Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 19-34

結合鑑別式訓練與模型合併於半監督式語音辨識之研究

Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition

羅天宏*、陳柏琳*

Tien-Hong Lo and Berlin Chen

摘要

近年來鑑別式訓練(Discriminative training)的目標函數 Lattice-free Maximum Mutual Information (LF-MMI)在自動語音辨識(Automatic speech recognition, ASR)上取得了重大的突破。儘管 LF-MMI 在監督式環境下斬獲最好的成果，然而在半監督式設定下，由於種子模型(Seed model)常因為語料有限而效果不佳。且由於 LF-MMI 屬於鑑別式訓練之故，易受到轉寫正確與否的影響。本論文利用兩種思路於半監督式訓練。其一，引入負條件熵(Negative conditional entropy, NCE)權重與詞圖(Lattice)，前者是最小化詞圖路徑的條件熵(Conditional entropy)，等同對 MMI 的參考轉寫(Reference transcript)做權重平均，權重的改變能自然地加入 MMI 訓練中，並同時對不確定性建模。其目的希望無信心過濾器(Confidence-based filter)也可訓練模型。後者加入詞圖，比起過往的只使用最佳辨識結果，可保留更多假說空間，進而提升找到參考轉寫(Reference transcript)的可能性；其二，我們借鑒整體學習(Ensemble learning)的概念，使用弱學習器(Weak learner)修正彼此的錯誤，分為假說層級合併(Hypothesis-level combination)和音框層級合併(Frame-level combination)。實驗結果顯示，加入 NCE 與詞圖皆能降低詞錯誤率(Word error rate, WER)，而模型合併(Model combination)則能在各個階段顯著提升效能，且兩者結合可使詞修復率(WER recovery rate, WRR)達到 60.8%。

關鍵詞：自動語音辨識、鑑別式訓練、半監督式訓練、模型合併、LF-MMI

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

Abstract

In recent years, the so-called Lattice-free Maximum Mutual Information (LF-MMI) criterion has been proposed with good success for supervised training of state-of-the-art acoustic models in various automatic speech recognition (ASR) applications. However, when moving to the scenario of semi-supervised acoustic model training, the seed models of LF-MMI are often show inadequate competence due to limited available manually labeled training data. This is because LF-MMI shares a common deficiency of discriminative training criteria, being sensitive to the accuracy of the corresponding transcripts of training utterances. This paper sets out to explore two novel extensions of semi-supervised training in conjunction with LF-MMI. First, we capitalize more fully on negative conditional entropy (NCE) weighting and utilize word lattices for supervision in the semi-supervised setting. The former aims to minimize the conditional entropy of a lattice, which is equivalent to a weighted average of all possible reference transcripts. The minimization of the lattice entropy is a natural extension of the MMI objective for modeling uncertainty. The latter one, utilizing word lattices for supervision, manages to preserve more cues in the hypothesis space, by using word lattices instead of one-best results, to increase the possibility of finding reference transcripts of training utterances. Second, we draw on the notion stemming from ensemble learning to develop two disparate combination methods, namely hypothesis-level combination and frame-level combination. In doing so, the error-correcting capability of the acoustic models can be enhanced. The experimental results on a meeting transcription task show that the addition of NCE weighting, as well as the utilization of word lattices for supervision, can significantly reduce the word error rate (WER) of the ASR system, while the model combination approaches can also considerably improve the performance at various stages. Finally, fusion of the aforementioned two kinds of extensions can achieve a WER recovery rate (WRR) of 60.8%.

Keywords: Automatic Speech Recognition, Discriminative Training, Semi-supervised Training, Model Combination, LF-MMI

1. 緒論 (INTRODUCTION)

近年來基於類神經網路的聲學模型(Deep neural network-hidden Markov model, DNN-HMM)取得重大的突破(Seide, Li & Yu, 2011) (Dahl, Yu, Deng & Acero, 2012)。傳統的 DNN-HMM 透過交互熵訓練(Cross-Entropy training, CE)和鑑別式訓練(Discriminative training) (Valtchev, Odell, Woodland & Young, 1996) (Valtchev, Odell, Woodland & Young, 1997) (Woodland & Povey, 2002) ，兩階段的訓練提升聲學模型的辨識效果。尤其是第二

階段的鑑別式訓練，由於提升效果顯著，吸引了許多研究者的目光。過往於鑑別式訓練的研究主題種類繁多，如 MMI (Bahl, Brown, de Souza & Mercer, 1986), MCE (Juang, Hou & Lee, 1997), MPE (Povey & Woodland, 2002), sMBR (Kaiser, Horvat & Kacic, 2000) (Gibson & Hain, 2006)和 bMMI (Povey *et al,* 2008)等。最近，隨著語料的增長，不透過第一階段的 CE 訓練，將鑑別式訓練做一階段訓練的端對端訓練(End-to-End)也越來越流行。目前兩種主流的端對端架構的目標函數為 CTC (Graves, Fernández, Gomez & Schmidhuber, 2006)和 Lattice-free MMI (LF-MMI) (Povey *et al*., 2016)。前者在語料非常充足(通常大於 500 小時)的情況下，表現可以媲美甚至超越傳統的二階段方法。而後者證明了在語料較為缺乏的情況下，儘管效能會下降，但仍可以勝過前者，因此成為了目前最具魅力的研究主題。在(Povey *et al*., 2016)的實驗中展示，基於 LF-MMI 的目標函數下，可從亂數初始化參數後，以鑑別式準則訓練類神經網路。實驗結果顯示 LF-MMI 效果更勝兩階段訓練的 sMBR 一籌，且還可結合 sMBR 進一步提升辨識結果。然而，在這樣的訓練準則下，仍受限於需大量訓練語料的問題(Data hungry)。進一步來說，便是在小語料庫上的表現(通常小於 100 小時)仍無法勝過在大語料庫的優異結果(Pundak & Sainath, 2016)。

在現實生活中，相對於高成本的人工轉寫語料，未轉寫語料十分容易取得。當我們沒辦法取得大量的轉寫語料時，就必須更有效地利用大量的未轉寫語料訓練模型。換句話說，探索存在於未轉寫語料的線索，並加入半監督式訓練的聲學模型就更顯重要。另一方面，半監督式訓練用途多元，不僅可用在自動語音辨識的訓練，也同樣適用於自動轉寫(Automatic labeling)及遷移學習(Transfer learning)、語者調適(Speaker adaptation)。過往的研究於半監督式聲學模型(Zavaliagkos, Siu, Colthurst & Billa, 1998)，最常見的訓練方法是自我訓練(Self-training) (Vesely, Hannemann & Burget, 2013) (Grezl & Karafiát, 2013) (Zhang, Liu & Hain, 2014)。自我訓練的架構主要分成兩階段，第一階段為利用轉寫語料訓練種子模型直到穩定，第二階段則是利用種子模型辨識未轉寫語料，並以此為答案重新訓練模型。在第二階段的辨識結果與真實答案難免會有誤差，因此會再加入信心過濾器(Confidence-based filter) (Lamel, Gauvain & Adda, 2002) (Chan & Woodland, 2004) (Liu, Chu, Lin & Chen, 2007) 挑選訓練語料，該動作可在不同層級上進行，分為音框層級(Vesely, Hannemann & Burget, 2013)、詞層級(Thomas, Seltzer, Church & Hermansky, 2013)以及語句層級(Grezl & Karafiát, 2013) (Vesely *et al*., 2013) (Zhang *et al*., 2014)。

最近 LF-MMI 訓練方法在 ASR 取得了重大的突破。有別於傳統的二階段訓練，LF-MMI 提供更快的訓練與解碼，同時在模型準度上取得目前最優異的表現。儘管 LF-MMI 在監督式環境下獲得最好的成果，但在半監督式環境下的研究成果仍然有限。在過往的研究中，鑑別式訓練的好壞很大層度地仰賴於訓練語句的正確性(Mathias, Yegnanarayanan & Fritsch, 2005) (Yu, Gales, Wang & Woodland, 2010) (Cui, Huang & Chien, 2011)，而屬於鑑別式訓練的 LF-MMI 也同樣對於正確性十分敏感。然而，在半監督式訓練過程中，由於第二階段訓練時無法保證語句的正確性，因此在過往研究常著重於二階段鑑別式訓練前的信心過濾器，如(Liu *et al*., 2007) (Mathias *et al*., 2005)將音框層

級的信心過濾器加入鑑別式訓練。而在(Walker, Pedersen, Orife & Flaks, 2017) 加入語句層級的信心過濾器以及後處理最佳辨識結果(One-best result)。本論文與(Manohar, Hadian, Povey & Khudanpur, 2018) (Manohar, Povey & Khudanpur, 2015)相同，是將詞圖的不確定性以條件熵(Conditional entropy)的形式加入，保留整個詞圖來做二階段的訓練。本論文與其不同的是，我們將這樣的方法做在更口語化的會議語料，以及基於這個方法之上，利用整體學習的觀念，進一步地探討模型合併帶來的成效。

整體的模型合併在自動語音辨識上能取得優於單一模型的成果(Fiscus, 1997) (Evermann & Woodland, 2000) (Deng & Platt, 2014) (Xu, Povey, Mangu & Zhu, 2011)，這樣效能的進步歸功於下列幾點，各別模型可以修正彼此的錯誤；減少選擇到較差模型的可能性；增加整體模型搜尋時的假說空間(Dietterich, 2000)，用以修正訓練時的問題。如語料選擇(Data selection)、目標函數(Objective function)、模型(Model)。這裡我們期待利用整體學習增加的假說空間，解決在半監督式訓練時有限語料造成效能降低的問題。訓練的過程為各別訓練每個模型，接著在訓練結束後的階段加入合併模型的技術，讓模型修正彼此的錯誤，進一步提升效能。這裡我們採用兩種不同層級的合併方法，音框層級的合併(Frame-level combination or score fusion) (Deng & Platt, 2014)，以及假說層級的合併(Hypothesis-level combination) (Fiscus, 1997) (Xu *et al*., 2011)。在(Senior, Sak, Quitry, Sainath & Rao, 2015)的研究中音框層級合併無助於 CTC 的表現，而 LF-MMI 被視為 CTC 的延伸，因此探討半監督式 LF-MMI 的合併結果是具有價值的事情。

本論文的實作目的便是在語料缺乏的半監督式環境下，使用負條件熵與詞圖輔助 LF-MMI 的訓練，並利用模型合併技術，進一步提升模型的辨識結果。我們希望即使在語料不足的情況下，仍能達到不錯的辨識效果，甚至媲美原先轉寫語料的訓練結果。

2. 相關文獻回顧 (RELATED WORK)

本篇論文於半監督式的環境研究 LF-MMI 的表現，並進一步利用模型合併的技術提升效能。其中相關研究可分為三大方向：半監督式聲學模型；MMI 與 LF-MMI 的改變；最後則是模型合併技術。

2.1 半監督式訓練於聲學模型 (Semi-supervised Acoustic Modeling)

半監督式聲學模型目的是解決下列問題：低資源的語料庫、大量的未轉寫語料、測試語料與訓練語料的不匹配。首先，充足的語料庫是讓目前最新穎的 ASR 系統可以表現優異的原因之一，但我們擁有的轉寫語料通常不大；其次，儘管取得足夠的轉寫語料十分困難，但取得未轉寫語料卻容易得多，要如何利用好大量的未轉寫語料便成了重要的問題；最後，也是最廣泛的問題，訓練與測試環境的不匹配。相關研究裡最常見的方法為自我訓練(Self-training) (Zavaliagkos *et al*., 1998) (Vesely *et al*., 2013) (Grezl & Karafiát, 2013) (Zhang *et al*., 2014)。自我訓練的步驟分為兩階段，首先使用轉寫語料訓練種子模型直到穩定(通常為 CE 訓練，但也可加入鑑別式訓練)，第二階段則利用種子模型辨識未轉寫語料，加入信心過濾器(Lamel *et al*., 2002) (Chan & Woodland, 2004) (Liu *et al*., 2007) 篩選

訓練語料，過濾可能會影響訓練的語料，再重新訓練模型。而信心過濾器(Confidence filter)可在音框層級(Vesely *et al*., 2013)、詞層級(Thomas *et al*., 2013)、語句層級(Grezl & Karafiát, 2013) (Vesely *et al*., 2013) (Thomas *et al*., 2013)多種層級上進行。由於鑑別式訓練對於訓練語句的正確性十分敏感(Mathias *et al*., 2005) (Yu *et al*., 2010) (Cui *et al*., 2011)，因此過往的研究著重於信心過濾器的選擇。在(Liu *et al*., 2007) (Mathias *et al*., 2005)中將音框層級的信心過濾器加入鑑別式訓練。而在(Walker *et al*., 2017)在鑑別式訓練中加入語句層級的信心過濾器以及後處理最佳辨識結果。(Manohar *et al*., 2018) 則將詞圖加入在半監督式 LF-MMI 的訓練。

2.2 Lattice-free Maximum Mutual Information

2.2.1 MMI

條件最大化可能性(Conditional maximum likelihood, CML) (Nadas, 1983)的目標函數是在給予聲學特徵 O 和模型參數下，估測轉寫(Transcript)的對數可能性。分子為正確轉寫(Reference transcript)的機率，而分母為所有可能答案的機率。因為一些歷史的原因，CML 成為了我們目前熟知的 MMI (Bahl *et al*., 1986)，式子如下：

$$\mathcal{F}^{\text{MMI}} = \sum_u \log P(S_u|O_u, \lambda) \tag{1}$$

式(1)的 u 為語句，S_u為語句 u 的正確狀態序列(Reference state sequences)，O_u為語句 u 的聲學特徵，λ為模型參數。MMI 的目標便是最大化上述的式子。詳細的計算可透過貝式定理(Bayes' theorem)拆解成式(2)：

$$\mathcal{F}^{\text{MMI}} = \sum_u \log \frac{P(O_u|S_u,\lambda)P(S_u)}{\sum_{S'} P(O_u|S',\lambda)P(S')} \tag{2}$$

在式(2)中，S'_u為語句 u 的競爭狀態序列(Competing state sequence)。可透過鑑別式訓練，將模型目標函數定義成接近正確狀態序列和遠離競爭狀態序列。若要計算語句 u 在時間 t，而輸出層為$\mathbf{y}(u,t)$，則可偏微分式(1)：

$$\frac{\partial \mathcal{F}^{MMI}}{\partial y(u,t)} = \delta_{S_u:\mathbf{y}(u,t)} - \gamma^{DEN}_{\mathbf{y}(u,t)} \tag{3}$$

式(3)中的 $\delta_{S_u:\mathbf{y}(u,t)}$為指示函數(Indicator function)，當$\mathbf{y}(u,t)$的輸出屬於正確狀態序列S_u時為 1，反之則為 0。$\gamma^{DEN}_{\mathbf{y}(u,t)}$則表示$\mathbf{y}(u,t)$為正確狀態序列的事後機率(Posterior)，可表示如下：

$$\gamma^{DEN}_{\mathbf{y}(u,t)} = \sum_S \delta_{S:\mathbf{y}(u,t)} P(S|O_u, \lambda)$$

$$= \frac{\sum_S \delta_{S:\mathbf{y}(u,t)} P(O_u|S,\lambda)P(S)}{\sum_{S'} P(O_u|S')P(S')} \tag{4}$$

式(3)裡最繁雜的問題便呈現在式(4)，式(4)為計算所有可能存在於假說的競爭序列。在較早期的研究裡，學者們利用 CE 作預先訓練限制假說空間的大小，使得 MMI 的競爭序列可由有限的詞圖中產生。這樣是二階段的訓練取得了不錯的成果，但為了產生可能序列的詞圖，不僅需要多餘的 CE 訓練，且受限於 CE 的訓練，第二階段的 MMI 訓練僅能找到一階段 CE 訓練結果的局部最佳解。LF-MMI 主要解決的是式(4)的計算，使得不用一階段 CE 預先訓練產生詞圖，即可直接計算所有可能的競爭訓練。

2.2.2 LF-MMI

近年來，(Povey *et al*., 2016)中提出 LF-MMI，避開需要 CE 訓練產生詞圖的冗餘步驟，可視為 CTC (Graves *et al.*, 2006)的延伸架構。主要改變有四種，利用 4 連音素語言模型(Four-gram phone LM) 且不會退化小於 3 連音素語言模型(Tri-gram phone LM)，取代傳統鑑別式訓練時的詞圖，使得搜尋的假說空間減少；提出多種避免過度擬合(Overfitting)的訓練技巧，如多任務架構的 CE 正則項(CE-based regularization)，讓訓練能同時最佳化 LF-MMI 和 CE；採用類似 CTC 的兩個左到右狀態 HMM (2-state left-to-right HMM)的拓樸架構，且第一個狀態沒有 self-loop，相似於 CTC 的空白輸出(Blank)；最後的假設則是類神經網路的輸出沒有軟式最大化(Softmax)，因此不是狀態的事後機率，而是偽對數可能性(Pseudo log likelihood)。前兩者的改變使得 MMI 的訓練可在一階段的聲學模型便加入訓練，且式(4)也不是計算在候選詞圖上，而是完整搜尋(Full search)所有的可能序列，最終效果可媲美甚至超越兩階段的聲學模型訓練。後兩者的改變則是模仿 CTC 的架構，因此 LF-MMI 也可視為 CTC 的延伸架構。

2.3 模型合併技術 (Model Combination)

整體模型可藉由多個模型互補的假說空間，用以修正單一模型難以解決的問題。如語料選擇(Data selection)、目標函數(Objective function)、模型(Model)。為了實現最大的組合增益，在整體系統裡的模型必須單獨且準確(Dietterich, 2000)。在 DNN-HMM 的模型中，可引入五種多樣性。特徵多樣性，如隨機特徵投影(Random feature projection)；架構多樣性，如 DNN、LSTM；模型參數多樣性，如隨機初始化(Random Initialization)；輸出目標多樣性，如隨機森林(Random forest) (Dietterich, 2000)；轉換模型(Transition model)和語言模型(Language model)的多樣性。過往在語音辨識的模型合併可分為兩種，假說層級合併(Hypothesis-level combination) (Evermann & Woodland, 2000) (Xu *et al*., 2011)和音框層級合併(Frame-level combination or score fusion) (Deng & Platt, 2014)。ROVER(Fiscus, 1997).利用多個 ASR 產生的可能轉寫(n-best)結果的聯集，透過詞頻(Word frequency)或信心分數(Confidence score)合併成單詞轉換網路(Word translation network)，自動重新計搜索生成的網路，選擇得分最高的輸出序列；而(Xu *et al*., 2011)則是將多個模型的解碼結果的詞圖取聯集，得到一個新的詞圖。結果證明可以在最小化貝式決策風險解碼(Minimum Bayes-risk decoding)中，改進貝式風險的界限；在(Deng & Platt, 2014)中結合聲學模型的網路輸出並進行解碼；(Evermann & Woodland, 2000)利用維特比(Viterbi)產生

的詞圖與混淆網路(Confusion network)，其提供最可能的單詞假設及其相關單詞的事後機率，實驗結果優於 ROVER 的性能。雖然上述的幾種模型合併皆證明可優於單一模型的表現。然而，在過往的研究中，由於 CTC 的輸出為高峰分佈(Peaky distribution)，音框層級合併不僅無助於 CTC 模型，甚至會惡化原先的表現(Senioret *et al*., 2015)。而 LF-MMI 被視為 CTC 的延伸，且音框層級合併比假說層級合併來得更高效，因此探討半監督式 LF-MMI 的音框合併是具有價值的事情。

3. 基於 LF-MMI 的半監督式訓練 (SEMI-SUPERVISED TRAINING USING LF-MMI)

3.1 半監督式LF-MMI (Semi-supervised LF-MMI)

在有參考轉寫的情況下，傳統 MMI 估測方式為 CML，計算的式子為式(2)。然而在半監督式的環境下，未轉寫語料的自動轉寫(分子項)未必正確。因此在半監督式環境下，我們可將原先的式(2)改寫如下：

$$\mathcal{F}^{\text{SemiMMI}} = \sum_{s_u \in \mathcal{H}} \log \frac{P(O_u|S_u,\lambda)P(S_u)}{\sum_{S'} P(O_u|S',\lambda)P(S')} \tag{5}$$

上式的 u 為語句，S_u為語句 u 的正確狀態序列，但在半監督式環境下的S_u來自於種子模型產生的假說$\mathcal{H}$，因此不能保證其正確性。O_u為語句 u 的聲學特徵。S' 為語句 u 的競爭狀態序列，早期的聲學模型透過 CE 第一階段的訓練限制產生競爭序列的假說空間，使得競爭的序列只能從 CE 訓練後的詞圖中產生。而 LF-MMI 透過一些實作上的機制避開上述冗餘的步驟，可以在訓練時直接計算所有的競爭序列。

當我們計算式(5)分子項的正確序列，與過往只取最佳辨識結果的計算方式不同，而是將整個詞圖加入計算，透過設定光束(Beam)保留搜尋時的數量。保留越多就越可能搜尋到最佳答案，但同時會增長計算複雜度。其餘實驗設定與(Povey *et al*., 2016)中一致。

3.2 條件熵 (Conditional Entropy)

前一段中提到正確序列的S_u來自於種子模型產生的假說$\mathcal{H}$，我們不能保證其分子項的正確性。因此直接加入第二階段訓練是危險的行為，甚至會惡化原先模型的表現。在過往的研究中為了解決此問題，最常見的便是在第一階段和第二階段中間，加入信心過濾器排除分數過低的語句，用以確保訓練語句的「品質」，但挑選過濾器的門檻值並不容易且非常浪費訓練時間。有別於以往的排除訓練語句，我們希望在訓練時仍保留分數較低的語句，並與分數高的語句一起訓練。這裡我們在原先的向前向後算法(Forward-backward algorithm)加入了權重機制，並將原先的式(1)改寫如下：

$$\mathcal{F}^{\text{NCE}} = \sum_{u \in \mathcal{H}} \sum_{s} P(S_u|O_u,\lambda) \log P(S_u|O_u,\lambda) \tag{6}$$

上式為未轉寫語料的估測方式。式(6)與式(1)相似，但在計算可能的正確序列S_u時，加入

了$P(S_u|O_u,\lambda)$的權重於詞圖中，用以改變詞圖中的分數矩陣。式(6)進一步化簡成下式：

$$\mathcal{F}^{\mathrm{NCE}} = -\sum_u H(S_u|O_u,\lambda) \tag{7}$$

式(7)便是 NCE(Grandvalet & Bengio, 2005) (Huang & Hasegawa-Johnson, 2010)。我們可以稱式(7)為給予模型參數λ和聲學特徵O_u條件下，參考轉寫序列S_u的條件熵$H(S_u|O_u,\lambda)$。式(7)的改變可利用資訊量對轉寫的「品質」建模，並且自然地加入 LF-MMI 目標函數，在不用信心過濾器的情況下也能提升訓練結果。

4. 模型合併技術應用於聲學模型 (MODEL COMBINATION OF ACOUSITIC MODELING)

模型合併的成果可透過修正各別模型的錯誤、減少較差選擇的可能性、增加模型搜尋時的假說空間來達到更好的模型效能。在聲學模型的合併可分為音框層級和假說層級。兩者的比較紀錄於表 1，前者因為是在聲學模型的輸出直接合併，因此具有較快的即時性。後者則是在模型產生詞圖後合併，與解碼標準更相關，有較好的辨識結果。

表 1. 合併方式比較
[Table 1. Frame combination vs. hypothesis combination]

面向	音框層級合併	假說層級合併
解碼詞圖	• 強制共享詞圖中的時間同步的狀態 • 僅需處理整體的單個詞圖	• 不需要時間同步的狀態 • 需先各別處理整體模型數的詞圖再合併
事後機率	• 旨在產生更好的音框事後機率或觀察可能性，從而產生更好的詞圖	• 旨在產生更好的假說事後機率，其與解碼標準更密切相關

4.1 音框層級合併 (Frame-level Combination)

音框層級合併是根據某個時間點中音框輸出的對數可能性(Log likelihood)，給予不同的權重後合併。因為合併的是音框，所以必須保持輸出時間的同步。聲學模型的對數可能性是類神經網路的輸出。式子如下：

$$P(S_{ut}|O_{ut},\lambda) = \sum_{m=1}^{M} \alpha_m P(S_{ut}|O_{ut},\lambda_m) \tag{8}$$

式(8)中S_{ut}為類神經網路的輸出，代表語句 u 在時間點 t 的狀態 S 的機率。M 為合併的模型總數，α_m為各別模型混和權重，且$\alpha_m \geq 0$, $\sum_{\mathrm{m=1}}^{\mathrm{M}} \alpha_m = 1$。α相似於在(Dietterich, 2000)中的利用對角線矩陣(Diagonal matrices)對各別模型的線性合併(Linear ensemble)。合併後的音框事後機率(Frame posterior)會當成隱藏式馬可夫模型(Hidden Markov model, HMM)的聲學特徵 O 的對數可能性，並進行標準的解碼程序。

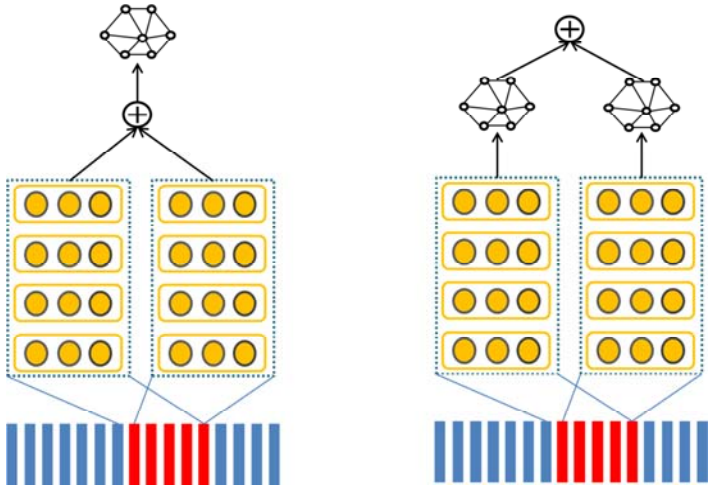

圖 1. 兩種層級的聲學分數合併。左方為音框層級，右方為假說層級。
[Figure 1. Frame-level vs. Hypothesis-level combination]

4.2 假說層級合併 (Hypothesis-level Combination)

假說層級合併則是利用 ASR 系統經過一般的解碼機制產生的詞圖，給予不同權重和損失函數進行合併。相較於音框層級合併，假說層級合併可允許非同步時間輸出，但因為需要合併各別 ASR 的輸出結果，因此較為費時。

$$h_u^* = \text{argmin}_{h'_u} \sum_{h_u} \mathcal{L}(h_u, h'_u) \sum_{m=1}^{M} \beta_m P(h_u|O_u, \lambda_m) \quad (9)$$

其中$\mathrm{h_u}$為各別 ASR 系統解碼時產生的詞序列。M為合併的模型總數。β_m為各別模型混和權重，且 $\beta_m \geq 0$,$\sum_{\mathrm{m=1}}^{\mathrm{M}} \beta_m = 1$。$L$為詞層級的損失函數，這裡使用編輯距離(Edit distance)。式(9)可理解為各別 ASR 產生詞圖的聯集，並透過最小化貝式決策風險對合併的詞圖解碼。音框層級和假說層級的示意圖可參考圖 1。

5. 實驗 (EXPERIMENTS)

5.1 實驗設定 (Experimental Setup)

實驗使用 Kaldi (Povey *et al.*, 2011)語音識別工具包。語料庫為 AMI (Augmented Multi-party Interaction) (McCowan *et al.*, 2005)。AMI 語料庫是來自歐盟發起的會議瀏覽(Meeting browser)計畫，其中包含情境會議 (Scenario meetings)和非情境的會議，情境會議是指明確的會議目標、會議間彼此有關連，如其中一個會議的主題為討論電視遙控器的設計；另一方面的非情境會議 (Non-scenario meetings)則反之，較沒有明確的主題，主要為英國愛丁堡大學、瑞士 Idiap 研究中心、荷蘭 TNO 人為因素研究所的學生或研究者組成討論的小型會議，如線性代數、微積分等。AMI 的語料庫也包含了影像、文字、語音，影像紀錄的是會議視角、投影機畫面和白板書寫記錄；文字有語音轉寫、對話特性，

表 2. AMI 會議之訓練、發展與測試集
[Table 2. The table shows that some basic statistics of the AMI corpus]

語料單位	訓練集	發展集	測試集 1	測試集 2	總計
小時數	70.09	7.81	8.71	8.97	95.79
語句數	97,222	10,882	13,059	12,612	133,775

可用於摘要、情緒與對話；最後是語音的部分，可分為耳掛式近距離麥克風、固定式遠距離麥克風。本實驗只用到了語音語料。表 2 為 AMI 的基本統計數據，由於原先 AMI 的訓練中並沒有用到發展集，因此實際訓練集為訓練集加發展集。我們用詞錯誤率(Word error rate, WER)和詞修復率(WER recovery rate, WRR)作為評估。WRR 如下：

$$\text{WRR} = \frac{BaslineWER - SemisupWER}{BaselineWER - OracleWER} \tag{10}$$

5.1.1 半監督式實驗流程與設定 (Semi-supervised Setup)

本實驗將 AMI 原先的訓練集切割成 16 小時的監督(轉寫)語料和 62 小時的非監督(未轉寫)語料，發展集和測試集。整體實驗的訓練為兩階段，第一階段為利用 16 小時的監督語料訓練種子模型，以及再使用 62 小時的非監督語料提升模型效能。整體實驗的詳細架構可參考圖 2。LF-MMI 的設定與(Povey *et al.*, 2016)一樣，特徵是 40 維 MFCC 和 100 維的 i-vector，類神經網路是使用時間延遲網路(Time-delay neural network, TDNN) (Peddinti, Povey & Khudanpur, 2015)。實驗分為訓練準則的有效性，以及後處理的模型合併。這裡需要注意的是合併時的權重皆為模型數量的倒數(e.g. M 個模型，權重為 1/M) 。

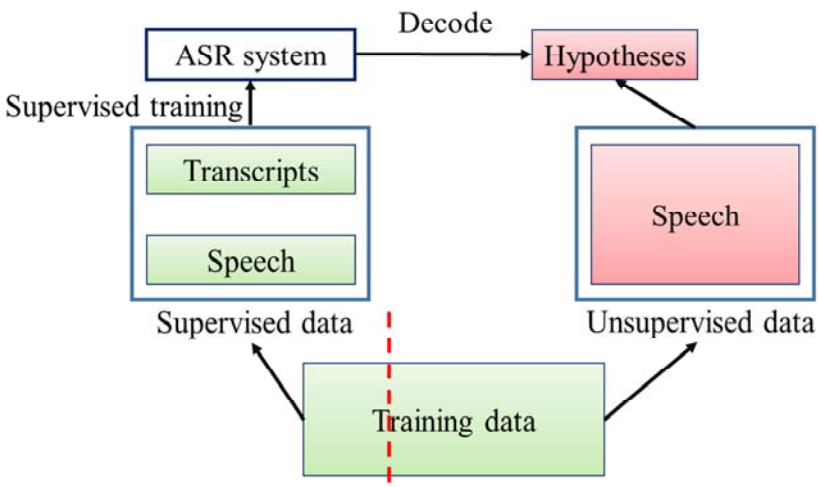

圖 2. 整體實驗架構
[Figure 2. A Flow chart of the experimental design.]

5.2 實驗結果與分析 (Results and Discussion)

5.2.1 加入NCE與詞圖的影響 (NCE and Lattice for supervision)

表 3 中呈現的是是否加入 NCE 權重和詞圖於訓練中的結果。第二欄中的 lm-scale 為第二階段的語音模型重新評分的縮放常數、Beam 為搜尋詞圖保留的種子個數、Tol 為在訓練時用到的詞圖的允許音框位移(Frame shift)，1 代表 30ms，這裡基於經驗上的設置，並無特地調動。第一欄由上到下，Baseline 是只用 16 小時訓練的聲學模型；No weight (NW)則是直接加入 62 小時未轉寫的語料；Best Phone Path(BPP)基於 NW 之上，加入 NCE 的權重的最佳辨識結果；Lattice for supervision (LS)則是基於 BPP 並加入詞圖。為了計算方便，LS 將原先的詞圖切成 1.5 秒的塊(Chunks)，再用向前向後算法計算；Oracle 則是將 78 小時的語料直接加入訓練。Dev 和 Eval 分別是測試集 1 和測試集 2。從實驗中的結果可以看出，直接進行傳統二階段訓練時，便可稍微提升 WRR 為 24%，這要歸功於良好的 LF-MMI 種子模型，只用 16 小時的訓練語料便達到尚可的辨識率，但 WRR 提升不夠可能是未轉寫語料相對使用太少導致修復率不佳。使用 NCE 可進一步地提升 WRR 至 33%，這也證明了加入音框層級的條件熵能有效輔助半監督式訓練，因此後續實驗探討皆會以 NCE 為主；最後則是加入整個詞圖後可將 WRR 提升至 45%，可看出多保留幾個搜尋的可能性後，增加的計算空間能輔助模型的訓練，進一步提升辨識結果。

表 3. 加入 NCE 與詞圖的影響
[Table 3. Negative conditional entropy and lattice for supervision]

Supervision	lm-scale	Beam	Tol	Dev	Eval	WRR
Baseline	-	-	-	27.2	27.8	-
NW	0	0	1	26.2	26.8	24%
BPP	0	0	1	26.0	26.2	33%
LS	0.5	4	1	**25.5**	**25.7**	**45%**
Oracle	-	-	-	23.5	23.1	-

5.2.2 模型合併於半監督式訓練 (Model combination in semi-supervised training)

這裡探討不同的訓練機制下，模型合併的成效。可分為音框層級的合併及假說層級的合併。雖然合併的增益主要來自模型的各別準確與多樣性，但這裡主要是探討整體學習合併的成效，因此採用簡單地調整訓練方式達成多樣性。實驗記錄於表 5，從第一列由左至右分別為不同方式訓練下的 LF-MMI 聲學模型，比較紀錄於表 4；FCOMB 和 HCOMB 則分別是音框層級和假說層級的合併；Test 則是 Dev 和 Eval 之 WER 的相加取平均。從上述實驗中可看出加入 Proportional shrink 和 L2-regularization 可比原先的訓練再降低詞錯誤率 1%，而調整初始化也會些微地影響 WER，也再一次證實了 LF-MMI 容易過度擬合的問題。另一方面，雖然在 WER 上看到成效，不過在 WRR 上則沒有顯著差異。這可

以看出這兩種方法的泛用性，不會受到不同的訓練準則影響進步成效。另一方面，從合併的觀點來看，音框層級合併在各個階段，比起單一系統的準度皆能提升 0.5 至 1.5 的 WER，證明了合併模型的技術應用於半監督式環境的有效性。另一方面，假說層級的合併效果更勝於音框層級的合併，這樣的結果歸功於假說合併是做在各個 ASR 的詞圖上，而音框合併則是類神經網路的輸出。比起音框合併，假說層級的合併更接近於辨識目標的詞，因此效果較好。但另一方面，音框合併雖然在效果上略輸假說合併，但由於可直接在音框階段就合併，不需要各別 ASR 產生詞圖，因此有更好的即時性。

表 4. 不同網路的設定差異
[Table 4. The table shows that different training criteria in the experiment. We combine four TDNN model generated by different random seed at both frame level and hypothesis level.]

	TDNN0	TDNN1	TDNN2	TDNN3
設定差異 (與 TDNN0)	基於表 2 的設定	+Proportional shrink	+L2-regularization	與 TDNN1 初始化不同

表 5. 音框和假說層級的聲學分數結合
[Table 5. Results on model combinations including frame-level and hypothesis-level combination.]

	TDNN1	TDNN1	TDNN2	TDNN3	FCOMB	HCOMB
	Test	Test	Test	Test	Test	Test
Baseline	27.5	26.7	26.5	26.5	25.6	**25.5**
BPP	26.1	25.2	25.5	25.1	24.5	**24.4**
LS	25.6	25.1	25.5	24.9	24.4	**24.2**
Oracle	23.3	22.5	22.8	22.5	21.5	**21.3**

5.2.3 不同半監督式準則的模型合併 (Model Combination and Semi-supervised Training)

表 6 中的第一列分別為音框層級合併與假說層級合併。第二欄由上至下為 NW、BPP 和 LS。而這裡合併的模型是表二的訓練結果。合併的方式為種子模型與 NW；種子模型、NW 和 BPP；種子模型、NW、BPP 和 LS。從實驗的結果中可看出基於音框層級與假說層級的合併十分有效，且假說層級的合併在大部分的情況下，仍勝過音框層級的合併，少部分是兩者持平。這裡我們可分析 NW、BPP 和 LS 這三種方法彼此的互補，最好的 WRR 為 60.8%。這些不同準則的合併雖有助於 WER 與 WRR 的提升，但可從實驗結果中觀察到 WER 進步的幅度約為 0.5，沒有比各別訓練多個模型並在同個半監督準則合併(表 5)來得更好。因此我們可得知，半監督準則在各別準確與多樣性上，相較使用 Proportional shrink 和 L2-regularization 來得小。儘管如此，不論是那種方式，在半監督式

環境下，我們皆可透過簡單地改變超參數，再以音框層級與假說層級的合併達到更好的辨識結果。

表 6. 不同半監督準則的模型合併
[Table 6. Results on model combination in conjunction with different semi-supervised criteria]

		F-COMB			H-COMB		
		Dev	Eval	WRR	Dev	Eval	WRR
TDNN-0	+NW	25.9	26.1	35%	25.7	26.0	39%
	+BPP	25.4	25.5	48%	25.3	25.5	50%
	+LS	25.1	25.1	57%	**24.9**	**25.0**	**60%**

6. 結論 (CONCLUSION AND FUTURE WORK)

本論文探討兩種思路於半監督式 LF-MMI。其一，利用 NCE 權重與詞圖模擬未轉寫語料的不確定性；其二，探討不同層級的合併，較快的音框層級合併和較準的假說層級合併。實驗結果得知，在無需信心過濾器的語料挑選下，這兩種思路可直接應用於半監督式 LF-MMI，並能有效地降低 WER 與提升 WRR 且相輔相成，最終 WRR 為 60.8%。未來的研究方向會針對有效性與即時性兩個方向繼續研究。根據這次的實驗結果，我們得知可透過模擬不確定性或更改參數的模型合併提升準度，未來會繼續朝如何利用未轉寫語料與互補多樣性的合併繼續研究，如 1) 利用不同模型種類，以產生更好的互補性。另一方面，2) 轉寫語料與未轉寫語料的比例，要到多少才能達成最好的 WRR；再者，儘管這次透過模型合併得到了不錯的結果，但同時也付出相較於單一 ASR 系統更高昂的運算資源，即便是相較於假說合併較為輕量的音框合併也是如此。因此 3) 未來會加入模型壓縮(Model combination)的技術，期許有一天能夠以少量的轉寫語料便達到有效且即時的辨識結果。

參考文獻 (REFERENCES)

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

Chan, H. Y., & Woodland, P. (2004). Improving broadcast news transcription by lightly supervised discriminative training. In *Proceedings of ICASSP 2004*. doi: 10.1109/ICASSP.2004.1326091

Cui, X., Huang, J., & Chien, J.-T. (2011). Multi-view and multiobjective semi-supervised learning for large vocabulary continuous speech recognition. In *Proceedings of ICASSP 2011*. doi: 10.1109/ICASSP.2011.5947396

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, *20*(1), 30-42. doi: 10.1109/TASL.2011.2134090

Deng, L., & Platt, J. C. (2014). Ensemble deep learning for speech recognition. In *Proceedings of Interspeech 2014*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of International workshop on MCS 2000*, 1-15.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139-157. doi: 10.1023/A:1007607513941

Evermann, G., & Woodland, P. C. (2000). Posterior probability decoding, confidence estimation and system combination. In *Proceedings of STW 2000*.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In *Proceedings of ASRU 1997*, 347-352. doi: 10.1109/ASRU.1997.659110

Gibson, M., & Hain, T. (2006). Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. In *Proceedings of Interspeech 2006*.

Grandvalet, Y., & Bengio , Y. (2005). Semi-supervised learning by entropy minimization. In *Proceedings of NIPS 2005*, 529-536.

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*. doi: 10.1145/1143844.1143891

Grezl, F., & Karafiát, M. (2013). Semi-supervised bootstrapping approach for neural network feature extractor training. In *Proceedingd of ASRU 2013*. doi: 10.1109/ASRU.2013.6707775

Huang, J.-T., & Hasegawa-Johnson, M. (2010). Semi-supervised training of gaussian mixture models by conditional entropy minimization. In *Proceedings of INTERSPEECH 2010*, 1353-1356.

Juang, B.-H., Hou, W., & Lee, C.-H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, *5*(3), 257-265. doi: 10.1109/89.568732

Kaiser, J., Horvat, B., & Kacic, Z. (2000). A novel loss function for the overall risk criterion based discriminative training of HMM models. In *Proceedings of ICSLP 2000*, 887-890.

Lamel, L., Gauvain, J.-L., & Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech & Language*, *16*(1), 115-129. doi: 10.1006/csla.2001.0186

Liu, S.-H., Chu, F.-H., Lin, S.-H., & Chen, B. (2007). Investigating data selection for minimum phone error training of acoustic models. In *Proceedings of ICME 2007*. doi: 10.1109/ICME.2007.4284658

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

Manohar, V., Povey, D., & Khudanpur, S. (2015). Semi-supervised maximum mutual information training of deep neural network acoustic models. In *Proceedings of Interspeech 2015*.

Mathias, L., Yegnanarayanan, G., & Fritsch, J. (2005). Discriminative training of acoustic models applied to domains with unreliable transcripts. In *Proceedings of ICASSP 2005*. doi: 10.1109/ICASSP.2005.1415062

McCowan, I., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., …Wellner, P. (2005). The ami meeting corpus. In *Proceedings of ICMTBR 2005*.

Nadas, A. (1983). A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *31*(4), 814-817. doi: 10.1109/TASSP.1983.1164173

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of Interspeech 2015*, 3214-3218.

Povey, D., & Woodland, P. C. (2002). Minimum phone error and i-smoothing for improved discriminative training. In *Proceedings of ICASSP 2002*. doi: 10.1109/ICASSP.2002.5743665

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for feature and model space discriminative training. In *Proceedings of ICASSP 2008*. doi: 10.1109/ICASSP.2008.4518545

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., 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

Pundak, G., & Sainath, T. N. (2016). Lower frame rate neural network acoustic models. In *Proceedings of Interspeech 2016*. doi: 10.21437/Interspeech.2016-275

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Proceedings of Interspeech 2011*.

Senior, A., Sak, H., Quitry, F. C., Sainath, T., & Rao, K. (2015). Acoustic modelling with CD CTC-SMBR LSTM RNNs. In *Proceedings of ASRU 2015*. doi: 10.1109/ASRU.2015.7404851

Thomas, S., Seltzer, M. L., Church, K., & Hermansky, H. (2013). Deep neural network features and semisupervised training for low resource speech recognition. In *Proceedings of ICASSP 2013*. doi: 10.1109/ICASSP.2013.6638959

Valtchev, V., Odell, J. J., Woodland, P. C., & Young, S. J. (1996). Lattice-based discriminative training for large vocabulary speech recognition. In *Proceedings of ICASSP 1996*. doi: 10.1109/ICASSP.1996.543193

Valtchev, V., Odell, J. J., Woodland, P. C., & Young, S. J. (1997). MMIE training of large vocabulary recognition systems. *Speech Communication*, *22*(4), 303-314. doi: 10.1016/S0167-6393(97)00029-0

Vesely, K., Hannemann, M., & Burget, L. (2013). Semi-supervised training of deep neural networks. In *Proceedings of ASRU 2013*. doi: 10.1109/ASRU.2013.6707741

Walker, S., Pedersen, M., Orife, I., & Flaks, J. (2017). Semi-supervised model training for unbounded conversational speech recognition. Retrieved from https://arxiv.org/abs/1705.09724

Woodland, P. C., & Povey, D. (2002). Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech & Language*, *16*(1), 25-47. doi: 10.1006/csla.2001.0182

Xu, H., Povey, D., Mangu, L., & Zhu, J. (2011). Minimum bayes risk decoding and system combination based on a recursion for edit distance. *Computer Speech and Language*, *25*(4), 802-828. doi: 10.1016/j.csl.2011.03.001

Yu, K., Gales, M., Wang, L., & Woodland, P. C. (2010). Unsupervised training and directed manual transcription for LVCSR. *Speech Communication*, *52*(7-8), 652-663. doi: 10.1016/j.specom.2010.02.014

Zavaliagkos, G., Siu, M., Colthurst, T., & Billa, J. (1998). Using untranscribed training data to improve performance. In *Proceedings of ICSLP 1998*.

Zhang, P., Liu, Y., & Hain, T. (2014). Semi-supervised DNN training in meeting recognition. In *Proceedings of SLT 2014*. doi: 10.1109/SLT.2014.7078564