Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 85-104

融合多任務學習類神經網路聲學模型訓練於會議語音辨識之研究

Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition

楊明翰*、許曜麒*、洪孝宗*、陳映文*、陳冠宇+、陳柏琳*

Ming-Han Yang, Yao-Chi Hsu, Hsiao-Tsung Hung, Ying-Wen Chen, Kuan-Yu Chen, and Berlin Chen

摘要

本論文旨在研究如何融合多任務學習(Multi-Task Learning, MTL)技術於聲學模型之參數估測，藉以改善會議語音辨識(Meeting Speech Recognition)之準確性。我們的貢獻主要有兩點：1)我們進行了實證研究以充分利用各種輔助任務來加強多任務學習在會議語音辨識的表現。此外，我們還研究多任務與不同聲學模型像是深層類神經網路(Deep Neural Networks, DNN)聲學模型及摺積神經網路(Convolutional Neural Networks, CNN)結合的協同效應，期望增加聲學模型建模之一般化能力(Generalization Capability)；2)由於訓練多任務聲學模型的過程中，調整不同輔助任務之貢獻(權重)的方式並不是最佳的，因此我們提出了重新調適法，以減輕這個問題。我們基於在台灣所收錄的中文會議語料庫

*國立台灣師範大學資訊工程學系
Department of Computer Science and Information Engineering, National Taiwan Normal University
E-mail: {mh_yang, ychsu, alexhung, cliffchen, berlin}@ntnu.edu.tw
+中央研究院資訊科學所
Institute of Information science, Academia Sinica
E-mail: kychen@iis.sinica.edu.tw

(Mandarin Meeting Recording Corpus, MMRC)建立了一系列的實驗。與數種現有的基礎實驗相比，實驗結果揭示了我們所提出的方法之有效性。

關鍵詞：多任務學習，深層學習，類神經網路，會議語音辨識。

Abstract

This paper sets out to explore the use of multi-task learning (MTL) techniques for more accurate estimation of the parameters involved in neural network based acoustic models, so as to improve the accuracy of meeting speech recognition. Our main contributions are two-fold. First, we conduct an empirical study to leverage various auxiliary tasks to enhance the performance of multi-task learning on meeting speech recognition. Furthermore, we also study the synergy effect of combing multi-task learning with disparate acoustic models, such as deep neural network (DNN) and convolutional neural network (CNN) based acoustic models, with the expectation to increase the generalization ability of acoustic modeling. Second, since the way to modulate the contribution (weights) of different auxiliary tasks during acoustic model training is far from optimal and actually a matter of heuristic judgment, we thus propose a simple model adaptation method to alleviate such a problem. A series of experiments have been carried out on the Mandarin meeting recording (MMRC) corpora, which seem to reveal the effectiveness of our proposed methods in relation to several existing baselines.

Keywords: Multi-Task Learning, Deep Learning, Neural Network, Meeting Speech Recognition.

1. 緒論

口語對話是人與人之間最自然的溝通方式，可以預期它也是人們與人工智慧助理等機器間最重要的互動方式。近六十年來，自動語音辨識的研究活動十分活躍，並且已取得了巨大的成功。在研究初期，語音辨識器只能在安靜的環境中識別一個單獨的詞彙。1980年代，以高斯混合模型-隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)做為聲學模型使得語音辨識有能力進行大詞彙量連續語音識別。由於GMM-HMM 的架構易於訓練模型和進行聲學解碼，因此近二十年來 GMM-HMM 是自動語音辨識系統的主流聲學模型，聲學模型的研究主要集中在以更好的模型結構與訓練演算法改良 GMM-HMM。顯著的成果包含狀態聯繫(State Tying) (Young & Woodland, 1993)、鑑別式訓練(Discriminative Training) (Povey, 2004)與最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT) (Gales, 1998)。在 GMM-HMM 模型主導語音界的時期內，研究學者們也探索了許多不同的聲學模型方法，然而卻沒有一種方法可以像 GMM-HMM 滿足建置成本和辨識效能的平衡。過去的五年內我們看見了深層學習架構和

技術在電腦視覺、語言及語言學習領域的巨大成功。深層類神經網路與其變體最終取代了 GMM，混合深層類神經網路-隱藏式馬可夫模型(Hybrid Deep Neural Networks-Hidden Markov Model, DNN-HMM)已成為大多數自動語音辨識系統的聲學模型。DNN 的竄起可歸功於以下六種因素：1)深層學習架構及演算法；2)通用計算圖形處理器(General Purpose Graphical Processing Units, GPGPU)的發展；3)數千小時的已轉寫語音訓練資料及更多的未標記資料；4)行動式網際網路和雲端計算；5)從生活到工作環境都廣泛地出現語音辨識技術需求。

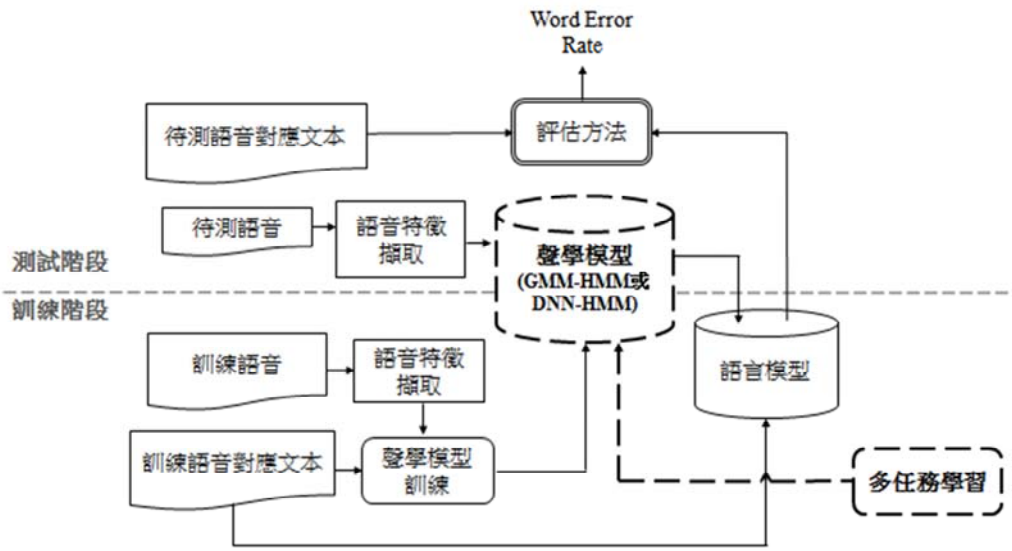

圖 1. 語音辨識流程與本論文嘗試改良之處(虛線標記部分)
[Figure 1. Illustration of the Enhanced Speech Recognition System]

雖然自動語音辨識技術已經是一項成熟的技術，但是在實際應用上仍有許多問題需要被解決。例如使用智慧型手機錄音時往往離手機麥克風較遠，錄音品質容易受環境影響。此外，現今語音辨識領域也面臨著海量詞彙、自由不受限的任務、吵雜的遠距離語音、自發性的口語及語言混雜情景的挑戰(Yu & Deng, 2014)。而會議語音辨識[1]正涵蓋了上述大部分的困境與挑戰，是一個相當困難的語音辨識任務。本論文將問題視為訓練與測試環境不匹配；除了語音和文字的使用不同，也包含了多語言的混用。簡而言之，上述問題是在考驗語音辨識之一般化能力。為克服此問題，我們首先比較各種輔助任務來加強多任務學習在會議語音辨識的表現。其次，藉由多樣異質模型的整合也是常見的方法，本論文透過實驗來驗證異質模型結合的效果。最後，我們提出了重新調適法以減輕傳統調整輔助任務權重方法的缺陷。所有結果是驗證於一在台灣所收錄的會議語料庫，內容多為中文和中英文混用語句。圖 1 所示為語音辨識流程與本論文嘗試改進的部分(以

[1]會議語音的內容如表 1 所示。不難發現語料中除了中英文交互使用外，也有自然對話所包含的語助詞、口頭禪、口吃與贅詞。專有名詞與時髦新詞也可能出現在語料中。

虛線標記)。

表 1. 會議語音範例
[Table 1. Meeting transcription examples extracted from MMRC]

語句編號	語句內容
U0006-002932	教材的 scenario 可能 target audience 如果是這群人可能是應該怎麼怎麼讓他 aggregate 出一套理論讓他再發揮更大
U0000-001623	呃考法律的法學院的碩士吼 然後呃就醫學院的碩士還有那個
U0000-001624	那個牙科的碩士吼那個那個補習班補得很兇啊那生意都很好
U0002-000118	呃 cover 到然後也講到這個 我們上次談的這個 megatrend 吼
U0002-000122	就是 knowledge formulation 呃 diversification 那個 is great

本論文後續章節安排如下：第二小節將簡介類神經網路的相關文獻探討，第三小節介紹多任務學習的發展演進以及我們想要探討的輔助任務，第四小節介紹我們提出的重新調整法，第五小節則解析基礎實驗及多任務學習的實驗結果，最後在第六小節進行結論與探討未來可能的研究方向。

2. 類神經網路之相關文獻探討

在機器學習的領域中，類神經網路的起源可以追溯到 1943 年的數學家 McCulloch，他設計了一套數學方法模擬神經元運作的模式，是開啟了類神經網路研究大門的先驅。接著在 1957 年，Rosenblatt 是第一個將類神經網路概念付諸實行的學者，提出了感知器(Perceptron)模型。1975 年，Werbos 提出倒傳導演算法(Backpropagation Algorithm) (Werbos, 1974)改善類神經網路參數更新的方式。終於在 1988 年，Rumelhart 等人發明了多層感知器(Multilayer Perceptron, MLP) (Rumelhart, Hinton & Ronald, 1988)，因為多層感知器適用於更多元的問題，使得類神經網路的研究熱潮再度熱絡起來。

在語音辨識領域中，從 1992 年起，就陸續有許多將類神經網路與隱藏式馬可夫模型結合(Hidden Markov Model, HMM)的研究。例如在 1998 年，Cook 等人(Cook *et al*., 1999)使用廣播新聞的語料，訓練多個遞迴神經網路(Recurrent Neural Network, RNN)與 MLP 的聲學模型，並透過 ROVER (Recognition Output Voting Error Reduction, ROVER) (Fiscus, 1997)的方法統整這些模型的辨識結果。2000 年時，學者們指出類神經網路也是合適的特徵擷取工具，例如 Bottleneck 特徵或 Tandem 特徵(Hermansky, Ellis & Sharma, 2000)。

早期的類神經網路研究受限於硬體計算資源的不足，且不易進行平行化處理，使得相關研究沒有顯著地突破。直到 2006 年開始，學者們在訓練演算法與架構上提出了一系列改進(Hinton, Osindero & Teh, 2006; Poultney, Chopra & Cun, 2006, Bengio, Lamblin, Popovici & Larochelle, 2007)，而後幾年的 GPGPU 運算設備發展迅速，使得深層類神經

網路模型計算成本問題大幅降低，也讓學者們願意投入此研究。現今主流的聲學模型設計即是利用深層類神經網路取代高斯混合模型(Hinton *et al*., 2012)。除了深層類神經網路之外，學者們也嘗試引入類神經網路變體，例如 CNN (Abdel-Hamid *et al*., 2014)與 RNN (Graves, Mohamed & Hinton, 2013)。這些新穎的深層模型在語音辨識領域也有顯著的成功(Sercu, Puhrsch, Kingsbury & LeCun, 2016)。

3. 多任務學習探討

多任務學習(Caruana, 1997)或者學會學習(Learning To Learn) (Thrun & Pratt, 1988)是一種機器學習的技術，其目的是希望藉由共同學習數個相關的輔助任務，以提升主任務的一般化能力。多任務學習大約在二十年前開始成為一項熱門的研究，有許多論文以理論的角度分析多任務學習的行為與一般化能力界限(Generalization Bound)，並提出了一系列有關多任務學習的統計理論，也進一步得知，透過相關輔助任務所產生的參數假設空間(Parameter Hypothesis Space)作為基礎，能夠提供更好的初始參數假設空間給其它新的輔助任務。近年來，多任務學習的研究開始探索自動地學習任務之間的關係，Zhang 等人將學習任務之間的關係視為求解凸函數(Convex Function)的過程(Zhang & Yeung, 2014)；假設數個線性回歸任務的參數具有相同的矩陣常態分佈事前機率(Matrix-Variate Normal Distribution Prior)，並由共變異數矩陣定義任務與任務之間的關係，模型訓練時能間接學習如何替正例任務關係(Positive Task Correlation)與負例任務關係(Negative Task Correlation) 的相關性建立模型。後來 Zhang 等人更融入了多任務特徵選取(Multi-Task Feature Selection) 與相關性學習(Relationship Learning)對高維度的輸入資料進行處理。據學者的研究證明，假設多個任務之間彼此相關，通過一起學習的方式來共享內在的表示資訊，就能夠達到知識轉移的效果。其實驗結果也證實此方法在模型遇到沒看過的資料(Unseen Data)時也能有不錯的成效。

3.1 語音辨識中的多任務學習

多任務學習結合深層類神經網路的架構(Multi-Task Deep Neural Network, MTL-DNN)如圖 2 所示。語音辨識領域中也有許多研究先進嘗試融合語音領域及多任務學習技術。例如 Parveen 等人(Parveen & Green, 2003)探討了 11 種不同的分類任務與語音增強任務的影響，例如語者的性別或情緒等。實驗結果發現在分類任務中，多任務訓練優於單任務訓練。Chen 等人(Chen & Mak, 2015)則是透過多任務學習的特性，使得資源豐富(Resource-Rich)的語言能夠在模型訓練的過程中，輔助資源貧乏(Resource-Poor)語言，提升它的辨識效果。Seltzer 等人的研究(Seltzer & Droppo, 2013)則是探討了以目前音框的音素標記、鄰近音素狀態標記(State Contexts)及鄰近音素的音素標記 (Phone Contexts) 做為輔助任務訓練聲學模型的效果，其文獻指出在英語音素語料(TIMIT) (Garofolo, Lamel, Fisher, Fiscus & Pallett, 1993)的語音辨識任務中有顯著地進步。然而，Seltzer 等人的研究只停留在單連音素(Monophone)，沒有使用到三連音素(Triphone)的資訊，也沒有探討若使用三連音素狀態標記取代單連音素狀態標記做為輔助任務的成效，這正是本論文想探

究的其中一項問題。另外，學者們嘗試將多種語言的語料混合在一起，共同訓練一個跨語言的聲學模型(Ghoshal,Swietojanski & Renals, 2013; Huang, Li, Yu, Deng & Gong, 2013)，證實這種做法確實能夠提升準確率。多語言的資料主要使用於訓練階段，所有語言的訓練資料皆會調整底層共享的隱藏層。每種語言有各自對應的輸出層，各個輸入語言的聲學特徵除了調整底層的隱藏層外，也會更新其所對應語言的輸出層，其它語言的輸出層將不會被 更新。如此一來，隱藏層就可被視為一層一層的特徵萃取器。總而言之，多任務學習允許學習多個任務時，以建設性(Constructive)或破壞性(Destructive)錯誤訊號梯度更新隱藏層。在多任務學習的框架下，模型將會同時學習：(1)主任務；(2)一個或數個相關的輔助任務；(3)任務間共享的隱藏層。

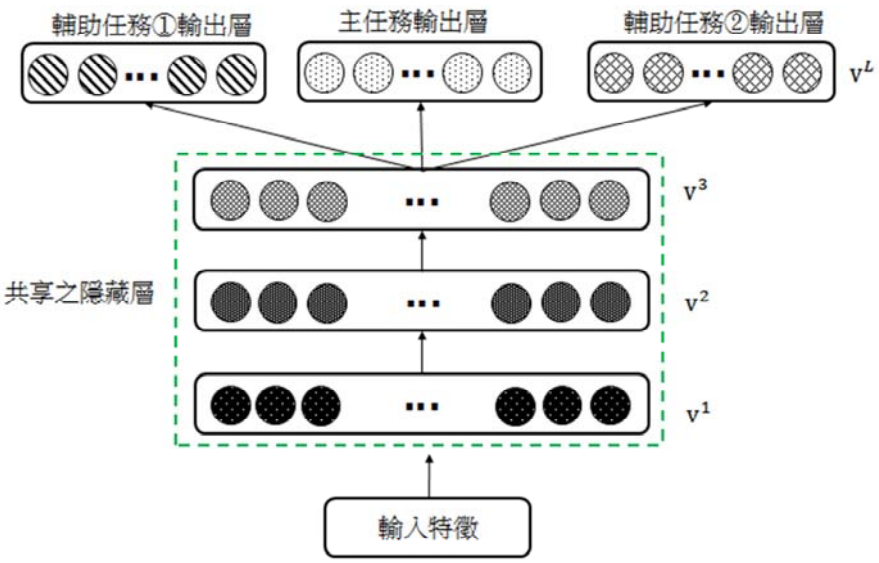

圖 2. 多任務深層類神經網路示意圖
[Figure 2. Illustration of the MTL-DNN system]

3.2 輔助任務探討

有學者的研究指出，多任務學習並不保證效能會提升，訓練的演算法及任務是否相關同樣也是重要的關鍵(Caruana, 1997)。有鑒於此，本論文從兩大類研究面向，篩選出 10 種輔助任務進行探討，如圖 3 所示。其中一個面向是語言與音韻學資訊，此類型的資訊主要分為 3 類：音框對應狀態標記、音框對應音素標記與多語言及跨語言資訊。另一個面向則是自動語音辨識回饋，我們採用的是模型壓縮技術(Model Compression) (Buciluǎ, Caruana & Niculescu-Mizil, 2006; Hinton, Vinyals & Dean, 2015)，從已訓練完成的強健模型中，將知識在訓練過程中轉移到待訓練模型。接下來將詳細介紹我們使用的輔助任務：

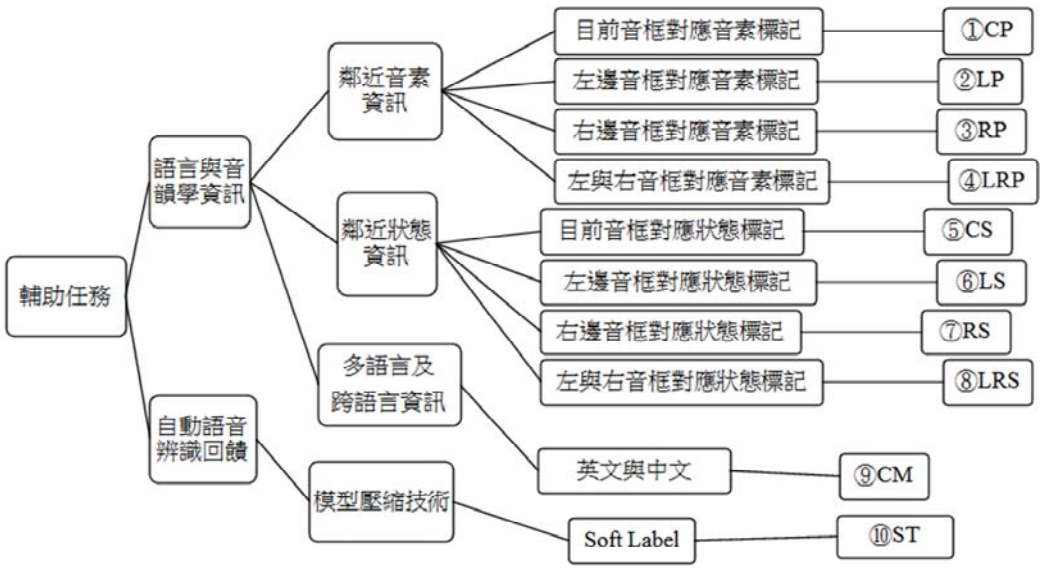

圖 3. 本論文使用的輔助任務一覽
[Figure 3. Auxiliry tasks used in this paper]

3.2.1 音框對應狀態標記：

音框對應狀態標記是以預測目前音框的前一個或後一個音框的 HMM 狀態標記做為輔助任務。由於以往類神經網路聲學模型的訓練方式，通常是以預測目前音框的 HMM 狀態標記為目標，這使得主任務的訓練目標並沒有鄰近音框的狀態資訊。這類輔助任務則是期望能夠提供模型訓練的過程中能夠加入這些額外資訊。以預測鄰近音框 t+1 的狀態標記為例，假設目前音框表示為$\mathbf{o}_t$，下一個時間點的狀態標記表示為s_{t+1}，則右邊音框狀態標記之目標函數表示為：

$$\mathcal{F}_{\text{Right-State}} = \sum_t \ln P_{\text{RS}}(s_{t+1}|\mathbf{o}_t) \tag{1}$$

3.2.2 音框對應音素標記：

這類輔助任務初始的設計理念跟 Triphone 模型類似，不同的地方在於以往產生 Triphone 模型所使用的 Decision Tree State Tying 技術(Young & Woodland, 1993)是靜態的二元表示，且與 DNN 的隱藏層無關。而融入多任務學習能夠在訓練的過程中，提供鄰近音素的資訊，達到動態更新，自動影響隱藏層的效果。我們除了遵循原有的設計理念外，也進一步延伸出不同的輔助任務。以預測目前音框所屬音素為例，假設時間 t 的音框所對應的音素標記表示為q_t，音框對應語音特徵向量表示為$\mathbf{o}_t$，則此輔助任務的目標函數可以表示為：

$$\mathcal{F}_{\text{Current-Phone}} = \sum_t \ln P_{\text{CP}}(q_t|\mathbf{o}_t) \tag{2}$$

3.2.3 多語言及跨語言資訊：

很自然地我們可以猜想，不同語言之間應該具有共同的發音模式。舉例來說，許多的子音和母音是跨語言共享的，運用語言之間共享的特性，來建立統計模型更優於僅使用單一語言建立的模型，這項優勢已經被許多研究報告證明。近年來，這類的研究透過深層類神經網路做為多語言及跨語言資訊的傳遞媒介也越來越火熱，其主要思路是認為低層靠近特徵的隱藏層，傾向於學習語言獨立(Language-Independent)的資訊；而較高層的隱藏層學習較多語言相關(Language-Dependent)的知識(Schultz & Waibel, 2001)。Swietojanski 等人提出以非監督式(Unsupervised)的方法，以多語言的資料目標語言的類神經網路模型進行初始化(Swietojanski, Ghoshal & Renals, 2012)。多語言訓練資料用於訓練一個多語言的 DNN 模型時，對每種語言來說，僅訓練特定語言的輸出層與共享底層隱藏層，也比每種語言重新訓練各自的 DNN 模型要容易得多。直到最近，仍然有許多研究先進們，追隨這樣的想法來進行改良。而在我們的任務中，由於會議語料具有中英文夾雜的特性，我們也嘗試希望透過與不同語言的語料一起訓練，使聲學模型更具一般化能力。

3.2.4 自動語音辨識回饋：

機器學習中，想要改進模型預測的準確率，最簡單且有效的方式，就是用同一組訓練資料訓練多個不同的模型，並且平均它們的預測結果。但是想要訓練多個模型在預測時結合它們預測的結果卻十分耗費計算與時間成本，尤其是當這些模型都屬於大規模的類神經網路時，所耗費的成本更是無法想像。因此，Buciluǎ 等人的研究顯示，把知識從這些已訓練的模型中擷取出來是可能的(Buciluǎ *et al.*, 2006)。

一般來說，我們都會認為用於訓練的目標函數應該盡可能地反映使用者的實際目標。儘管如此，模型的目標函數常常設計成要在訓練資料集上有最佳的辨識效能為準則。這樣的訓練方式使得模型盡可能的讓訓練資料所屬的類別機率越大越好，反而忽略了錯誤答案之間可能隱含的關係。以影像辨識為例，雖然高級跑車的圖片可能會被預測成不同的物體。但是以經驗來看，高級跑車被誤認為垃圾車的機率，應該比被誤認為胡蘿蔔的機率高。假設模型本身對於不同類別的輸出機率已經偷偷告訴我們這些知識，那麼在訓練時若能加入這些資訊，應該有助於提升模型的一般化能力。因此，我們嘗試融合前人的方法，從訓練有素(Well-Trained)的模型中蒸餾出有用的知識，這些知識又被稱為柔性標記(Soft Label) (Hinton *et al.*, 2015)。以 Soft Label 取代傳統非 1 及 0 表示法(Hard Label)，做為訓練模型的標記，這種做法的優點是將不同類別之間的排序資訊也融入訓練的過程中。假設現有已訓練完成的模型，則知識的蒸餾可透過加高輸出層 Softmax 函數的溫度 T，產生 Soft Label，而訓練新的模型時，將 Soft Label 做為輔助任務來進行訓練。Softmax 加上溫度 T 如下式所示：

$$v_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{3}$$

隨著溫度 T 上升，會使得函數輸出值較平緩(Smooth)。此外，Soft Label 在訓練過程中得到的錯誤訊號較小，也較容易滿足目標函數設定的目標。假設目前音框表示為$\mathbf{o}_t$，目前時間點的狀態標記表示為s_t，則 Soft Label 之目標函數表示為：

$$\mathcal{F}_{\text{Soft-Label}} = \sum_t \ln P_{\text{Soft}}(s_t|\mathbf{o}_t) \quad (4)$$

4. 模型重新調適法

主任務與輔助任務的影響力界限該如何拿捏，一直是個重要的研究議題，假使所選擇的輔助任務與主任務不相關，或沒有適當地調整任務的權重，辨識的效果便會大打折扣。為了減輕這個問題，我們受到的學者的研究啟發(Huang *et al*., 2013)，提出了重新調適法。這個方法的核心概念為：在訓練模型時，模型參數較容易擬合最後幾次迭代所送入的訓練資料。因此我們可以將 MTL-DNN 訓練所萃取出的共享隱藏層(Shared Hidden Layers, SHLs)，視為是一個含有額外知識的特徵擷取模組，以此模組為基礎，對主任務重新進行調適訓練，進一步突顯主任務的影響力。重新調適法的流程十分簡單：首先，先以多任務的方式訓練類神經網路聲學模型。待訓練結束後，保留底層 SHLs 的部分，在 SHLs 上層加上新的 Softmax 層，使用主任務的訓練資料與標記進行調適訓練。這種做法將多任務學習視為監督式預訓練(Supervised Pre-Training)，輔助任務在預訓練中扮演正則項(Regularizer)的角色，將模型定位在參數空間中較好的位置，使得後續的微調(Fine Tuning)容易找到良好的局部最小值。實驗也證實了我們提出的重新調適法，確實能夠提升會議語料的辨識率。

5. 實驗

5.1 實驗環境簡介

本論文主要使用的語料庫為國內某大公司錄製的中文會議語料庫(MMRC)，其中收錄了約 43.18 小時的會議語料。語料庫文本經由專家進行轉寫與標記。會議參與人員有 23 位語者。共有 40,022 句。本實驗將會議語料庫分為訓練集、發展集與測試集，如表 2 所示。其中訓練集有 36.02 小時，35,769 句；發展集有 3.52 小時，3,269 句；測試集有 3.64 小時，984 句。

表 2. 中文會議語料庫
[Table 2. Statistics of the Mandarin Meeting Recording Corpus]

MMRC	訓練集	發展集	測試集	總計
小時數	36.02	3.52	3.64	36.02
語句數(句)	35,769	3,269	984	35,769

中文會議語料庫中，所有會議的談話內容及參與人員的對話方式並沒有經過設計，較貼近一般科技公司實際開會的流程。例如聊到專業技術時，通常會出現中英文夾雜的

對話；發表談話時可能有贅詞或口吃的現象，甚至根據開會氣氛的變化，語速和音量也可能產生差異；會議進行時也可能受到外部出現不可預期的噪音干擾；對話過程中，主題可能斷斷續續不連貫；加上不同會議可能位於不同的地點，錄音品質、所使用的麥克風都可能不同。例如有些會議室只有近距離麥克風，有些則是只有遠距離麥克風，抑或是會議室可能有回音干擾等等。因此會議語音是一種十分具備挑戰性的語料。

本論文的實驗使用美國約翰霍普金斯大學學者所發展出的一套大詞彙連續語音辨識的發展工具軟體 Kaldi (Povey *et al*., 2011)，以及 Python 程式語言上的函式庫等提供機器學習或是深層學習與 GPGPU 運算結合的開發環境。此外，多語言及跨語言資訊的輔助任務中，我們用於訓練聲學模型的語料庫如表 3 及表 4 所示。英文的語料為英文音素語料語料庫(TIMIT)，其中訓練集有 3.14 小時，3,696 句；發展集有 0.34 小時，400 句。中文的語料為中文廣播新聞語料庫(Mandarin Chinese Broadcast News Corpus, MATBN) (Wang, Chen, Kuo & Cheng, 2005)，訓練集有 25.60 小時，34,672 句；發展集有 1.3 小時，292 句。實驗所採用的語言模型為以 MMRC 語料所訓練的 *N*-連詞(*N*-gram)語言模型。詞彙的數量有 33,814 詞。音素方面，語料庫含有中文與英文的音素。中文音素切分為聲母及韻母，所含總音素數量為 247 個。

表 3. TIMIT 語料庫
[Table 3. Statistics of the TIMIT corpus]

TIMIT	訓練集	發展集	測試集	總計
小時數	3.14	0.34	(未使用)	3.48
語句數(句)	3,696	400	(未使用)	4,096

表 4. MATBN 語料庫
[Table 4. Statistics of the MATBN corpus]

MATBN	訓練集	發展集	測試集	總計
小時數	25.60	1.36	(未使用)	26.96
語句數(句)	34,672	292	(未使用)	34,964

5.2 基礎實驗

本小節主要想比較不同的語音特徵以及不同的聲學模型對中文會議語料辨識字錯誤率的影響，所使用的聲學模型高斯混合模型(Gaussian Mixture Model, GMM)與深層類神經網路模型。GMM 使用的語音特徵有三種，包含梅爾倒頻譜特徵(MFCC)、線性鑑別分析加上最大化相似度線性轉換(LDA+MLLT)與線性鑑別分析加上最大化相似度線性轉換與語者調適訓練(LDA+MLLT+SAT)，所有的特徵皆使用倒頻譜正規化法，為了表示方便，特

徵與代號的對應表示如下：*tri2* 表示使用梅爾倒頻譜特徵(MFCC)、*tri3* 表示為線性鑑別分析加上最大化相似度線性轉換(LDA+MLLT)的特徵，線性鑑別分析加上最大化相似度線性轉換與語者調適(LDA+MLLT+SAT)則表示為 *tri4*。

在會議的情景下，如果能將不同語者的說話方式與習慣，也就是語者的資訊，加入聲學模型的訓練，應該會有不錯的成效。從實驗結果也可以驗證我們的想法，在 GMM 的基礎實驗中，效果最好的特徵為線性鑑別分析加上最大化相似度線性轉換與語者調適(*tri4*)，所以接下來以類神經網路為主的實驗，都是以 *tri4* 特徵訓練 GMM 以產生類神經網路訓練資料的標記。類神經網路的語音特徵則是梅爾濾波器組特徵(Mel-Frequency Filterbanks, FBANK) 40 維加上 3 維的聲調特徵(Pitch) 共 43 維。鄰近音窗的大小設定為 11 個音框(取目前音框前後各 5 個音框)。並對 43 維語音特徵取相對的一階差量係數(Delta Coefficient)和二階差量係數(Acceleration Coefficient)，輸入特徵的維度為 1419 維。DNN 實驗設定使用 6 層隱藏層，每層 2048 個神經元，隱藏層的活化函數(Activation Function)使用 Sigmoid。CNN 實驗使用的設定是沿著頻率軸掃描的 CNN，架構以摺積層(Convolution Layer)-池化層(Pooling Layer)-摺積層-全連接隱藏層 2 層的順序排列。第一層摺積層的摺積核(Filter)大小為 8 維，共 128 個。第二層摺積層的摺積核大小為 4 維，共 256 個。池化層採用最大池化法(Max Pooling)運算，池化窗的大小為 3 維，池化步伐(Pooling Step)為 3。為了聚焦在輔助任務與重新調適法的效果，類神經網路的權重我們沒有透過預訓練調整。

基礎實驗的步驟流程如下：我們先透過訓練集的語音特徵訓練單連音素的 GMM 聲學模型。根據單連音素模型，訓練三連音素的 GMM 聲學模型。基於三連音素模型再使用不同的特徵(例如 *tri2*、*tri3* 及 *tri4*)分別訓練出三組不同的 GMM。接著我們利用上述三組 GMM，對訓練集的語音資料進行強制對齊(Forced Alignment)，取得每個音框對應的機率密度函數編號(不同的 GMM 所求取的機率密度函數之編號會有差異)，做為訓練資料的標記。最後，我們分別保留這三組 GMM 計算出來的初始機率、轉移機率與強制對齊的資訊(標記)，以最小化交叉熵(Minimum Cross-Entropy, MCE)的目標函數，重新訓練三組 DNN，取代原有的 GMM 來產生每個音框所對應 HMM 狀態的機率。訓練 DNN 與 MTL-DNN 時我們使用小批次隨機梯度下降法(Mini-Batch Stochastic Gradient Descent)，每次 mini-batch 抽樣 256 筆訓練語料特徵輸入類神經網路，微調(Fine-Tuning)使用倒傳導演算法進行網路參數的調整。實驗結果如表 5 所示，可以發現聲學模型改成使用類神經網路後，字錯誤率從 GMM 的 51.88%降低到 38.30%。使用摺積神經網路效果更加顯著，能從 51.88%降低到 38.16%。

表 5. 不同類神經網路模型用於 MMRC 的字錯誤率(%)
[Table 5. Recognition Results achieved by various systems (in Character Error Rate(%))]

模型	特徵	測試集
GMM_*tri4*	MFCC	51.88
DNN6※	FBANK	38.30
CNN-DNN2※	FBANK	38.16
CNN-DNN4	FBANK	35.60
LSTM	FBANK	36.48

※未使用預訓練的類神經網路模型

5.3 多任務學習之實驗結果

本論文的輔助任務可分成 2 大類，一類是語言與音韻學資訊，此類型的資訊主要分為 3 種：音框對應音素標記(代號①到④)、音框對應狀態標記(代號⑤到⑧) 與多語言及跨語言資訊(代號⑨)。另一類則是自動語音辨識回饋(代號⑩)，自動語音辨識回饋有許多研究面向，二 我們所採用的是模型壓縮技術，從已訓練完成的強健模型中產生 Soft Label，提供待訓練模型進行訓練。接下來我們將詳細介紹實驗中輔助任務的設定：

1)音框對應音素標記：音框對應音素標記又可分為 4 種：前一個時間點(左邊)之音框對應的音素標記、目前音框對應的音素標記、下一個時間點(右邊)之音框對應的音素標記與同時預測前一個時間點(左邊)。

2)音框對應狀態標記：概念與前項輔助任務相同，音框對應狀態標記也可分為 3 種：前一個時間點(左邊)之音框對應的狀態標記、目前音框對應的狀態標記與下一個時間點(右邊)之音框對應的狀態標記。值得注意的是，目前音框對應的狀態標記想要預測的目標是不同的特徵所訓練之高斯混合模型產生的三連音素模型的狀態，舉例來說，若主任務以 *tri1* 的狀態標記為目標進行訓練時，輔助任務的目標則是以 *tri2* 的狀態標記為目標進行訓練。

3)多語言及跨語言資訊：我們分別使用 TIMIT 語料庫與 MATBN 語料庫，做為英文和中文的輔助資訊，語料庫統計資訊如表 3 及表 4 所示。我們的實驗可分為兩種，一種是輔助任務只使用 TIMIT 語料庫進行訓練。另一種則是使用 TIMIT 語料庫與 MATBN 語料庫。

4)自動語音辨識回饋：Soft Label 的實驗我們針對不同的溫度、已訓練模型與待訓練模型是否同質、訓練標記精確與否及不同的任務權重比例進行實驗。

上述輔助任務的權重除了特別標註外，皆設定為 1。我們先探討預測目前的音框屬於哪一種音素之輔助任務是否對辨識有幫助，可以從表 6 中觀察到，輔助任務預測音素 *tri3* 有較好的效果，反而預測音素 *tri4* 效果卻不如 *tri3* 明顯。可能是因為經過語者調適訓練的音素 *tri4*，與主任務的 *tri4* 標記性質過於相近，導致效果不明顯。因此，我們可

以發現選擇輔助任務時，除了任務需與主任務相關之外，選擇異質性的任務會有較佳的辨識效果。

表 7 為其它輔助任務在中文會議語料庫的字錯誤率。首先，我們可以先從音素對應標記的輔助任務(②與③)觀察到：預測下一個(右邊)音框所屬的音素標記的辨識效果(37.90%)較預測上一個(左邊)音框的音素標記的效果(38.58%)明顯。而從狀態對應標記的輔助任務(⑥與⑦)觀察到：預測上一個(左邊)音框所屬的狀態標記的辨識效果(36.79%)較預測上一個 (左邊)音框的狀態標記的效果(39.19%)明顯，但同時預測左邊及右邊的音素標記或狀態標 記的辨識率並不如預期，分別為 38.33%與 38.83%，原因應該是由於左右音框標記的輔助任務所佔的權重十分難選擇，需要仰賴經驗來調整。

表 6. 不同音素標記在 MMRC 會議的字錯誤率(%)
[Table 6. Recognition results achieved by MTL-DNN6 with different phoneme labels (in Character Error Rate(%))]

模型	任務編號	輔助任務預測音素	測試集 1
DNN6	Baseline	-	38.30
MTL-DNN6	①	*mono*	37.76
MTL-DNN6	①	*tri2*	37.60
MTL-DNN6	①	*tri3*	36.98
MTL-DNN6	①	*tri4*	37.14

表 7. 不同輔助任務在 MMRC 的字錯誤率(%)
[Table 7. Recognition results achieved by MTL-DNN6 with different auxiliary tasks (in Character Error Rate(%))]

模型	任務代號	備註		測試集
DNN6	Baseline	-		38.30
MTL-DNN6	②	Phone	Left	38.58
MTL-DNN6	③		Right	**37.90**
MTL-DNN6	④		Both	38.33
MTL-DNN6	⑥	State	Left	**36.79**
MTL-DNN6	⑦		Right	39.19
MTL-DNN6	⑧		Both	38.83

表 8 為多語言及跨語言資訊於中文會議語料庫之辨識結果，可以發現 MMRC 語料與 TIMIT 語料一起進行訓練的字錯誤率為 38.06%，可以使得聲學模型在訓練時能夠額外獲取英文音素的知識，在中英文轉換頻繁的語料確實有幫助。而 MMRC、TIMIT 與 MATBN 一起訓練的模型辨識率卻些微上升到 38.23%。我們認為原因可能有二：其一，輔助任務與主任務皆以中文為主，因此幫助並不明顯。其二，輔助任務貢獻(權重)的選擇十分關鍵，也需要經過不斷地嘗試才能找到適合此任務的權重。

表 9 為 Soft Label 的實驗，我們先探討不同的輸出層溫度的影響。以產生 Soft Label 的模型為以狀態標記 *tri4* 所訓練的模型(DNN6)為例，從實驗中可以發現，溫度較高的效果較佳：溫度設定為 5 的辨識率(35.91%)優於溫度設定為 2(36.58%)。而使用較精確的狀態標記(*cnn_ali*)訓練的狀況，則是溫度 2 的辨識效果 (36.20%)優於溫度 5 的辨識效果(36.53%)。因為輸出層溫度較高，表示類別間的排序資訊較豐富。模型訓練時如果以較精確的標記表示(*cnn_ali*)時，不需要過多的排序資訊就能夠訓練得不錯。而當使用較模糊的狀態標記(*tri4*) 時，則需要更多的排序資訊才有較佳的效果。最後是產生 Soft Label 的模型與待訓練模型屬於同質模型的情景。從數據可以發現，當兩者都是摺積神經網路時，字錯誤率為 37.30%，進步的幅度並不大，這表示同質的模型產生的 Soft Label 幫助有限。

表 8. 多語言及跨語言資訊在 MMRC 的字錯誤率(%)
[Table 8. Recognition results obtained by using multi/cross-lingual corpora (in Character Error Rate(%))]

模型	任務代號	共同訓練之語料庫	測試集
DNN6	Baseline	-	38.30
MTL-DNN6	⑨	+TIMIT	38.06
MTL-DNN6	⑨	+TIMIT+MATBN	38.23
MTL-CNN2-DNN2	⑨	+TIMIT	38.17
MTL-CNN2-DNN2	⑨	+TIMIT+MATBN	37.64
MTL-CNN2-DNN2※	⑨	+TIMIT+MATBN	**37.31**

※輔助任務權重為 0.7

表 9. Soft Label 在 MMRC 的字錯誤率(%)
[Table 9. Recognition results obtained by integrating the soft label technique with various DNN models (in Character Error Rate(%))]

模型	溫度	目標標記	任務權重比例 (主:輔)	測試集
DNN6	-	*tri4*	-	38.30
MTL-DNN6 [1]	2	*tri4*	0.5:1	36.58
MTL-DNN6 [1]	5	*tri4*	0.5:1	**35.91**
MTL-CNN2-DNN4 [2]	5	*tri4*	0.5:1	37.30
MTL-DNN6 [2]	2	*cnn_ali*	0.5:1	36.20
MTL-DNN6 [2]	5	*cnn_ali*	0.5:1	36.53

[1]產生 Soft Label 的模型為 DNN6;
[2]產生 Soft Label 的模型為 CNN2-DNN4

最後，重新調適法的實驗結果如表 10 所示。在我們的任務中，調整所有網路的參數之辨識效果優於僅調整輸出層的參數之辨識效果，因此表 10 的數據列出的是重新調整所有網路參數的辨識錯誤率。從實驗中可以發現重新調適的方法對 Soft Label 的模型較無效果，可以推斷 Soft Label 訓練的模型效果較穩定。在多語言及跨語言的任務中，即使模型並沒有調整到最佳的權重(38.23%)，但是在經過調適後，辨識效果提升十分明顯(36.33%)。進步幅度較大的主因可能是因為用於預訓練時，所使用到的訓練語料較多的緣故。另外，預測鄰近音框的音素標記與狀態標記的任務經過重新調適法調整訓練後，預測左邊音框音素標記的辨識率進步到 37.73%，而預測右邊音框音素標記的辨識率進步到 37.13%。而預測左邊音框狀態標記的辨識率進步到 37.94%，預測右邊音框狀態標記的辨識率則可以進步到 37.97%。總結來說，重新調適法並不直接受限輔助任務的優劣，可以嘗試在更多元的設定。

表 10. 重新調適法在 MMRC 的字錯誤率(%
[Table 10. Recognition results achieved by the proposed method (in Character Error Rate(%))]

重新調適之模型	任務代號	備註	測試集
DNN6	baseline	-	38.3
MTL-DNN6	②	Left Phone	37.76
MTL-DNN6	③	Right Phone	**37.13**
MTL-DNN6	⑥	Left State	37.94
MTL-DNN6	⑦	Right State	37.97
MTL-DNN6	⑨	+TIMIT	37.13
MTL-DNN6	⑨	+TIMIT+MATBN	**36.33**
MTL-DNN6	⑩	溫度 2， $tri4$標記	36.96
MTL-DNN6	⑩	溫度 5，cnn_ali標記	37.14

6. 結論與未來展望

聲學模型在會議語音辨識的研究上扮演著十分重要的角色，本論文旨在研究如何融合多任務學習技術於聲學模型之參數估測，藉以改善會議語音辨識之準確性。研究成果與貢獻可分成兩點：1)在多任務學習技術中，我們探究 10 種不同輔助任務在類神經網路聲學模型的成效。其中以使用 Soft Label 及多語言及跨語言資訊做為輔助任務的進步幅度最大。以 Soft Label 做為輔助任務可以使得字錯誤率從 38.30%降低到 35.91%。而以多語言及跨語言資訊做為輔助任務也能使得字錯誤率從 38.30%下降到 38.23%。另外，我們除了結合多任務學習與類神經網路聲學模型之外，也嘗試融合多任務學習於新穎的摺積神經網路上。以跨語言與多語言的輔助任務為例，多任務學習摺積神經網路可以使字錯誤率從 38.23%進一步下降到 37.64%；2) 我們提出重新調適法，使得未調整到最佳輔助任務貢獻(權重)的模型，經過重新調整後，其辨識準確率能夠提升。在多語言及跨語言的任務中，有著最佳的進步幅度，辨識錯誤率從 38.23%降低到 36.33%。

未來可近一步探討五大面向：1)輔助任務選擇：多任務學習的研究大多還是停留在使用語音學及音韻學的資訊(音素，音框狀態)做為輔助任務，未來我們希望能夠探究更有效的輔助任務(例如透過詞、句子或更高層次的特徵)；2)輔助任務間的關係：目前在語音辨識中，多任務學習的研究尚未探討任務之間彼此的關係，例如不同粗細程度的輔助任務之間可能具有階層式的關係，若能將這些關係融入類神經網路訓練中，或許也是一個研究方向；3)輔助任務的影響深度：由於現在多任務學習中，輔助任務只放在類神經網路輸出層，或許可以融入課程學習(Curriculum Learning) (Bengio, Louradour, Collobert

& Weston, 2009)的概念，針對類神經網路不同深度的隱藏層設計不同難度的任務；4)融合更新穎的模型：現今遞迴式神經網路與其改良之長短期記憶網路(Long Short-Term Memory) (Li, Mohamed, Zweig & Gong, 2016)在語音辨識中也取得了不錯的成果，我們也希望善用這些新穎的聲學模型來改進辨識的正確率；5)最後，由於大多數的辨識錯誤常發生於中文詞與英文詞發音很類似的狀況，例如『size』被辨識為『才是』或 bottleneck 被辨識成『把那個』。辨識錯誤之詞在句子中的位置，又會間接影響到句子後面其它詞的辨識結果，導致一連串的辨識錯誤連鎖發生。如果想要使得會議語音辨識更加地符合實際需求，其中一項重要的目標，應該是提升關鍵詞的辨識準確率。如果能夠依據關鍵詞辨識正確與否，以最佳化辨識關鍵詞效能的訓練準則調整聲學模型也是一個值得研究的方向。

致謝

本論文之研究承蒙教育部-國立臺灣師範大學邁向頂尖大學計畫(104-2911-I-003-301)與行政院科技部研究計畫(MOST 104-2221-E-003-018-MY3 和 MOST 105-2221-E-003-018-MY3)之經費支持，謹此致謝。

Reference

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

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*, 153.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 41-48.

Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 535-541.

Caruana, R. (1997). *Multitask learning* (Doctoral dissertation, University of Carnegie Mellon).

Chen, D., & Mak, B. K. W. (2015). Multitask learning of deep neural networks for low-resource speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(7), 1172-1183.

Cook, G., Christie, J., Ellis, D., Fosler-Lussier, E., Gotoh, Y., Kingsbury, B., Morgan, N., Renals, S., Robinson, T., & Williams, G. (1999). An overview of the SPRACH system for the transcription of broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 347-354.

Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, *12*(2), 75-98.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM*. NIST speech disc 1-1.1. NASA STI/Recon technical report n, 93.

Ghoshal, A., Swietojanski, P., & Renals, S. (2013). Multilingual training of deep neural networks. In P*roceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 7319-7323.

Graves, A., Mohamed, A. R., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 6645-6649.

Hermansky, H., Ellis, D. P., & Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 1635-1638.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & 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.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. Retrived from https://arXiv preprint arXiv:1503.02531.

Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 7304-7308.

Li, J., Mohamed, A. R., Zweig, G., & Gong, Y. (2016). Exploring multidimensional LSTMs for large vocabulary ASR. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 4940-4944.

Parveen, S., & Green P. D. (2003). Multitask learning in connectionist ASR using recurrent neural networks. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 1813-1816.

Poultney, C., Chopra, S., & Cun, Y. L. (2006). Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems*, 1137-1144.

Povey, D. (2004). *Discriminative training for large vocabulary speech recognition* (Doctoral dissertation, University of Cambridge).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer G., & Vesely, K. (2011). The

Kaldi speech recognition toolkit. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Rumelhart, D. E., Hinton, G. E., & Ronald, J. W. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, *5*(3).

Schultz, T., & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, *35*(1), 31-51.

Seltzer, M. L., & Droppo, J. (2013). Multi-task learning in deep neural networks for improved phoneme recognition. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 6965-6969.

Sercu, T., Puhrsch, C., Kingsbury, B., & LeCun, Y. (2016). Very deep multilingual convolutional neural networks for LVCSR. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 4955-4959.

Swietojanski, P., Ghoshal, A., & Renals, S. (2012). Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proceedings of the International Conference on Spoken Language Technology Workshop (SLT)*, 246-251.

Thrun, J. S., & Pratt, L. (1988). *Learing to learn*. Norwell, MA : Kluwer Academic Publishers.

Wang, H. M., Chen, B., Kuo, J. W., & Cheng, S. S. (2005). MATBN: a Mandarin Chinese broadcast news corpus. *Journal of Computational Linguistics and Chinese Language Processing*, *10*(2), 219-236.

Werbos, P. J. (1974). *Beyond regression: new tools for prediction and analysis in the behavioral sciences* (Doctoral dissertation, University of Harvard).

Young, S. J., & Woodland, P. C. (1993). The use of state tying in continuous speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 2207-2219.

Yu, D., & Deng, L. (2014). *Automatic speech recognition: a deep learning approach*. London, England: Springer-Verlag.

Zhang, Y., & Yeung, D. Y. (2014). A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *8*(3), 12.