# 非監督式學習於中文電視新聞自動轉寫之初步應用

郭人瑋　蔡文鴻　陳柏琳

國立台灣師範大學資訊工程研究所

{rogerkuo, louis, berlin}@csie.ntnu.edu.tw

**摘要.** 本論文探討非監督式學習於中文電視新聞自動轉寫之初步應用。在聲學模型訓練上，我們提出以發音確認(Utterance Verification)技術來克服訓練語料沒有正確人工轉寫的問題，所謂的發音確認是使用候選詞信心度評估(Candidate Word Confidence Measure)來對某語句及其轉寫進行篩選的動作，用以決定此語句及轉寫是否有足夠的可靠度，進而成爲訓練語料。我們先使用大詞彙連續語音辨識器對龐大且無人工轉寫的語料進行自動轉寫，再使用發音確認(Utterance Verification)針對辨識後的語料進行篩選，從中擷取較正確可靠的語料片段，以供聲學模型訓練使用，此舉不僅可大大節省人力成本，在效果上，經訓練過的聲學模型也和單純以人工轉寫結果所訓練出來的模型相距不遠；同時，較正確可靠的文字語料片段，則用於語言模型調適，以增進辨識效能。同樣地，候選詞信心度評估也被應用到非監督式聲學模型調適上，我們初步將它與「最大相似度線性迴歸」(Maximum Likelihood Linear Regression)聲學模型調適技術作結合，以語音辨識所產生之詞圖(Word Graph)作爲調適標的。我們以公共電視台的新聞語料爲研究題材，結果顯示非監督式聲學模型訓練與調適的結合的確可有效降低字錯誤率(Character Error Rate)，驗證了此作法之可行性。

## 1　序論

隨著科技快速發展，日常生活中能取得的多媒體影音資訊愈來愈多，如廣播電視節目、演講稿和數位典藏等。這些多媒體資訊早已成爲傳統文字資訊外，可供社會大眾廣泛使用之資訊。例如在廣播及電視新聞語音辨識和資訊檢索技術的發展上，近年來已有許多的研究和令人鼓舞的成果陸續被發表出來[1]-[4]。但爲了要以語音辨識技術來自動轉寫這些影音資訊，尤其當我們想在新的應用領域建立一套語音辨識系統時，通常必須仰賴大量經由人工轉寫(Manually Transcribed)的語料來供給聲學模型訓練使用，才可達到不錯的辨識效果，但這往往既耗人力又費時間。有鑑於此，近幾年來開始有一些研究，嘗試發展以近乎非監督式(Lightly Supervised)的方式，結合廣播或電視新聞節目對應字幕(Closed Caption)，嘗試從大量龐雜的語料中擷取較可靠的語句片段供聲學模型訓練，使語音辨識系統的準確性更爲提升，或能於新的應用領域中迅速建立起新的雛形系統，目前也有一些初步的成果被發表出來[5]-[6]。但是，上述作法的前提爲廣播或電視新聞節目對應字幕必須事先提供才能進行。從先前的研究中[7]，我們提出以發音確認(Utterance Verification)的技術來克服訓練語料沒有正確人工轉寫的問題。先使用大詞彙連續語音辨識器對龐大且無人工轉寫的語料進行語音辨識，利用信心度評估(Confidence Measure)對自動轉寫的語料進行篩選，擷取較爲正確可靠的自動轉寫語料片段，達到非監督式(Unsupervised)聲學模型訓練的目的。嚴格來說，先前的研究中，所使用的語料都是錄製於數個相同的廣播電台[4]，但仍屬於同一領域內的非監督式聲學模型訓練。因此，本論文嘗試作跨領域的非監督式聲學模型訓練之研究，希望以少量含有正確人工轉寫的廣播新聞語料爲基礎[4]，利用非監督式聲學模型訓練方式，建立一個語音辨識雛形系統，處理公共電視台的新聞語音資料(簡稱公視新聞語料或MATBN)[8]-[9]。公視新聞語料爲中央研究院資訊所口語小組與公共電視台所合作完成[10]，現在也開始廣爲國內各大學及研究機構所使用，如台大、交大、成大等學校都已有初步的研究成果。我們將對公視新聞語料作整理及統計，並定義一些訓練語料及測試語料，對本論文中所提出的方法加以實驗評估。

另一方面，聲學模型調適(Acoustic Model Adaptation)在語音辨識中一直扮演著相當重要的角色，爲的就是要補償聲學模型訓練環境與測試環境不匹配所造成的問題，進而提高辨識率。在聲學模型調適方法中最常被使用的調適技術爲「最大事後機率(Maximum *a Posteriori*, MAP)」[11]與「最大相似度線性迴歸(Maximum Likelihood Linear Regression, MLLR)」[12]。前者(MAP)視聲學模型參數爲一組隨機變數，並爲它假設一組對應的事前機率分佈(Prior Distributions)加以限制。當調適語料量多時，調適的效果漸近於「最大相似度估測(Maximum Likelihood Estimation, MLE)訓練」[13]；若調適語料不足時，調適後的模型

參數愈接近原始的模型參數。因此提供了良好的強健性。後者(MLLR)試著爲聲學模型中的高斯分佈，因統計特性相近所形成的迴歸群集(Regression Classes)求取一共同的轉換矩陣，再藉由調適語句的參與，使群集內的高斯分佈參數經由轉換矩陣旋轉平移後，對所屬之調適語句得到最大的相似度，就算無調適語料的聲學模型參數，也能藉由共享相同群集的轉換矩陣來做調適，當調適語料不多時，效果則較「最大事後機率」來的顯著。同樣地，聲學模型調適也有監督式與非監督式之分，它們的最大差別在於，後者的調適語句沒有對應的正確人工轉寫(Manual Transcription)，語句的自動轉寫(Automatic Transcription)需先經由一次語音辨識產生，再以此進行聲學模型調適。倘若第一次的轉寫存在大量的錯誤，將連帶地將影響後續的聲學模型調適，錯誤的累積會使辨識率的進展愈來愈受侷限。在應用上，非監督式聲學調適技術較具實際應用價值，但在論文的發表上，大部分仍是以監督式聲學模型調適的實驗爲主。故在本論文中，我們將候選詞信心度評估應用於非監督式聲學模型調適的研究，初步配合「最大相似度線性迴歸」的模型調適方法，以信心度評估對自動轉寫的調適語料作適當的加權，並使用語音辨識過程中產生的詞圖(Word Graph)作爲聲學模型調適的環境[14]，嘗試研究如何從詞圖的豐富資訊中擷取出適合於非監督式聲學模型調適的資訊，增進聲學模型的準確性。最後，我們也將結合語言模型調適，兩種以最大事後機率(Maximum *a Posteriori*, MAP)爲基礎之語言模型調適(Language Model Adaptation)技術：詞頻數混和(Count Merging)和語言模型插補(Language Model Interpolation)[15]-[16]，也初步作爲應用自動轉寫用於語言模型調適的方法。

本論文的安排如下：第二節將描述台灣師範大學資工所的新聞語音辨識系統及實驗的語料，第三節將說明我們所使用的發音確認技術，第四節將提出非監督式學習的架構，包含非監督式聲學模型訓練、自動轉寫用於語言模型調適及非監督式聲學模型調適等方法，第五節將報告我們的實驗數據及討論，最後將作總結，並敘述我們正進行的各項研究。

## 2 新聞語音辨識系統與實驗語料

### 2.1 台師大資工所新聞語音辨識系統

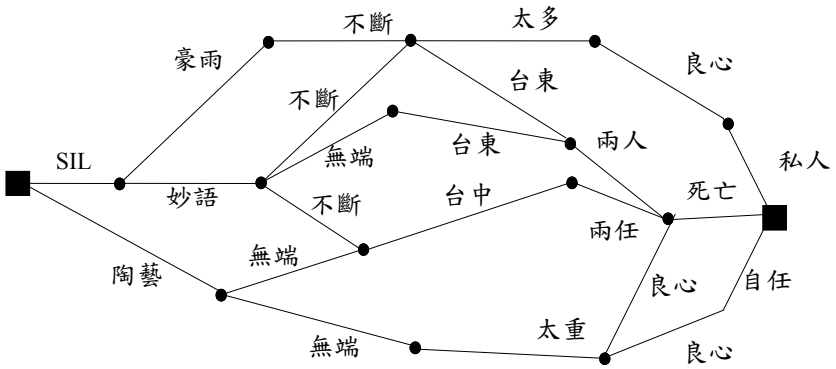
在本節中，我們將扼要介紹台灣師範大學資工所目前所發展的新聞語音辨識系統，它基本上是一套大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition)系統，主要包括前端處理(Front-end Processing)、聲學模型訓練(Acoustic Model Training)、詞典的建立(Lexicon Construction)、語言模型訓練(Language Model Training) 和詞彙樹複製搜尋(Tree-Copy Search)等部分。同時，我們也將介紹與分析本論文中所使用的廣播新聞語料與公視新聞語料。

#### 2.1.1 前端處理與聲學模型訓練

在本論文中我們使用梅爾倒頻譜特徵向量(Mel-frequency Cepstral Coefficients，簡稱MFCC特徵向量)作爲語音訊號的特徵參數。在求取MFCC特徵向量時，我們將語音資料切割成一連串部分重疊的音框，每一個音框(Frame)由13維的梅爾倒頻譜特徵加上其一階與二階的時間軸導數(Time Derivatives)所形成的39維特徵向量所組成。其中13維的梅爾倒頻譜特徵是由18個梅爾頻譜上濾波器組(Filter Banks) 的輸出經餘弦轉換求得。同時，爲了降低通道效應對語音辨識的影響，我們使用倒頻譜平均消去法(Ceptral Mean Subtraction，簡稱CMS)。另外，在辨識所需的聲學模型訓練上，考慮了中文語音結構，聲學模型由22個INITIAL模型、38 FINAL模型(每個中文的音節都是由一個INITIAL及一個FINAL所組成)及一個靜音(Silence) 模型組成，其中INITIAL模型會因其右邊可能接的FINAL模型種類而進一步細分成112個INITIAL模型[4]。我們最後總共使用了151個隱藏式馬可夫模型(Hidden Markov Models)來作爲這些INITIAL-FINAL聲學模型的統計模型。在隱藏式馬可夫模型中，每個狀態則依據其對應到的訓練語料多寡，以2到128個高斯統計分佈來表示，不管男女性別都使用同一套聲學模型。

#### 2.1.2 詞典建立及語言模型訓練

在中文裡約有7000個單字詞，新詞可由此7000個單字詞合併產生，我們可根據字詞在語料中的統計特性，以自動化的方式產生新的複合詞(Compound Words)。新增複合詞的自動產生方式如下面所述：對於語料中任意相鄰的兩個詞$(w_i, w_j)$，我們分別計算它們的前雙連(Forward Bigram)機率$P_f(w_j \mid w_i)$，與後雙連




圖一、詞彙複製搜尋所產生之詞圖，爲所有可能候選詞的簡潔表示。

(Backward Bigram)機率 $P_f(w_i \mid w_j)$，並以前後雙連(Forward and Backward Bigrams)的機率幾何平均 $FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}$，作爲 $(w_i, w_j)$ 是否合併的依據[7]。文字語料先經由一個含有一至四字詞約六萬八千個詞的詞典來斷詞，然後利用上述的公式，經數次的疊代以及不同的基準閥值(Threshold)設定，產生約五千個二至十字詞的複合詞，使得最後的語音辨識詞典約含有七萬二千個一至十字詞。在語言模型的使用上，我們使用了詞雙連以及詞三連語言模型(Word Bigram and Trigram Language Models)，並以從中央通訊社(Central News Agency)2000與2001年所收集到的約一億七千萬個中文字語料作爲背景語言模型訓練時的訓練資料[17]。在本論文中的語言模型使用了Katz語言模型平滑技術[18]，在訓練時是採用SRI Language Modeling Toolkit (SRILM)，它是一套相當方便且容易使用的語言模型研究工具軟體[19]。

### 2.1.3 詞彙樹複製搜尋

我們發展的大詞彙連續語音辨識方法是採用由左至右(Left-to-right)、音框同步(Frame-synchronous) 的詞彙樹搜尋方式[20]。在詞彙樹中每個分枝(Arc)代表一個INITIAL或FINAL的隱藏式馬可夫模型，由樹根(Root)到任一個樹梢(Leaf)的路徑代表一個詞或一些發音相同的詞，路徑上的分枝就是代表這個詞或這些詞會使用到的隱藏式馬可夫模型。具體來說，我們採用所謂的詞彙樹複製搜尋演算法(Tree-copy Search)，搜尋時每個音框會同時存在數棵詞彙樹複製(Tree Copies)，每個詞彙樹代表不同的語言模型歷史或限制(Language Model History or Constraint)。實際上，搜尋時產生的不完全路徑(Partial Paths)如果擁有相同的語言模型歷史會被歸類在同一棵詞彙樹複製裡，進行隱藏式馬可夫模型狀態層次(State-level)維特比動態規劃搜尋(Viterbi Dynamic Programming Search)。在每個音框中，若有不完全路徑已抵達樹梢時，代表一個完整詞已可被產生；同時，不同棵詞彙樹複製間已抵達樹梢的不完全路徑，若具有相同的語言模型歷史， 則會進行再結合(Recombination)，保留最大分數者，並以它們的語言模型歷史爲標註，產生新的一棵詞彙樹複製，或加入到一棵已存在且具有相同語言模型歷史的詞彙數複製中。值得注意的是，在實作時並不需要真的建立如此多的詞彙樹複製，僅需建立一棵詞彙樹作爲搜尋時路徑展開參考之用即可，並分別紀錄搜尋時存活下來的隱藏式馬可夫模型狀態節點(也就是不完全路徑目前拜訪到的節點)的相關資訊。另一方面，由於存活的隱藏式馬可夫模型狀態節點可能會隨音框數呈指數倍增加，因此必須以光束剪裁(Beam Pruning)技術適當地剪裁分數較低的狀態節點或不完全路徑。在執行剪裁動作時會同時考量每一個詞彙樹複製內部狀態節點(Internal Node)下涵蓋的可能拜訪樹梢節點代表之所有詞對應的語言模型機率，並以其中最大者當做每一個詞彙樹複製內部狀態節點的語言模型前看分數(Language Model Look-ahead Score)[20]，再加上內部狀態節點本身搜尋時所累積的解碼分數(Decoding Score)當成剪裁比較的依據。在本研究中，我們採用的是詞單連語言模型前看(Word Unigram Language Look-ahead)，對每一個詞彙樹複製內部狀態節點，我們會以其所在分枝(或隱藏式馬可夫模型)之可能拜訪樹梢節點中具最大詞單連語言模型機率，做爲該內部狀態節點的語言模型前看分數。此外，在每個音框，我們會紀錄存活的詞彙樹複製樹梢節點中分數較高者的相關資訊(這些樹梢節點本身代表著可能的候選詞)，諸如它們的語言模型歷史、對應候選詞開始與結束的音框以及搜尋時聲學解碼的分數(Acoustic Decoding Scores)，然後再依此資訊建立起一個詞圖(Word Graph)，如圖一所示。並且在這詞圖上使用更高階的語言模型，如

表一、公視新聞語料(NTNU_SA-2)訓練集統計資訊。長度小於了2秒的主播語句在本研究中被排除，男女生語料長度的比約爲1:8。

| 訓練語料(主播部分) | 總時間(小時) | 句數(句) | 平均句長(秒/句) | 最長句長(秒) | 最短句長(秒) | 佔比例(%) | 性別 |
|---|---|---|---|---|---|---|---|
| 林建成 | 1.47 | 422 | 12.53 | 55.91 | 2.01 | 9.71 | 男 |
| 馬紹 | 0.13 | 35 | 13.30 | 26.58 | 5.29 | 0.86 | 男 |
| 葉明蘭 | 12.98 | 2,860 | 16.34 | 68.92 | 2.08 | 85.85 | 女 |
| 洪蕙竹 | 0.48 | 127 | 13.66 | 30.94 | 2.70 | 3.19 | 女 |
| 蘇怡如 | 0.06 | 17 | 12.58 | 34.21 | 5.55 | 0.39 | 女 |
| 總計 | 15.12 | 3,461 | - | - | - | 100.00 | 2男3女 |
| 平均 | - | - | 15.73 | 68.92 | 2.01 | - | - |

詞三連、詞四連語言模型等，重新進行一次動態規劃搜尋，找出最佳的詞句。在本研究中，我們在詞彙樹複製搜尋階段是使用詞雙連語言模型，而在詞圖搜尋(Word Graph Rescoring)階段是使用詞三連語言模型。

### 2.2 廣播及電視新聞語料

本研究所使用的的中文廣播新聞語音語料總共有176小時以上，全是透過收音機收錄，爲1998年11月至2004年4月之間由台北地區數家廣播電台所播送之新聞節目。所有的語料都經由人工切割爲一則一則的新聞語音檔，每一則新聞均由一個主播所播報，性別上男女都有。某些檔案因錄音的關係，含有相當大的背景雜訊。這些廣播新聞語料僅有少部分有對應的正確人工轉寫，其中有大約4小時語料收錄於1998至1999年，用來作爲初始的聲學模型訓練。而電視新聞語料則全爲公視新聞語料(MATBN)，爲中央研究院資訊所口語小組耗時三年與公共電視台合作錄製完成，預計將收錄220小時的廣播新聞，所有的新聞語料都有正確的人工轉寫以及其它的標註資訊(如：停頓、語助詞、呼吸、強調語氣、反覆、不適當的發音)，所有的人工轉寫與標註均使用DGA&LDC的轉寫器(Transcriber)來完成。每天的新聞約含有二十多則報導，每則報導爲一完整主題。除了語音資料，文字語料在其它應用上也有很大的價值(如資訊檢索、主題偵測與文章分段)。公視新聞語料大致上可分內場及外場兩個部份，內場部分主要爲主播(Studio Anchors)的語料，外場部分主要爲記者(Field Reporters)與受訪者(Interviewees)的語料。經由統計，MATBN2002與MATBN2003共120小時的語料內，只含有五位主播，由於本語料以新聞內容爲主，主播不會有大幅度的變動，其中以「葉明蘭」主播的語料佔絕大多數，約85%，使得要在內場中定義出一套較具代表性的訓練及測試語料，顯得有些困難，希望未來能經由國內各相關研究機構及人士的集思廣益與討論，爲這套資訊豐富的新聞語料，定義出有實驗價值的訓練及測試語料，作爲技術開發的比較平台。我們由MATBN2002與MATBN2003中選擇了內場約16小時的語料作爲本實驗的語料(NTNU_SA-2)[21]，包含了約15小時的內場主播語料供訓練與約44分鐘(0.74小時)的測試語料，統計資料如表一及表二所示。訓練語料中，佔有85%語料的主播葉明蘭，也是測試語料內唯一的語者，使得本實驗之聲學模型有著語者相依(Speaker-dependent)的缺失，但本論文強調於完全非監督模式下進行學習，包含聲學模型訓練、聲學模型調適及語言模型調適，相較於初始系統，辨識率上仍有明顯的進步。

## 3 發音確認

發音確認利用候選詞信心度評估(Candidate Word Confidence Measure) (3.3節介紹)來決定某語句是否予以挑選成爲非監督式訓練的語料，候選詞信心度評估包含了候選詞事後機率(Candidate Word Posterior Probability) (3.1節介紹)與聲學信心(Acoustic Confidence Measure) (3.2節介紹)兩個部份。

表二、公視新聞語料(NTNU_SA-2)測試集統計資資訊。長度小於了2秒的主播語句在本研究中被排除，語料共約44分鐘。

| 測試語料(主播部分) | 總時間(小時) | 句數(句) | 平均句長(秒/句) | 最長句長(秒) | 最短句長(秒) | 佔比例(%) | 性別 |
|---|---|---|---|---|---|---|---|
| 葉明蘭 | 0.74 | 163 | 16.28 | 38.50 | 2.57 | 100.00 | 女 |
| 總計 | 0.74 | 163 | - | - | - | 100.00 | 1女 |
| 平均 | - | - | 16.28 | 38.50 | 2.57 | - | - |

### 3.1 候選詞事後機率

由詞彙樹複製搜尋所產生的詞圖(Word Graph)是存放語音辨識過程中所有可能候選詞(Candidate Word Hypotheses)的簡潔表示[14]，包含了分數較高的樹梢節點相對應的分支，每一分支即代表一個詞，包含起始時間、結束時間及搜尋時聲學解碼的分數。候選詞的事後機率則可利用不同階層的語言模型，利用Forward-Backward演算進行詞圖搜尋(Word Graph Rescoring)，候選詞的事後機率的估測如下[22][23]：

$$CM_{Posterior}(w_{t_s}^{t_e} \mid X) = p(w_{t_s}^{t_e} \mid X) = \frac{p(w_{t_s}^{t_e}, X)}{p(X)} = \frac{\sum_{W_1^{t_s-1}} \sum_{W_{t_e+1}^{T}} p(W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^{T}, X)}{\sum_{W_1^{T}} p(W_1^{T}, X)}, \quad (1)$$

其中 $w_{t_s}^{t_e}$ 爲起始時間 $t_s$ 和結束時間 $t_e$ 的候選詞；

$X$ 爲起始時間 $1$ 及結束時間 $T$ 的聲學特徵向量序列；

$W_{t_1}^{t_2}$ 則爲起始時間 $t_1$ ，結束時間 $t_2$ 的候選詞序列(Candidate Word Hypothesis Sequence)；

$p(w_{t_s}^{t_e} \mid X)$ 爲給定聲學特徵向量序列 $X$ ，候選詞 $w_{t_s}^{t_e}$ 的事後機率，由於此事後機率也常被用來表示詞的信心度，我們以 $CM_{Posterior}(w_{t_s}^{t_e} \mid X)$ 來表示 $w_{t_s}^{t_e}$ 的事後機率；

$p(W, X)$ 爲候選詞序列 $W$ 與 $X$ 的聯合機率，包含了聲學及語言模型解碼(Decoding)分數。

### 3.2 聲學信心

在另一方面，我們可對 $w_{t_s}^{t_e}$ 求出其聲學信心，設 $SUB = \{sub_1, \cdots, sub_{N_w}\}$ 爲 $w_{t_s}^{t_e}$ 內的次詞序列(Subword Sequence)， $N_w$ 爲 $SUB$ 內次詞單位(Subword Unit,在本研究中爲INITIAL或FINAL)之個數。設 $sub_i$ 爲起始時間 $t_{i,s}$ ，結束時間 $t_{i,e}$ 之次詞，則聲學信心所使用的公式如下：

$$CM_{Acoustic}(w_{t_s}^{t_e}) = \frac{1}{N_w} \sum_{i=1}^{N_w} \frac{2}{1 + \exp\left[-\tau \cdot LLR(sub_i) + \eta\right]}, \quad (2)$$

$$where \;\; LLR(sub_i) = \log \frac{p(X_{t_{i,s}}^{t_{i,e}} \mid sub_i)}{\max_{sub} p(X_{t_{i,s}}^{t_{i,e}} \mid sub)},$$

其中 $CM_{Acoustic}(w_{t_s}^{t_e})$ 爲給定 $X$ 時， $w_{t_s}^{t_e}$ 的聲學信心， $\tau$ 及 $\eta$ 分別用來調整指數函數的成長率與平移； $p(X_{t_{i,s}}^{t_{i,e}} \mid sub)$ 爲在 $X_{t_1}^{t_2}$ 下， $sub$ 的相似度(Likelihood)； $LLR(sub)$ 爲 $sub$ 與擁有最大相似度的第一名次詞單位之對數相似度比值(Likelihood Ratio)。

### 3.3 候選詞信心度評估

候選詞信心度評估包含了候選詞事後機率 (3.1)及聲學信心(3.2)，就前者而言，雖然對詞圖上每一候選詞都能求其事後機率，但根據觀察，以愈高階的語言模型進行詞圖搜尋，候選詞之間的事後機率差異愈是懸殊，例如以三連語言模型進行詞圖搜尋時，第一名詞序列(Top1 Word Sequence)中的候選詞往往佔有超過0.95的事後機率，換句話說，語言模型所用的階層(Order)愈高，則候選詞事後機率愈受語言模型所影響，第一名詞序列的事後機率會出奇的高。若以此事後機率作爲信心度評估，難免對第一名詞序列產生偏頗。有鑑於此，我們引入信心度比例係數(Confidence Scale Factor)，將原先候選詞事後機率的刻度(Scale)加以調整，使之成爲合理的候選詞事後機率。候選詞事後機率經修正後如下：

$$CM^{\alpha}_{Posterior}(w^{t_e}_{t_s} \mid X) = p(w^{t_e}_{t_s} \mid X) = \frac{\left[\sum_{W_1^{t_s-1}} \sum_{W_{t_e+1}^{T}} p(W_1^{t_s-1} \cdot w^{t_e}_{t_s} \cdot W_{t_e+1}^{T}, X)\right]^{\alpha}}{\sum_{W_1^{T}} \left[p(W_1^{T}, X)\right]^{\alpha}}. \tag{3}$$

公式(3)符號定義與公式(1)相同，其中$\alpha$爲信心度比例係數(Confidence Scale Factor)，$\alpha$介於0與1之間，表示對聯合機率施以壓縮，使候選詞間的事後機率差異變小。當$\alpha$等於1時，則表示刻度不變；當$\alpha$等於0時，事後機率爲均勻機率(Uniform Probability)。其中，$CM^{\alpha}_{Posterior}(w^{t_e}_{t_s} \mid X)$爲給定聲學特徵向量序列$X$，信心度比例係數爲$\alpha$時，候選詞$w^{t_e}_{t_s}$的事後機率。

候選詞信心度評估則包含了候選詞事後機率及聲學信心，公式如下：

$$CM(w_n \mid X) = c_1 \cdot CM_{Acoustic}(w_n \mid X) + c_2 CM^{\alpha}_{Posterior}(w_n \mid X), \tag{4}$$

其中$c_1$與$c_2$爲權重參數，在以下的非監督式聲學模型訓練中，我們將設$c_1 = c_2 = 0.5$。

### 3.4 發音確認

我們提出發音確認(Utterance Verification)之技術，來決定某語句是否予以挑選成爲非監督式訓練的語料。發音確認可視爲一個決斷函數$V(X,W,Thr) \in \{accept, reject\}$，根據平均候選詞信心度評估，來決定自動轉寫產生的第一名詞序列$W = \{w_1, \cdots, w_N\}$是否能成爲訓練語料。決斷函數$V$定義如下：

$$V(X,W,Thr) = \begin{cases} accept & if \quad \frac{1}{N}\sum_{n=1}^{N} CM(w_n \mid X) \geq Thr \\ reject & otherwise \end{cases}, \tag{5}$$

其中，$X$爲對應的聲學特徵向量序列，$W$爲自動轉寫產生的第一名詞序列，$Thr$爲篩選基準閥值。當平均信心度評估大於篩選基準閥值時，則決斷函數輸出爲*accept*，表示$X$值得我們採用爲非監督式訓練的語料，$W$爲其對應的自動轉寫；若輸出爲*reject*，則表示不予採用。

## 4 非監督式學習

### 4.1 非監督式聲學模型訓練

我們先使用大詞彙連續語音辨識系統(聲學模型由四小時廣播新聞語料來訓練)對十五小時的公視訓練語料(共3,461句)進行自動轉寫，根據辨識結果，每句可再藉由靜音(Silence)切成數個子句，少於五個中文字的子句將被排除，最後有15,473個子句被留下。對每個子句，我們以辨識結果第一名的詞序列當作此子句對應之詞序列，進行發音確認決定此子句是否予以採用。若此子句被留下來作爲非監督式聲學模型訓練的語料。則其對應的自動轉寫片段，也將被留下作爲自動轉寫用於語言模型調適(4.3節將會介紹)的文字語料。

### 4.2 非監督式聲學模型調適

大多數的非監督式聲學模型調適僅取第一次辨識所產生的第一名詞序列來做聲學模型調適的依據。然而語音辨識的錯誤可能會對聲學模型調適造成影響，使得調適效果有限[24]。本研究中，我們嘗試使用候

選詞信心度評估爲詞圖上的候選詞進行加權，使得每一個候選詞依其信心度評估分數對模型調適都有不同程度的貢獻。但由於計算詞圖上所有候選詞聲學信心的計算量相當大，因此，在聲學模型調適中，我們只使用了事後機率。我們初步地將它與「最大相似度線性迴歸」(Maximum Likelihood Linear Regression, MLLR)聲學模型調適技術做結合。最大相似度線性迴歸的調適技術需先爲聲學模型中的高斯分佈加以分群，因統計特性相近而形成的群集稱爲迴歸群集(Regression Classes)，根據相似度最大的估測法則對每一迴歸群集求取轉換矩陣，使群集內的高斯分佈參數經此轉換矩陣旋轉平移後，相對應的調適語句能得到最大的相似度，就算調適語料無涵蓋所有的聲學模型，迴歸群集內的高斯分佈也能藉此轉換矩陣來得到調適。

由於非監督式調適沒有正確的人工轉寫，我們須先經由一次的語音辨識來產生語句的相關資訊。實驗中對於測試語句進行非監督式聲學模調適的步驟如下：

1. 測試語句經由詞彙樹複製搜尋(Tree-Copy Search)，產生詞圖(Word Graph)。
2. 利用Forward-Backward演算法在詞圖上進行詞圖搜尋(Word Graph Rescoring)，爲詞圖上的每一候選詞求出其對應的事後機率$CM^{\alpha}(w_{t_s}^{t_e}|X)$，其中$\alpha$爲信心度比例係數。
3. 針對每一候選詞語音段落，再使用一次狀態層次(State Level) Forward-Backward演算法，爲每一音框(Frame) $t$及狀態(State) $i$求其事後機率 $\gamma_t(i|w_{t_s}^{t_e})=\Pr(s_t=i|X_{t_s}^{t_e},w_{t_s}^{t_e})$。
4. 最後，將 $\gamma_t(i|w_{t_s}^{t_e})$乘上所屬候選詞的事後機率$CM^{\alpha}(w_{t_s}^{t_e}|X)$，並對所有候選詞語音段落加總。可得音框$t$時，狀態$i$的事後機率$\gamma_t(i)=\Pr(s_t=i|X_1^T)=\sum_{w_{t_s}^{t_e}} CM^{\alpha}(w_{t_s}^{t_e}|X)\gamma_t(i|w_{t_s}^{t_e})$。

重覆上述步驟，收集MLLR模型調適時所需的統計量，並進行MLLR模型調適。

### 4.3 自動轉寫用於語言模型調適

統計式語言模型(Statistical Language Models)旨在以統計的方法分析及模擬自然語言的規律特性，並以機率量化的方式來決定一個詞串的接受程度。在過去二十年間，一直是語音及語言處理領域中重要的課題。在統計式語言模型中，N連語言模型(N-gram Language Models)是最常被使用的(尤其是二連及三連語言模型)，它主要根據前面的N-1詞歷史(Word History)來決定下一個詞可能出現的機率[25][16]。N-連語言模型的機率表示，通常由最大相似度(Maximum Likelihood Estimation, MLE)來估測，然而，在特定領域下訓練N-連語言模型時，爲了解決統計模型訓練時資料稀疏的問題(Data Sparseness Problems)，過去幾年，已經有一些像平滑(Smoothing)或插補(Interpolation)等方法陸續被提出，達到不錯的效果[18]。但另一方面，在處理一些較複雜困難的語音辨識課題上如廣播及電視新聞自動轉寫，由於新聞播報的主題和語言內容的詞彙使用具多變性與時效性，會使得統計式語言模型往往很難做到準確的估測，於是便有了所謂的語言模型調適(Language Model Adaptation)的研究[16]。語言模型調適通常會結合背景文字語料庫(Background Corpus)與測試語音同一時期(Contemporary)或者是同一領域(In-domain)的文字語料庫來訓練出較具強健性的調適後語言模型，以得到較佳的詞接連預測能力，而在過去已有一些不錯研究被發表出來[26]-[27]。在本研究我們嘗試研究使用語音辨識產生的電視新聞自動轉寫用於語言模型調適(Unsupervised Language Model Adaptation)的可行性，直接以非監督式聲學模型性訓練時經發音確認篩選過後的語音片段對應的自動轉寫文字(參見4.1節)，當成同一領域文字語料來做語言模型的調適。我們初步使用兩種常用的語言模型調適技術：語言模型插補(Language Model Interpolation)及詞頻數混合(Count Merging)[15][7]，並比較這兩種由傳統貝氏估測所發展出來的語言模型調適技術。詞頻數混合和語言模型插補的調適公式分別如下(以三連語言模型爲例)：

$$\widetilde{P}_{Adapt\text{-}1}(w_i|w_{i-2}w_{i-1})=\frac{m_1\cdot C_{d,Cont}(w_{i-2}w_{i-1}w_i)+m_2\cdot C_{d,Back}(w_{i-2}w_{i-1}w_i)}{m_1\cdot C_{Cont}(w_{i-2}w_{i-1})+m_2\cdot C_{Back}(w_{i-2}w_{i-1})}, \tag{6}$$

及

$$\widetilde{P}_{Adapt\text{-}2}(w_i|w_{i-2}w_{i-1})=\gamma\cdot P_{Cont}(w_i|w_{i-2}w_{i-1})+(1-\gamma)\cdot P_{Back}(w_i|w_{i-2}w_{i-1}). \tag{7}$$

在第(6)式中，$C_{d,Cont}(w_{i-2}w_{i-1}w_i)$與$C_{d,Back}(w_{i-2}w_{i-1}w_i)$分別代表調適語料中及背景訓練語料中的三連減值詞頻(Trigram Discounted Count)，而 $C_{Cont}(w_{i-2}w_{i-1})$ 與 $C_{Back}(w_{i-2}w_{i-1})$ 則分別代表調適語料中及背景訓練語料中的二連詞頻，$m_1$與$m_2$則爲可調整的權重參數。在第(6)式中，$P_{Cont}(w_i|w_{i-2}w_{i-1})$與$P_{Back}(w_i|w_{i-2}w_{i-1})$分別代表由調適語料及背景訓練語料所估測的三連機率，$\gamma$爲可調整的參數，公式(6)及公式(7)的詳細推導可參考[15]。詞頻數混合是在詞機率估測前，將領域內(In-domain)文字語料與背景(Background)文字語料在詞頻

表三、基礎實驗與非監督式聲學模型調適之語音辨識結果：嘗試改變信心度比例係數$\alpha$與計算候選詞事後機率時語言模型的階層。 MLLR(Top1)爲傳統只取用第一名辨識結果詞序列來做MLLR調適；MLLR(CM)爲引入信心度評估的MLLR調適。字錯誤率減少百分比爲相對於無聲學模型調適之字錯誤率。

| 計算候選詞事後機率時所用的語言模型階層 | 三連語言模型 | | 二連語言模型 | |
|---|---|---|---|---|
| 聲學模型調適 | 字錯誤率 (%) | 相對字錯誤率減少百分比(%) | 字辨識率 (%) | 相對字錯誤率減少百分比(%) |
| 無 | 27.67 | - | 27.67 | - |
| MLLR(Top1) | 25.93 | 6.29 | 25.93 | 6.29 |
| MLLR(CM), $\alpha = 1$ | 25.80 | 6.76 | 26.12 | 5.60 |
| MLLR(CM), $\alpha = 1/4$ | 25.69 | 7.16 | 25.92 | 6.32 |
| MLLR(CM), $\alpha = 1/8$ | 25.80 | 6.76 | 25.95 | 6.22 |
| MLLR(CM), $\alpha = 1/12$ | 25.37 | 8.31 | 25.49 | 7.88 |
| MLLR(CM), $\alpha = 1/16$ | 25.26 | 8.71 | 25.54 | 7.70 |
| MLLR(CM), $\alpha = 1/20$ | 25.14 | 9.14 | 25.73 | 7.01 |
| MLLR(CM), $\alpha = 1/24$ | 25.38 | 8.28 | 25.82 | 6.69 |
| MLLR(CM), $\alpha = 1/28$ | 25.51 | 7.81 | 25.93 | 6.29 |

數空間(Frequency Space)上給予權重加總，進而估測機率；而語言模型插補則是估測個別模型之機率後，才根據權重於機率空間(Probability Space)上相加。

## 5 實驗結果與討論

### 5.1 實驗環境與非監督式聲學模型調適基礎實驗

我們使用台師大資工所發展的新聞語音辨識系統，並以普遍被使用的梅爾倒頻譜特徵向量作爲語音特徵參數。初始的聲學模型由四小時的廣播新聞語料所訓練而成，初始背景語言模型則由從中央通訊社收集的新聞語料訓練而得。這一小節的基礎實驗有三個目的：第一、計算候選詞事後機率時，比較不同階層語言模型帶來的影響；第二、計算候選詞事後機率時，改變信心度比例係數，討論它們對字錯誤率所帶來的影響；第三、比較傳統只取用第一名辨識結果詞序列(Top1)來作調適與使用信心度評估(CM)來作調適的結果。表三爲本節基礎實驗的結果。我們在3.3節中曾提到，語言模型所用的階層(Order)愈高，則候選詞事後機率愈受語言模型所影響，第一名詞序列的事後機率會出奇的高。若以此事後機率作爲信心度評估，難免對第一名詞序列產生偏頗。這是當語言模型階層愈高，會使得特定詞彙擁有較大的機率(根據訓練語料的特性)，使得聯合機率差距愈趨懸殊，連帶影響候選詞的事後機率。由表三中可見，使用三連語言模型計算候選詞事後機率時，最佳的字錯誤率(25.14%)出現在$\alpha = 1/20$時，若使用二連語言模型時，最佳的字錯誤率(25.49%)出現在$\alpha = 1/12$時。這也驗證了使用較高階語言模型時，將會造成特定詞彙事後機率刻度(Scale)過高的不合理現象，需要使用較小的信心度比例係數加以調整。由於使用三連語言模型明顯比使用二連語言模型要來的好，故往後的實驗中，均使用三連語言模型來計算候選詞事後機率時。信心度比例係數的決定，和系統的語音辨識率有密切的關係，當辨識率較高時，應有較大的信心度比例係數，信任第一階段辨識產生的結果；反之，則信心度比例係數應較小。使用三連語言模型計算候選詞事後機率時，雖然$\alpha = 1/20$時，我們可得較佳的結果，但考量往後的實驗，我們將加上非監督式聲學模型訓練，系統的語音辨識率會再提升，同時爲了兼顧一般化，我們在往後的實驗中，信心度比例係數均設爲1/16。在表三中，傳統只取用第一名辨識結果詞序列來作調適MLLR(Top1) 之後，可得到6.29%的相對字錯誤率減少百分比。而在引入信心度評估以詞圖資訊來作調適MLLR(Top1)之後，則可達到9.14%的

表四、非監督式聲學模型訓練在使用不同基準閥值下的語音辨識結果。Thr爲非監督式聲學模型訓練用以選取語句之基準閥值，MLLR(CM)爲引入信心度評估的MLLR調適，$\alpha$ 在此設爲1/16。同一列中，MLLR括弧內的數據爲相對於無聲學模型調適時字錯誤率減少百分比。最後一列的「監督式訓練」爲對照組。

| | 字錯誤率(%) (相對字錯誤率減少百分比(%)) | | |
|---|---|---|---|
| | 無聲學模型調適 | MLLR(Top1) | MLLR(CM) |
| 原來四小時訓練之聲學模型 | 27.67 | 25.93 (6.29) | 25.26 (8.71) |
| + 3.80小時(Thr=0.9) | 21.37 | 21.00 (1.73) | 20.97 (1.87) |
| +11.57小時(Thr=0.8) | 20.09 | 20.00 (0.45) | 19.56 (2.64) |
| +13.30小時(Thr=0.7) | 20.25 | 20.01 (1.19) | 19.71 (2.67) |
| +13.61小時(Thr=0.6) | 20.18 | 19.94 (1.19) | 19.59 (2.92) |
| +13.67小時(Thr=0.5) | 20.21 | 20.01 (0.99) | 19.69 (2.57) |
| +13.70小時(Thr=0.0) | 20.32 | 20.07 (1.23) | 19.76 (2.76) |
| +15.12小時(監督式訓練) | 16.26 | 16.29 (-0.18) | 16.47 (-1.29) |

相對字錯誤率減少百分比( $\alpha = 1/20$)，有近3%的改善，顯示本論文所提出結合信心度評估和詞圖資訊的非監督式聲學模型調適方法，的確能有效的降低字錯誤率。

### 5.2 非監督式聲學模型訓練實驗結果

我們根據不同基準閥值(Threshold Values)進行語句的篩選，進行非監督式聲學模型訓練。訓練時我們使用HTK Toolkit [28]，進行三次的嵌入式訓練(Embedded Training)。實驗結果如表四所示，在非監督式訓練下，使用發音確認，特別在基準閥值爲0.8時，我們可得最佳的字錯誤率20.09%，再經由非監督式聲學調適之後，更可達到19.56%的字錯誤率。由於聲學模型若經監督式訓練爲非監督式訓練的上限，但仍有16.26%的字錯誤率，使得非監督式訓練與監督式訓的差距不到4%，說明了非監督式聲學模型訓練有其利用的價值。基準閥值的設定與訓練語料的多寡必須加以妥協，當基準閥愈高，則留下的訓練語料愈少，模型參數則無法有效的估測，如基準閥值爲0.9時，錯誤率不降反升；若基準閥值太低，留下的語料雖多，但錯誤標註的語料反而會影響了模型參數估測的正確性。當基準閥值爲0.8時，訓練語料總時間還留有11個小時，僅有16%不到的語料被篩除，這表示在0.8之上應存在更佳的基準閥值。在對照組「監督式訓練」的聲學模型中，可發現MLLR的調適反而帶來負面的影響，經過信心度評估的MLLR調適之後，字錯誤率攀升至16.47%，我們嘗試將信心度比例係數 $\alpha$ 調小至1/4，則字錯誤率能降低至16.02%，驗證了在高辨識率的系統上應使用較小的信心度比例係數 $\alpha$ 。

### 5.3 語言模型調適實驗結果

#### 5.3.1 自動轉寫用於語言模型調適

我們將語音辨識產生的電視新聞自動轉寫用於語言模型的調適，進行了一些初步的語音辨識實驗。在模型插補的方法中，調適語言模型與背景語言模型的權重各爲0.5(公式(7)中之 $\gamma = 0.5$)。詞頻數混合的實驗中我們根據訓練語料的大小，調適語言模型與背景語言模型的詞頻數加權比爲250:1(公式(6)中之 $m_1 = 250$ 、 $m_2 = 1$)。實驗結果如表五所示，我們可觀察出自動轉寫中的詞連接規則資訊對語言模型仍能有不錯的貢獻，如當基準閥值爲0.8時，相對於無語言模型調適，詞頻數混合可達到1.74% 的相對字錯誤率減少百分比。雖然以自動轉寫爲基礎的適語料過於稀疏(Sparse)，(當基準閥值爲0.9僅有約66K個字，即使在篩選閥值爲0.0時，也只有約250K的字)，使得整體的字錯誤率下降幅度並不顯著。未來研究中，我們希望能藉由詞圖所提供豐富資訊來加以改善，並且以信心度評估爲每一個候選詞的詞頻作加權，俾使詞圖上的候選詞均能對語言模型調適有所貢獻。

表五、自動轉寫用於語言模型調適的語音辨識結果。Thr爲非監督式聲學模型訓練用以選取語句之基準閥值，括弧內之數據爲相對於無語言模型調適之字錯誤率減少百分比。

| 聲學模型 | 調適語料字數 | 字錯誤率(%) (相對字錯誤率減少百分比(%)) | | |
|---|---|---|---|---|
| | | 無語言模型調適 | 語言模型插補 | 詞頻數混合 |
| + 3.80小時(Thr=0.9) | 66,540 | 21.37 | 21.85 (-2.25) | 21.08 ( 1.36) |
| +11.57小時(Thr=0.8) | 209,489 | 20.09 | 19.97 ( 0.60) | 19.74 ( 1.74) |
| +13.30小時(Thr=0.7) | 242,630 | 20.25 | 20.06 ( 0.94) | 20.27 (-0.10) |
| +13.61小時(Thr=0.6) | 248,701 | 20.18 | 20.04 ( 0.69) | 20.06 ( 0.59) |
| +13.67小時(Thr=0.5) | 249,880 | 20.21 | 20.05 ( 0.79) | 20.23 (-0.10) |
| +13.70小時(Thr=0.0) | 250,640 | 20.32 | 20.02 ( 1.48) | 20.18 ( 0.69) |

### 5.3.2 領域內之語言模型調適

在這個實驗中，我們從公視新聞網[8]所收集的2001年與2002年文字語料(約五百萬個中文字)來做爲語言模型調適語料，這些語料大多是新聞節目對應字幕(Closed Caption)。本研究訂立兩套調適語言模型來加以實驗：PTS_LM_1由2001年1月至2002年12月的公視新聞網語料所訓練，由於此語言模型訓練語料涵蓋的日期包含測試語料的那五天(2002年8月6日到2002年8月9日及2002年9月26日)，故我們稱之爲偏差語言模型(Biased Language Model)。PTS_LM_2則排除2002年8月(含)之後的語料，由2001年1月至2002年7月的語料來進行訓練。其目的主要在於觀察領域內訓練語料的時效性對語言模型調適的影響。我們在此初步以使用監督式聲學訓練的聲學模型(15小時新聞語音資料)來進行實驗；在語言模型插補的方法中，調適語言模型與背景語言模型的權重各爲0.5(公式(7)中$\gamma = 0.5$)；在詞頻數混合的方法中，PTS_LM_1的加權比約爲20:1(公式(6)中之$m_1 = 20$、$m_2 = 1$)、PTS_LM_2的加權比約爲25:1(公式(6)中之$m_1 = 25$、$m_2 = 1$)。結果如表六所示，其中我們也顯示出當使用非監督式聲學模型調適後的辨識結果。在PTS_LM_1下，經過非監督式的聲學模型調適，辨識率可達92.67%(字錯誤率7.23%，語言模型插補)及92.77%(字錯誤率7.33%，詞頻數混合)；PTS_LM_2下，經過非監督式的聲學模型調適的辨識率僅有84.68%(字錯誤率15.32%，語言模型插補)及84.55%(字錯誤率15.45%，詞頻數混合)。由此可見，對電視新聞語音辨識來說，時效性對於語言模型的影響甚鉅。

## 6 結論與未來展望

本論文探討非監督式的聲學模型訓練與調適於中文電視新聞自動轉寫之初步應用。由實驗結果可觀察出發音確認能有效地挑選較爲可靠的語料來進行訓練，節省大量的人力進行人工轉寫，使龐大的語料能被運用，篩選基準閥值的取決，影響了訓練的品質，如何在語料量與信心度評估找到平衡點，仍是一個課題；信心度評估也使得詞圖上更多的資訊能應用在非監督式聲學模型調適上，不再只侷限於Top1辨識的路徑，因此能解決非監督式調適時使用含有錯誤資訊的自動轉寫以及所需調適語料統計量過少的問題，但信心度比例係數$\alpha$的調整則需考慮辨識率及語言模型的階層。自動轉寫用於語言模型調適能解決新聞辨識主題和語言內容的詞彙使用具多變性的問題，由於資料稀疏，使得字錯誤率的進步並不大，但由於詞圖上含有大量的資訊，我們甚至可根據詞圖上的信心度評估爲每一個候選詞的詞頻作加權，俾使詞圖上的候選詞均能對語言模型調適有所貢獻。在寫此論文的同時，我們正將前端特徵值抽取部份，改用更有鑑別力的特徵向量，如線性鑑別分析(Linear Discriminant Analysis, LDA)及異質性鑑別分別(Heteroscedastic Discriminant Analysis, HDA)[29] ，也試著將最大交互資訊(Maximum Mutual Information, MMI)訓練[30]與最小音素錯誤(Minimum Phone Error, MPE)訓練[31]等方法結合詞圖(Word Graph)的豐富語音辨識資訊，應用在非監督式聲學模型訓練上，以期能得到更好的語音辨識率。

表六、領域內語言模型調適的語音辨識結果。MLLR(CM)爲引入信心度評估的MLLR調適，$\alpha$ 在此設爲1/16。

| | 字錯誤率(%) | | |
|---|---|---|---|
| | 無聲學模型調適 | MLLR(Top1) | MLLR(CM) |
| 無語言模型調適 | 17.83 | 17.67 | 17.51 |
| PTS_LM_1(語言模型插補) | 7.46 | 7.32 | 7.23 |
| PTS_LM_1(詞頻數混合) | 7.47 | 7.39 | 7.33 |
| PTS_LM_2(語言模型插補) | 15.08 | 14.93 | 15.32 |
| PTS_LM_2(詞頻數混合) | 15.94 | 15.72 | 15.45 |

## 誌謝

本研究承蒙國科會「中文語音資訊辨識與檢索之研究」，編號：（91-2218-E-003-002-）及「中文語音資訊摘要技術之研究」，編號（92-2213-E-003-008-）等計畫補助。並感謝中研院口語小組提供公視新聞實驗語料及台大語音實驗室提供廣播新聞實驗語料。另外，也感謝三位審查委員所提供之意見。

## 參考文獻

[1] P. Beyerlein et al., “Large Vocabulary Continuous Speech Recognition of Broadcast News – The Philips/RWTH Approach,” *Speech Communication*, May 2002.
[2] P.C. Woodland, “The development of the HTK Broadcast News transcription system: An overview,” *Speech Communication*, May 2002.
[3] J. L. Gauvain, L. Lamel, G. Adda, “The LIMSI Broadcast News Transcription System,” *Speech Communication*, May 2002.
[4] B. Chen, H-M Wang, and L-S Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese", *IEEE Trans. on Speech and Audio Processing*, July 2002.
[5] L. Nguyen, B. Xiang, “Light Supervision in Acoustic Model Training,” in *Proc. ICASSP 2004*.
[6] L. Chen, L. Lamel and J. L. Gauvain, “Lightly Supervised Acoustic Model Training Using Consensus Networks,” in *Proc. ICASSP 2004*.
[7] B. Chen, J. W. Kuo, W. H. Tsai. “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” in *Proc. ICASSP 2004*.
[8] 財團法人公共電視文化事業基金會-公共電視台. http://www.pts.org.tw/.
[9] H. M. Wang. “MATBN 2002: A Mandarin Chinese Broadcast News Corpus,” in *Proc. SSPR’03*, Tokyo, Japan.
[10] 中央研究院資訊所中文組口語小組. http://sovideo.iis.sinica.edu.tw/SLG/.
[11] J.-L. Gauvain, C.-H. Lee, “Maximum a Posteriori Estimation for Multivate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Processing*, April 1994.
[12] M. J. F. Gales and P. C. Woodland (1996). “Mean and Variance Adaptation within the MLLR Framework,” *Computer Speech and Language*, pp.249-264, Vol. 10, 1996.
[13] L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, February 1989.
[14] S. Ortmanns, H, Ney, X Aubert, “A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language*, Vol. 11, 1997.
[15] M. Bacchiani, B. Roark, “Unsupervised Language Model Adaptation,” in *Proc. ICASSP 2003*.
[16] J. R. Bellegarda, “Statistical Language Model Adaptation: Review and Perspectives,” *Speech Communication*, Vol. 42, 2004.
[17] 中央通訊社. http://www.cna.com.tw/.
[18] S. F. Chen, J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling,” *Computer Speech and Language*, Vol. 13, 1999.

[19] A. Stolcke, “SRI language Modeling Toolkit, ” version 1.3.3, http://www.speech.sri.com/projects/srilm/.
[20] X. L. Aubert, “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language*, January 2002.
[21] 公視新聞語料整理與分析(台師大資工所). http://speech.csie.ntnu.edu.tw/MATBN_SetDefinition/.
[22] F. Wessel, R. Schluter, K. Macherey, H. Ney, “Confidence Measures for Large Vocabulary Continuous Speech Recognition,” *IEEE Trans. on Speech and Audio Processing*, March 2001.
[23] W. Chou (editor),. B.H. Juang (editor). *Pattern Recognition in Speech and Language Processing*. Chapter 2, CRC Press, 2003.
[24] M. Padmanabhan, G. Saon and G. Zweig, “Lattice-Based Unsupervised MLLR for Speaker Adaptation,” in *Proc. ISCA ITRW ASR2000*.
[25] R. Rosenfeld, “Two Decades of Statistical Language Modeling: Where Do We Go from Here,” *Proc. IEEE*, 88 (8), 2000.
[26] M. Federico, N. Bertoldi, “Broadcast News LM adaptation Using Cotemporary Texts,” in *Proc. Eurospeech 2001*.
[27] W. Kim, S. Khudanpur, “Cross-Lingual Latent Semantic Analysis for Language Modeling,” in *Proc. ICASSP 2004*.
[28] S. Young et al. *The HTK Book*. Version 3.2, 2002. http://htk.eng.cam.ac.uk/.
[29] Nagendra Kumar. *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. Ph.D dissertation, Johns Hopkins University, 1997.
[30] P. C. Woodland, D. Povey, “Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition,” *Computer Speech and Language*, pp.25-47, Vol. 16, 2002.
[31] P. C. Woodland, D. Povey, “Minimum Phone Error and I-Smoothing for Improved Discriminative Training,” in *Proc. ICASSP 2002.*