為國立臺北科技大學和國家教育廣播電台合作錄製之語料庫，主要目的為大量轉寫教育電台之節目，產生節目逐字稿，以建置大規模台灣腔之語料庫，詳細內容如表 4 所示，內容大部份為談話性節目，多為自發性(Spontaneous)語音，僅少部分為新聞報導之朗讀式(Reading)語音。

此語料庫依照 1. 為錄音室內或為錄音室以外之場所錄製，2. 有無任何背景襯樂或非人聲之噪音兩項條件分為兩個部分：乾淨語料(Clean，約 19.4 小時，共 5106 個檔案)及其他語料(Other，約 107.4 小時，共 15983 個檔案)合計共約 126.8 小時，21089 個檔案，取樣頻率為 16000 Hz，取樣位元數為 16 位元，聲道數為 1(mono)。

語料庫中逐字稿來源由國立臺北科技大學之雙語語音辨識器進行初步轉寫逐字稿，後經由人工校正以及切割，並移除有版權疑慮之音樂段落後產生。

本實驗亦進一步將此語料庫分為訓練語料及測試語料，詳細資訊如下：

- 訓練語料：約為 111.5 小時，共 18710 句發音，1715091 個音節數。
- 測試語料：
  - Clean：約為 1.9 小時，共 549 句發音，33660 個音節數。
  - Other：約為 9.0 小時，共 1322 句發音，133746 個音節數。

表 4. NER 語料庫資訊
[Table 4. NER corpus information]

環境類型	節目名稱	代碼	總時數	音節總數	檔案總數
Clean	創設市集	CS	14.4	235052	4028
	技職最前線	JZ	1.8	34352	438
	國際教育心動線	GJ	3.2	55057	640
Other	多愛自己一點點	DA	13.6	212821	2347
	科學 SoEasy	KX	1.8	23415	208
	青年故事館	QG	17.3	260116	3202
	不太乖學堂	BG	9.5	143138	1586
	星期講座	WK	8.4	113202	1102
	遇見幸福幼兒園	YX	5.6	90419	826
	收藏人生	SR	16.5	280074	2670
	雙語新聞	SY	34.5	434851	4015

3.3 AIShell語料庫 (AIShell Corpus)

AIShell 語料庫(Bu, Du, Na, Wu & Zheng, 2017)，是由北京希爾貝殼科技有限公司釋放之開源語音資料庫，錄製內容如表 5，涉及智能居家、無人駕駛等 11 項領域，錄製過程皆在安靜的室內環境。

使用高效能麥克風錄製而成，取樣頻率為 44100 Hz，後降低取樣頻率至 16000 Hz，取樣位元數為 16 位元，由 400 名來自中國不同口音地區的參與者錄製而成，語者資訊如表 6、表 7 所示，此語料庫文本經人工校正過，正確率為 95%以上。

本實驗進一步將此語料庫分為訓練語料及測試語料：

- 訓練語料：約為 162.4 小時，共 129341 句發音，1862171 個音節數。
- 測試語料：約為 16.6 小時，共 12259 句發音，178041 個音節數。

表 5. AIShell 語料庫文本內容
[Table 5. AIShell corpus text contents]

主題	語句數
智能居家	5
地理訊息	30
音樂播放指令	46
數字串	29
電視與電影播放指令	10
金融	132
科學與科技	85
體育	66
娛樂	27
新聞	66
英文拼寫	4

表 6. AIShell 語料庫語者資訊
[Table 6. AIShell corpus speaker information]

年齡範圍	語者數	地區	語者數
16 - 25	316	北方	333
26 – 40	71	南方	56
> 40	13	其他	11
合計	400	合計	400

表 7. AIShell 語料庫資訊
[Table 7. AIShell corpus information]

語者總數		音節總數		檔案總數	
男	186	男	939132	男	65205
女	214	女	1101080	女	76395
合計	400	合計	2040212	合計	141600

4. 深層類神經網路模型配置 (Deep Neural Network Model Configuration)

CLDNN 為近年來被提出(Sainath, Vinyals, Senior & Sak, 2015)適合用來建立聲學模型的一種架構，其名稱來源為卷積類神經網路(CNN)加上長短期記憶(LSTM)後再接上深層類神經網路(DNN)，普遍認為，CNN 能夠學習特徵參數在頻域上的變化程度，LSTM 則擅長時域上的模型建立，最後 DNN 適合將特徵映射至更可分離的空間上。

此主要模型亦使用 TCC300 作為訓練語料，特徵參數的抽取也是 40 維之 Fbank，本研究使用的 LSTM 帶有映射層及窺視孔，詳細架構參見圖 1，虛線連結部分即為窺視孔作用之途徑，目的在於讓閘門做決定時能同時考慮短期記憶與長期記憶，而映射層之目的在於降低 LSTM 輸出或遞迴的神經元數量，降低模型總參數量，幫助網路訓練更為快速，和 CNN 後連接的降維全連接層有異曲同工之妙。

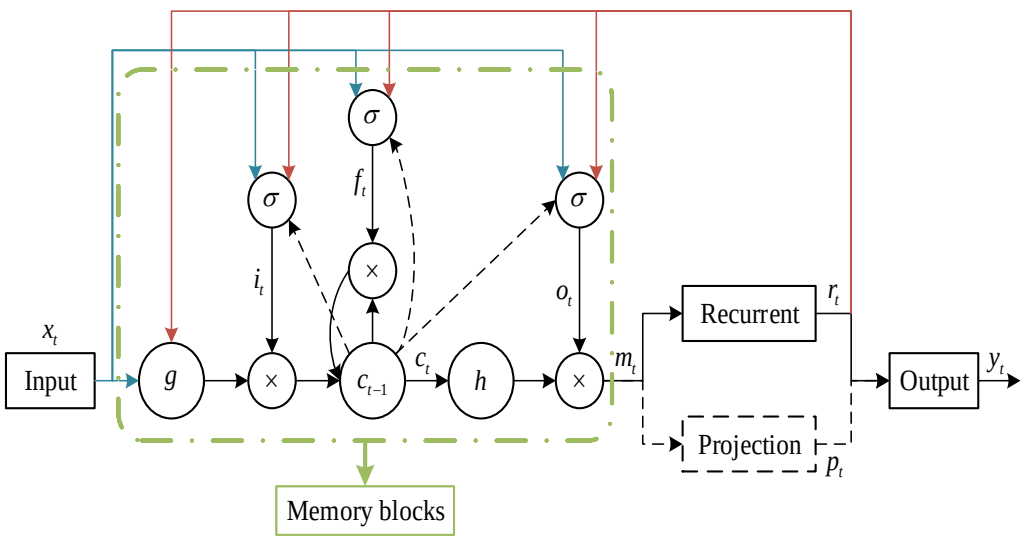

圖 1. 長短期記憶內部結構圖
[Figure 1. Long Short-Term Memory internal structure]

每一層 LSTM 中心細胞數目皆為 512，映射(Projection)層及遞迴(Recurrent)層數目皆為 256，而在訓練過程中，為了避免梯度產生爆炸，在遞迴的過程中會設定限幅閥值

(Clipping-threshold)為 30，即當梯度大於此閥值時，將梯度設定為 30，如此一來便解決 梯度在反向傳播的時候數值過大的問題。

另外，在 DNN 部分，為了解決層數過多導致學習困難的問題，有研究提出批次正規化(Batch normalization, BN)方法(Ioffe, & Szegedy, 2015)，將每一層 DNN 之輸出依照小型批次數(mini-batch)進行正規化，如此一來就可以大幅增加訓練學習率 讓模型訓練加速以及避免層數過深而造成的過度擬合(Over-fitting)的問題。

5. 語言模型之建立 (Language Model Establishment)

本研究目的於建立一中文大詞彙辨識系統，因此需要建立語言模型，並加入至系統中，辨識出中文詞彙序列。如圖 2 所示，建立流程為：將文字語料經國立交通大學語音處理實驗室王逸如老師撰寫之繁體中文斷詞器進行斷詞，後將文字進行正規化、移除冗餘贅字、取代同義異字詞(Variant Word, VW)等前處理，接著依照詞頻(Term Frequency, TF)及檔案頻率(Document Frequency, DF)進行選詞，一般來說，語音辨識系統之語言模型需要 TF 高及 DF 亦高之詞彙，本研究選擇了八萬詞、十萬詞及十二萬詞分別建立三個 3-gram 語言模型，而最後須將前處理置換的同義異字詞置換回來，詳細內容在五之(二)章節解說，最後以有限狀態轉換機表示此語言模型。

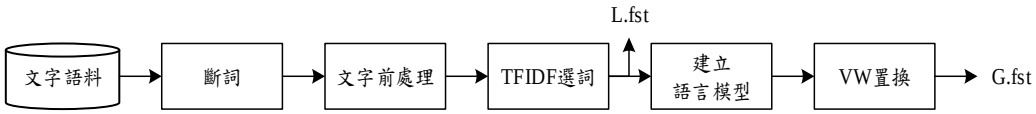

圖 2. 中文語言模型建立流程圖
[Figure 2. Chinese language model establishment flow chart]

5.1 文字語料庫簡介 (Introduction to the Text Corpus)

本研究用於訓練與研模型之文字語料庫共約 4.4 億個詞彙，包含以下：

- 光華雜誌(Sinorama)：內容為一般雜誌之文章，資料年份介於 1976 至 2000 年。
- NTCIR：為一個建立資訊檢索系統的標竿測試集，其內容由數種不同學科領域文章構成。
- 中研院平衡語料庫(Sinica)：由中研院收集，內容包含多種主題，以語言分析研究為目的之資料庫。
- Chinese Gigaword：由 Linguistic Data Consortium (LDC)整合發行，內容包括台灣中央社、北京新華社等國際新聞。
- 中文維基百科語料(Wiki)：中文維基百科內容廣泛，且資訊較新，能使語言模型更為多元，且增加資料庫。
- TCC300：包含詞、短句、長句，內容由中研院 500 萬詞標示語料庫中選取。

5.2 形音義分合詞前處理 (Preprocess of Variant Words)

漢字具有三大要素：「形、音、義」，其中字義為語文之核心，字形、字音皆因字義而存在。而在文字的演進當中，有些字形變的不一致，或者因為沒有創立而借用，甚至可能是錯用，各種複雜的因素導致漢字形成了「多形、歧音、異義」的狀況，因此目前漢字在使用上呈現字形不一、字音分歧、且字義寬廣的特性。

中文語音辨識因形音義之不同，有些詞類是可以合併的，如表 8 所示，大致上可以分為三類，即同形異音、異形同音及異形異音，置換原則是建立在字義相同之上，目的是為了讓文章中同義詞正規化，以利選詞時容納更多詞彙，但是在語言模型建立後，辨識端會產生一個狀況：異音類的同義字無法被搜尋到，如範例中的「禮拜一」被置換成「週一」，因此在語言模型中無法找到「禮拜一」這個詞彙，因此我們在語言模型建置的最後一步，需要處理不同發音之同義異字詞的置換，將「週一」展開成「週一」、「星期一」及「禮拜一」。本研究使用之同義異字詞表(variant word table)為 4261 詞。

表 8. 形音義分合詞範例
[Table 8. Example of variant words]

形音義分合詞類型	置換前文字	置換後文字
同形異音	爸	爸爸
	媽	媽媽
異形同音	手表	手錶
	瓩	千瓦
異形異音	禮拜一	週一
	星期一	週一

6. 實驗結果分析與討論 (Analysis and Discussion of Experimental Results)

本章節將進行實驗結果的分析與探討，其中包含使用無文法(Free-grammar)之語言模型測試音節錯誤率(Syllable Error Rate, SER)，以及加入不同詞典大小之語言模型測試詞錯誤率(Word Error Rate, WER) 與其即時係數(Real-Time Factor, RTF)。

即時係數如式(1)所示，表示平均一個音框需要解碼(decode)之時間，又因為本研究設定音框之間隔為 10ms，因此可以解釋為每秒辨識系統所需之解碼時間，若建立之系統為即時系統(Real-time System)，則 RTF 須小於 1.0；辨識錯誤的分析則可以分為以下三種錯誤：取代型錯誤(Substitution)、插入型錯誤(Insertion)及刪除型錯誤(Deletion)；而辨識錯誤率計算方式如式(2)所示。

解碼過程我們使用維特比演算法(Viterbi algorithm)，透過神經網路輸出狀態序列，搜尋計算找出最佳路徑，但是一般維特比算法過於耗時，因此加入光束搜尋演算法(Beam searching algorithm)，設定最大存活狀態數(Max-active states)及光束值(Beam)，找出當下

音框所有可能路徑(Hypotheses)，並刪除分數之光束臨界值，即當下路徑與最高分差大於光束值，則刪除該路徑，最後將狀態數控制於最大存活狀態數下，如此雖然會犧牲些許辨識率，但能大幅提升辨識速度，本研究設定最大存活狀態數為 7000、光束值為 15.0。

$$RTF = \frac{Seconds}{Frames} \times 100 \tag{1}$$

$$ER = \frac{S + I + D}{N} \times 100\% \tag{2}$$

6.1 各式類神經網路聲學模型辨識結果 (Various Types of Neural Network Acoustic Model Recognition Results)

為了探討遞迴式類神經網路對於聲學模型之影響，本實驗設計四組模型，使用的訓練語料皆為 TCC300，CDNN 為一般卷積類神經網路(CNN)結合深層類神經網路(DNN)，輸入之間彼此獨立，並沒有記憶特性；LDNN 為長短期記憶(LSTM)結合深層類神經網路，多了時間軸之資訊，過去的隱藏層狀態被保留，且透過閘門篩選控制，避免發生過擬現象；而 CLDNN 則結合以上三種類神經網路，詳細架構如圖 3 所示。

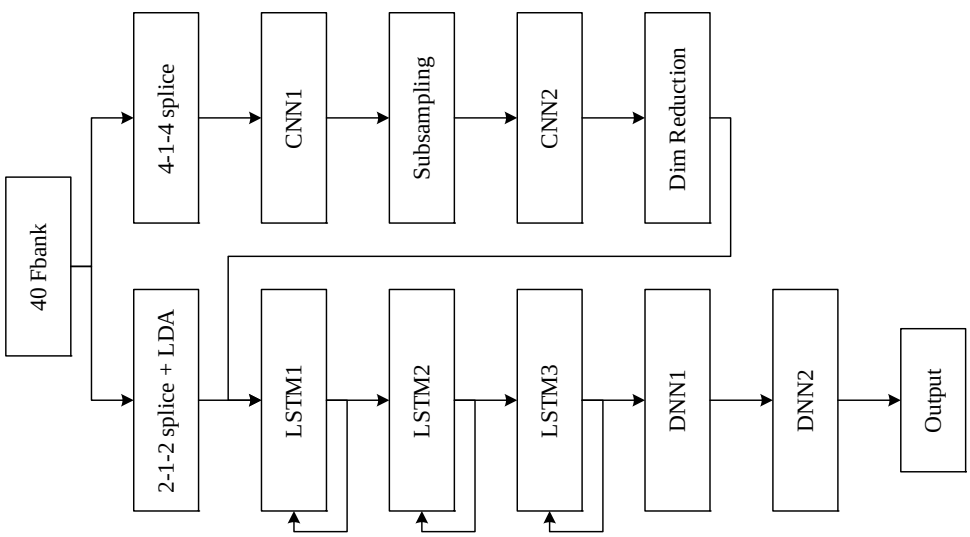

圖 3. CLDNN 模型架構圖
[Figure 3. CLDNN model architecture diagram]

測試語料部分亦選擇使用 TCC300，音節數為 26357，辨識結果如表 9 所示，卷積類神經網路對比傳統類神經網路而言，對於特徵的學習是有幫助的，相對改善率約 5%，而長短期記憶模型在聲學模型的建立上則是非常有益處的，相對改善率高達約 25%，但是使用遞迴式類神經網路在解碼時相對耗時，對照 CDNN 及 CLDNN 之 RTF，將近多出兩倍的時間。

表 9. 各式類神經網路模型之音節錯誤率
[Table 9. Syllable error rate of various neural network models]

Model	SER (%)	RTF
DNN	21.17	0.04
CDNN	19.52	0.05
LDNN	15.72	0.10
CLDNN	15.23	0.15

另外，本實驗亦使用鏈式模型(Chain model) (Povey *et al*., 2016)建構聲學模型，一般聲學模型的訓練使用的是最大化相似度(Maximum Likelihood, ML)，用於最大化模型及其特徵參數之相似度；鏈式模型則是使用最大交互資訊法則(Maximum Mutual Information, MMI)進行訓練如式(3)，其中 $P(W)$ 表示給定逐字文本(Transcription)中序列 W 之語言模型機率，而交互資訊可以拆成兩項相減，M^{num} 表示參考文本序列，M^{den} 表示所有可能之文本序列，最大化 F_{MMI} 表示讓參考文本的路徑機率 $P(O_r|W_r)$ 在所有路徑中最為突出，但是一般語言模型皆建立在詞彙(word)上，這會導致訓練過程效率不彰，因此鏈式模型在訓練上，會先以音素(phone)為單位，建立一個 4-gram 之語言模型，作為訓練時參考用。另外參考圖 4 及圖 5，鏈式模型使用降低 3 倍之音框速率 (Sak, Senior, Rao & Beaufays, 2015)，即一次觀察 30ms 之音框，以及更為簡單的 HMM 拓樸圖，一個音素(phone)僅用一個 HMM 描述，因此鏈式模型在解碼時比一般類神經網路模型加速三倍左右，實驗結果如表 10 所示，Chain-CLDNN 在音節錯誤率以及 RTF 都表現較 CLDNN 模型佳。

$$F_{MMI} = \sum_{r=1}^{R} \log \frac{P(O_r|W_r)P(W_r)}{\sum_W P(O_r|W)P(W)} = \log P(O|M^{num}) - \log P(O|M^{den}) \tag{3}$$

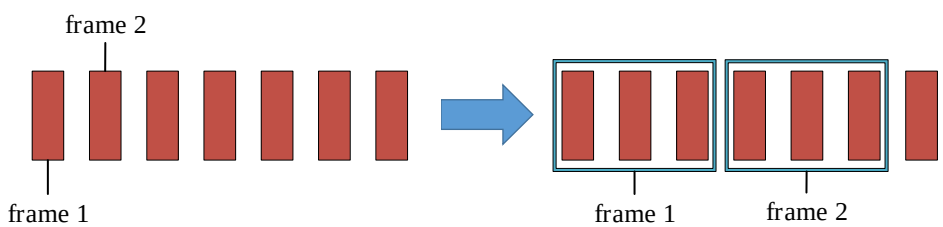

圖 4. 鏈式模型音框速率示意圖
[Figure 4. Chain model frame rate diagram]

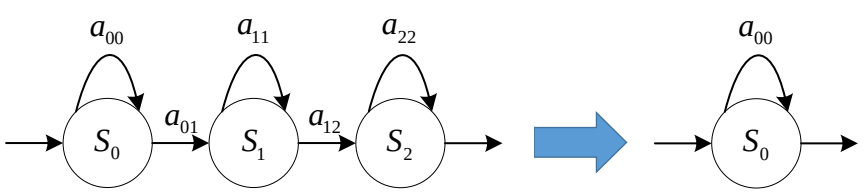

圖 5. 鏈式模型使用之 HMM 拓樸
[Figure 5. HMM topology used by chain models]

表 10. Chain/Non-chain 模型音節錯誤率
[Table 10. Chain/Non-chain model syllable error rate]

Model	SER (%)	RTF
CLDNN	15.23	0.15
Chain-CLDNN	13.66	0.05

6.2 加大訓練語料對辨識率的影響 (Impact of Increasing Training Corpus on Recognition Rates)

在深度學習(Deep learning)或是機器學習中，常常會遭遇模型過度擬合(Over-fitting)之問題，若能順利解決，則能使得模型訓練得更為深層，方法除了本研究所使用的批次正規化之外，另一個方法就是直接增加訓練語料，然而在資料有限的情形下，可以透過資料轉換技術來增加訓練資料，此概念在影像處理(Image Processing)領域已被實現(Krizhevsky, Sutskever & Hinton, 2012)，圖片可以透過旋轉(Rotation)、翻轉(Flip)、縮放(Zoom)、平移(Shift)、尺度轉換(Rescale)等方法產生新的圖片。

語音辨識方面，亦能使用類似的方法，比如改變音檔之音高(Pitch)、節奏(Tempo)、語速(Speed)等產生出假造之資料，擴充語料庫，本研究除了使用 TCC300、NER 及 AIShell 語料庫，亦利用上述方法產生語速 1.1 及 0.9 之擾動語料(Speed perturbation data)，並加入訓練語料。

實驗結果如表 11 所示，首先針對一般 CLDNN 模型，加入 AIShell 語料庫，訓練語料由原本的 24 小時增加到 186.4 小時，雖然 AIShell 語料庫來自中國各地口音，但是對於音節辨識率之相對改善率仍高達約 15.5%，若以語者個別分析，如圖 6 所示，則可以發現到，主要降低音節錯誤率之貢獻來自原本音節錯誤率高之語者，對於錯誤率低之語者無太多改善，換句話說，辨識系統更具強健性(Robustness)。

接著加入屬於台灣腔調之 NER 自發性語料，訓練語料增加至 297.9 小時，並使用上述之擾動語速方法，增加至 900.7 小時，逐一訓練出 CLDNN 鏈式模型，實驗結果如表 12、圖 7 所示。

表 11. 使用不同訓練語料之 CLDNN 模型比較
[Table 11. Comparison of CLDNN models using different training corpora]

Model	Training data	SER (%)
CLDNN	TCC300	15.23
	TCC300+AIShell	12.87

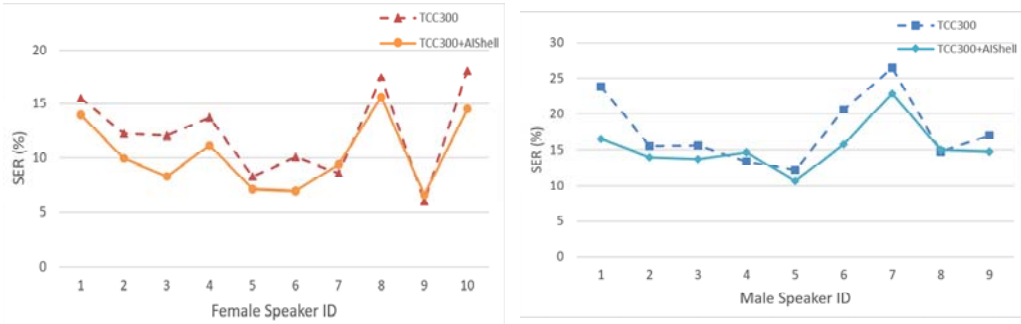

圖 6. TCC300 女性/男性測試語者之音節辨識率
[Figure 6. Syllable error rate of TCC300 female/male testers]

表 12. 使用不同訓練語料之 Chain-CLDNN 模型比較
[Table 12. Comparison of Chain-CLDNN models using different training corpora]

Model	Training data	SER (%)
Chain-CLDNN	TCC300	13.66
	TCC300+AIShell	11.97
	TCC300+AIShell_sp	11.49
	TCC300+AIShell+NER_sp	8.92

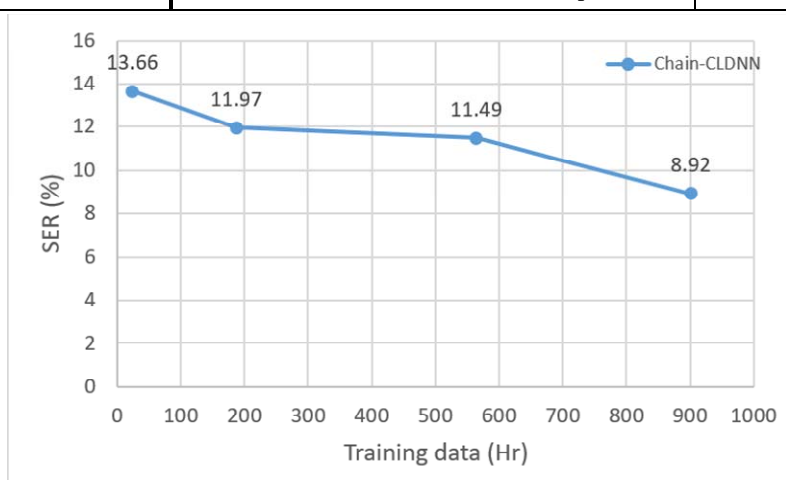

圖 7. 訓練語料量對音節辨識率之影響
[Figure 7. The effect of the amount of training corpus on the syllable recognition rate]

6.3 加入語言模型並探討不同環境對辨識率之影響 (Add Language Models and Explore the Impact of Different Environments on Recognition Rates)

本節我們選擇使用透過 900.7 個小時訓練語料建立之 Chain-CLDNN 模型作為聲學模型，測試語料部分選擇 1. 朗讀語速之 TCC300、2. 自發性語音且背景無噪音之 NER-clean、3. 自發性語音且具雜訊之 NER-other，音節辨識率如表 13，後加入三組 Tri-gram 語言模型，分別將詞典大小設定為：八萬詞、十萬詞及十二萬詞，分別測試其最佳辨識結果(Oracle)，即當聲學模型輸出之音素序列皆完全正確情況下，語言模型辨識出之詞錯誤率，以及 EDO(Error Due to OOVs)，即 OOV 造成之錯誤率，平均一個 OOV 影響 2.103 個詞，最後結合聲學模型解碼計算出詞錯誤率(WER)，如表 14 至表 16 所示。RTF 部分對於三個測試語料分別為 0.27、0.48 及 0.59。然而本實驗室之文字語料庫大多取自新聞文章，domain 相對偏向 TCC300 測試集，而 NER 測試集則多為談話性節目，因此本實驗利用 NER 之訓練語料逐字稿進行語言模型之調適，如式(4)所示，實驗結果如表 17 所示，WER 獲得大幅度的改善。

$$LM_{adapt} = 0.3LM_{ori} + 0.7LM_{ner} \tag{4}$$

表 13. Chain-CLDNN 模型對於各測試集之音節錯誤率
[Table 13. The syllable error rate of the Chain-CLDNN model for each test set]

Model	Test data	SER (%)
Chain-CLDNN [TCCAINER-sp]	TCC300	8.92
	NER-clean	16.89
	NER-other	22.14

表 14. 八萬詞語言模型辨識結果
[Table 14. 80K word LM recognition results]

Model	Test data	80K-LM		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	7.73	6.74	5.85
	NER-clean	24.95	9.39	2.48
	NER-other	31.92	11.92	3.91

表 15. 十萬詞語言模型辨識結果
[Table 15. 100K word LM recognition results]

Model	Test data	100K-LM		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	7.12	6.06	5.19
	NER-clean	24.80	9.27	2.25
	NER-other	31.69	11.57	3.26

表 16. 十二萬詞語言模型辨識結果
[Table 16. 120K word LM recognition results]

Model	Test data	120K-LM		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	6.56	5.34	4.52
	NER-clean	24.72	9.05	2.02
	NER-other	31.61	11.42	2.92

表 17. 十二萬詞調適語言模型辨識結果
[Table 17. 120K word adaptation LM recognition results]

Model	Test data	120K-LM-Adapt		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	7.79	5.87	4.52
	NER-clean	15.12	4.00	2.02
	NER-other	21.66	4.74	2.92

7. 結論與未來展望 (Conclusion and Future Prospects)

本論文使用 Kaldi speech recognition toolkit 來實現結合卷積類神經網路、長短期記憶及深層類神經網路的聲學模型(CLDNN)，經過各式類神經網路模型的比較後，確定長短期記憶對於聲學模型的建構上助益良多，也確定卷積類神經網路對於特徵的學習對整體模型有幫助，且加入大量不同來源之訓練語料(NER、AIShell)，並使用資料增強轉換技術，能使模型之強健度提升，最後再以實驗室 4.4 億詞彙量文本訓練 Tri-gram 語言模型，以建構中文大詞彙語音辨識系統，從實驗結果顯示，系統之詞辨識率與語言模型有非常密切的關聯，也就是說，測試語料之領域依存性(domain dependence)相當高。

本實驗建構之中文辨識系統雖然在朗讀語速及自發性且無噪音環境下的辨識率(6.56%、15.12%)有不錯的表現，但是在自發性且環境雜訊高的環境下，詞錯誤率仍高達 21.66%，因此在聲學模型方面如何抗噪，亦是一個研究課題，另外許多研究加入 I-vector

作為特徵參數進行聲學模型訓練 (Madikeri, Dey, Motlicek & Ferras, 2016)，目的為了學習語者特性，增加模型強健性，訓練語料不足的方面，可以使用半監督式學習 (semi-supervised learning) (Manohar, Hadian, Povey & Khudanpur, 2018)，蒐集無轉寫文本之語料，透過辨識結果之信心分數決策是否加入為訓練語料；至於語言模型部分，解決人名造成 OOV 之問題，且將文本進行分類，以建構出不同 domain 之語言模型，以及快速進行調適語言模型之建立與轉換。

參考文獻 (References)

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533-1545. doi: 10.1109/TASLP.2014.2339736

Abdel-Hamid, O., Mohamed, A., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *Proceedings of ICASSP 2012*, 4277-4280. doi: 10.1109/ICASSP.2012.6288864

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *Proceedings of 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. doi: 10.1109/ICSDA.2017.8384449

Ioffe, S., & Szegedy. C. (2015). Batch normalization:Accelerating deep network training by reducing internal covariate shift. In *Proceedings of ICML' 15*, *37*, 448-456.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, *1*, 1097-1105.

Madikeri, S., Dey, S., Motlicek, P., & Ferras, M. (2016). *Implementation of the standard i-vector system for the kaldi speech recognition toolkit* (Idiap- RR Idiap-RR-26-2016). Retrieved from IDIAP Research Institute website: http://publications.idiap.ch/downloads/reports/2016/Madikeri_Idiap-RR-26-2016.pdf

Manohar, V., Hadian, H., Povey, D., & Khudanpur, S. (2018). Semi-supervised training of acoustic models using lattice-free MMI. In *Proceedings of ICASSP 2018*. doi: 10.1109/ICASSP.2018.8462331

Mohamed, A. (2014). *Deep Neural Network Acoustic Models for ASR* (Doctoral dissertation). Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/44123/1/Mohamed_Abdel-rahman_201406_PhD_thesis.pdf

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., …Vesely´, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of IEEE 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*, 2751-2755. doi: 10.21437/Interspeech.2016-595

Reynolds, D. A. (2009). Gaussian mixture models. In S. Z. Li (Eds.), *Encyclopedia of Biometrics* (pp. 659-663) 2009. doi: 10.1007/978-0-387-73003-5_196

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional Long Short-Term Memory Fully Connected Deep Neural Networks. In *Proceedings of 2015 IEEE International Conference on Acoustics Speech and Signal Processing*. doi: 10.1109/ICASSP.2015.7178838

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

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., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. In *Proceedings of Sixteenth Annual Conference of the International Speech Communication Association*.

Zhang, X., Trmal, J., Povey, D., & Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *Proceedings of ICASSP 2014*. doi: 10.1109/ICASSP.2014.6853589