The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 46-60

會議語音辨識使用語者資訊之語言模型調適技術

On the Use of Speaker-Aware Language Model Adaptation Techniques for Meeting Speech Recognition

陳映文 Ying-wen Chen, 羅天宏 Tien-hong Lo, 張修瑞 Hsiu-jui Chang, 趙偉成 Wei-Cheng Chao, 陳柏琳 Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{cliffchen, teinhonglo, 60647028S,60647061S, berlin}@ntnu.edu.tw

摘要

本論文試圖減緩會議語音辨識時語者間用語特性不同所造成的問題。多個語者的存在可能代表有多種的語言模式；更進一步地說，人們在講話時並沒有嚴格地遵循文法，而且通常會有說話延遲、停頓或個人慣用語以及其它獨特的說話方式。但是，過去會議語音辨識中的語言模型大都不會針對不同的語者進行調整，而是假設不同的語者間擁有相同的語言使用模式，將包含多個語者的文字轉寫合成一個訓練集，藉此訓練單一的語言模型。為突破此假設，本研究希望針對不同語者為語言模型的訓練和預測提供額外的資訊，即是語言模型的語者調適。本論文考慮兩種測試階段的情境—「已知語者」和「未知語者」，並提出了對應此兩種情境的語者特徵擷取方法，以及探討如何利用語者特徵來輔助語言模型的訓練。我們分別在中文和英文會議語音辨識任務進行一系列語言模型的語者調適實驗，其結果顯示本論文所提出的語言模型無論是在已知語者，還是未知語者情境下都有良好的表現，並且比基礎類神經網路語言模型有較佳的效能。

Abstract

This paper embarks on alleviating the problems caused by a multiple-speaker situation occurring frequently in a meeting for improved automatic speech recognition (ASR). There are a wide variety of ways for speakers to utter in the multiple-speaker situation. That is to say, people do not strictly follow the grammar when speaking and usually have a tendency to stutter while speaking, or often use personal idioms and some unique ways of speaking. Nevertheless, the existing language models employed in automatic transcription of meeting

recordings rarely account for these facts but instead assume that all speakers participating in a meeting share the same speaking style or word-usage behavior. In turn, a single language model is built with all the manual transcripts of utterances compiled from multiple speakers that were taken holistically as the training set. To relax such an assumption, we endeavor to augment additional information cues into the training phase and the prediction phase of language modeling to accommodate the variety of speaker-related characteristics, through the process of speaker adaptation for language modeling. To this end, two disparate scenarios, i.e., "known speakers" and "unknown speakers," for the prediction phase are taken into consideration for developing methods to extract speaker-related information cues to aid in the training of language models. Extensive experiments respectively carried out on automatic transcription of Mandarin and English meeting recordings show that the proposed language models along with different mechanisms for speaker adaption achieve good performance gains in relation to the baseline neural network based language model compared in this study.

關鍵詞：會議語音辨識、語言模型、語者調適、遞迴式類神經網路

Keywords: speech recognition, language modeling, speaker adaptation, recurrent neural networks.

一、緒論

語音辨識技術越趨成熟，生活中隨處可見其應用；自動語音辨識技術(Automatic Speech Recognition, ASR)讓電腦能聽得懂人類的語言，也就是試圖理解人類在發音上和用語上的規則與內容。語言模型是一種將文字文本模型化的技術，在語音辨識任務上，語言模型可以判斷一條詞序列是否符合訓練文本的所隱含的規則或規律性。常用於語音辨識的 N 連詞模型是利用 N 連詞的發生機率模型化文本中詞彙共同出現的關係。但是 N 連詞模型在面臨資料過於稀疏(Data Sparseness)時，便有難以估測的問題。原因是 N 連詞模型假設每個詞有各自獨立的語意，僅從統計的觀點來考慮彼此共同出現關係並估測模型[1]。類神經網路便可以解決這個問題，因為類神經網路語言模型可以習得詞語的分布式表示(Distributed Representation) [2]，使得詞彙間的語意關係能被表示出來。據此，透過接續串接的前饋式神經網路(Feedforward Neural Networks, FNN)或是遞迴式神經網路(Recurrent Neural Networks, RNN)可以來做預測。近幾年已有許多的研究嘗試改良前饋式神經網路，並用應用到不同的自然語言處理和語音辨識任務上[3-8]。

會議語音語料和一般新聞、朗讀等，較為嚴謹的語料有非常大的差異[9]，會議語

料通常有著冷門用詞、短語句、語言混雜使用和個人用語習慣等特性；有鑑於此，本研究嘗試針對不同語者有著不同的說話習慣的特性，來發展改良的語言模型架構與訓練方式。從統計的觀點來看，每種語言都有一套文法，但是實際上人們說話時，並不會嚴格遵守文法，且會擁有習慣用語或是口吃等獨特的說話方式；但是現今常見的用於語音辨識的語言模型，並不會針對不同語者做不同的調整，而是將整份訓練資料當作一種語言模式。所以我們希望根據不同的語者，對語言模型的訓練與預測提供額外的資訊，也就是對語言模型作語者調適(Speaker Adaptation)。為此，本論文考慮兩種測試階段的情境—「已知語者」和「未知語者」，並提出了對應此兩種情境的語者特徵擷取方法，以及探討如何利用語者特徵來輔助語言模型的訓練。

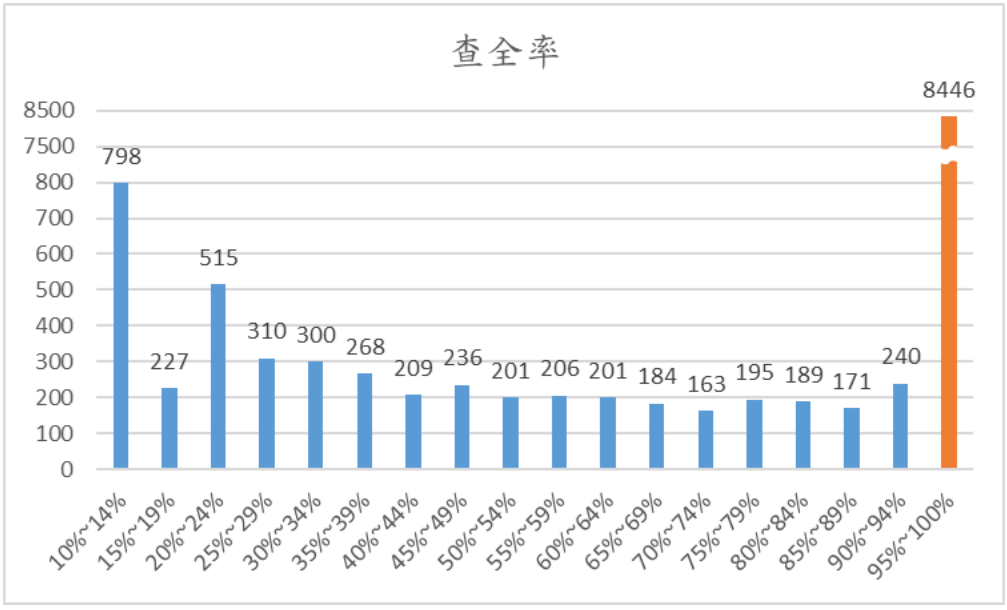
二、類神經網路語言模型相關研究

至今，類神經網路應用在語言模型變化繁多，最早的應用可回溯於 Yoshua Bengio 在 2003 年提出的類神經網路語言模型，在[2]中將 N 連詞的估測交由類神經網路來計算。另一方面，為了改善 N 連詞無法應付資料太過稀疏的缺點，他也將一個重要的概念—詞嵌入(Word Embeddings)，應用在類神經網路語言模型中。在 2010 年，有學者提出了遞迴式類神經網路語言模型(Recurrent Neural Network Language Model, RNNLM)[10]，讓語言模型不再受到 N 連詞的限制，歷史詞不再只能是 N-1 個詞。但是這個方法的缺點是模型難以訓練，並且容易遭受梯度消失或爆炸(Gradient Vanishing or Exploding)的問題[8]。Martin Sundermeyer 在 2012 年提出了利用長短期記憶(Long Short-Term Memory, LSTM)語言模型解決這個問題[11, 12]。從此，LSTM 語言模型一直是被視為最好的語言模型架構之一；但是也有些學者試圖使用其它架構建模，例如 Yann N. Dauphin 在 2016 年提出了在摺積式類神經網路上面加上閥門(Gate)，能稍微的改善語言模型效能，但是也因為它的網路過於複雜而產生訓練不易等問題[6]。

類神經網路語言模型因執行效率導致兩個問題，其一，難以用在語音辨識的第一階段解碼(First Pass Decoding)，所以通常用在第二階段的語言模型重新排序(Rescoring)；其二，只能將類神經網路應用在 M 最佳候選詞序列(M-best List)的重新排序，而不能應用在詞圖(Lattice)。為了解決這個問題，Xunying Liu 提出藉由減少詞圖的分支以加速詞圖中候選詞序列的重新排序。它屬於一種近似的方法，所以會使得結果略遜於候選詞序列重新排序，但卻能使類神經網路語言模型也可有效應用在詞圖重新排序[13]。

三、類神經語言模型應用於語音辨識

RNN 或 LSTM 語言模型在做預測時，需要經過數個矩陣運算，而不像傳統 *N* 連詞語言模型僅需透過查表來完成。如前面所提到，因為它們在執行效能上的限制，所以多半是使用在第一階段解碼過產生詞圖，並將詞圖上可能的候選詞序列重新計分。另一方面，同樣因為執行效率，詞圖需經過候選詞序列的刪減(Pruning)才能直接重新計分，缺點是會喪失一些候選詞序列的部分路徑。或是藉由 *N* 連詞語言模型產生 *M* 最佳候選詞序列(*M*-best List)，再將之重新排序。圖一是 AMI 英文會議語音辨識任務測試集(請參見第五節)中 1,000 候選詞序列被詞圖包含的查全率(Recall)。



圖一、1,000 候選詞序列被詞圖包含的查全率

可以由圖一發現，1,000 候選詞序列中有許多詞序列並沒有在詞圖重新排序時被計算，95~100%只佔全部的 64.68%。另一方面，從表一所示(AMI 英文會議語音辨識任務)詞錯誤率數據也可以發現，就此會議語音語料庫，詞圖重新排序並不會比 *M* 最佳候選詞序列重新排序的結果來的好。驗證過去研究的實驗一致 [13]。依據此結果，我們的語言模型調適實驗將會在 *M* 最佳候選詞序列重新排序時進行。

表一、詞圖重新排序與 *M* 最佳候選詞序列重新排序(*M*=1,000)之比較

詞錯誤率	發展集	測試集
1,000 候選詞序列重新排序	21.17%	20.41%
詞圖重新排序	21.53%	20.75%

四、語者調適於會議語音辨識所使用之語言模型

4.1 問題解析

在會議語音辨識任務中，待轉寫的語音紀錄常會包含多個語者，不同語者間其實存在用語和講話習慣的差異，但是過去用於會議語音辨識的語言模型，並不會考慮不同語者所造成語言使用行為不同的問題。以下將探討如何運用「訓練語料中的語者資訊」來輔助語言模型的訓練以達到語者調適的效果。

在第一階段的語音辨識的過程，通常基於當語音訊號X發生時詞序列W的事後機率來進行語言解碼，並可化簡成依$P(X|W)P(W)$來決定詞序列W是最終語音辨識輸出的可能性(或所謂的排序分數)，如式(1)所示。

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \propto P(X|W)P(W) \tag{1}$$

其中，$P(W)$通常以 N 連詞語言模型來估計，$P(X|W)$則可透過聲學模型來估計。藉由這兩個模型的相乘可算出候選詞序列W的分數，並經由語言解碼過程，適當地修剪候選詞序列而形成詞圖(Lattice)或 M 最佳候選詞序列(M-best List)。接著，在第二階段語言模型重新排序時，我們可使用類神經網路語言模型重新估測$P(W)$，而達到候選詞序重新排序的目的(以 M 最佳候選詞序重新排序為例)：

$$W^* = \underset{W' \in M-best}{\operatorname{argmax}} \; P(X|W')P(W') \tag{2}$$

接著，我們將說明如何運用語者資訊輔助語言模型的訓練，以達到語者調適的效果。

(一) 已知每一條詞序列對應的語者為k時(假設總共有K位語者)：

$$P(W') = \sum_{k' \in K} P(W'|k')P(k') \tag{3}$$

若其中 $P(k') := \begin{cases} 1, & k' = k \\ 0, & k' \neq k \end{cases}$，則

$$P(W') = \sum_{k' \in K} P(W'|k')P(k') = P(W'|k) \tag{4}$$

依據全機率定理(Total Probability Theorem)，$P(W')$可以寫成$\sum_{k' \in K} P(W'|k')P(k')$，而當

已知詞序列對應的語者為k時，便可得到具有語者資訊的機率式(也就是語者調適過後的語言模型)$P(W'|k)$來估測詞序列W'的機率。

(二) 未知每一條詞序列對應的語者時：

$$P(W') = \sum_{k' \in K} P(W'|k')P(k') = \sum_{k' \in K} P(W'|k') \sum_{W''} P(k'|W'')P(W'') \tag{5}$$

若其中 $P(W'') = \begin{cases} 1, & W'' = W' \\ 0, & W'' \neq W' \end{cases}$，則

$$P(W') = \sum_{k' \in K} P(W'|k')P(k'|W') \tag{6}$$

當未知詞序列的語者時，同樣依據全機率定理將$P(W')$展開；但是此時$P(k)$因為是未知語者，所以再次依據全機率公式將$P(k')$拆解成$\sum_{W''} P(k'|W'')P(W'')$。因為詞序列已知是$W'$，所以當$W''$為$W'$時，$P(W'') = 1$，其餘為 0。更具體地說，進行第二階段語言模型重新排序時，我們先對每個候選詞序列W'估測各語者k'發生的機率$P(k'|W')$，並使語者調適過後的語言模型$P(W'|k')$來共同估測$P(W')$。在下一節，我將說明如何估測$P(k|W')$和$P(W'|k)$，其中$P(k|W')$涉及到語者特徵的擷取，而$P(W'|k)$是探討如何將語者特徵運用在具語者調適性的語言模型訓練和測試。

4.2 語者特徵的擷取

本篇論文提出兩種情境的語者特徵擷取方法。第一種情境是假設語者已知，可先用所有語者各自對應的訓練文本擷取出專屬的特徵；每個候選詞序列的語者特徵便是直接使用特定語者已擷取好的語者特徵，來做為語言模型額外的輸入資訊。此種預先擷取特徵的方法稱為「語者用詞特徵模型(Speaker Word-Usage Characteristics Model)」。而第二種情境是假設語者未知，所以必須動態地從每個候選詞序列擷取出隱藏的語者特徵，再該語者特徵來輔助語言模型。為此，我們提出了動態產生語者特徵的方法，即「語者慣用語模型(Speaker Slang Model)」。以下會對兩種情境的模型方法做進一步的介紹。

4.2.1 語者用詞特徵模型(Speaker Word-Usage Characteristics Model)

我們希望能夠擷取出不同語者使用詞彙的特性或頻率資訊。為此，我們提出三種產生語者特徵的單詞模型：分別為，詞頻模型(TF-based Model)、基於機率式潛在語意分析模

型(PLSA-based Model) [14]和語者特殊用詞模型(Speaker Specific Model, SSM)。

(一) 基於詞頻模型(TF-based Model)

此模型希望表現出某一位語者經常使用詞彙的頻率資訊，將此語者所說過的所有語句s基於其詞頻建模成語言模型：

$$P(t|s) = \frac{c(t,s)}{\sum_{t' \in s} c(t',s)} \tag{7}$$

其中s是語句，$c(t,s)$是計算詞t在語句s的出現次數。然後再將每一個語句的語言模型線性結合(Linear Combination)，其中每一個語句模型的權重相等。「基於詞頻模型」雖能表現每位語者不同的用詞，但是會有以下兩項缺點：

1. 此模型的維度是詞典的大小，過於龐大且稀疏。為了解決這個問題，我們提出第二種語者用詞模型，「基於機率式潛在語意分析的模型(PLSA-based Model)」，此模型可將語者詞彙使用資訊投影至潛在的語意空間，以達到降維的效果。
2. 此模型因為只計算詞頻，所以背景詞(Background Word)會使語者模型間差異不大。有鑒於此，我們提出「特殊用詞模型」，藉濾掉頻繁出現的背景詞，提升語者模型之間的鑑別度。

(二) 基於機率式潛在語意分析的模型(PLSA-based Model)

有別於基於詞頻模型，PLSA[14]是藉由找出潛在語意空間，以重新估測語者模型

$$P(t|s) = \sum_{k=1}^{K} P(t|z_k)P(z_k|s) \tag{8}$$

其模型參數可藉由期望值最大化(Expectation Maximization, EM)演算法來估測，其目標函數(Objective Function) F_{PLSA}表示如下：

$$F_{\text{PLSA}} = \sum_{s \in S} \sum_{t \in V} c(t,s) \log \left(\sum_{k=1}^{K} P(t|z_k)P(z_k|s) \right) \tag{9}$$

式(9)的目標是找出能最大化F_{PLSA}的$P(t|z_k)$與$P(z_k|\text{s})$。為了達到降維的目的，所以我們取用估測得到的模型參數$p(z_k|\text{s})$做為語者的特徵。

(三) 語者特殊用詞模型(Speaker Specific Word Model, SSWM)

為了提升語者模型之間的鑑別度，我們提出意在減少背景詞彙影響的特殊用詞模型。假設每位語者模型間，詞彙使用規律可由語者特殊用詞模型和背景詞模型的組合表示：

$$P(t|s) = \sum_{x\in\{\mathrm{BG,SSWM}\}} \lambda_x P(t|\theta_x) \tag{10}$$

其中$P(t|\theta_{\mathrm{BG}})$代表背景詞模型，$P(t|\theta_{\mathrm{SSWM}})$代表語者特殊用詞模型。而語者特殊用詞模型的訓練目標函數F_{SSWM}可表示為

$$F_{\mathrm{SSWM}} = \sum_{s\in S}\sum_{t\in V} c(t,s)\, log\left(\sum_{x\in\{bg,sswm\}} \lambda_x P(t|\theta_x)\right) \tag{11}$$

根據式(11)的訓練目標函數，我們同樣可使用期望值最大化(EM)演算法來估測參數。在 **E** 步驟以現有的模型參數求得$P(\theta_x)$的期望值，基於 E 步驟得到的期望值，在 **M** 步驟最大化目標函數，重複直到收斂，便可以得到特殊用詞模型$P(t|\theta_{\mathrm{SSWM}})$。

E 步驟(Expectation Step)：

$$P(\theta_x) = \frac{\lambda_x P(t|\theta_x)}{\sum_{x'\epsilon\{\mathrm{BG,SSWM}\}} \lambda_{x'} P(t|\theta_{x'})} \tag{12}$$

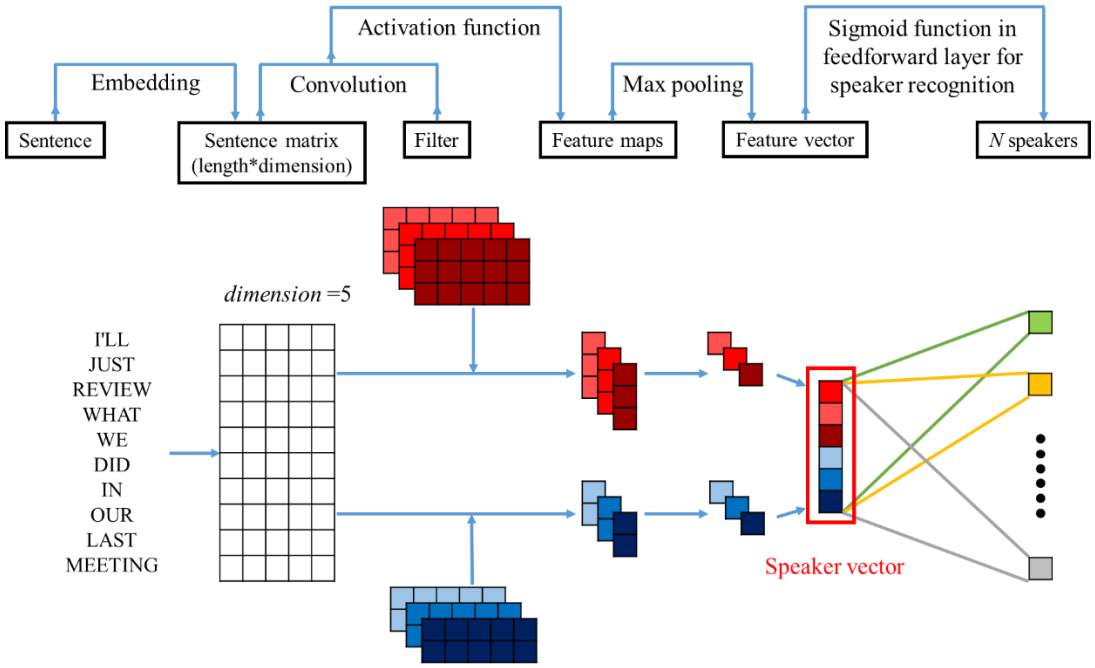
M 步驟(Maximization Step)：

$$P(t|\theta_{\mathrm{SSWM}}) = \frac{\sum_{s\epsilon S} c(t,s)\, P(\theta_{\mathrm{SSWM}})}{\sum_{t'\epsilon V}\sum_{s\epsilon S} c(t',s)\, P(\theta_{\mathrm{SSWM}})} \tag{13}$$

儘管單詞模型能夠表現語者的用詞習慣，但是這類單詞模型的方法具有兩項缺點：(1)無法表現語者的前後文用語習慣；(2)測試語句(候選詞序列)也必須有對應的語者資訊。

4.2.2 語者慣用語模型(Speaker Slang Model, SSM)

上述的方法注重的是語者的用詞特徵，用單詞模型的結構描述語者。但是除了用詞外，說話時人們也常會有習慣性的用語，且並不限於單一詞彙，例如：有的人說「對啊」時會習慣性的講兩次變成，「對啊、對啊」。為此，我們希望設計出能表示慣用語特徵的語者模型。在本研究，我們使用摺積式類神經網路(Convolutional Neural Network, CNN)對



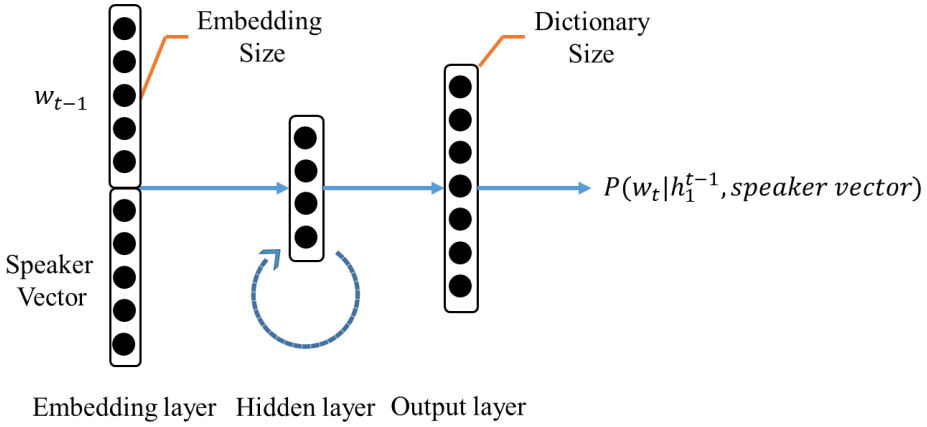
圖二、以摺積式類神經網路擷取語者慣用語特徵

每一個語句進行特徵擷取，並使用其隱藏層來做為語者慣用語特徵表示。

摺積式類神經網路的任務是語者識別。因為每個語句的內容並不一定會只來自某位語者(舉例來說，我們能確定某 A 句是由某語者所述，但是不能肯定 A 句不會出自其他語者)，所以在輸出層的設計上，我們不是選用分類的歸一化指數函數(Softmax)，而是針對每位語者對應各自的 S 函數(Sigmoid)。而在正例和反例上，正例為屬於該語者的語句；另一方面，我們藉由查詢相似度估計(Query Likelihood Estimation, QLE)計算與其相距最遠的語者，從中隨機挑選語句作為該語者的反例語句。圖二為以摺積式類神經網路擷取語者慣用語特徵的示意圖。

4.2.3 語者特徵用於語言模型調適

在獲得語者詞彙或慣用語特徵後，我們便可將之運用在語言模型的調適。過往常見於類神經網路的調適架構主要可分成兩類。第一類是添加輔助特徵到主任務的隱藏層（主任務為語言模型訓練），另一類則將特徵用於多任務學習(Multi-task Learning)的副任務。另一方面，在[15]的實驗結果卻指出，在輸入層添加輔助特徵可獲得更好的效果。其他神經網絡模型調適的相關研究也表明，將輔助特徵直接附加到主要特徵可帶來最佳效能，例如使用 i-vector 進行聲學模型語者調適[16]。因此在本研究中，我們採用的架構是將



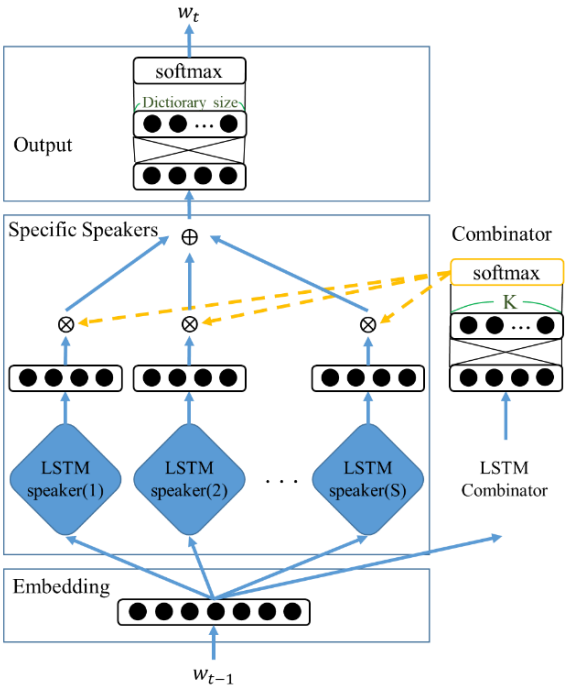
圖三、語者調適資訊融入 RNN(或 LSTM)為基礎的語言模型

語者特徵作為輔助特徵與主要特徵(一般詞彙特徵)共同輸入到隱藏層(如圖三所示)。

4.2.4 語者調適混和模型(Speaker Adaptive Mixture Model, SAMM)

相較於上述兩階段式方法，語者調適混和模型(SAMM)則是在訓練語言模型時，動態地擷取語者特徵，所以可直接使用於第二階段語言模型重新排序。SAMM 的主要想法是讓語言模型可自行動態地估測目前的語者，先訓練特定語者或具代表性語者(Specific or Representative Speakers)的各自語言模型，接著由組合器(Combinator)動態地為每個語句決定特定語者語言模型的權重，圖四為其示意圖。值得一提的是，在模型訓練時，特定語者的選取是基於語者們各自的 N 連詞語言模型(使用對應到此語者的訓練語句訓練而成)與背景 N 連詞語言模型(使用所有語者的訓練語句訓練而成)的差異來決定。在此研究，我們是選取差異最大的前 L 位語者來訓練 L 套特定語者語言模型。

從圖四可看出，在模型的測試階段，語者調適混和模型共用嵌入層和輸出層。當前一個時間點的詞輸入到此模型，先經過嵌入層投影至空間向量，接著經過各個特定語者語言模型和組合器的 LSTM 模型，然後輸出每位特定語者的隱藏層輸出以及組合權重，並再線性組合每位特定語者的隱藏層輸出和組合器輸出的權重。最後經過一個全連接層(Fully Connected Layer)與歸一化指數函數(Softmax)輸出下一個詞的機率。語者調適混和模型訓練階段可進一步分成下列幾個步驟：



圖四、SAMM 語者特徵擷取

第一步：使用所有訓練語料來訓練背景 LSTM 語言模型。

第二步：以背景 LSTM 語言模型為基礎，使用其參數做為每一套特定語者語言模型的初始化參數。接著，對於每一套特定語者語言模型固定其嵌入層和輸出層、保持這些參數不變，並僅對應到此語者的訓練語句來訓練每一套特定語者 LSTM 語言模型的隱藏層間網路的參數。

第三步：獲取所有特定語者 LSTM 語言模型參數，輸入前一階段的嵌入和輸出參數以初始化最終組合器模型，保持所有特定語者的 LSTM 語言模型參數以及嵌入層參數不變，並在混合器 LSTM 上訓練所有數據，同時微調輸出層參數。

此方法中的組合器輸出可視為一種語者特徵，來輔助動態地產生最終的語言模型，因為此模型的輸出是下一個詞的機率，所以也可以直接當作語音辨識第二階段語言模型重新排序所需之語言模型。

五、實驗結果與分析

5.1 實驗語料與設定

表二、華語會議語料語言模型之訓練、發展與測試集

語料型別	訓練集	發展集	測試集	總計
小時數(小時)	44.2	1.5	1.1	46.8
語句數(句)	42,998	1,267	1,019	45,284
語者數(位)	20	9 (1 無出現在訓練集)	6 (1 無出現在訓練集)	21

表三、AMI 會議之訓練、發展與測試集

語料型別	訓練集	發展集	測試集	總計
小時數(小時)	70.29	7.81	8.71	95.79
語句數(句)	97,222	10,882	13,059	133,775
語者數(位)	155	21 (19 無出現在訓練集)	16 (16 無出現在訓練集)	173

我們所使用的語料庫為「華語會議語料」以及「AMI 會議語料」[9]。其中，華語會議語料庫為國內企業所整理的語料庫。華語會議語料對於會議談話內容與參與人員的對話方式並沒有經過設計，而是貼近一般公司在實際開會中將會面臨的問題。例如聊到專業技術時，常會出現中英文夾雜的對話；發表談話時可能有停頓、口齒不清或口吃的現象。相較於 AMI 語料庫，華語會議語料更具挑戰性；表二是華語會議語料的詳細統計資訊。AMI 會議語料是由歐盟資助開發，AMI 團隊致力於研究和開發輔助團體互動的技術，其主要的目的是開發會議瀏覽器，使得會議記錄易於索引。該團隊收集了 AMI 會議語料，一系列已記錄的會議現在已提供給大眾做為研究開發使用，雖然數據集是專門為該工作所設計的，但它可用於語言學、組織和社會心理學、語音和語言工程、影音處理和多模式系統等多種不同目的，表三是 AMI 的詳細統計資訊。本研究語音辨識系統的發展是使用 Kaldi 工具；聲學模型是 Lattice-free Maximum Mutual Information(LF-MMI) [17]，第一階段中的語言模型是三連詞語言模型。類神經網路語言模型是由 Pytorch 實現。

5.2 實驗結果與探討

本論文第一組的實驗是實作在華語會議語料；表四是在此語料上的語言模型複雜度(Perplexity)和語音辨識字錯誤率(Character error rate, CER)結果。首先，從訓練和測試語句正確轉寫(Reference Transcription)的語言模型複雜度的實驗結果可看出，使用基礎 LSTM 語言模型結合傳統三連詞語言模型(或單獨使用基礎 LSTM 語言模型，如括號內數值所示)均能較使用傳統三連詞語言模型有較低的語言模型複雜度，也就是說有較佳

表四、華語會議語料上語言模型複雜度和語音辨識結果

華語會議語料	發展集		測試集	
	複雜度	CER	複雜度	CER
第一階段語音辨(三連詞語言模型)	205.11	20.19	210.26	17.23
第二階段語言模型重新排序(基礎 LSTM 語言模型)	161.20 (184.44)	16.89	165.44 (191.97)	15.91
+基於 TF 的語者特徵	158.99 (202.35)	16.84	163.26 (208.35)	15.84
+基於 PLSA 的語者特徵	156.20 (188.00)	16.75	160.93 (194.16)	15.86
+基於 SSWM 的語者特徵	158.41 (199.51)	16.84	162.94 (210.31)	15.91
+基於 CNN 的語者慣用語特徵	161.20 (255.85)	16.88	165.44 (264.11)	15.94
語者調適混和模型(SAMM)	158.52 (184.45)	16.75	161.71 (187.68)	15.89

的語言模型預測能力。尤其使用基礎 LSTM 語言模型結合傳統三連詞語言模型能較僅使用傳統三連詞語言模型，不論是在發展集或是測試集均能提供超過 20%的相對複雜度降低。其次，從語音辨識第二階段使用基礎 LSTM 語言模型結合傳統三連詞語言模型來對於第一階段產生的 1000 最佳候選詞序列重新排序的實驗，這樣的結合能對於發展集和測試集的字錯誤率分別為 16.89%與 15.91%的改進。

再者，我們觀察以語者用詞特徵來輔助訓練語言模型在華語會議語料上的實驗結果；在表四呈現出分別使用基於 TF 的語者特徵、基於 PLSA 的語者特徵、基於 SSWM 的語者特徵和基於 CNN 的語者慣用語特徵在語言模型複雜度和 CER 降低的表現。融入這些語者相關的輔助特徵的 LSTM 語言模型均能較基礎 LSTM 語言模型在語言複雜度的表現上有些微的提升，其中基於 PLSA 的語者特徵有最好的表現；而經期望值最大化(EM)演算法估測所得之基於 SSWM 的語者特徵也會較基於 TF 的語者特徵的表現來的好，說明了使用期望值最大化演算法降低背景詞影響的重要性。接著，在辨識的結果，三種語者相關特徵在發展集上都能帶來些微的改善，尤其以 PLSA 的效果最好。而在測試集上，是使用基於 TF 的語者特徵的表現最好；使用基於 CNN 的語者慣用語特徵反而使

表五、AMI 會議語料實驗結果

AMI 英文會議語料	**發展集**		**測試集**	
	複雜度	CER	複雜度	CER
第一階段語音辨識(三連詞語言模型)	85.19	23.25	76.44	23.02
第二階段語言模型重新排序(基礎 LSTM 語言模型)	68.02 (73.40)	21.17	60.61 (65.40)	20.41
+基於 CNN 的語者慣用語特徵	66.28 (104.1)	21.07	59.05 (93.12)	20.32
語者調適混和模型(SAMM)	67.61 (99.80)	21.04	60.43 (93.60)	20.33

得 CER 有些許的上升，推測應是在 CNN 模型訓練時，反例語句檢索並沒有找到足以代表反例的語句，導致依其所建立之具語者調適性的語言模型沒有展現出期望中的效能。另一方面本論文所提出的語者調適混和模型(SAMM)不管是在複雜度或是 CER，均與上述方法旗鼓相當。

本論文第二組的實驗在 AMI 英文會議語料；表五是在此語料上的語言模型複雜度和 CER 結果。由於 AMI 英文會議語料的發展集與訓練集的語者重複極少、測試集與訓練集的語者完全不同，在表五僅呈現使用基於 CNN 的語者慣用語特徵和語者調適混和模型的實驗結果。使用基於 CNN 的語者慣用語特徵的語言模型在 AMI 英文會議語料的語言複雜度降低上並沒有預期中的表現，但是在語音辨識實驗卻有不錯的效能，可以降低約 0.5%的詞錯誤率。另一方面，語者調適混和模型在語言複雜度上，沒有預期中的表現，僅在 CER 些微下降；這可能是 AMI 英文會議語料語者較多，而本研究所選用的特定語者數(七位)占總語者數的比例太小，導致其表現沒有像在華語會議語料來的好。

六、結論與未來展望

本論文考慮了「已知語者」和「未知語者」兩種的情境，也以不同的角度考慮語者特徵擷取模型的建立，我們提出了三種語者特徵擷取模型，「語者用詞特徵模型」、「語者慣用語特徵模型」、「語者調適混和模型」，其中「語者慣用語特徵模型」以及「語者調適混和模型」適用於未知語者的測試階段，結果顯示語者調適混和模型不管在「已知語者」還是「未知語者」的情境都能達到一定的效果，但是語者慣用語特徵模型的表現較其它方法差，原因是所選取的反例語句無法很好的表現該語者的相反用語特性。

參考文獻

[1] M. Bacchiani and B. Roark, “Unsupervised language model adaptation,” IEEE International Conference on Acoustics, Speech and Signal Processing, 2003.

[2] Y. Bengio et al., “A neural probabilistic language model,” Journal of Machine Learning Research, 2003.

[3] X. Chen et al., “Future word contexts in neural network language models,” IEEE Automatic Speech Recognition and Understanding Workshop, 2017.

[4] X. Chen et al., “Recurrent neural network language model adaptation for multi-genre broadcast speech recognition,” The Annual Conference of the International Speech Communication Association, 2015.

[5] J. Chung et al., “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv, 2014.

[6] Y. N. Dauphin et al., “Language modeling with gated convolutional networks,” arXiv:1612.08083, 2016.

[7] Y. Kim, et al., “Character-aware neural language models,” AAAI Conference on Artificial Intelligence, 2016.

[8] S. Kombrink et al., “Recurrent neural network based language modeling in meeting recognition,” The Annual Conference of the International Speech Communication Association, 2011.

[9] J. Carletta et al., “The AMI meeting corpus: A pre-announcement,” The International Workshop on Machine Learning for Multimodal Interaction, 2005.

[10] T. Mikolov et al., “Recurrent neural network based language model,” The Annual Conference of the International Speech Communication Association, 2010.

[11] S. Hochreiter et al., “Long short-term memory,” Neural Computation, 1997.

[12] M. Sundermeyer et al., “LSTM neural networks for language modeling,” The Annual Conference of the International Speech Communication Association, 2012.

[13] X. Liu et al., “Efficient lattice rescoring using recurrent neural network language models,” IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.

[14] T. Hofmann, “Probabilistic latent semantic analysis,” The Conference on Uncertainty in Artificial Intelligence, 1999.

[15] M. Ma et al., “Modeling non-linguistic contextual signals in LSTM language models via domain adaptation,” IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.

[16] T. Tan et al., “Speaker-aware training of LSTM-RNNs for acoustic modelling,” IEEE International Conference on Acoustics, Speech and Signal Processing, 2016.

[17] D. Povey et al., “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” Annual Conference of the International Speech Communication Association, 2016.