

國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis



使用查詢詞擴展與自動習得之聲學組型強化語音數位
內容之語意檢索

Enhanced Semantic Retrieval of Spoken Content with Query
Expansion and Automatically Discovered Acoustic Patterns

李昀樵

Yun-Chiao Li

指導教授：李琳山 教授

Advisor: Lin-Shan Lee, Ph.D.

中華民國一百零三年六月

June, 2014

誌謝



兩年來的碩士生活，隨著這份論文的誕生而告了一個段落。回思過去兩年來的生活，有做出成果的喜悅、也有處處碰壁時的苦悶、也有跟同學共同奮鬥的記憶，自己在這兩年中實在成長了許多，而這些都要感謝實驗室的大家長：李琳山教授。教授在實驗室營造了自由研究的氛圍，讓實驗室的同學都能按自己喜好自由發展研究方向，並從旁關心協助同學的研究。我在這樣的氛圍下也受益許多，學習到了許多做研究與做人處事的方法。

這些研究能夠順利完成要由衷地感謝我的家人，包含我的父母和妹妹，他們無論何時都支持我的決定，並且從旁給我協助，在我最忙碌而都很晚回家的那段時間，他們也是很包容我，並給予我支持。還有我的女朋友，她也在我最辛苦的時候支持和幫忙我，並常常跟我討論未來的規劃，讓我的決定都做的更好。

接著要感謝的是從大四下就帶領我研究的李宏毅學長。學長在我對語音什麼都不懂的狀況下教我基礎知識，並教導我如何尋找研究方向與撰寫 Paper，一直到最後順利地發表這些論文，都必須大力感謝學長的幫忙。

與實驗室的同學相處的這段時光將是碩士生活中最難忘的一段日子，博士班的學長小安、阿邦、瑪雅、Aaron、青峰哥、宏毅哥在我們還是菜鳥時教導我們很多知識；上一屆的學長蘇培豪、溫宗憲、周宥宇、林博智、陳泰元常常跟我們在實驗室討論跟一起聊天，也恭喜你們最近都有很好的發展；同屆的向思蓉、周伯威、余典翰、鍾承道、蘇嘉雄，我們無論是修課或是研究上都是彼此的好戰友，我從你們身上都學到很多；下一屆的楊子毅、劉元銘、吳全勳、曾柏翔、熊信寬、蔡政昱、魏承寬，你們將是實驗室下一代的主力，祝你們未來研究順利！

最後要感謝我的朋友、同學、以及伙伴們，不論是平常一起吃飯聊天、有正事時的一起奮鬥、或一起出去玩，你們都是我平常生活上最大的支持！

摘要



本論文之主軸在探討語音數位內容之語意檢索 (Semantic Retrieval of Spoken Content)。由於近年來網路日新月異，使得網路上包含語音資訊的多媒體數位內容 (Multimedia Content) 如線上課程、電影、戲劇、會議錄音等日漸增加，因此，語音數位內容之檢索也隨之受到重視。但以前的語音數位內容檢索多半著重於口述語彙偵測 (Spoken Term Detection)，而本篇論文將把目標放在語意檢索（指找到語意相關的語音文件，但未必包含查詢詞 (Query Terms)），實現的方法主要是借助查詢詞擴展 (Query Expansion)，並另外加入了一套自動習得之聲學組型 (Automatically Discovered Acoustic Patterns) 用以解決以往語音數位內容語意檢索之困難。

首先，由於傳統的語音數位內容語意檢索是先將語音文件辨識為以文字構成的詞圖後，再於詞圖上進行查詢詞擴展，但有許多聲學上的資訊會在辨識之中流失，或是有辨識錯誤與辭典外辭彙也會使檢索系統的成效下降，因此本論文在文字的查詢詞擴展之外，再加入一套自動習得之聲學組型的查詢詞擴展，並結合兩套查詢詞擴展之結果回傳給使用者。

此外，使用聲學組型也可以直接達成非監督式 (Unsupervised) 語音文件的語意檢索。傳統的語意檢索必須依賴文字才知語意，故需將語音文件辨識成詞圖，但是這樣需要已訓練得很好的聲學模型和語言模型，而這兩者的訓練需要有妥為標注 (annotated) 並和數位內容適度匹配 (matched) 的訓練語料。通常是非常昂貴的，因此我們將所有語音文件辨識為聲學組型的序列之後，在這些聲學組型的序列上進行查詢詞擴展，進而達到無需標注語料的非監督式語音數位內容之語意檢索。

另一方面，由於聲學組型在訓練時並不知道聲音和詞之間的關聯，所以會將



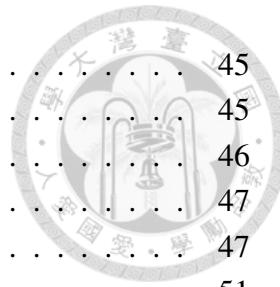
所有同音詞的聲音歸類到同一個聲學組型中，這會使得檢索的成效下降。所以本論文進一步使用遞迴式類神經網路語言模型 (Recurrent Neural Network Language Model) 的詞表示法 (Word Representation) 將同一個聲學組型按照句法 (Syntactics) 和語意 (Semantics) 的不同進一步分群為不同的聲學組型，以便提升檢索系統成效。

最後，由於行動裝置日益重要，也使得行動裝置上的語音輸入漸受重視，因此本論文在 Google 眼鏡上開發了兩個應用程序：雲端個人化語言翻譯系統和雲端個人化新聞查詢系統，幫助使用者在行動裝置上快速地取得想要的資訊。



Contents

誌謝	i
中文摘要	ii
一、導論	1
1.1 研究動機	1
1.2 研究方向	3
1.3 章節安排	4
二、背景知識	5
2.1 語音數位內容檢索	5
2.1.1 簡介	5
2.1.2 口述語彙偵測與語意檢索	5
2.1.3 詞圖與唯一最佳序列	7
2.1.4 辭典外詞彙 (Out of Vocabulary)	9
2.1.5 檢索系統	10
2.1.6 片段式動態時間校準 (Segmental DTW)	11
2.1.7 資訊檢索評估機制	15
2.2 相關回饋	17
2.2.1 外顯回饋	17
2.2.2 隱含回饋	18
2.2.3 虛擬回饋	18
2.2.4 查詢詞擴展與文件擴展	19
2.3 自動習得之聲學組型	22
2.4 遷迴式類神經網路語言模型	25
2.5 本章總結	29
三、以自動習得之聲學組型加強監督式語意檢索	30
3.1 簡介	30
3.2 傳統監督式語音文件語意檢索	31
3.2.1 第一次檢索結果	32
3.2.2 查詢詞擴展與第二次檢索	33
3.3 以聲學組型改善監督式語意檢索	34
3.3.1 前處理	35
3.3.2 第一次檢索結果	37
3.3.3 查詢詞擴展	37
3.4 實驗設定	41
3.5 實驗結果及分析	41
3.6 本章總結	43
四、以自動習得之聲學組型實現非監督式語意檢索	44
4.1 簡介	44



4.2	基於聲學組型之語意檢索	45
4.2.1	系統架構	45
4.2.2	前處理	46
4.2.3	第一次檢索結果	47
4.2.4	語意檢索	47
4.3	實驗設定	51
4.4	N 連聲學組型分析	51
4.5	實驗結果及分析	52
4.5.1	聲學組型語意檢索能力分析	57
4.6	本章總結	57
五、	利用遞迴式類神經網路語言模型加強非監督式語音文件檢索	59
5.1	簡介	59
5.2	基於遞迴式類神經網路語言模型之詞表示法	60
5.2.1	基於遞迴式類神經語言模型之詞表示法	60
5.3	以詞表示法改善非監督式語意檢索	61
5.4	實驗基礎架構	63
5.5	實驗結果	63
5.6	本章總結	65
六、	在Google Glass上實作個人化的語音翻譯與新聞檢索系統	66
6.1	簡介	66
6.2	個人化的語言翻譯系統簡介	68
6.3	個人化的語音文件檢索系統簡介	70
6.4	系統展示	71
6.4.1	個人化的語音翻譯系統展示	71
6.4.2	個人化的語音文件檢索系統展示	71
6.5	本章總結	72
七、	結論與展望	74
7.1	本論文主要的研究貢獻與未來展望	74
7.1.1	使用聲學組型加強語音文件檢索	74
7.1.2	實作雲端語音辨識與應用程式於 Google 眼鏡	75
參考文獻		76

圖目錄



2.1	詞圖示意圖	8
2.2	片段式動態時間校準示意圖 [1]	13
2.3	示意如何從兩個對角片段中找到相關分數最大的假設區域，其中 $R = 2$	14
2.4	準確率、召回率和平均準確率之關係	16
2.5	相關回饋的基本架構	18
2.6	在類詞片段上進行的Watershed轉換，紅線部分為可能的次詞邊界 .	24
2.7	遞迴式類神經網路語言模型示意圖	25
2.8	將潛藏層向前展開三層的示意圖	28
3.1	系統架構示意圖	35
3.2	將文字形式和聲學組型形式的查詢詞擴展疊加後的平均準確率 ($\lambda = 800$)	42
3.3	將文字形式和聲學組型形式的查詢詞擴展疊加後的平均準確率 ($N = 10$)	43
4.1	系統架構示意圖	46
4.2	將第一次檢索結果和聲學組型形式的查詢詞擴展檢索結果疊加後的 平均準確率 ($N = 800$)	56
4.3	將第一次檢索結果和聲學組型形式的查詢詞擴展檢索結果疊加後的 平均準確率 ($\lambda = 300$)	56
5.1	基於遞迴式類神經網路語言模型之詞表示法示意圖	61
5.2	演算法示意圖	62
5.3	實驗結果：K分群法的K設定為4，潛藏層長度設為100	64
5.4	實驗結果：K分群法的K設定為5，潛藏層長度設為100	64
6.1	Google 眼鏡	67
6.2	雲端個人化語言翻譯系統架構	68
6.3	iBrille 首頁	72
6.4	iBrille 使用示範	72
6.5	新聞隨手查首頁	73
6.6	新聞隨手查使用示範	73

表目錄



4.1	一些單連和雙連聲學組型的聲音與其對應到的中文詞	53
4.2	系統在語意檢索和口述語彙偵測時的平均準確率	54
4.3	系統找回與查詢詞語意相關的文件數量，包括含查詢詞與不含查詢詞的文件數。	55
4.4	查詢詞為”學校(/xue-xiao/)"時，擴展後查詢詞模型 ϕ_{qe} 中機率最大的五組聲學組型	58

第一章 導論



1.1 研究動機

現在是一個資訊爆炸的時代，每天資訊的增長量是十分驚人的，因此，我們需要一套好的檢索系統 (Retrieval System) 幫助我們快速地瀏覽資訊，並找到其中有用的部分，在過去已有許多文字檢索系統與演算法被開發出來並應用於產業中，如 Google Search、Microsoft Bing Search、Yahoo Search 等。但近年來隨著科技與網際網路的興起，語音文件量正蓬勃地增加當中，隨著線上影片、會議錄音、線上課程等網站的興起，語音資料量越來越多，因此如何在其中找到使用者感興趣的資料便成為重要的議題，即為語音數位內容檢索 (Spoken Content Retrieval) [2,3]。相較於文字資訊檢索，語音資訊檢索面臨到更多的挑戰，如辨識錯誤、辨識訓練資料不足等問題，使得此問題更形困難。

近年來更由於智慧型手機的崛起與使用者需要在移動裝置上取得資訊的強烈需求，促使許多網路公司一一推出了自家的用語音輸入來檢索文字資訊的系統，如 Google 公司推出的語音檢索功能即可讓使用者在手機或瀏覽器的介面上以語音輸入，由 Google 將其辨識成文字後再於 Google 的搜尋引擎上檢索資訊。Apple 公司推出的個人語音助理 Siri，也讓使用者能以十分自然的方式對 Siri 說出想要查詢的查詢詞 (Query)，由 Siri 辨識後在網路上檢索，並將檢索結果分門別類整理好後呈現給使用者看。如上述所說的這類檢索系統是用語音輸入的查詢詞去檢索大量的文字資訊，此方法稱為人聲檢索 (Voice Search)，和本論文所探討的語音數位內容檢索 (Spoken Content Retrieval) 完全不同。

本論文所探討的語音數位內容檢索，是指由於網路上有大量的多媒體文件，如線上影片、會議錄音、線上課程、電視連續劇、演講等，而使用者也有搜尋這



些多媒體文件的需求，此類允許使用者用文字或聲音輸入查詢詞並搜尋語音數位內容 (Spoken Content) 的系統稱為語音數位內容檢索 (Spoken Content Retrieval)，如 TED (美國著名的演講網站) 會將網站上的演講內容轉寫 (Transcript) 成為文字，並允許使用者於網站上輸入文字檢索這些影片的文字稿。Youtube 也會於離線時將其網站上的影片辨識成文字，但目前尚不支援直接輸入查詢詞檢索影片轉寫的方式，可以期待未來 Youtube 會開放這方面的功能。只是上述兩個例子仍要倚賴人工的轉寫，要完全只靠機器自動辨識仍不容易做到。這種語音數位內容檢索將是本論文主要的研究主軸。

由於以上所述的語音數位內容檢索系統大部分都是回傳給使用者有出現查詢詞的語音文件，但如此一來使用者必須完美地輸入有出現在語音文件中的查詢詞，如果使用者心中想的概念與語音文件中的詞彙不匹配，則檢索系統的成效就會大大地降低。使用者通常期待的是系統會查到所有與查詢詞「語意上相關」的文件，比如查詢詞為「東京旅遊」的話，使用者想要查詢到的文件通常包含與「東京住宿」、「東京景點」等有關的結果，而不只是那些包含了「東京旅遊」的文件。此即為語意檢索 (Semantic Retrieval) 的系統。語意檢索一個很常見的實作方法是查詢詞擴展 (Query Expansion)，查詢詞擴展的精神是先進行第一次檢索，得到第一次檢索結果 (First-pass Retrieval Result)，並從其中最相關的前幾篇中找出常常出現卻不是在所有文章中都常常出現的詞，再加入到原查詢詞中成為擴展後查詢詞 (Expanded Query)，再使用擴展後查詢詞進行第二次檢索 (Second-pass Retrieval)。

本論文想要探討的主題將是針對語音數位內容之語意檢索 (Semantic Retrieval of Spoken Content)，是指系統在接受到口語形式或文字形式的查詢詞之後，儘可能回傳給使用者所有與查詢詞語意上相關的語音文件。由於傳統的語音數位內容

之語意檢索系統主要的實作方法為先將語音文件辨識為文字檔後，對文字做檢索，但在辨識之中，會遇到如詞典外詞彙 (OOV)、辨識錯誤等情況，更甚者，許多語音中珍貴的資訊如韻律 (Prosody)、語速 (Speaking Rate)和語者特徵 (Speaker Characteristic)等在辨識後就消失了，十分地可惜。因此本論文試圖結合一套自動習得的聲學組型 (Automatically Discovered Acoustic Patterns) 至語音數位內容之語意檢索當中，以期改善傳統的檢索系統。

1.2 研究方向

本論文之研究方向為使用自動習得之聲學組型強化語音數位內容之語意檢索，主要包含以下幾點：

- 傳統的語意檢索系統是先將語音文件辨識為文字後，將輸入的文字查詢詞進行查詢詞擴展 (Query Expansion)，再用擴展後查詢詞對辨識後的文字進行檢索，但如此一來許多語音訊號中的珍貴的聲學資訊就消失了。因此本章中在文字的查詢詞擴展之外，再加入一套自動習得之聲學組型的查詢詞擴展，並結合兩套查詢詞擴展之結果回傳給使用者。
- 更進一步地，本論文希望能處理口語形式的查詢詞，可以用口語形式的查詢詞進行語音數位內容之語意檢索。另一方面，語意檢索通常需要自動語音辨識系統 (Automatically Speech Recognition) 將聲音辨識成文字，進而得到語意上的資訊，但自動語音辨識系統的訓練是很昂貴的，需要大量標注完善的語料才能訓練出很好的聲學與語言模型。因此本章中會將聲音辨識成聲學組型，並在聲學組型上進行查詢詞擴展，進而達到非監督式語音數位內容之語意檢索 (Unsupervised Semantic Retrieval of Spoken Content)。



- 由於聲學組型在訓練時是盡量將聲音很像之片段盡量分群 (Clustering) 在一起，但如此一來會使得同一個聲學組型中包含了大量同音但對應到不同字詞的聲學組型，會使得檢索系統的成效大幅下降。因此本章使用基於遞迴式類神經網路語言模型 (Recurrent Neural Network Language Model, RNNLM) 之詞表示法 (Word Representation) 將這些聲學組型按照句法 (Syntactics) 和語意 (Semantics) 進一步分群為不同的聲學組型，進而提升檢索系統之成效。
- 最後，由於近年來行動裝置與穿戴式裝置日漸流行，使用者也漸漸習慣於在行動裝置上用語音輸入並取得資訊，因此本論文基於 Google 眼鏡 (Google Glass) 上推出了一套語音翻譯系統與語音數位內容檢索系統，讓使用者能夠隨時隨地用最方便的方法取得新資訊。

1.3 章節安排

本論文之章節安排如下：

- 第二章：介紹本論文相關背景知識。
- 第三章：介紹如何以聲學組型改善監督式語意檢索。
- 第四章：介紹如何以聲學組型實現非監督式語意檢索。
- 第五章：介紹如何以遞迴式類神經網路語言模型產生之詞向量改善第四章的非監督式語音文件檢索。
- 第六章：介紹如何將本論文之語音檢索系統與語音翻譯系統實作到 Google Glass 上。
- 第七章：本論文之結論與未來研究方向。

第二章 背景知識



2.1 語音數位內容檢索

2.1.1 簡介

資訊檢索 (Information Retrieval) 系統一直以來都是研究人員與產業界關心的重點，但直到近年來由於網際網路、雲計算 (Cloud Computing) 的發達，使得網路上越來越多含語音資訊的多媒體檔案，我們稱之為語音數位內容 (Spoken Content)，語音數位內容檢索的基本使用流程為：當使用者輸入了一段查詢詞 Q (Query)，系統會進行檢索並回傳按照相關性 (Relevance) 排序後的語音文件 x (Spoken Documents)。此處每一篇文件與查詢詞的相關性被定義為 $S(Q, x)$ ，這個相關性函式通常是按照檢索系統的需求去決定的，可以是用機器學習 (Machine Learning) 的方法學習出來，也可以是由系統設計者決定。當系統接收到查詢詞 Q 後，系統會計算每一篇文件 x 與查詢詞 Q 的相關性 $S(Q, x)$ 並排序後回傳給使用者，本篇論文中所使用的查詢詞 Q 形式有二：文字構成的詞串、或口述形式。本篇論文使用的 x 則都是語音文件。

2.1.2 口述語彙偵測與語意檢索

語音數位內容的檢索可以分為兩類：「口述語彙偵測 (Spoken Term Detection)」與「語意檢索 (Semantic Retrieval)」，口述語彙偵測回傳有出現查詢詞的語音文件，語意檢索則是回傳概念上相關的語音文件，而不一定要出現查詢詞，在此將兩者簡介如下：



口述語彙偵測

口述語彙偵測的目的是檢索出所有包含查詢詞的語音文件，主要有兩種檢索情境。第一種是先將語音文件辨識為詞圖 (Lattice) 後，當使用者輸入文字的查詢詞後，系統會在詞圖上進行查詢詞的檢索。這種檢索情境需要有訓練好的自動語音辨識系統 (Automatic Speech Recognition, ASR)，由於語音辨識系統並無法保證在所有情況下都有很好的辨識率，因此由辨識錯誤所導致的檢索性能下降是在此最需要解決的問題。過去的文獻有用聲學特徵如梅爾倒頻譜係數 (Mel-Frequency Cepstral Coefficients (MFCC)) 幫助分類器 (Classifier) 在分類一篇語音文件是否相關時的判斷，也有利用相關回饋 (Relevance Feedback)、圖論 (Graph) 與隨機漫步 (Random Walk) 解決這些問題。本論文中會應用到這類的檢索系統，計算此情境下的 $S(Q, x)$ 本章用到的方法為語言模型檢索法 (Language Modelling Approach)，詳細方法列於 2.1.5。

第二種是非監督式 (Unsupervised)，的方法，查詢詞與語音文件都是語音形式的，也有人稱之為依例查詢 (Query-by-Example)，系統直接利用如動態時間校準 (Dynamic Time Warping) 等方法在信號上比對語音文件中是否有某一段聲音與查詢詞很相像，過去的方法通常是為了解決不同文件間語速上的差異提出如有斜率限制的動態時間校準 (Slope-Constraint DTW)，或是為了解決動態時間校準逐一比對所有文件庫所花時間過多的問題。本論文中會應用到這類的檢索系統，計算此情境下的 $S(Q, x)$ 為利用片段式動態時間規劃 (Segmental Dynamic Time Warping)，簡介於 2.1.6。

語意檢索

時至今日，大部分的語音數位內容檢索的研究是專注於口述語彙偵測的，但是



這是不夠的，因為使用者通常希望查詢結果與自己輸入的查詢詞是概念上匹配 (Concept Matching) 的，而不是只回傳有出現查詢詞的文件，比如當使用者查詢「東京旅遊」，他可能期待查到包含「東京旅館」、「東京鐵塔」的文件，而不是與查詢詞完全一樣的文件。文字領域的資訊檢索已經有「概念匹配」的做法來達到語意檢索，但是由於文字領域的資訊檢索是在文字為完全正確的狀況下，不像語音數位文件的檢索往往會有辨識錯誤的問題，因此本論文提出了利用自動尋找的聲學組型來解決這個問題，語意檢索也是本論文的主軸。一個較常見的實作方法通常是使用相關回饋 (Relevance Feedback) 與查詢詞擴展 (Query Expansion)，分別介紹於 2.2.3 和 2.2.4。

2.1.3 詞圖與唯一最佳序列

當給定了一段語音文件，自動語音辨識系統 (Automatic Speech Recognition, ASR) 能夠將這段語音文件辨識成兩種格式：詞圖 (Lattice) [4] 與唯一最佳序列 (One-Best Transcription)，本論文中兩種形式的辨識格式都會使用，故介紹如下：

詞圖像張網一樣，如圖 2.1，將每個時間所有可能的詞都呈現出來，而唯一最佳序列只呈現了詞圖上一條最可能的詞串當作辨識結果而已。通常在語音檢索時會使用詞圖來搜尋，因為語音辨識通常沒有辦法做到完全準確，可能會辨識成音很像的其他詞彙，比如「美國」可能會被辨識成「沒過」等等，如果使用唯一最佳序列的話，選中了正確的詞彙固然很好，但如果沒選中就會完全搜尋不出来了，而使用詞圖的話，即使唯一最佳序列被辨識錯了，通常正確的詞彙都還是會被表示在詞圖上，所以一般來說會使用詞圖來表示辨識的結果。

如圖 2.1 所示，詞圖主要由 N, A 所構成，其中 N 是所有節點 (Node) 的集合，而 A 則是所有的詞弧 (Word Arc) 的集合。節點上存有各節點的時間資訊，而詞

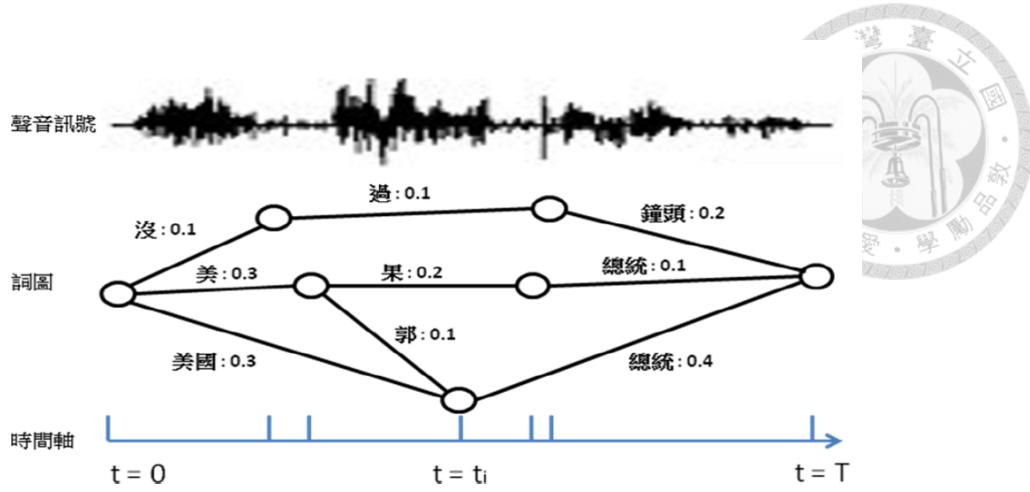


圖 2.1: 詞圖示意圖

弧上除了有起始節點和終止節點外，還包含了在這個詞弧上的假設詞 (Word Hypothesis) 與此假設詞的信心分數 (Confidence Score)，而此假設詞和信心分數都是由聲學模型和語言模型所計算出來的。

語音辨識系統的輸入是聲學特徵，並找出最有可能的詞彙串。因此相當於最大化以下的式子：

$$w_{seq}^* = \operatorname{argmax}_{w_{seq} \in W_{seq}} P(w_{seq}|O) \quad (2.1)$$

O 為輸入訊號之聲學特徵， w_{seq} 為某個詞串， W_{seq} 為所有 w_{seq} 之組合， argmax 則是尋找一個 w_{seq} 使得 $P(w_{seq}|O)$ 最大，此 w_{seq} 即為 w_{seq}^* 。但 $P(w_{seq}|O)$ 無法直接計算，所以根據貝氏定理 (Bayes' Theorem)，可以表示如下：

$$P(w_{seq}|O) = \frac{P(O|w_{seq})P(w_{seq})}{P(O)} \quad (2.2)$$

由於 $P(O)$ 對固定的 O 是常數，可以不考慮，上式可簡化成：

$$P(w_{seq}|O) = P(O|w_{seq})P(w_{seq}) \quad (2.3)$$

可以看得出來， $P(O|w_{seq})$ 可以由聲學模型求得， $P(w_{seq})$ 可以從語言模型中求得。求出來的 w_{seq}^* 即為唯一最佳序列。有時也會使用N最佳序列 (N-Best List)，概念與唯一最佳序列很像，只是找出前N個讓 $P(w_{seq}|O)$ 最大的 w_{seq} 。



2.1.4 辭典外詞彙 (Out of Vocabulary)

由於自動語音辨識系統可以辨識出的字彙在自動語音辨識系統被訓練時就被決定好了，因此當系統要辨識新的詞彙的時候，如果這套系統在辨識時遇到辭典中沒有的詞彙時，系統就無法辨識出來，此問題稱為辭典外詞彙問題 (Out of Vocabulary Problem)。如此一來，如果檢索系統發現查詢詞是辭典外詞彙，則這段查詢詞就無法被辨識，進而使得檢索系統的成效下降，更糟糕的是，由於查詢詞往往會是較少見的字彙，因此查詢詞是辭典外詞彙的機率有時會超過15%。

一個常見的解決辦法是使用次詞單位 (Subword Units)。先建立一套自動語音辨識系統，而這套系統是將聲音辨識成次詞單位的唯一最佳序列或是詞圖，當查詢詞進來後，也是將查詢詞轉換為次詞單位，之後系統再比對次詞單位的辨識結果與查詢詞的次詞單位版本，而在這套系統下，通常會需要字形轉音素的系統 (Grapheme-to-phoneme)，使得這個作法變得較為困難。

以詞為基礎的檢索 (Word-based) 和以次詞單位為基礎的檢索 (Subword-based) 各有優缺，詞的檢索會遇到很多辭典外詞彙的問題，使得檢索系統的成效大幅下降，而次詞單位為基礎的檢索則能提高召回率 (Recall)，但由於次詞單位不是代表字義的最小單位，因此檢索回來的文件準確率則會下降。所以比較好的方法是同時使用詞為基礎的檢索和以次詞單位為基礎的檢索，系統同時利用詞和次詞單位進行檢索，並將兩者的檢索結果進行疊加 (Interpolation)，此方法需要決定疊加時的權重，而此權重通常可以利用一套查詢詞的訓練集完成，本論文中的檢索也是同時使用詞與次詞單位為基礎的檢索。



2.1.5 檢索系統

計算文字查詢詞與詞圖的相關性 $S(Q, x)$ (使用語言模型檢索)

檢索系統的主要目的是當使用者輸入查詢詞 Q 時，系統能夠計算出每個文件 x 與查詢詞間的相關分數 $S(Q, x)$ ，進而將此相關分數排序後回傳給使用者。當系統接收到語音訊號時，系統會先對其抽取出聲學訊號，再經過聲學模型和語言模型辨識成唯一最佳序列或詞圖，一般來說，語音檢索系統使用詞圖的表現會比較好，所以以下主要就如何計算出查詢詞和詞圖的相關分數 $S(Q, x)$ 加以介紹。

由於本章使用的檢索方式是語言模型檢索 (Language Modelling Retrieval) [5, 6]，因此首先要把辨識所得的詞圖轉換為語言模型。對每個詞圖中的詞 t ，它在詞圖中的期望出現次數 (Expected Count) 可以如此計算：

$$E[t|x] = \sum_{\mu \in L(x)} N(t, \mu) P(\mu|x) \quad (2.4)$$

$L(x)$ 是 x 的詞圖中所有的路徑， μ 是 $L(x)$ 中的一條路徑， $N(t, \mu)$ 是 t 在 μ 中出現的次數， $P(\mu|x)$ 是路徑 μ 的事後機率 (Posterior Probability)。

有了 $E[t|x]$ 之後，就能把詞圖表示成單連詞 (Uni-gram) 語言模型 θ_x ：

$$P(t|\theta_x) = \frac{E[t|x]}{\sum_t E[t|x]} \quad (2.5)$$

因為 θ_x 裡沒有包含每個詞，為了讓 θ_x 中每個詞都有一點機率，會再把 θ_x 與一個背景語言模型 (Background Language Model) θ_b 做線性疊加，此過程稱為平滑化 (Smoothing) [7]， θ_b 可以如此估計：

$$P(t|\theta_b) = \frac{\sum_{x \in C} E[t|x]}{\sum_t \sum_{x \in C} E[t|x]} \quad (2.6)$$



C 是所有文件 x 的集合。

同樣地，查詢詞 Q 也可以被表示成語言模型 θ_Q ：

$$P(t|\theta_Q) = \frac{N(t, Q)}{|Q|} \quad (2.7)$$

$N(t, Q)$ 是詞 t 出現在 Q 中的次數，而 $|Q|$ 是 Q 中詞的總數。

有了 $\theta_x, \theta_b, \theta_Q$ 之後，就可以計算 θ_x 與 θ_Q 之間的相關分數 $S(x, Q)$ 了，由於 $\theta_x, \theta_b, \theta_Q$ 都是機率分布 (Probability Distribution)，所以在這裡選擇了KL散度 (Kullback–Leibler divergence) 用來計算兩個語言模型之間的距離，KL散度的計算方式如下：

$$KL(\theta_x|\theta_Q) = \Pi_{w \in V} P(w|\theta_x)^{P(w|Q)} \quad (2.8)$$

於是定義文件 x 和查詢詞 Q 之間的相關分數 $S(x, Q)$ 如下：

$$S(x, Q) = -[(1 - w_1)KL(\theta_q^w|\bar{\theta}_x^w) + w_1KL(\theta_q^s|\bar{\theta}_x^s)] \quad (2.9)$$

$\bar{\theta}_x$ 是將 θ_x 與 θ_b 疊加後的語言模型，上標 w 代表的是用以詞為基礎的詞圖產生的語言模型，上標 s 代表的是用以次詞 (Subword) 為基礎的詞圖產生的語言模型，上式是分別對詞為基礎的語言模型和次詞為基礎的語言模型做檢索，再將兩者得到的相關分數用 w_1 做線性疊加，最後再將此分數排序後回傳給使用者。

2.1.6 片段式動態時間校準 (Segmental DTW)

口述語彙偵測的目的是要搜尋整個語料庫後，找出其中有出現查詢詞的語音文件，並且找出這些語音文件中可能有出現查詢詞的假設區域 (Hypothesized Region)



，並給這個假設區域一個相關分數，最後系統再根據相關分數進行排序後回傳給使用者(分數較高為較相關)。

這裡考慮的口述語彙偵測是查詢詞也是語音形式的情境。此時所用的方法為片段式動態時間校準 (Segmental Dynamic Time Warping) [8,9]。假設輸入的查詢詞的特徵為 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{|\mathbf{X}|})$ ，語音文件的特徵為 $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_{|\mathbf{Y}|})$ ，有了這兩者之後，就可以建立一個距離表格 (Distance Table) $D(i, j) = \rho(\mathbf{x}_i, \mathbf{y}_j)$ ，通常這裡使用的特徵是 高斯事後機率 (Gaussian Posteriorgrams) [1]或是梅爾倒頻譜係數 (MFCC)。如果使用高斯事後機率的話，兩個音框之間的距離為 $\rho(\mathbf{x}_i, \mathbf{y}_j) \equiv -\log(\mathbf{x}_i \cdot \mathbf{y}_j)$ 。如果使用梅爾倒頻譜係數的話，兩個音框之間的距離為兩點之間的歐幾里得距離 (Euclidean Distance)，即 $\rho(\mathbf{x}_i, \mathbf{y}_j) \equiv \sqrt{|\mathbf{x}_i - \mathbf{y}_j|^2}$

動態時間校準的目的是要在 $D(i, j)$ 上找一條距離總合最短的路徑從 $(1, s)$ 到 $(|\mathbf{X}|, e)$ 表示從 \mathbf{X} 對應到 $(\mathbf{y}_s, \dots, \mathbf{y}_e)$ (因為查詢詞一定要被完全對應，而語音文件不一定要被完全對應到)。由於假設區域可以出現在語音文件中的任何地方，因此片段式動態時間校準將距離表格 $D(i, j)$ 切成數個重疊的對角片段 (寬度為 R)，所以說每個片段的起始點分別為 $(1, 1), (1, 1 + R), (1, 1 + 2R), \dots$ 如圖 2.2 所示，每個對角片段都代表了一個可能出現查詢詞的區域，所以要在每個對角片段上找出距離總合最短的路徑，即這個對角片段上的假設區域。片段式動態時間校準會從每個片段中找出一段距離總合最短的路徑，在每個片段中所有對應的路徑 (Warping Path) 都必須要完整地待在片段內，不可超出片段。

假設在每個片段中的對應路徑為：

$$\phi = (i_t, j_t), t = 1, \dots, |\phi|$$

代表著如下的對應關係：

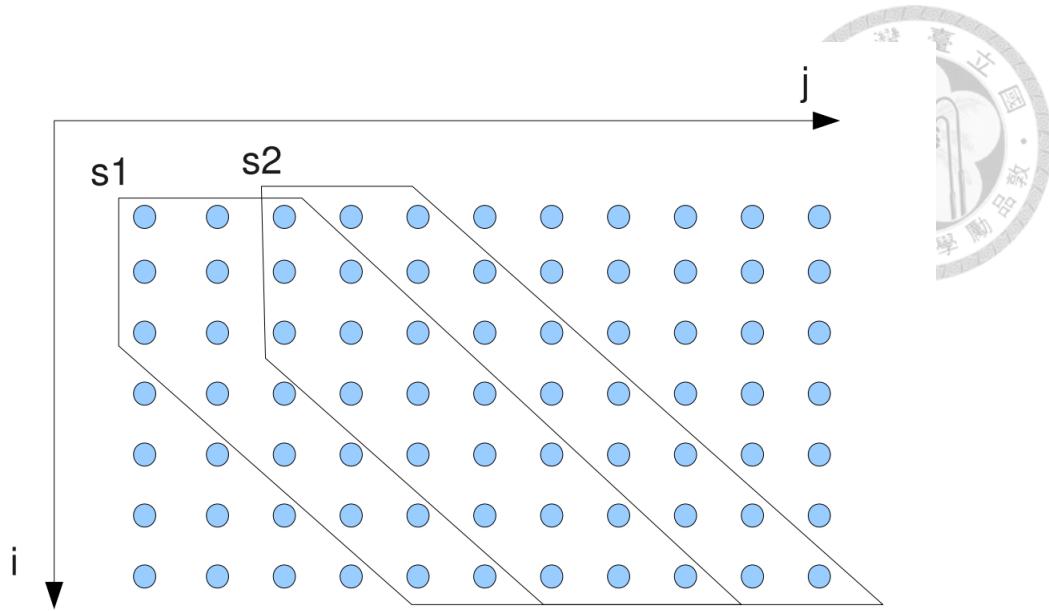


圖 2.2: 片段式動態時間校準示意圖 [1]

$$\mathbf{x}_{i_1} \leftrightarrow \mathbf{y}_{j_1}, \mathbf{x}_{i_2} \leftrightarrow \mathbf{y}_{j_2}, \dots, \mathbf{x}_{|\phi|} \leftrightarrow \mathbf{y}_{|\phi|},$$

而邊界條件是 $i_1 = 1, i_{|\phi|} = |\mathbf{X}|$ ， j_1 為 $1 + kR$ ，對應路徑中所有的點都要在片段內，即：

$$|(i_t - i_1) - (j_t - j_1)| \leq R$$

而片段式動態時間校準的目標是要在每個片段內找到一條路徑 ϕ 使得下式的距離總和最小：

$$C_\phi(\mathbf{X}, \mathbf{Y}) = \sum_{t=1}^{|\phi|} \rho(\mathbf{x}_{i_t}, \mathbf{y}_{j_t}) \quad (2.10)$$

每個片段中使得 $C_\phi(\mathbf{X}, \mathbf{Y})$ 最小，即使得相關分數 $-C_\phi(\mathbf{X}, \mathbf{Y})$ 最大的那條 ϕ 即為每個片段中的假設區域 $(\mathbf{y}_{j_1}, \dots, \mathbf{y}_{j_{|\phi|}})$ ，而在每個片段中找到最大相關分數的路徑可以使用動態規劃 (Dynamic Programming) 求解。圖 2.3 中顯示了兩個對角片

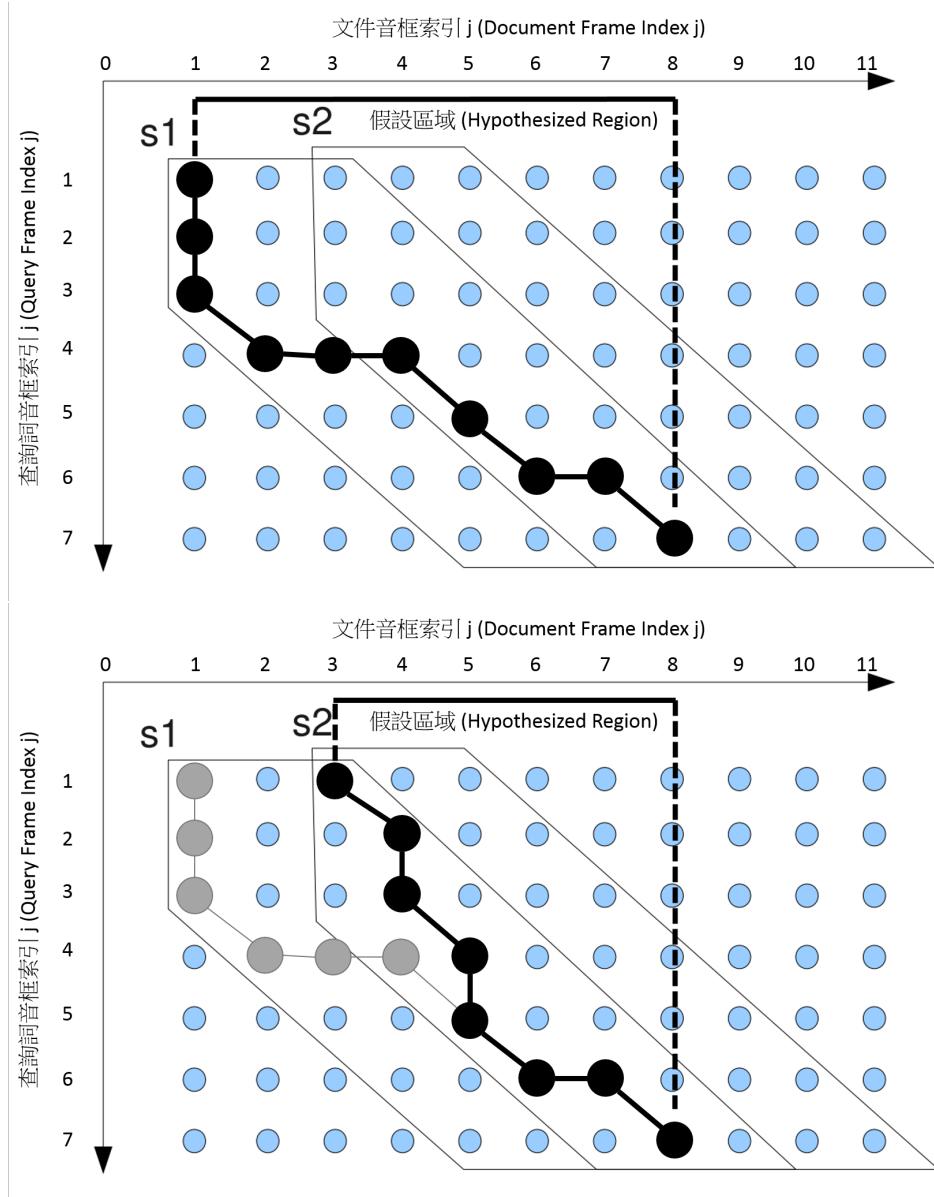
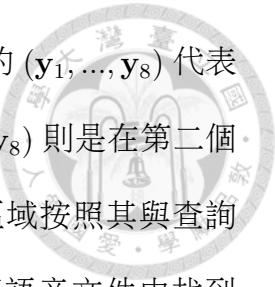


圖 2.3: 示意如何從兩個對角片段中找到相關分數最大的假設區域，其中 $R = 2$



段與其對應的最大相關分數的路徑與假設區域。圖中上半部中的 (y_1, \dots, y_8) 代表在第一個對角片段中找到的假設區間，圖中下半部中的 (y_3, \dots, y_8) 則是在第二個對角片段中找到的假設區域。找到所有假設區域後，將假設區域按照其與查詢詞的相關分數進行排序後，即為口述語彙偵測的結果。在一篇語音文件中找到所有可能的假設區域需要 $O(|X||Y|)$ 的計算量來計算點與點之間的距離 ρ ，並需要 $O(|X||Y|)$ 的計算量來找到相關分數最大的路徑。

2.1.7 資訊檢索評估機制

為了讓研究人員能夠比較彼此系統之成效，制定資訊檢索評估機制的標準是很重要的一環，本節將介紹此篇論文使用的評估機制。

準確率(Precision)與召回率(Recall)

準確率越高代表所找出的檢索結果越可靠，而召回率越高的話代表系統找回越多相關的檢索目標，通常會為系統設定一個閥值(Threshold)，文件的分數若高於閥值，則視為相關，反之若文件的分數低於閥值，則視為不相關。準確率和召回率的定義如下：

$$\text{準確率} = \frac{\text{檢索到的相關檢索對象數}}{\text{檢索到的檢索對象數}}$$

$$\text{召回率} = \frac{\text{檢索到的相關檢索對象數}}{\text{所有的相關檢索對象數}}$$

通常這兩個值彼此之間的關係為負相關。調高閥值的話準確率會上升，但召回率則會下降；反之若調低閥值，準確率會因此下降，召回率則會很高。可以考慮一個極端例子：當閥值非常低時，幾乎所有的文件都是相關文件，此時的召回

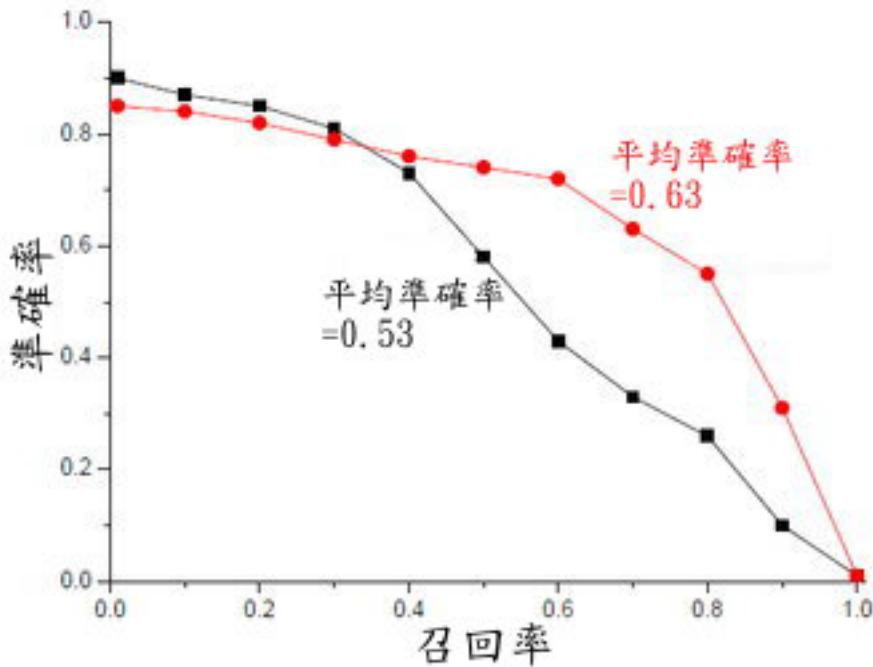


圖 2.4: 準確率、召回率和平均準確率之關係

率相當於1，但準確率就會很低了。因此單看準確率或召回率是無法準確地評估系統的優劣的，必須要兩者一起評估。

P@N

通常使用者最重視的是檢索系統傳回的前幾名結果，所以就發展出了 $P@N$ 這個評估機制。 $P@N$ 就是只看前 N 個檢索結果的正確率。例如：前五個檢索結果中有一個是相關的，那 $P@5$ 就是 20%。

$P@N$ 的定義如下：

$$P@N = \frac{\text{前 } N \text{ 個文件裡的相關文件數}}{N}$$

平均準確率 [10]

因為準確率和 $P@N$ 都需要事先決定，當查詢詞和條件不同時，很難準確地評估



兩個系統的效能。因此有人提出了平均準確率(Mean Average Precision, MAP)的概念，如圖 2.4，平均準確率就是準確率和召回率曲線下面積的平均值。平均準確率的定義如下：

$$MAP = \frac{1}{|Q|} \sum_Q \frac{\sum_{d \in D^R} precision(d)}{|D^R|} \quad (2.11)$$

其中Q代表查詢詞的集合， $|Q|$ 為查詢詞的總數， D^R 為和查詢詞Q相關的文件d的集合， $|D^R|$ 代表和查詢詞Q相關的文件數量。precision(d)代表系統檢索出文件d時的準確率。

2.2 相關回饋

相關回饋 (Relevance Feedback) 是資訊檢索的一項重要的技術 [11]，通常可以顯著地提升系統的成果。相關回饋基本的架構如圖 2.5，使用者輸入查詢詞之後，系統會先根據查詢詞與所有文件的相關分數 $S(Q, d)$ 排序出第一次檢索結果 (First-pass Result)。然後把第一次檢索結果中部分文件標注為與查詢詞相關的正例 (Positive Example)，部分文件標注為與查詢詞非相關的反例 (Negative Example)，系統會再根據標注的結果重新計算查詢詞與文件的關係，把調整後的結果呈現給使用者看。

把部分文件標注成正例與反例的方法可以大致分為以下幾種：

2.2.1 外顯回饋

外顯回饋 (Explicit Feedback) 的做法通常是由使用者告訴系統文件是否為正例或反例，進而增進檢索的成果。外顯回饋可分為使用者回饋 (User Feedback) 和積極回饋 (Active Feedback)，使用者回饋是由使用者依據第一次檢索結果按照文件相關性進行標注 [12–14]；積極回饋則是由系統主動詢問使用者某文件為正例或反例

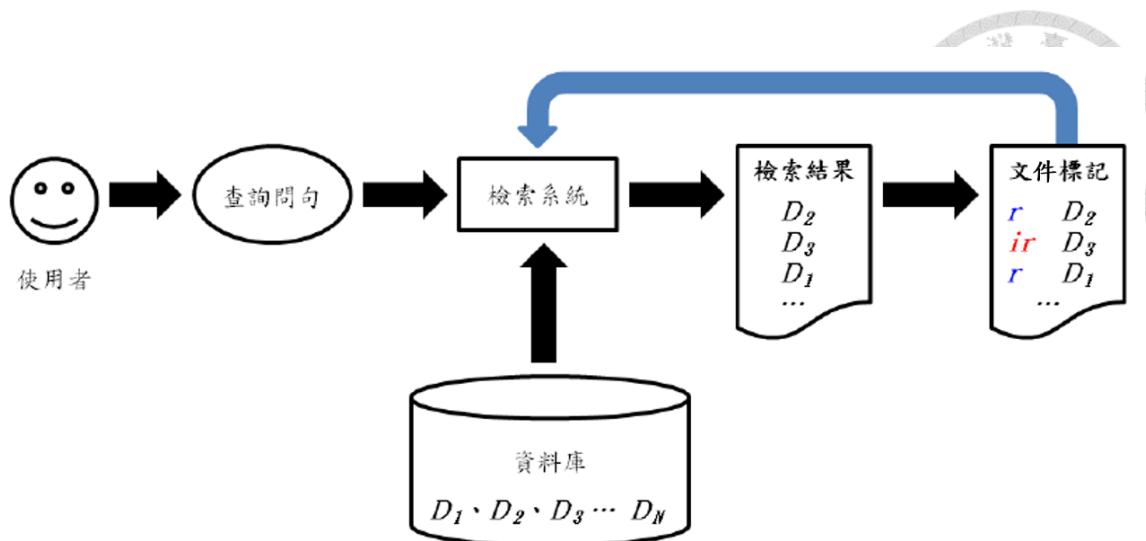


圖 2.5: 相關回饋的基本架構

(通常是系統無法決定的文件)，此方法可以盡量減少使用者的標注量以改進系統的成果 [15–17]。

2.2.2 隱含回饋

隱含回饋 (Implicit Feedback) 是不由使用者直接提供正反例的資訊，而且透過觀察使用者的行為 (Behavior) 而分析文件的相關性，因此使用者並不知道自己正在回饋給系統。最常用的方法是點擊數據 (Click-through Data)，比如檢索系統的前幾名如果是 d_1, d_2, d_3, \dots ，而使用者沒有檢視 d_1 ，而是直接檢視 d_2 ，如此一來系統就可以假設 d_1 是不相關的文件，而 d_2 是相關的文件。除此之外，還可以利用查詢記錄 (Query Log)、網頁捲動 (Scrolling) 和滑鼠軌跡 (Mouse Movements) 等 [18]。

2.2.3 虛擬回饋

虛擬回饋 (Pseudo Feedback) 不需要使用者的參與，在系統產生第一次檢索結果後，系統會直接假設其中某些部分為正例，某些部分為反例（通常是直接假設相

關分數最高的 N 篇為正例，相關分數最低的 N 篇為反例）。系統會再根據這些假設進行進一步的檢索，在這個過程中，系統雖然沒有得到額外的資訊，但系統的成效通常能有一定的提升 [19–26]，而虛擬回饋也是此篇論文的主要研究主題之一。

2.2.4 檢索詞擴展與文件擴展

語意檢索最大的困難點是由於最理想的檢索結果不一定要包含查詢詞，假設查詢詞是「東京旅遊」，則包含「東京鐵塔」的文件的 $S(Q, d)$ 則會非常低，常見的解決方法是使用查詢詞擴展 (Query Expansion) 和文件擴展 (Document Expansion)，一個是將許多與查詢詞語義上相關的詞加入原查詢詞中成為新的查詢詞模型，另一是將文件加入更多語義上相關的詞，這裡使用的是潛藏語意分析 (Probabilistic Latent Semantic Analysis, PLSA) 來找到與語義上相關的詞。這兩個方法都能解決查詢詞與文件中詞彙不匹配的問題，以下將分別介紹：

查詢詞擴展 (Query Expansion)

這裡使用正規化查詢詞混合模型 (Query-Regularized Mixture Model) 作為查詢詞擴展的實作方法，這套方法假設每篇文件都是由查詢詞相關詞彙 (Query-Related Terms) 和一般詞彙 (General Words) 所組成，而這兩者的比例在每篇文件中都是不一樣的。舉例來說，當一篇實際上不相關的文件被錯誤地假設成虛擬相關文件時，這個比例應該要很低，而反之亦然。然而實際上這個比例是不知道的，不過可以從虛擬相關文件中估計出來這個比例，估計完之後新的查詢詞模型稱為 θ'_Q ，用來取代原本的 θ_Q 。

假設所有的文件為集合 D ，其中包含了文件 (已經過第一次檢索排序後)

$d_1, d_2, \dots, d_n, \dots$ ，因此虛擬相關的文件為 $d_1, d_2, \dots, d_m, \dots, d_M$ ，而 M 是虛擬相關文件的總數。每一篇文件中的詞都可以看作是由背景語言模型 (Background Language Model) θ_b 產生，或是由新的查詢詞模型 θ'_Q 產生， α_{d_m} 是 d_m 這篇文章中詞彙由 θ'_Q 產生的機率，反之 $(1 - \alpha_{d_m})$ 則是由背景模型 θ_b 產生的機率。透過最大概似估計最大化下式即可以估計出 θ'_Q 和每一篇 d_m 的 α_{d_m} ：

$$F_1(\theta'_Q, \alpha_{d_1}, \dots, \alpha_{d_M}) = \prod_{m=1}^M \Pi_w (\alpha_{d_m} P(w|\theta'_Q) + (1 - \alpha_{d_m}) P(w|\theta_b))^{P(w|\theta_{d_m})} \quad (2.12)$$

上式中，產生詞彙 w 的機率為 $\alpha_{d_m} P(w|\theta'_Q) + (1 - \alpha_{d_m}) P(w|\theta_b)$ 。因此上式可視為由新的查詢詞模型產生這些虛擬相關文件的可能性 (Likelihood)。然而，如果只最大化式 2.12， θ'_Q 中主要會包含這些虛擬相關文件的主題，而不一定是查詢詞相關的詞彙，為了要解決這個問題，比較好的方法是用原本的查詢詞模型 θ_Q 正規化 (Regularize) θ'_Q ，因此定義 $F_2(\theta'_Q)$ 如下：

$$F_2(\theta'_Q) = \Pi_w P(w|\theta'_Q)^{P(w|\theta_Q)} \quad (2.13)$$

當 θ'_Q 與 θ_Q 越近時， $F_2(\theta'_Q)$ 的值會越大。

有了式 2.12 和式 2.13 後，只要最大化下式即可估計 θ'_Q 和 α_{d_m} ：

$$F(\theta'_Q, \alpha_{d_1}, \dots, \alpha_{d_M}) = F_1(\theta'_Q, \alpha_{d_1}, \dots, \alpha_{d_M}) F_2(\theta'_Q)^\lambda \quad (2.14)$$

λ 是個用來控制 $F_2(\theta'_Q)$ 影響力的參數， λ 越大，新的查詢詞模型 θ'_Q 就會跟原本的查詢詞模型 θ_Q 越像。最大化式 2.14 的好處是新的查詢詞不會過適 (Overfit) 到虛擬文件，而會保留與原本的查詢詞模型 θ_Q 一定程度的相似性。

最大化式 2.14 可以採用 EM 演算法 (Estimation-Maximization Algorithm)：

E step: 對於 d_1, d_2, \dots, d_M 中的每篇文件 d_m 中的每一個詞彙 w :



$$P(R|w, d_m) = \frac{\alpha_{d_m} P(w|\theta'_Q)}{\alpha_{d_m} P(w|\theta'_Q) + (1 - \alpha_{d_m}) P(w|\theta_b)} \quad (2.15)$$

其中 $P(R|w, d_m)$ 為當給定文件 d_m 中的詞彙 w 時，此文件 d_m 中的詞彙 w 與查詢詞是語意上相關的機率。

M step : 對於 d_1, d_2, \dots, d_M 中的每篇文件 d_m ，其中每篇語音文件 d_m 的語言模型 (使用詞單位) 為 θ_d :

$$\alpha_{d_m} = \sum_w P(R|w, d_m) P(w|\theta_d) \quad (2.16)$$

對每個詞彙 w :

$$P(w|\theta'_Q) = \frac{\lambda P(w|\theta_Q) + \sum_{m=1}^M P(w|\theta_d) P(R|w, d_m)}{\lambda + \sum_w \sum_{m=1}^M P(w|\theta_d) P(R|w, d_m)} \quad (2.17)$$

重覆地執行 E Step 和 M Step 即可得到新的查詢詞模型 θ'_Q

文件擴展 (Document Expansion)

這裡使用潛藏語意分析來進行文件擴展，潛藏語意分析使用了一組潛藏主題變數 (Latent Topic Variables) $Z_t, t = 1, 2, \dots, T$ ，其中 T 是主題 (Topics) 的數目。所以當給定了所有的語音文件，潛藏語意分析訓練後就能夠提供 $P(w|Z_t)$ ，即在給定每一個潛藏主題 Z_t 下看到詞 w 的機率，與 $P(Z_t|d)$ ，即對於每一篇文件 d 的潛藏主題分布 (Topic Distributions)。因此根據潛藏主題分析所得之 $P(w|Z_t)$ 和 $P(Z_t|d)$ ，可以估計出在每篇文件 d 中出現詞 w 的機率為：

$$P_{plsa}(w|d) = \sum_{t=1}^T P(w|Z_t) P(Z_t|d) \quad (2.18)$$

其中 $P(w|Z_t)$ 和 $P(Z_t|d)$ 的訓練過程是藉由 EM 演算法來最大化下式的目標函

數：

$$L = \sum_{d \in \mathcal{C}} \sum_w P(w|\theta_d) \log P_{plsa}(w|d) \quad (2.19)$$

其中 θ_d 是語音文件 d 的語言模型，式 2.19 可以視為是尋找一組 $P(w|Z_t)$ 和 $P(Z_t|d)$ 使得語音文件的語言模型 θ_d 與 $P_{plsa}(w|d)$ 的 KL 分歧度最小。

得到 $P_{plsa}(w|d)$ 後，我們就可以利用 P_{plsa} 為每一篇語音文件 d 產生一組新的背景語言模型，而這套新的背景語言模型是基於它的潛藏主題產生的。如下式所示，新的背景語言模型是將 $P_{plsa}(w|d)$ 與原本的背景語言模型 θ_b 進行線性疊加所得：

$$P(w|\theta_b^d) = b_d P_{plsa}(w|d) + (1 - b_d) P(w|\theta_b) \quad (2.20)$$

其中 b_d 是根據每一篇語言文件所得的疊加權重，定義為 $\frac{L_d}{L_d+b}$ ， L_d 是文件 d 的長度， b 為一常數。接著當系統在計算查詢詞 Q 與每一篇文件 x 的相關性 $S(Q, x)$ 時，即可使用新的 θ_b^d 當作 d 的背景語言模型來作平滑化，如 2.1.5 中所示，如此一來，經過平滑化的 θ_d 即會與 d 的主題呈高度的相關，即達到我們想要將與文件 d 概念上相同的詞加入的目的。

2.3 自動習得之聲學組型

近年來，如何尋找非監督式的信號組型 (Unsupervised Signal Pattern) 是很熱門的研究主題之一，這些信號組型可以被應用於語音數位內容分類、口述語彙偵測、音樂檢索、影片檢索和語音數位內容檢索等等，但尚未被用於語音數位內容的語意檢索。

本論文中使用一套自動習得之聲學組型 (Automatically Discovered Acoustic Patterns) ，是一套兩層式 (Two-level) 的聲學組型，一層是類似詞單位的聲學組型 (Word-like acoustic patterns) (以下稱類詞聲學組型)，另一層是類似次詞單位的聲學組型 (Subword-like acoustic patterns) (以下稱類次詞聲學組型)，其中類詞聲學組型是由一連串的類次詞聲學組型組合而成，並得到一套辭典 (Lexicon) 用以表示類詞聲學組型是如何由類次詞聲學組型組合而成的，也同時得到類詞聲學組型的語言模型。每個類次詞聲學組型都是訓練成一個隱藏式馬可夫模型 (Hidden Markov Model, HMM) ，所有的參數包括類次詞聲學組型的隱藏式馬可夫模型的參數、類次詞聲學組型的數目、類詞聲學組型的數目和類詞聲學組型的 N 連語言模型的參數都是自動地在非監督式的情形下從目標語料集習得的。

整套聲學組型的訓練可以主要可以分為四個步驟：初始化、聲學最佳化、語言最佳化、詞彙最佳化；在每個階段結束後，對每個聲音片段都會有一套假定的標識符 (Label) W_i ，並在每一個階段中進行最佳化的動作後產生新假定的標識符 W_{i+1} 後進入下一個階段：

- 初始 (Initialization)：初始化採用的是由上而下的方法，先將語句利用能量 (Energy) 上的不連續點切成類詞的片段，並在這些類詞片段上進行 Watershed 轉換 [27]，如圖 2.6，將其切成數個類次詞的片段。接著在這些類次詞的片段上抽取特徵，並進行 K 平均分群演算法，之後給每一個群標注不同的標識符 (ID)，稱為 W_0 ，如此一來即可得到最原始的聲學組型，此時類詞聲學組型如何由類次詞組成的關係也決定了最原始的辭典。
- 聲學最佳化 (Acoustic Optimization)：此階段中，會使用上一個階段產生的 W_i 訓練隱藏式馬可夫模型，從 W_i 中也可以得到一套辭典用以表示類詞聲學組型如何由類次詞聲學組型組合而成，有了聲學模型和辭典後，就可以辨識

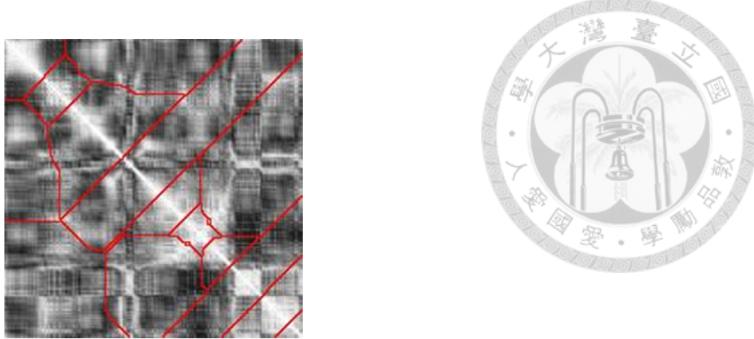


圖 2.6: 在類詞片段上進行的Watershed轉換，紅線部分為可能的次詞邊界

- (Decode) 出新的標識符 W_{i+1} 進入下一個階段。
- 語言最佳化 (Linguistic Optimization): 此階段和前一階段十分類似，唯一的差別是在此用前一階段的標識符 W_i 估算了一組類詞聲學組型的 N 連語言模型用在辨識當中，並辨識成新的標識符 W_{i+1} ，這個 N 連語言模型能幫助辨識出更好的 W_{i+1} ，特別是在類詞聲學組型常常重覆出現的情況下。
 - 詞彙最佳化 (Lexical Optimization): 此階段用來建立新的類詞聲學組型，首先將原有的類詞聲學組型都拆散成類次詞聲學組型，並重新尋找適合組合成類詞聲學組型的類次詞聲學組型串列，尋找的方法主要是找到某一串類次詞聲學組型其左和右的上下文變化 (left and right context variation) 夠大，並且出現足夠多次，則這一串類次詞聲學組型就會被組合成新的類詞聲學組型，而這個過程可以使用 PAT 樹在原有的標識符 W_i 上訓練而成。新的辨識結果標識為 W_{i+1} 。
- 結束以上的訓練過程後，我們就可以得到聲學組型的聲學模型、語言模型與辭典，於是這套聲學/語言模型與辭典可以被用來建立一套聲學組型辨識器 (Acoustic Pattern Decoder)，並藉由這個辨識器將聲音辨識成聲學組型的最佳序列以供後面的實驗使用。

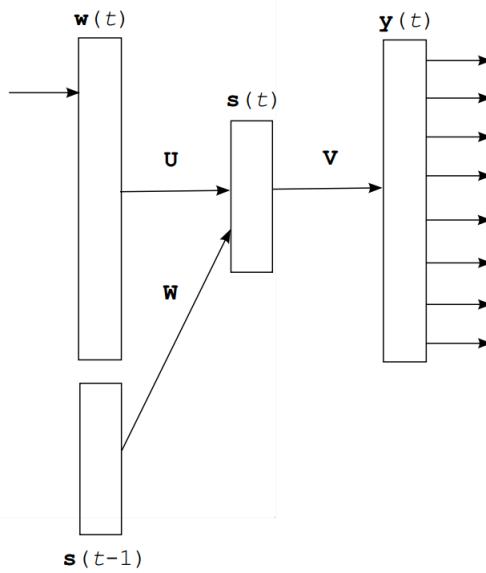


圖 2.7: 遞迴式類神經網路語言模型示意圖

2.4 遞迴式類神經網路語言模型

模型結構

如圖 2.7 所示，遞迴式類神經網路語言模型主要可以分為三個層級：

1. 輸入層 (Input Layer)：是一個與辭典大小同長度的陣列，採用的是1-of-N編碼 (1-of-N Encoding)形式，輸入為前一個詞在辭典裡的索引 $w(t)$ ，而只有該索引的位置的值為一，其餘皆為零。
2. 輸出層 (Output Layer)：同樣是一個與辭典大小同長度的陣列，也是採用1-of-N編碼形式，代表的是該模型對下一個字出現機率分布的預測 $y(t)$ 。
3. 潛藏層 (Hidden Layer)：或稱上下文層(Context Layer)，通常維度較小，代表的是該時間點此模型保存的上下文資訊 $s(t)$ 。此層與輸入層、輸出層各有一組權重矩陣 (Weight Matrix) ， \mathcal{U} 和 \mathcal{V} ，與上一個時間點的潛藏層向量 $s(t - 1)$ 也存在一組權重矩陣 \mathcal{W} ，其用意是為了模擬上下文間的相依關係。

有了以上三層與矩陣後，可以將輸入層、潛藏層及輸出層的關係表示如下：



$$s(t) = f(\mathcal{U} \times w(t) + \mathcal{W} \times s(t-1)) \quad (2.21)$$

$$y(t) = g(\mathcal{V} \times s(t)) \quad (2.22)$$

其中的 $f(\cdot)$ 和 $g(\cdot)$ 分別為邏輯函式 (Logistic Function) 與 Soft-Max 函式：

$$f(z) = \frac{1}{1 + e^{-z}} \quad (2.23)$$

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \quad (2.24)$$

最佳化演算法-沿時間反向傳播

沿時間反向傳播的核心精神是基於反向傳播演算法 (Back Propagation) 而來，反向傳播演算法是訓練類神經網路 (Neural Network) 的最佳化演算法，而反向傳播演算法主要的步驟包含了以下三個部分：

1. 紿定一筆訓練範例 (Training Example) 及現階段的模型參數組合，從輸入層向輸出層做順向傳遞 (Forward Pass)，輸出現階段模型的預測結果 y_j ， $0 < j < |V|$ 。
2. 根據第一步中模型的輸出結果 y_j ，與真正的結果 d_j 間計算誤差值，在語言模型中誤差函數 (Error Function) 為交叉熵 (Cross Entropy)：

$$E = \sum_j d_j \log y_j \quad (2.25)$$

3. 根據第二步算出來的誤差函數，計算其一階倒數 (First Derivative)，調整進來的權重 (Incoming Weight)，並重複將這個誤差訊號利用連鎖法則 (Chain Rule) 往



前傳遞，直到傳回輸入層為止。

假設 y_j 與 z_j 分別是輸出節點 j 的輸出訊號與輸入訊號， y_i 是潛藏節點 i 的輸出訊號。其中 $y_j = f(z_j)$ 經過一個激活函數 (Activation Function) $f(\cdot)$ 轉換。假設已知誤差函數為式 2.25，我們就可以計算它對 y_j 的一次偏微 $\frac{\partial E}{\partial y_j} = \frac{d_j}{y_j}$ ，根據連鎖法則，即可推得：

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j} \frac{y_j}{z_j} = f'(z_j) \frac{\partial E}{\partial y_j} \quad (2.26)$$

且：

$$\frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial z_j} \frac{\partial z_j}{\partial y_i} = \sum_j w_{ij} \frac{\partial E}{\partial z_j} \quad (2.27)$$

故我們可以推得誤差函數 E 對權重 w_{ij} 的一階導數：

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial z_j} \frac{\partial z_j}{\partial w_{ij}} = y_i \frac{\partial E}{\partial z_j} \quad (2.28)$$

而此值就被拿來當作梯度下降法 (Gradient Decent) 中更新權重 w_{ij} 的依據。

沿時間反向傳播演算法 [28]是由反向傳播演算法而來，其唯一不同之處是在做模型訓練之前，必須根據所設定的展開時間層數，將潛藏層往前展開 N 層，圖 2.8 中顯示的是將潛藏層往前展開三層的結果。展開之後其實相當於訓練兩個前饋式類神經網路，一為圖中的上半部，包含輸出層、原始潛藏層及輸入層，另一為圖中的下半部，包含所有沿時間展開的潛藏層及其對應的輸入層。另一方面，由於模型有隱含序列的特性，因此在訓練的過程中必須保持序列的順序一筆一筆餵入作訓練。訓練後各時間點的潛藏層與潛藏層間的權重矩陣將被取平均，以讓各時間點的權重相同。

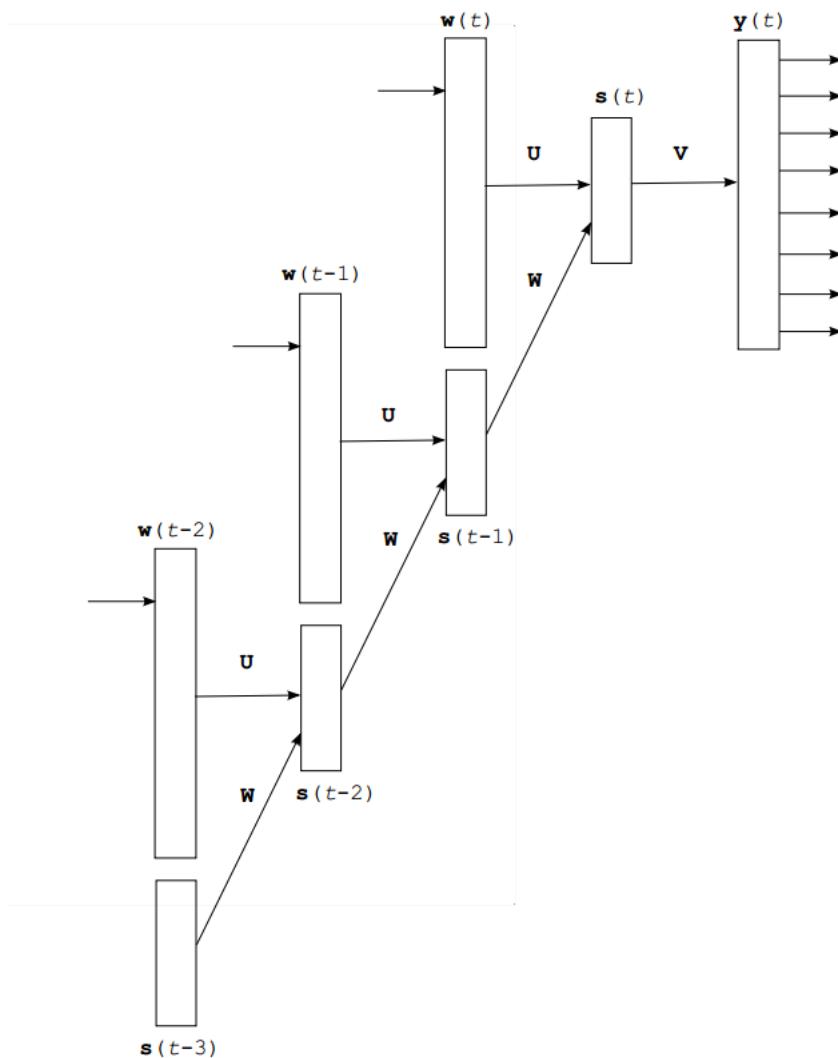


圖 2.8: 將潛藏層向前展開三層的示意圖



2.5 本章總結

本章介紹了資訊檢索的背景，包含了基礎的資訊檢索架構、口述語彙偵測與語意檢索的差別，以及相關回饋的基本概念，並介紹了語音檢索系統的兩個主要的部分：辨識系統與檢索系統，最後介紹了動態時間校準做為口述語彙偵測的方法。

第三章 以自動習得之聲學組型加強監督式

語意檢索



3.1 簡介

傳統的語意檢索是為了檢索純文字的文件，一個很成功的解決方式是查詢詞擴展 (Query Expansion) [25,29,30]。當系統試圖對語音文件實作查詢詞擴展時，通常的實作方法是先將語音文件辨識成文字，再對這些文字文件套用查詢詞擴展，然而這種做法的缺點是辨識過程中不可避免地會出現辨識錯誤、辭典外詞彙而導致辨識結果不準確，進而影響到檢索結果。而且許多語音文件本身帶有的語音資訊如音高、聲調等等在被辨識成文字後就消失了且再也找不回來，過去已有許多文獻使用聲學中的資訊以幫助語音文件檢索 [31–35]，而這也是本章想探討的方向。本章中提出的解決方法使用一套自動尋找的聲學組型 (Automatically Discovered Acoustic Patterns) [36–44] 並將語音文件辨識成這些聲學組型的序列，而這些聲學組型是直接根據輸入信號的特性去尋找的，所以可以彌補文字查詢詞擴展在語音辨識階段損失的資訊，在過去已有文獻將聲學組型應用於語音文件分類 (Spoken Document Classification) [45–48]、口述語彙偵測 [36,49–51]、音樂檢索 [52] 和影片檢索 [53]。

查詢詞擴展是一套基於虛擬回饋的架構，當使用者輸入文字形式的查詢詞後，系統會產生第一次檢索結果，並假設前 N 篇都是與查詢詞虛擬相關 (Pseudo Relevant) 的，稱為虛擬相關文件 (Pseudo Relevant Document) 再根據最大概似估計 (Maximum Likelihood Estimation) 計算出最有可能與查詢詞語意相關的詞彙加入到查詢詞中。本章中將在這個架構上加上聲學組型的資訊用以解決以上所說的語音



數位內容檢索會因為辨識錯誤、辭典外詞彙而導致檢索性能下降等問題。我們在此提出的架構將把語音文件同時辨識成文字的詞圖與聲學組型的唯一最佳序列，並同時於其上進行查詢詞擴展。根據文字上的第一次檢索結果可以得知哪些語音文件與查詢詞是虛擬相關，並假設那些常出現在虛擬相關文件中的聲學組型為與查詢詞語意相關的，再用這些聲學組型去檢索所有的聲學組型序列得到另一份檢索結果，最後由於聲學組型成效不如文字穩定，因此將此檢索結果與文字檢索結果疊加後可得穩定進步的成效。如此一來，即使文件中包含了辭典外詞彙或是辨識錯誤，也有可能可以從聲學組型的檢索中找出來。

3.2 傳統監督式語音文件語意檢索

語意檢索的目的是要找出所有與查詢詞語意上相關的文件，其中一種解決方式是查詢詞擴展 (Query Expansion)，查詢詞擴展可以找出在目標文件集合中與查詢詞語意上相關的詞彙，並將其加入到原查詢詞的集合中，成為新的一組查詢詞，稱為擴展後的查詢詞 (Expanded Query)，系統會自動使用這組新的查詢詞進行一次新的檢索，這次檢索回傳的文件就會包含擴展後的查詢詞，因此這些文件就會是與原查詢詞語意上相關的文件。

查詢詞擴展使用虛擬回饋的框架，檢索系統在使用者輸入文字形式的查詢詞後，會先檢索出第一次檢索結果，再假設排序結果的前 N 篇都是與查詢詞虛擬相關的文件，再利用查詢詞擴展找出這些虛擬相關文件中較常共同出現的詞彙，卻又不是常常出現在所有文件中的詞彙，並將這些詞彙加入原查詢詞中，成為新的一組查詢詞，稱為擴展後查詢詞 (Expanded Query)，系統再利用擴展後查詢詞進行下一次檢索後將結果回傳給使用者，即為傳統監督式語音文件語意檢索的做法。



方法細節簡介如下：

前處理

本章使用的檢索方式主要是基於語言模型的檢索 (Language Modelling Retrieval)，因此首先要把辨識所得的詞圖轉換為語言模型。對語音文件 x 中的每個詞 t ，它在詞圖中的期望出現次數 (Expected Count) 可以如此計算：

$$E[t|x] = \sum_{\mu \in L(x)} N(t, \mu) P(\mu|x) \quad (3.1)$$

$L(x)$ 是 x 的詞圖中所有的路徑， μ 是 $L(x)$ 中的一條路徑， $N(t, \mu)$ 是 t 在 μ 中出現的次數， $P(\mu|x)$ 是路徑 μ 的事後機率(Posterior Probability)。

有了 $E[t|x]$ 之後，就能把詞圖表示成單連詞語言模型 θ_x ：

$$P(t|\theta_x) = \frac{E[t|x]}{\sum_t E[t|x]} \quad (3.2)$$

因為 θ_x 裡沒有包含每個詞，為了讓 θ_x 中每個詞都有一點機率，會再把 θ_x 與一個背景語言模型 (Background Language Model) θ_b 做線性疊加，這個過程稱為平滑化，平滑化後的語音文件模型為 $\bar{\theta}_x$ 。 θ_b 可以如此估計：

$$P(t|\theta_b) = \frac{\sum_{x \in C} E[t|x]}{\sum_t \sum_{x \in C} E[t|x]} \quad (3.3)$$

其中 C 是所有語音文件 x 的集合。

3.2.1 第一次檢索結果

在此，我們將使用者輸入的文字查詢詞表示成語言模型 θ_q ：

$$P(t|\theta_q) = \frac{N(t, q)}{|q|} \quad (3.4)$$

$N(t, q)$ 是詞彙 t 在查詢詞 q 中出現的次數， $|q|$ 是查詢詞 q 中所有的詞彙總數。第一次檢索結果是由排序以下分數 $S_0(q, x)$ 得到：



$$S_0(q, x) = -[(1 - w_1)KL(\theta_q^w | \bar{\theta}_x^w) + w_1KL(\theta_q^s | \bar{\theta}_x^s)] \quad (3.5)$$

上式同時使用了詞單位的語言模型 $\theta_q^w, \bar{\theta}_x^w$ 和次詞 (Subword) 單位的語言模型 $\theta_q^s, \bar{\theta}_x^s$ 。 w_1 是線性疊加兩者時的權重，通常我們使用 $w_1 = 0.95$ 。並取其中的前 N 篇為虛擬相關文件，進行下一步的查詢詞擴展。

3.2.2 查詢詞擴展與第二次檢索

這裡使用的查詢詞擴展訓練方法同 2.2.4，經過查詢詞擴展後，即可得到擴展後查詢詞模型 θ_{qe} ，在擴展後查詢詞模型中，會包含許多常一起出現在虛擬相關文件中，但不會常出現在所有文件中的詞，而這些詞通常會是與原查詢詞成語意相關的，因此再用這套擴展後查詢詞模型進行第二次檢索 (Second-pass Retrieval)，即可檢索到更多與原查詢詞語意相關的文件，第二次檢索的相關分數 $S(q, x)$ 為：

$$S_0(q, x) = -[(1 - w_1)KL(\theta_{qe}^w | \bar{\theta}_x^w) + w_1KL(\theta_{qe}^s | \bar{\theta}_x^s)] \quad (3.6)$$

這裡也是同時使用了詞單位的語言模型 $\bar{\theta}_x^w$ 和次詞 (Subword) 單位的語言模型 $\bar{\theta}_x^s$ 。 w_1 是線性疊加兩者時的權重，通常我們使用 $w_1 = 0.95$ 。

將所有語音文件 x 按照其與查詢詞 q 的相關分數 $S(q, x)$ 排序後即可得到第二次檢索結果 (Second-pass Retrieval Result)，再將其呈現給使用者。



3.3 以聲學組型改善監督式語意檢索

上一節提到的傳統監督式語意檢索確實可以很好地找出與查詢詞語意上相關的文件，但由於語音文件被辨識成文字文件時很有可能會出現辨識錯誤，或某些詞彙可能是辭典外詞彙，進而使得查詢詞擴展遇到困難。事實上當系統將語音文件辨識成文字文件時，就已經損失了很多資訊，並且這些資訊是再也找不回來的了，因此這節的目的是要利用這些聲學的資訊彌補查詢詞擴展。

本論文提出了一個系統解決以上的問題，如圖 3.1。在圖 3.1 的左下角系統會自動從所有文件的集合中學出一套類詞單位和類次詞單位的聲學組型的聲學模型和語言模型，再將語音文件用這套聲學組型辨識出來。因此圖 3.1 中的下半部將每個語音文件都表示成兩種格式：由語音辨識系統辨識而成的詞圖（包含詞與次詞兩種單位的詞圖）、由聲學組型組成的聲學組型串列。

當使用者輸入一個文字的查詢詞時，系統會將查詢詞與詞圖作比對(圖中的右邊偏下)並產生第一次檢索結果，系統並不會將這套第一次檢索結果回傳給使用者，而是假設其中的前 N 篇是與查詢詞虛擬相關的，系統再利用這些虛擬相關文件作查詢詞擴展並用新的查詢詞模型檢索出一套新的檢索結果。接下來用聲學組型再做一次查詢詞擴展，即可得到聲學組型的查詢詞模型，聲學組型的查詢詞擴展假設其虛擬相關文件與文字的查詢詞擴展時相同，於是可得另一份新的查詢詞模型並用其檢索出另一套新的檢索結果，最後再把兩份檢索結果做線性疊加再排序後回傳給使用者。如此一來，即使在文字的查詢詞擴展時有些詞彙被錯誤地辨識或是有辭典外詞彙，也可以透過聲學組型的查詢詞擴展將對應到這些詞的聲學組型加到聲學組型的查詢詞模型中，於是仍然可以找到包含有這些詞彙的文件。

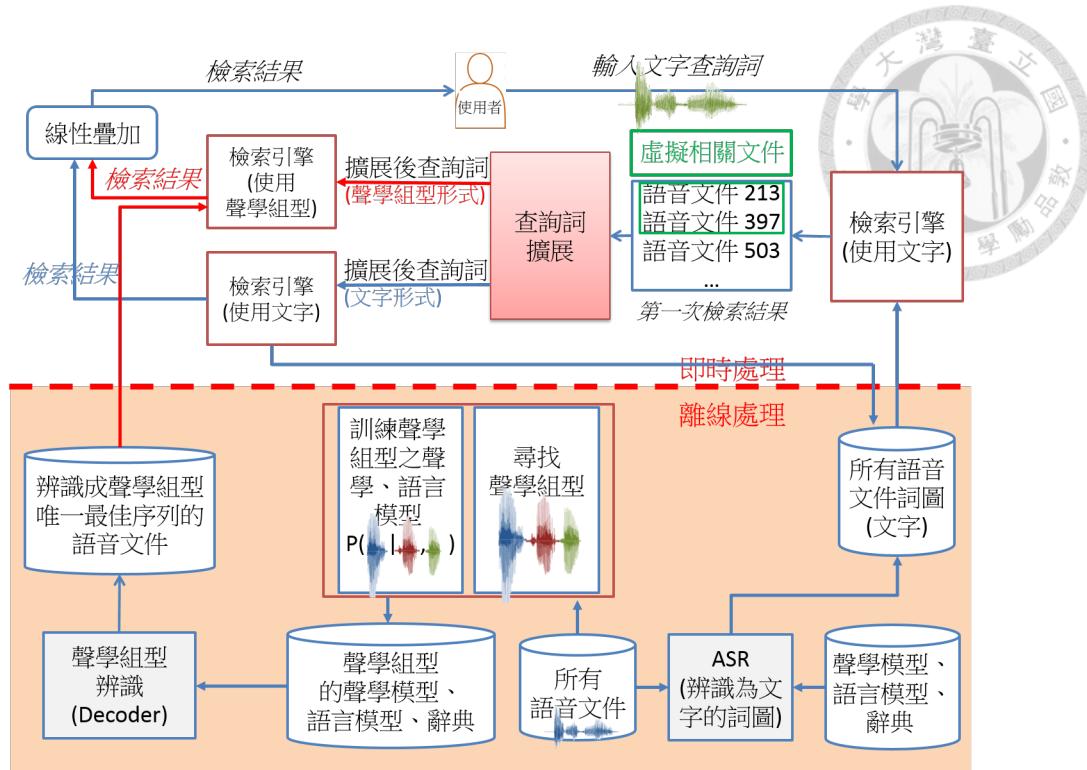


圖 3.1: 系統架構示意圖

3.3.1 前處理

文字語言模型

此處文字形式的前處理方式同 3.2。因此要先把語音文件 x 辨識為詞單位的詞圖與次詞單位的詞圖後，再將詞圖轉換為詞單位的語言模型 $\bar{\theta}_x^w$ 與次詞單位的語言模型 $\bar{\theta}_x^s$ 。以下舉詞單位的詞圖為例，對語音文件 x 中的每個詞 t ，它在詞圖中的期望出現次數 (Expected Count) 可以如此計算：

$$E[t|x] = \sum_{\mu \in L(x)} N(t, \mu) P(\mu|x) \quad (3.7)$$

$L(x)$ 是 x 的詞圖中所有的路徑， μ 是 $L(x)$ 中的一條路徑， $N(t, \mu)$ 是 t 在 μ 中出現的次數， $P(\mu|x)$ 是路徑 μ 的事後機率 (Posterior Probability)。

有了 $E[t|x]$ 之後，就能把詞圖表示成單連詞語言模型 θ_x^w ：



$$P(t|\theta_x^w) = \frac{E[t|x]}{\sum_t E[t|x]} \quad (3.8)$$

因為 θ_x^w 裡沒有包含每個詞，為了讓 θ_x^w 中每個詞都有一點機率，會再把 θ_x^w 與一個背景語言模型 (Background Language Model) θ_b^w 做線性疊加，這個過程稱為平滑化，平滑化後的語音文件模型為 $\bar{\theta}_x^w$ 。 θ_b^w 可以如此估計：

$$P(t|\theta_b^w) = \frac{\sum_{x \in C} E[t|x]}{\sum_t \sum_{x \in C} E[t|x]} \quad (3.9)$$

其中 C 是所有語音文件 x 的集合。

以上的過程將同時對詞單位與次詞單位的詞圖計算，即可得詞單位的語言模型 $\bar{\theta}_x^w$ 與次詞單位的語言模型 $\bar{\theta}_x^s$ 。

聲學組型語言模型

聲學組型指的是在某個特定語料中時常重複出現的聲音，比如像中文的音節、字等單位，這些聲學組型可以被用在口述文件分類、口述語彙偵測、語音文件檢索等。這裡使用的是本實驗室過去提出的雙層式聲學組型 (Two-Level Acoustic Pattern) [54] (細節列於 2.3)，其中包含了類詞的聲學組型（類詞聲學組型是由數個類次詞聲學組型所組成）和類次詞的聲學組型、類詞聲學組型如何由類次詞聲學組型組合而成的辭典、類詞聲學組型的 N 連語言模型 (N -gram Language Model)。每個類次詞聲學組型就是一個隱藏式馬可夫模型，所有的隱藏式馬可夫模型的參數、類次詞聲學組型的數目、類詞聲學組型的數目和類詞聲學組型的 N 連語言模型都是在非監督式的狀況下自動地從語料庫學習出來的。學習出聲學組型以後，這些聲學模型、語言模型和辭典可以用來建立一個辨識系統，並將這些語音文件辨識成聲學組型的序列，如此即可將語音文件 x 表示成聲學組型的語言模型 ϕ_x ：



$$P(v|\phi_x) = \frac{C(v,x)}{\sum_v C(v,x)} \quad (3.10)$$

$C(v,x)$ 為聲學組型 v 在語音文件 x 中出現的次數。 ϕ_x 會再與聲學組型背景語言模型 ϕ_b 做平滑化(即線性疊加)，平滑化的語音文件模型為 $\bar{\phi}_x$ 。 ϕ_b 表示如下：

$$P(v|\phi_b) = \frac{\sum_{x \in C} C(v,x)}{\sum_v \sum_{x \in C} C(v,x)} \quad (3.11)$$

其中 C 是所有語音文件 x 的聲學組型序列， v 是任一聲學組型， $C(v,x)$ 為聲學組型 v 在語音文件 x 的聲學組型序列中出現的次數。

3.3.2 第一次檢索結果

由於查詢詞 q 是文字形式的，因此第一次檢索結果只能用查詢詞模型 θ_q 與文字的文件模型 $\bar{\theta}_x^w$ 和 $\bar{\theta}_x^s$ 做比對，而不能使用聲學組型的文件模型 $\bar{\phi}_x$ 。所以計算第一次檢索結果的相關分數 $S_0(q,x)$ 同式 3.5 為：

$$S_0(q,x) = -[(1-w_1)KL(\theta_q^w|\bar{\theta}_x^w) + w_1KL(\theta_q^s|\bar{\theta}_x^s)] \quad (3.12)$$

並取其中的前 N 篇為虛擬相關文件，進行下一步的查詢詞擴展。

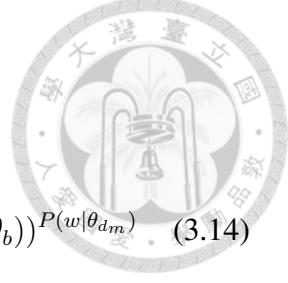
3.3.3 查詢詞擴展

以下簡介文字型式的查詢詞擴展與聲學組型形式的查詢詞擴展如下：

文字形式的查詢詞擴展

此處的查詢詞擴展與 2.2.4 中類似，計算方法與式 2.14 相似，只要最大化下式：

$$F(\theta_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M}) = F_1(\theta_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M})F_2(\theta_{qe})^\lambda \quad (3.13)$$



其中 $F_1(\theta_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M})$ 定義如下：

$$F_1(\theta_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M}) = \prod_{m=1}^M \Pi_w (\alpha_{d_m} P(w|\theta_{qe}) + (1 - \alpha_{d_m}) P(w|\theta_b))^{P(w|\theta_{d_m})}. \quad (3.14)$$

上式中，產生詞彙 w 的機率為 $\alpha_{d_m} P(w|\theta_{qe}) + (1 - \alpha_{d_m}) P(w|\theta_b)$ 。因此上式可視為由新的查詢詞模型產生這些虛擬相關文件的可能性 (Likelihood)。然而，如果只最大化式 3.14， θ_{qe} 中主要會包含這些虛擬相關文件的主題，而不一定是查詢詞相關的詞彙，為了要解決這個問題，比較好的方法是用原本的查詢詞模型 θ_q 正規化 (Regularize) θ_{qe} ，因此定義 $F_2(\theta_{qe})$ 如下：

$$F_2(\theta_{qe}) = \Pi_w P(w|\theta_{qe})^{P(w|\theta_q)} \quad (3.15)$$

當 θ_{qe} 和 θ_q 越近時， $F_2(\theta_{qe})$ 就會越大，因此如果我們同時最大化兩者的乘積即式 3.13，就能夠同時將虛擬相關文件中的詞彙加進擴展後查詢詞，也能同時正規化這個過程，使得擴展後查詢詞不致於與原查詢詞相差太遠。正規化的強度由 3.13 中的 λ 決定， λ 越大，擴展後查詢詞 θ_{qe} 與 θ_q 就會被強制靠得越近， λ 越小則反之。

θ_{qe} 即為擴展後的文字查詢詞模型，最大化 3.13 的方法採用 EM 演算法：

E step: 對於 d_1, d_2, \dots, d_M 中的每篇語音文件 d_m 中的每一個詞彙 w :

$$P(R|w, d_m) = \frac{\alpha_{d_m} P(w|\theta_{qe})}{\alpha_{d_m} P(w|\theta_{qe}) + (1 - \alpha_{d_m}) P(w|\theta_b)} \quad (3.16)$$

其中 $P(R|w, d_m)$ 為當給定文件 d_m 中的詞彙 w 時，此文件 d_m 中的詞彙 w 與查詢詞是語意上相關的機率。

M step : 對於 d_1, d_2, \dots, d_M 中的每篇語音文件 d_m ，其中每篇語音文件 d_m 的語言模型為 θ_d :



$$\alpha_{d_m} = \sum_w P(R|w, d_m)P(w|\theta_d)$$

(3.17)

對每個詞彙 w :

$$P(w|\theta_{qe}) = \frac{\lambda P(w|\theta_q) + \sum_{m=1}^M P(w|\theta_d)P(R|w, d_m)}{\lambda + \sum_w \sum_{m=1}^M P(w|\theta_d)P(R|w, d_m)}$$

重覆地執行 E Step 和 M Step 即可得到新的查詢詞模型 θ_{qe} ，以上的步驟必須同時對詞版本的詞圖與次詞版本的詞圖各進行一次，故會得到兩種版本的擴展後查詢詞，分別為詞版本的擴展後查詢詞 θ_{qe}^w 與次詞版本的擴展後查詢詞 θ_{qe}^s 。

聲學組型形式的查詢詞擴展

此處的查詢詞擴展也與 2.2.4 中相似，但由於我們沒有聲學組型形式的原查詢詞，故此時不需要如上一小節乘上 $F_2(\phi_{qe})$ ，所以是最大化下式：

$$F(\phi_{qe}, \alpha'_{d_1}, \dots, \alpha'_{d_M}) = F_1(\phi_{qe}, \alpha'_{d_1}, \dots, \alpha'_{d_M})$$

由於聲學組型形式的 α 與文字形式的 α 不同，故以 α' 表示聲學形式的 α 。其中 $F_1(\phi_{qe}, \alpha'_{d_1}, \dots, \alpha'_{d_M})$ 定義如下：

$$F_1(\phi_{qe}, \alpha'_{d_1}, \dots, \alpha'_{d_M}) = \prod_{m=1}^M \prod_v (\alpha'_{d_m} P(v|\phi_{qe}) + (1 - \alpha'_{d_m}) P(v|\phi_b))^{P(v|\phi_{d_m})} \quad (3.20)$$

上式中，產生聲學組型 v 的機率為 $\alpha'_{d_m} P(v|\phi_{qe}) + (1 - \alpha'_{d_m}) P(v|\phi_b)$ 。因此上式可視為由新的查詢詞模型產生這些虛擬相關文件的可能性 (Likelihood)。

ϕ_{qe} 即為擴展後的文字查詢詞模型，最大化 3.19 的方法採用 EM 演算法：

E step: 對於 d_1, d_2, \dots, d_M 中的每篇語音文件 d_m 中的每一個聲學組型 v :



$$P(R|v, d_m) = \frac{\alpha'_{d_m} P(v|\phi_{qe})}{\alpha'_{d_m} P(v|\phi_{qe}) + (1 - \alpha'_{d_m}) P(v|\phi_b)} \quad (3.21)$$

其中 $P(R|v, d_m)$ 為當給定文件 d_m 中的聲學組型 v 時，此文件 d_m 中的聲學組型 v 與查詢詞是語意上相關的機率。

M step：對於 d_1, d_2, \dots, d_M 中的每篇語音文件 d_m ，其中每篇語音文件 d_m 的聲學組型語言模型為 ϕ_d ：

$$\alpha'_{d_m} = \sum_v P(R|v, d_m) P(v|\phi_d) \quad (3.22)$$

對每個聲學組型 v ：

$$P(v|\phi_{qe}) = \frac{\sum_{m=1}^M P(v|\phi_d) P(R|v, d_m)}{\sum_v \sum_{m=1}^M P(v|\phi_d) P(R|v, d_m)} \quad (3.23)$$

重覆地執行 E Step 和 M Step 即可得到聲學組型形式的擴展後查詢詞模型 ϕ_{qe}

線性疊加與第二次檢索

有了文字的擴展後查詢詞模型 θ_{qe} 與聲學組型的擴展後查詢詞模型 ϕ_{qe} 後，由於聲學組型形式的擴展後查詢詞模型不如文字形式的擴展後查詢詞模型穩定，故需要線性疊加文字形式的第二次檢索與聲學組型形式的第二次檢索，因此整體的相關分數 $S(q, x)$ 計算如下：

$$S(q, x) = -(1 - w_2)[(1 - w_1)KL(\theta_{qe}^w | \bar{\theta}_x^w) + w_1KL(\theta_{qe}^s | \bar{\theta}_x^s)] + w_2KL(\phi_{qe} | \bar{\phi}_x) \quad (3.24)$$

其中 w_2 為線性疊加時聲學組型形式的第二次檢索結果所佔的比重， w_1 為使用詞版本的詞圖與次詞版本的詞圖時的比例，通常 w_1 使用 0.95。



3.4 實驗設定

本章的實驗使用的語料為2001年間從電台廣播中錄下的4小時新聞，並手動切成5034篇語音文件，每篇語音文件大約包含了13句的語句。用來辨識的語言模型是用1999年間收集的新聞文章(包含4000萬個詞彙)訓練而成，辭典中包含了62000個詞彙，聲學模型是用2000年間收集的8小時廣播新聞訓練而成的音節內(Intra-syllable)右方資訊相依(Right-context-dependent)聲韻母模型(Initial-Final models)。辨識後的唯一最佳序列的字元正確率為(Character Accuracy)為75.27%。總共測試了29組查詢詞，每組查詢詞都有人工標注對應的語意相關文件，而這些語意相關文件中「並不一定要」包含查詢詞。本章中使用平均準確率做為評量標準。

3.5 實驗結果及分析

圖 3.2 中顯示的是式 3.24 疊加的結果，其中 λ 固定為 800， N 為 5, 10, 15, 20, 25，紅線則是第一次檢索結果，是不使用查詢詞擴展時的結果。縱軸是平均準確率(MAP)，橫軸則是式 3.24 的 w_2 ，為聲學組型與文字的擴展後查詢詞疊加時的權重， w_1 在實驗中都被固定為 0.95。首先來比較第一次檢索結果與文字的查詢詞擴展後的結果 ($w_2 = 0$) 時的狀況，可以觀察到文字的查詢詞擴展比第一次檢索結果好上一些，即使它可能有受到辨識錯誤或辭典外詞彙的影響。接下來看有使用聲學組型時的狀況：當 w_2 大於零時，可以注意到整體的平均準確率相對 $w_2 = 0$ 時都是進步的(只要 $w_2 < 0.5$)，這代表了使用聲學組型確實能夠幫助查詢詞擴展找到更多聲學方面的資訊，進而使平均準確率更為進步。另一方面也可以看到，當 N 增加，平均準確率會先成長後再下降，大約在 $N = 10$ 的時候為最佳的狀況，這

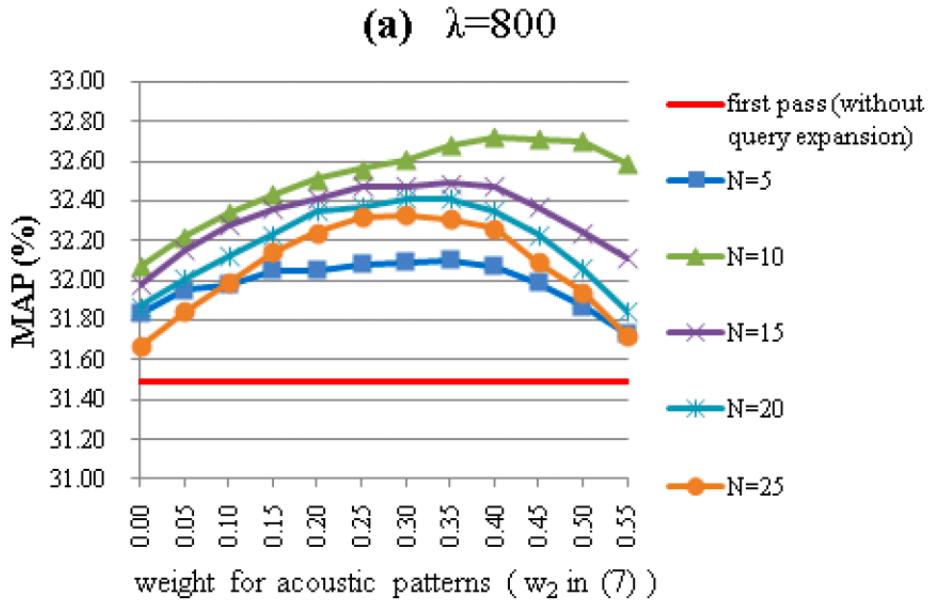


圖 3.2: 將文字形式和聲學組型形式的查詢詞擴展疊加後的平均準確率 ($\lambda = 800$)

個原因是由於當 N 變多時，被假設為虛擬相關的文件數也變多，因此更多的資訊被用來進行虛擬回饋，但另一方面如果假設了太多的虛擬文件時，會使用太多不相關資訊進行虛擬回饋，使得系統的平均準確率下降，最好的平均準確率是在 $N = 10$ 而且 $w_2 = 0.40$ 的狀況，而這代表說文字的查詢詞擴展是比聲學組型的查詢詞擴展來得可靠的，因此文字查詢詞擴展的權重必須要設得高一些，由於這是最好的結果，所以在之後的實驗中會將 N 設為 10。

圖 3.3 和圖 3.2 類似，只是圖 3.3 中將 N 固定為 10，並且測試不同的 λ ，在圖中可以發現當 λ 很低的時候 (如 100)，系統的平均準確率甚至是比第一次檢索結果還要差的，然而當 λ 夠大的話 ($\lambda \geq 400$)，系統的平均準確率都是比第一次檢索結果和 $w_2 = 0$ 的狀況還要好的。這證明了式 3.15 的重要性，如果 λ 太小的話就會導致系統在訓練時對於虛擬文件過適而使得成果更差，而 λ 夠大時 ($\lambda > 400$) 就幾乎都能看到不錯的進步。

圖 3.2 和圖 3.3 中表示了將文字形式和聲學組型形式的查詢詞擴展疊加後的

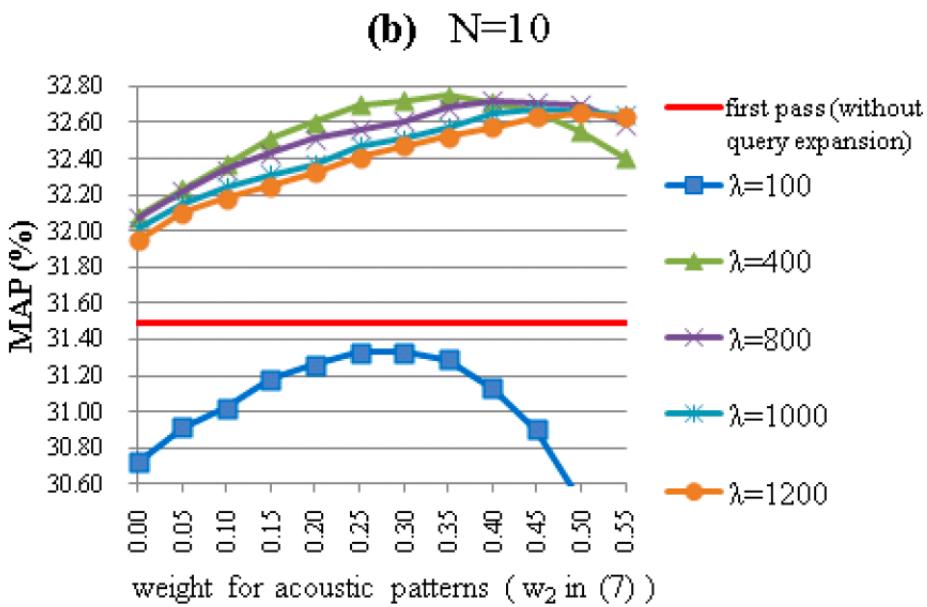


圖 3.3: 將文字形式和聲學組型形式的查詢詞擴展疊加後的平均準確率 ($N = 10$)

平均準確率，縱軸為平均準確率(MAP)，橫軸為線性疊加中聲學組型的權重。圖中的紅色橫線是第一次檢索結果的平均準確率（不含查詢詞擴展）。

3.6 本章總結

在本章中使用了聲學組型加強監督式語音文件語意搜索的成效，如此一來可以解決諸如辨識錯誤、語音文件中包含詞典外詞彙的問題，可以有效提升語意檢索的平均準確率。

第四章 以自動習得之聲學組型實現非監督

式語意檢索



4.1 簡介

在上一章中，我們嘗試了使用聲學組型加強語意檢索的成果，使得系統能夠在接收到文字形式的查詢詞後，進行語意檢索回傳給使用者語意相關的語音文件。這樣的框架需要使用者在每次查詢詞都要對系統輸入文字形式的查詢詞，但是對使用者來說，使用口語形式的輸入才是最自然的輸入方法，尤其在行動裝置與穿戴式裝置日益盛行的情況下，如果能讓使用者不需要使用行動裝置的鍵盤就進行檢索，將能大幅提高使用者體驗與使用者的滿意度。另一方面，如果使用者輸入的是口語形式的查詢詞，那系統就能直接使用語音去直接檢索語音數位內容，如此就能直接在語音訊號上進行比對，這樣一來就能略過許多使用自動語音辨識系統(ASR)的困難：如辨識錯誤，辭典外詞彙等辨識困難，或是很難訓練一套好的聲學/語言模型與辭典，因為自動語音辨識系統的訓練過程需要很多人標注過的檔案，而這會導致這個訓練過程十分地昂貴，因此，在本章提出的架構中希望不用使用自動語音辨識系統。

本章的目標是達成「非監督式的語音內容語意檢索」，在這個狀況下，使用者輸入口語形式的查詢詞，系統直接進行語意檢索並回傳給使用者檢索結果，但如果沒有使用自動語音辨識系統，通常是很難實作語意檢索的，因為系統需要知道這些文字之間的關係才能進行語意檢索，這也是為什麼過去幾乎所有的非監督式檢索方法都只實作在口述語彙偵測(系統只回傳包含查詢詞的文件)，而沒有實作在語意檢索之中。但本章試圖使用之前提到的自動習得之聲學組型解決這個問



題：系統先將所有語音文件轉寫成聲學組型序列，當使用者口述一段查詢詞給系統後，系統會使用動態時間校準 (Dynamic Time Warping, DTW) 直接在語音訊號上比對相似性，並按照語音文件中是否有出現此段查詢詞排序後產生第一次檢索結果，並假設第一次檢索結果的前 N 篇為虛擬相關文件，再將虛擬相關文件中很常一起出現的聲學組型當作和查詢詞語意相關的詞彙並將其當作新的查詢詞模型。如此一來，雖然系統沒辦法將語音文件轉寫成文字，但是系統仍然可以藉由其在聲學組型上的關聯性進行查詢詞擴展，並在查詢詞擴展的過程中找到與查詢詞語意相關的聲學組型，進而達成語意檢索的目的。

4.2 基於聲學組型之語意檢索

4.2.1 系統架構

系統架構如圖 4.1，圖中的下半部為離線處理 (Offline Processing) 的部分，包括聲學組型的聲學模型、語言模型、辭典都是離線的時候自動從語料庫中學習出來的 [54]，有了聲學、語言模型和辭典後，即可用其建造出一個辨識系統 (圖 4.1 中左下角) 將語音文件辨識成聲學組型形式的唯一最佳序列。圖 4.1 中的上半部為在線處理 (Online Processing) 的部分，當使用者「口述」了一段查詢詞後，系統會使用片段式動態時間校準 (在圖 4.1 中的檢索引擎1) 產生第一次檢索結果，並假設其中最相關的前 N 篇為虛擬相關，虛擬相關文件中時常出現的聲學組型即可視作是與查詢詞語意相關的聲學組型，這些即為擴展後的查詢詞模型，接下來圖中的檢索引擎2 會利用擴展後查詢詞模型尋找之前得到的聲學組型形式的唯一最佳序列，如果這些唯一最佳序列中包含了這些與查詢詞語意相關的聲學組型，即是與原查詢詞語意相關的文件。

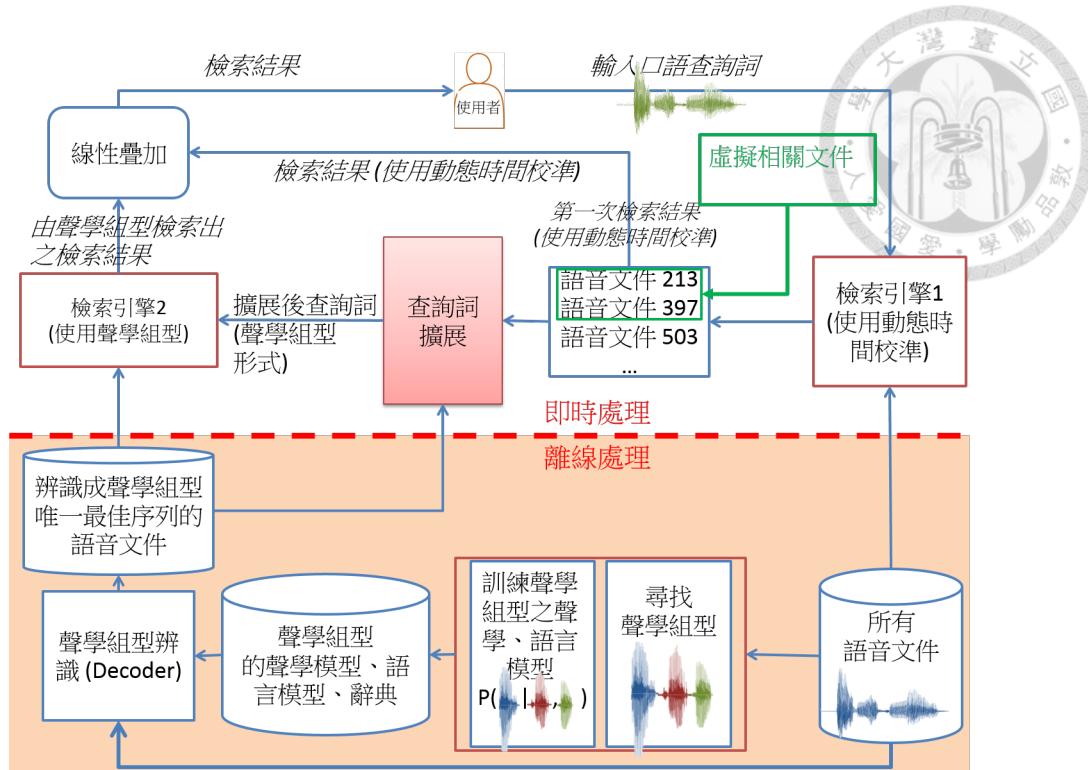


圖 4.1: 系統架構示意圖

4.2.2 前處理

聲學組型語言模型

聲學組型指的是在某個特定語料中時常重複出現的聲音，比如像中文的音節、字等單位，這些聲學組型可以被用在口述文件分類、口述語彙偵測、語音文件檢索等。這裡使用的是本實驗室過去提出的雙層式聲學組型 (Two-Level Acoustic Pattern) [54] (細節列於 2.3)，其中包含了類詞的聲學組型（類詞聲學組型是由數個類次詞聲學組型所組成）和類次詞的聲學組型、類詞聲學組型如何由類次詞聲學組型組合而成的辭典、類詞聲學組型的N連語言模型 (N-gram Language Model)。每個類次詞聲學組型就是一個隱藏式馬可夫模型，所有的隱藏式馬可夫模型的參數、類次詞聲學組型的數目、類詞聲學組型的數目和類詞聲學組型的N連語言模型都是在非監督式的狀況下自動地從語料庫學習出來的。學習出聲

學組型以後，這些聲學模型、語言模型和辭典可以用來建立一個辨識系統，並將這些語音文件辨識成聲學組型的序列，如此即可將語音文件 x 表示成聲學組型的語言模型 ϕ_x ：



$$P(v|\phi_x) = \frac{C(v, x)}{\sum_v C(v, x)} \quad (4.1)$$

$C(v, x)$ 為聲學組型 v 在語音文件 x 中出現的次數。 ϕ_x 會再與聲學組型背景語言模型 ϕ_b 做平滑化(即線性疊加)，平滑化的語音文件模型為 $\bar{\phi}_x$ 。 ϕ_b 表示如下：

$$P(v|\phi_b) = \frac{\sum_{x \in C} C(v, x)}{\sum_v \sum_{x \in C} C(v, x)} \quad (4.2)$$

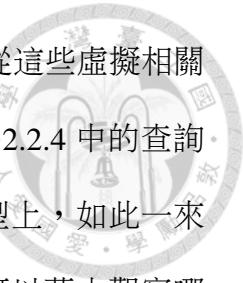
其中 C 是所有語音文件 x 的聲學組型序列， v 是任一聲學組型， $C(v, x)$ 為聲學組型 v 在語音文件 x 的聲學組型序列中出現的次數。

4.2.3 第一次檢索結果

由於使用者在此是以口述的方式輸入查詢詞，因此查詢詞是以訊號形式存在，故無法直接使用 3.2.1 提到的文字形式查詢詞的檢索，故在此處我們使用第一次檢索為 2.1.6 提到的片段式動態時間校準。當使用者輸入查詢詞後，系統即使用片段式動態時間校準在所有待檢索文件中找到所有的假設區域並按照相關分數排序後得到 h_1, h_2, \dots, h_n ，分別屬於語音文件 d_1, d_2, \dots, d_m (注意 d_i 可以等於 d_j ($i \neq j$)，因為一個語音文件中可以有多個假設區域)。將 d_1, d_2, \dots, d_m 去除重覆的語音文件即得第一次檢索結果。

4.2.4 語意檢索

系統至此已經得到了由片段式動態時間校準所得的第一次檢索結果，並且可以假



設其中的前 N 篇為與口述查詢詞 q 虛擬相關的文件，故在此試圖從這些虛擬相關的文件中找出與查詢詞 q 相關的聲學組型，所採取的作法是類似 2.2.4 中的查詢詞擴展的方法，只是這次不是使用在文字上，而是使用在聲學組型上，如此一來即使系統不知道那些聲學組型在文字中代表了什麼意思，但是它可以藉由觀察哪些聲學組型常常和查詢詞的聲學組型一起出現，即可將這些聲學組型當作是與查詢詞相關聯的意思。比如說如果查詢詞是口述的「美國總統」，系統不知道「美國總統」的聲音是什麼意思，但是系統會觀察到在虛擬文件中，有一些聲學組型例如「白宮」、「歐巴馬」時常與「美國總統」的聲學組型共同出現，因此即可認為「白宮」、「歐巴馬」的聲學組型是與查詢詞語意相關的，並利用「白宮」、「歐巴馬」的聲學組型來檢索。

利用第一次檢索結果得到查詢詞的聲學組型語言模型

由於查詢詞是口述形式，因此需要將查詢詞轉為聲學組型形式的語言模型，但如果直接用 4.2.2 中提到的聲學組型辨識系統辨識的話，辨識結果會很差，這是由於查詢詞通常很短，而且會遇到像不同語者、不同語速和不同上下文等不匹配的問題，所以此處不會直接使用聲學辨識系統辨識，而是使用第一次檢索結果。由於第一次檢索結果會回傳查詢詞 q 在虛擬相關文件中出現的假設區域，因此即可將這些假設區域對應到的唯一最佳序列中的聲學組型取出，當作口述形式查詢詞的聲學組型。

因此查詢詞模型可以如下估算：

$$P(v|\phi_q) = \frac{\sum_{n=1}^N C'(v, x_n)}{\sum_v \sum_{n=1}^N C'(v, x_n)} \quad (4.3)$$

$C'(v, x_n)$ 為聲學組型 v 在虛擬相關文件 x 的假設區域中出現的次數。如果 v



完全落在假設區域中， $C'(v, x_n)$ 即為 1，如果只是部分落在假設區域中， $C'(v, x_n)$ 即為它落在假設區域中的時間除以它自己的總長。如此一來即可得到查詢詞模型。 ϕ_q 。

查詢詞擴展

此處的查詢詞擴展與 2.2.4 相似，即最大化下式：

$$F_1(\phi_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M}) = F_1(\phi_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M}) F_2(\phi_{qe})^\lambda \quad (4.4)$$

ϕ_{qe} 即為擴展後的聲學組型查詢詞模型。

其中 $F_1(\theta_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M})$ 定義如下：

$$F_1(\phi_{qe}, \alpha_{d_1}, \dots, \alpha_{d_M}) = \prod_{m=1}^M \Pi_v (\alpha_{d_m} P(v|\phi_{qe}) + (1 - \alpha_{d_m}) P(v|\phi_b))^{P(v|\theta_{d_m})} \quad (4.5)$$

上式中，產生詞彙 v 的機率為 $\alpha_{d_m} P(v|\phi_{qe}) + (1 - \alpha_{d_m}) P(v|\phi_b)$ 。因此上式可視為由新的查詢詞模型產生這些虛擬相關文件的可能性 (Likelihood)。然而，如果只最大化式 4.5， ϕ_{qe} 中主要會包含這些虛擬相關文件的主題，而不一定是查詢詞相關的詞彙，為了要解決這個問題，比較好的方法是用 4.2.4 中得到的原查詢詞模型 ϕ_q 正規化 (Regularize) ϕ_{qe} ，因此定義 $F_2(\phi_{qe})$ 如下：

$$F_2(\phi_{qe}) = \Pi_v P(v|\phi_{qe})^{P(v|\phi_q)} \quad (4.6)$$

其中 ϕ_q 即為 4.2.4 中由第一次檢索結果得出的原查詢詞模型。當 ϕ_{qe} 和 ϕ_q 越近時， $F_2(\phi_{qe})$ 就會越大，因此如果我們同時最大化兩者的乘積即式 4.4，就能夠同時將虛擬相關文件中的詞彙加進擴展後查詢詞，也能同時正規化這個過程，使得擴展後查詢詞不致於與原查詢詞相差太遠。正規化的強度由 4.4 中的 λ 決定， λ 越



大，擴展後查詢詞 ϕ_{qe} 與由第一次檢索結果得出的原查詢詞模型 ϕ_q 就會被強制靠得越近， λ 越小則反之。

最大化 4.4的方法是採用 EM 演算法：

E step: 對於 d_1, d_2, \dots, d_M 中的每篇語音文件 d_m 中的每一個聲學組型 v :

$$P(R|v, d_m) = \frac{\alpha_{d_m} P(v|\phi_{qe})}{\alpha_{d_m} P(v|\phi_{qe}) + (1 - \alpha_{d_m}) P(v|\phi_b)} \quad (4.7)$$

其中 $P(R|v, d_m)$ 為當給定文件 d_m 中的聲學組型 v 時，此文件 d_m 中的聲學組型 v 與查詢詞是語意上相關的機率。

M step : 對於 d_1, d_2, \dots, d_M 中的每篇語音文件 d_m ，其中每篇語音文件 d_m 的語言模型為 ϕ_d :

$$\alpha_{d_m} = \sum_v P(R|v, d_m) P(v|\phi_d) \quad (4.8)$$

對每個聲學組型 v :

$$P(v|\phi_{qe}) = \frac{\lambda P(v|\phi_q) + \sum_{m=1}^M P(v|\phi_d) P(R|v, d_m)}{\lambda + \sum_v \sum_{m=1}^M P(v|\phi_d) P(R|v, d_m)} \quad (4.9)$$

重覆地執行 E Step 和 M Step 即可得到新的查詢詞模型 ϕ_{qe} 。

線性疊加

有了擴展後的查詢詞模型 ϕ_{qe} 後，即可計算查詢詞 q 與文件 x 的相關分數 $S(q, x)$ ，但由於聲學組型形式的擴展後查詢詞模型的表現通常不如動態時間校準來得穩定，所以我們需要線性疊加用擴展後查詢詞檢索的相關分數 $R_{QE}(q, x)$ 與使用動態時間校準檢索的相關分數 $R_{DTW}(q, x)$ （兩者都被正規化 (Normalize) 至0到1之間），並排序後回傳給使用者：



$$S(q, x) = -[w_1 R_{DTW}(q, x) + (1 - w_1) R_{QE}(q, x)] \quad (4.10)$$

4.3 實驗設定

本章的實驗使用的語料為2001年間從電台廣播中錄下的4小時新聞，並手動切成5034篇語音文件，每篇語音文件大約包含了13句的語句。用來辨識的語言模型是用1999年間收集的新聞文章(包含4000萬個詞彙)訓練而成，辭典中包含了62000個詞彙，聲學模型是用2000年間收集的8小時廣播新聞訓練而成的音節內(Intra-syllable)右方資訊相依(Right-context-dependent)聲韻母模型(Initial-Final Models)。辨識後的唯一最佳序列的字元正確率為(Character Accuracy)為75.27%。總共測試了30組查詢詞，每組查詢詞都有人工標注對應的語意相關文件，而這些語意相關文件中並「不一定要」包含查詢詞，每組查詢詞也都有人工標注的口述語彙偵測相關文件，這些文件中「一定要」包含查詢詞。本章中使用平均準確率做為評量標準。

4.4 N連聲學組型分析

前面提過了如何尋找聲學組型，此節將對找到的聲學組型進行分析，使讀者能對找到的聲學組型有更實際的了解。在這章中，我們將5034句的新聞語音文件辨識成兩階段的聲學組型，包括類詞與類次詞的聲學組型，系統總共從目標語料庫中找到了208個類次詞的聲學組型，並使用這些類次詞單位聲學組型的單連(Unigram)、雙連(Bigram)和三連(Trigram)組合(總共85534個組合)來表示每一個語音文件模型(ϕ_x)與查詢詞模型(ϕ_q)並用在檢索系統當中。聲學組型在訓練時



長短上是沒有太多限制的，不過當我們實際上去聽時會發現其中大多數的類次詞聲學組型接近中文的音節，所以雙連、三連組合則接近於中文的雙音節詞或是三音節詞。

表 4.1 中列了一些聲學組型與其對應到的聲音和含義，比如說第 (1) 列中的是編號為 (106) 的聲學組型，當我們去聽這個聲學組型時，發現它的聲音大部分是 /dian/（長度不會切得這麼剛好就是/dian/，但聽起來很接近這個聲音），而它實際對應的中文轉寫為「店」、「點」、「電」等，因此同一個聲學組型會對應到許多不同的中文文字。第 (2) 列中的是一個雙連聲學組型(由兩個編號為 (106) 和 (27) 的聲學組型組成)，當我們實際去聽這個雙連聲學組型時，會發現其對應到的聲音大約是「電腦」、「電能」（由於聲學組型訓練時的自由性，其實不會這麼剛好就是這個詞，要聽得很仔細才聽得出來它對應到的聲音）等，由於這裡是將類似聲音的字分群在一起，因此有些實際上讀音未必一樣但很像的詞也會被分群在一起。

4.5 實驗結果及分析

表 4.2 中列出了本章提出的檢索方法的結果。列 (1) 中的結果是只有第一次檢索結果(即圖 4.1 中的檢索引擎1)，列(2) 中的結果是圖 4.1 中的檢索引擎2，為只使用查詢詞擴展後的結果，列 (3) 中顯示了將兩者以不同權重 w_1 線性疊加後的結果。欄 (A) 中是用語意檢索的答案計算出的平均準確率，也就是說欄 (A) 中的答案都是與查詢詞語意上相關而且不一定要包含查詢詞的文件，欄 (B) 是用有出現查詢詞的文件當作答案，因此欄 (B) 的平均準確率都會比欄 (A) 高上許多。雖然欄 (A) 中的平均準確率都低於 10%，但這是由於語意檢索與生俱來的困難性和聲學組型模稜兩可的特性所致。首先在欄 (A) 中可以觀察到本章提出的查詢詞擴展



v (N 連聲學組型): (IDs) 聲學組型舉例	
	店(/dian/),
(1) 單連: (106)	點(/dian/), 電(/dian/)
	電腦(/dian-nau/),
(2) 雙連: (106)-(27)	電能(/dian-neng/)
	手(/shou/),
(3) 單連: (93)	收(/shou/), 熟(/shou/)
	受傷(/shou-shang/),
(4) 雙連: (93)-(145)	首相(/shou-shiang/)

表 4.1: 一些單連和雙連聲學組型的聲音與其對應到的中文詞

能夠有效地幫助到非監督式語意檢索 (列(3) 對比列(1))，多半是由於查詢詞擴展能夠有效地從第一次檢索結果中找出與查詢詞語意相關的資料。很有趣的是本章的查詢詞擴展也能有效地幫助到口述語彙檢索的結果(欄(B))，可能是因為動態時間校準在比對查詢詞與文件時，很有可能會因為信號的不匹配 (Signal mismatch) 如語速、語者、發音等聲學差異而使得動態時間校準比對錯誤，但查詢詞很有可能會常與某些詞一起出現，一旦系統透過查詢詞擴展將這些常與查詢詞一起出現的詞加入擴展後查詢詞中，系統即可利用這些常一起出現的詞去找到擁有查詢詞的文件。由表 4.2 中可以注意到系統對於線性疊加的權重 w_1 並不是很敏感，在 $w_1 = 0.7$ 和 $w_1 = 0.9$ 都可以觀察到進步，由於 $w_1 = 0.7$ 是最好的結果，因此在後續的實驗會將 w_1 設為 0.7。

同時我們也想比較監督式的查詢詞擴展(3.2節)在語意檢索上和口述語彙偵測的結果，這些結果呈現在表 4.2 的下半部，這結果是先將所有的語音文件辨識成



文字轉寫，再利用文字形式的查詢詞進行查詢詞擴展的檢索，在這邊可以觀察到查詢詞擴展在監督式的查詢下能有效地使系統進步(列(6)(7))。可以觀察到監督式的語意檢索進步量約為1.28%，而非監督式的語意檢索可以達到0.94%的進步量，而這進步量是接近於已被驗證許久的監督式查詢詞擴展了。

平均準確率		(A)	(B)
		語意檢索	口述 詞彙偵測
非監督式檢索	(1) 第一次檢索結果(DTW) ($w_1 = 1.0$)	8.76%	28.30%
	(2) 查詢詞擴展 ($w_1 = 0.0$)	6.03%	7.82%
	(3) 線性疊加	$w_1 = 0.9$	9.28%
		$w_1 = 0.7$	9.70%
	(4) 進步量 ($w_1 = 0.7$)	0.94%	2.01%
	(5) 第一次檢索結果 ($w_1 = 1.0$)	29.49%	70.07%
	(6) 查詢詞擴展 ($w_1 = 0.0$)	30.30%	68.86%
	(7) 線性疊加 ($w_1 = 0.7$)	30.77%	76.80%
監督式檢索	(8) 進步量	1.28%	6.73%

表 4.2: 系統在語意檢索和口述語彙偵測時的平均準確率

接著我們想觀察本章提出的方法是否真的能找到更多語意上相關的文件。首先，在使用動態時間校準產生第一次檢索結果時，對於每個查詢詞都取出其相關分數最高的200篇文件，由於總共有30個查詢詞，因此總共有6000篇文件，此為表4.3中的列(1)，利用本章方法找出的6000篇文件為列(2)，欄(A)中為這6000篇文件中與查詢詞語意相關的文件數，欄(B)為欄(A)中有出現查詢詞的文件數，欄



(C)為欄 (A)中沒有出現查詢詞的文件數，從表中可以看到，本章的方法確實能夠找到更多語意相關但是不包含查詢詞的文件，進步量為15.07%，而這些文件通常是很難被找到的，另外欄 (A) 和欄 (C) 中也各有 13.41% 和 11.36% 的進步。

MAP	(A) 所有與查詢詞語意相關的文件數	(B) (A) 中不含查詢詞的文件數	(C) (A) 中含查詢詞的文件數
(1) 第一次檢索結果 (DTW) $(w_1 = 1.0)$	589	325	264
(2) 本章提出的方法 $(w_1 = 0.7)$	668	374	294
(3) 進步量	13.41%	15.07%	11.36%

表 4.3: 系統找回與查詢詞語意相關的文件數量，包括含查詢詞與不含查詢詞的文件數。

接著來觀察系統的參數 N (虛擬相關文件數)、 w_1 (式 4.10 中線性疊加的權重)和 λ (式 4.4 中原查詢詞模型對於查詢詞擴展時的影響)對於系統的影響。圖 4.2 和圖 4.3 中顯示了使用不同的線性疊加權重 w_1 時系統的平均準確率，縱軸為平均準確率 (MAP)，橫軸為不同的線性疊加權重 w_1 ，圖中使用的答案均為語意檢索的答案，圖中的紅色虛線為檢索系統的基準 (Baseline) 檢索結果，為使用動態時間校準產生的第一次檢索結果。圖 4.2 中固定 N 為 12，將 λ 改變為0, 300, 900, 1500，可以觀察到當 $\lambda = 0$ 時，系統的表現是特別差的，因為此時系統會過適到虛擬相關文件的主題中，當 $\lambda > 300$ 後，系統的表現並不隨 λ 改變而影響太多，而最好的表現出現於 $\lambda = 300$ ，另外可以注意到系統在 w_1 從0.45到0.95時都有進步，因

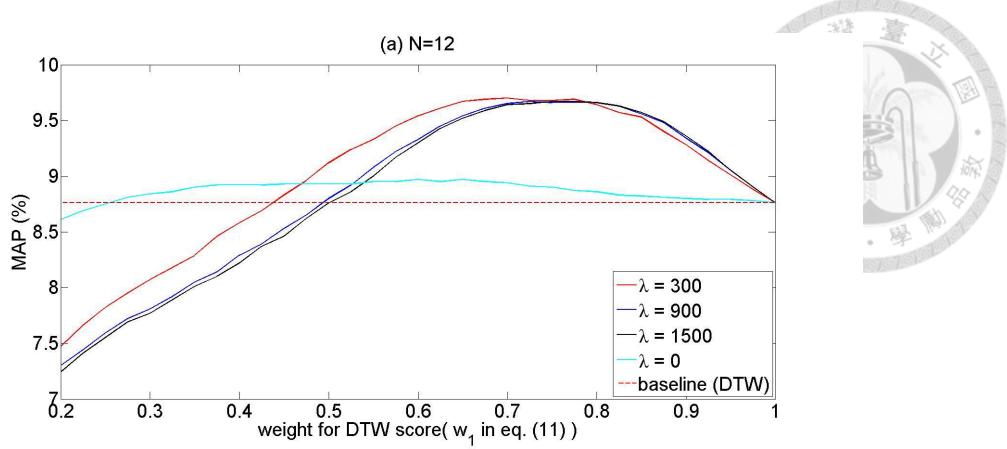


圖 4.2: 將第一次檢索結果和聲學組型形式的查詢詞擴展檢索結果疊加後的平均準確率 ($N = 800$)

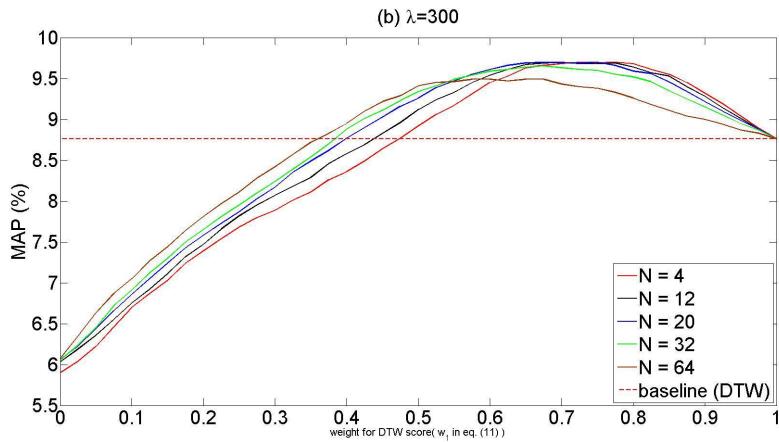


圖 4.3: 將第一次檢索結果和聲學組型形式的查詢詞擴展檢索結果疊加後的平均準確率 ($\lambda = 300$)

此本系統對於 w_1 的改變並不是太敏感。圖 4.3 中將 λ 固定為 300，並改變 N 為 4, 12, 20, 32, 64。同樣地可以觀察到 w_1 從 0.45 到 0.95 都有進步，並且當 N 上升時，平均準確率會先上升後下降，這原因是當 N 上升時，有更多文件被當作虛擬相關文件，因此有更多的訓練資料，但當 N 太大時，系統會用過多的不相關資料訓練，進而使得系統的表現下降。



4.5.1 聲學組型語意檢索能力分析

表 4.4 中的實驗顯示了本章提出的方法確實能夠在聲學組型的查詢詞中加入與查詢詞語意相關的聲學組型，即使系統並不知道這些聲學組型的含義。表 4.4 中是「學校」這個查詢詞在經過本章的查詢詞擴展後機率 $P(v|\theta_Q)$ 最高的五個 N 連聲學組型（由於聲學組型訓練時並沒有強制它一定是對應到中文的詞或是字，所以以下聲學組型對應到的聲音是經過很仔細的聆聽之後寫下它最接近的中文聲音）。列(1)(2) 中的聲學組型為 4.2.4 中提到的利用第一次檢索結果得到的原查詢詞的聲學組型，很明顯地列(1)(2) 中的聲學組型在擴展後的查詢詞模型中也佔了大部分的機率。列(3)(4)(5) 中的 N 連聲學組型則是在查詢詞擴展的過程中被自動加進去的，列(3)是由列(1)(2) 中的單連聲學組型組合而成的雙連聲學組型，列(4) 中的單連聲學組型對應到的中文聲音其實是與列(2) 中非常相似，但由於訓練時因為聲學上的特徵不一樣而被分群為不同的聲學組型，在此本章的方法也能將這些片段加進來進而提升檢索結果。而列(5) 加進來的是雙連聲學組型，當我們去聽這個聲學組型對應到的聲音時，發現聲音很像中文的「學生」，而這是與查詢詞「學校」有語意相關的詞，因此被系統加進來後也能提升檢索的結果。由以上的觀察可以發現本章提出的方法確實能把與查詢詞語意相關的聲學組型加進查詢詞模型中，並達成非監督式語意檢索。

4.6 本章總結

本章嘗試了使用聲學組型完成了非監督式的語意檢索，使用此種方法即可在不需要人為標注的情況下達成語意檢索。如此一來即不需要花費很多時間、金錢完成人為標注，不過目前非監督式語意檢索的表現比起監督式仍有一段落差，仍然需



$v(n$ 連聲學組型): (IDs)	$P(v \theta_Q)$	聲學組型舉例
(1) 單連: (87)	0.4280	校(/xiao/)
(2) 單連: (56)	0.3880	學(/xue/)
(3) 雙連: (56)-(87)	0.0040	學校(/xue-xiao/)
(4) 單連: (129)	0.0030	學(/xue/)
(5) 雙連: (129)-(23)	0.0016	學生(/xue-sheng/)

表 4.4: 查詢詞為”學校(/xue-xiao/)”時，擴展後查詢詞模型 ϕ_{qe} 中機率最大的五組聲學組型

要更多的研究才能夠使其應用在產業中。

第五章 利用遞迴式類神經網路語言模型加

強非監督式語音文件檢索



5.1 簡介

在上一章中，我們探討了如何使用聲學組型進行非監督式語意檢索。但另一方面，我們觀察到使用聲學組型時的問題：由於聲學組型在訓練時是採用非監督式方法的，因此訓練時系統並無法知道聲音對應到的文字是什麼，導致在分群時無法知道同音的字的意思是什麼，進而使得聲學組型在訓練時會盡量將同樣發音的字群組在一起。以上所述狀況會使檢索結果變差許多，因為同一個聲學組型事實上是對應到很多不同的字，舉其中一個聲學組型為例，其中包含了「網、王、亡、望、往」等字。因此本章試圖解決聲學組型的模糊性，比如說如果能將「網、王、亡、望、往」分為「網」、「王」、「亡」、「望」、「往」等五群，即可大幅地提升檢索系統的功效了。

在此我們使用了其於遞迴式類神經網路語言模型 (Recurrent Neural Network Language Model) [55–57] 之詞表示法 (Word Representation)，此一詞表示法可以將每一個詞或字表示為一固定長度的向量，並被認為能有效地表示詞或字在句法 (Syntactic) 和語意上的相似度 [58]。由於同一個聲學組型中包含了許多實際上含義不同的字，因此我們的目的是將聲學組型表示為詞表示法，並基於詞表示法為這些聲學組型做分群，以期在句法和詞義的基礎上將原本同音的聲學組型分為不同的聲學組型。



5.2 基於遞迴式類神經網路語言模型之詞表示法

5.2.1 基於遞迴式類神經語言模型之詞表示法

傳統的詞表示法通常將字表示為 1-of-N 編碼，即每個詞都被表示成一個辭典大小的陣列，只有那個詞對應到的維度是一，其它都是零。這種詞表示法的缺點是將每個詞都視為獨立的詞，即詞與詞之間是無相關性的。舉例來說，如果辭典中有「Hotel」、「Hostel」兩個字。1-of-N 編碼會將這兩個字表示為：

$$hotel = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$$

$$hostel = [0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$$

於是「Hotel」與「Hostel」兩字的詞表示法取內積後為零，但這是不合理的狀況，因為這兩個字有很強烈的語意關係，因此就有人提出了使用連續空間的詞表示法 (Continuous Space Word Representation) [59,60]，此一方法的目的是用一個高維度的實數陣列代表一個詞，並且當兩個詞在句法上和含義上相近的時候會讓兩個詞表示法很接近。類神經網路語言模型的一個特色是它們將字表示成高維度的實數陣列，而這些詞表示法甚至可以用在許多自然語言處理的工作上。基於遞迴式類神經網路語言模型的詞表示法是目前被廣泛使用的詞表示法之一 [55]，模型中的 \mathcal{U} 是一個 $\mathcal{H} \times \mathcal{L}$ 的矩陣，其中 \mathcal{H} 是隱藏層的維度， \mathcal{L} 則是辭典的長度，而 \mathcal{U} 中的每一欄都代表一個詞的詞表示法。雖然類神經網路語言模型在訓練時完全沒有句法(Syntax)、構詞(Morphology)、或語意(Semantics)相關的資訊，但很令人驚訝地，靠著沿時間反向傳遞演算法的最佳化就能夠非常好地在句法和語意上描述這些詞表示法 [58]。

圖 5.1 中示意了基於遞迴式類神經網路語言模型之詞表示法，這是使用了



圖 5.1: 基於遞迴式類神經網路語言模型之詞表示法示意圖

公視新聞訓練的詞表示法，隱藏層維度為100，並將這些100維的詞表示法降為兩維後畫在圖上，由圖中可以看出詞表示法表示詞的能力：如在圖中右方，「今天」、「明天」、「昨天」，「晚上」、「晚間」、「上午」、「下午」都被畫在很接近的位置，而這些詞都是用來表示時間的詞彙，又例如圖中下方的「布希」、「陳水扁」被畫得非常接近，而這是由於在這些語料錄製的年份，這兩個人分別是美國與台灣的總統，由這些例子可以看出，雖然遞迴式類神經網路語言模型之詞表示法訓練的過程中是沒有任何關於語意的知識的，但是它卻可以很好地捕捉到詞彙在語意上的資訊，因此很適合用在自然語言處理。

5.3 以詞表示法改善非監督式語意檢索

演算法之示意如圖 5.2，圖上方的紫色框框中為訓練資料，其中的每一行代表

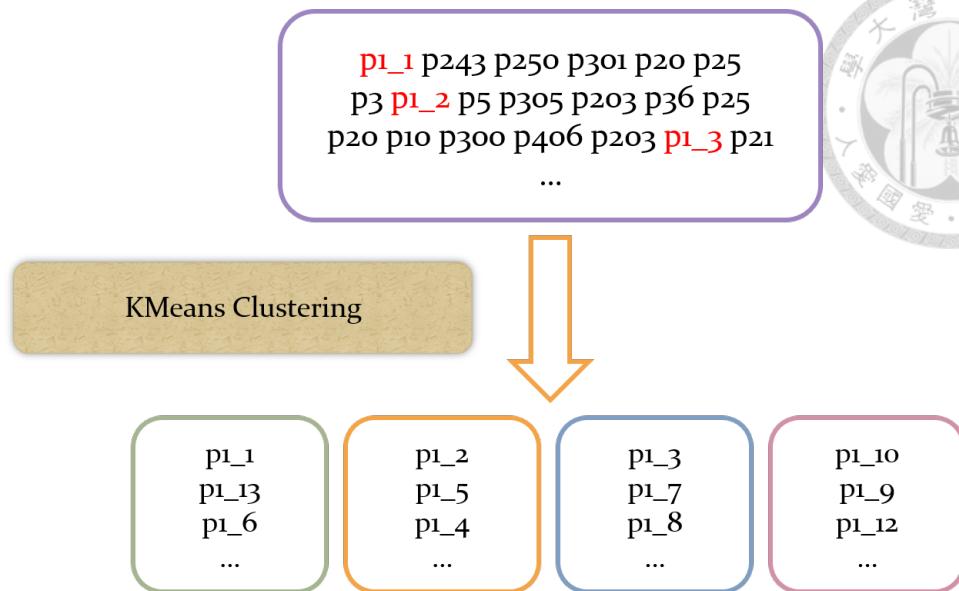


圖 5.2: 演算法示意圖

一句由聲學組型組合而成的句子，由於其中每一個聲學組型實際上都可能代表了很多不同的詞或字，因此我們在每一次的迴圈當中要對一個聲學組型進行分群。我們的做法是在每一個迴圈中選定一個聲學組型進行分群，如圖 5.2 中的迴圈中選定了 p_1 這個聲學組型，並將所有的 p_1 視為是不同的聲學組型，如 $p_{1.1}, p_{1.2}, p_{1.3}, \dots$ ，並進行K平均分群演算法 (K-Means Clustering) 基於這些詞和字的詞表示法進行分群，由於這些詞表示法攜有了許多的句法和詞義的資訊，因此這個分群的結果會將這些聲學組型按照句法和語意分開，以期在進行K平均分群後，每一群中包含的聲學組型對應到的中文詞彙會盡量一致。

演算法的步驟為在每個迴圈都從所有的聲學組型中選一個聲學組型（選過的不再選），稱為 p ，並在這個迴圈中重複以下的步驟：

1. 將訓練語料中所有的 p 都視為不同的聲學組型，亦即將所有的 p 按照出現順序改為 p_1, p_2, p_3, \dots 。
2. 將修改後的訓練語料做為遞迴式類神經網路語言模型的訓練語料，並訓練一套遞迴式類神經網路語言模型。

3. 取出遞迴式類神經網路語言模型的 \mathcal{U} 做為所有聲學組型的詞表示法，並用K平均分群法將這些詞表示法分為K群，這K群的聲學組型即為不同的聲學組型，標示為 $p_{k1}, p_{k2}, p_{k3}, \dots$ 。



5.4 實驗基礎架構

本章實驗所使用的設定同 4.3，使用的語料為2001年間從電台廣播中錄下的4小時新聞，並手動切成5034篇語音文件，每篇語音文件大約包含了13句的語句。用來辨識的語言模型是用1999年間收集的新聞文章(包含4000萬個詞彙)訓練而成，辭典中包含了62000個詞彙，聲學模型是用2000年間收集的8小時廣播新聞訓練而成的音節內(Intra-syllable)右方資訊相依(Right-context-dependent)聲韻母模型(Initial-Final Models)。辨識後的唯一最佳序列的字元正確率為(Character Accuracy)為75.27%。本章中使用的聲學組型設定為將5034句的新聞語音文件辨識成208個字單位的語音組型。總共測試了30組口語查詢詞，這些口語查詢詞都使用了4.2.4 轉為對應的聲學組型查詢詞，每組查詢詞都有人工標注對應的相關文件，而這些相關文件中「一定要」包含查詢詞。

5.5 實驗結果

實驗結果如圖 5.3 和 5.4 所示。橫軸為迴圈的次數，每次迴圈都從中找出一個聲學組型進行 5.3 的演算法，由於總共有208個聲學組型，因此橫軸到208後即為對所有的聲學組型都執行一次演算法。縱軸為檢索結果的平均準確率，紅線為檢索系統的基準(Baseline)檢索結果，同 4.2 中的紅線，為使用動態時間扭曲產生的第一次檢索結果。

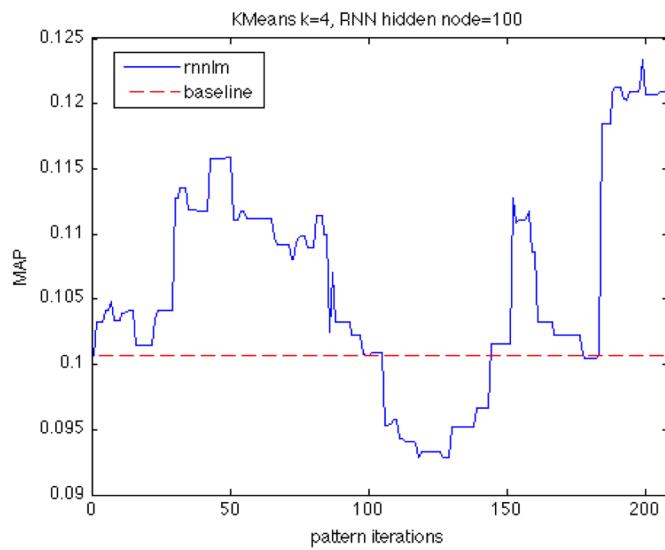


圖 5.3: 實驗結果：K分群法的K設定為4，潛藏層長度設為100

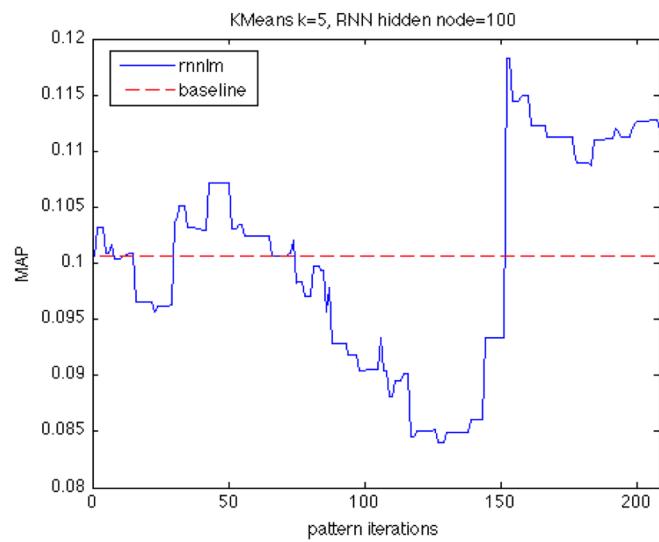


圖 5.4: 實驗結果：K分群法的K設定為5，潛藏層長度設為100



由圖中可以觀察到，在大部分的時間點，此方法都能使檢索的成效增加，在最高點的平均準確率可以到達12.23%，進步量相對於基準檢索結果為21.4%，證明此方法確實能有效地將同音的聲學組型切為對檢索結果更有貢獻的聲學組型，並給予聲學組型更多地語意資訊，但此方法的缺點目前的強健性 (Robustness) 還不夠，無法讓平均準確率在所有的迴圈中都穩定上升，但能夠讓某些聲學組型有很大幅的成長。未來可能的改進方向如下：

1. 嘗試用不同的分群演算法，並使用開發語料 (Development Set) 調整分群演算法的參數。
2. 觀察為何有些聲學組型能獲得大幅的進步，並試圖找出判斷依據，如此一來就可以只分會獲得進步的聲學組型，並不要對會退步的聲學組型做這套演算法。

5.6 本章總結

本章試圖解決聲學組型在訓練時只考慮同音的問題，並試圖為聲學組型加入更多的句法與語意資訊，再將加入額外資訊的聲學組型應用於語音文件檢索系統中，並獲致了初步的進步。

第六章 在Google Glass上實作個人化的語

音翻譯與新聞檢索系統



6.1 簡介

隨著近年來行動裝置與雲端運算的興起，行動裝置逐漸成為了業界的兵家必爭之地，許多公司都喊出了「行動優先」口號，足見業界對於行動裝置平台之重視，如Google、Facebook、Apple等科技大廠皆紛紛搶入行動裝置平台。由於行動裝置小、鍵盤輸入介面不方便、往往需要在移動時輸入等特性，使得語音輸入在行動裝置上相形重要，因為語音相對於傳統的鍵盤輸入是更自然也更貼近人性的輸入方法。但行動裝置上的語音輸入不比傳統的語音輸入，有許多與傳統語音輸入不同的特性，由於這些特性將導致一些困難的挑戰和更多有趣的性質可供研究，以下將簡介這些特性。

第一個特性為自發性語音 (Spontaneous) 含發音差異 (Pronunciation Variation)、不流暢性 (Disfluency)、語速變化 (Speaking Rate Variation) [61]及背景雜訊 (Background Noise)，由於使用者使用行動裝置上的語音輸入時往往是很自發性地 (Spontaneous) 說話，因此這些語速是很不固定的，將隨著使用者的使用情景、當下的情緒等等而改變，另一方面行動輸入通常會在有背景雜音的狀況下輸入，如馬路上、百貨公司中，會附有很多的背景雜音，因此語速變化與背景雜音可以說是行動語音輸入最主要的困難之一。

第二個特性為個人化 (Personalization)，這是由於行動裝置通常只屬於某一特定的使用者，因此行動裝置可以搜集大量的個人資訊對使用者進行個人化，使得使用者的語音辨識結果最接近他有可能說出的話。這其中包含了語音模型、語言

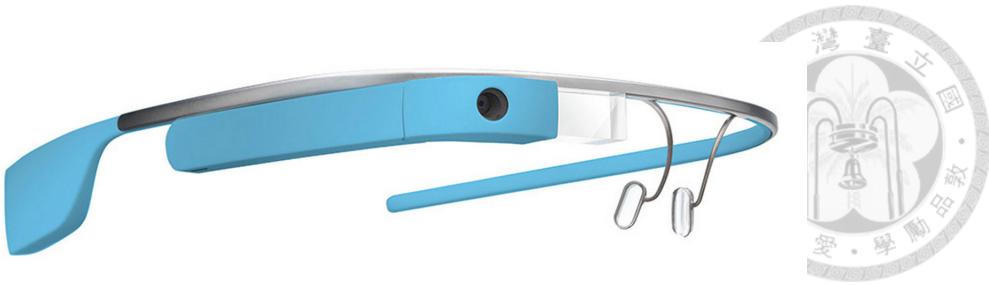


圖 6.1: Google 眼鏡

模型與辭典的個人化 [62–64]，語音模型可以調適 (Adaptation) 至接近使用者語音特性的語音模型，語言模型與辭典則可以藉由收集使用者常說與常用的詞（可以從過去的語音輸入中收集，或是從社群網路如 Facebook、Twitter 上收集）進行調適。如此一來即可以將使用者的行動語音輸入個人化成最適合使用者使用的行動語音輸入。

第三個特性為行動裝置上的感測器 (Sensor)，這是行動裝置上特有的特性，由於行動裝置上通常帶有全球定位系統 (Global Positioning System, GPS)，而未來甚至可能帶有個人身體狀況的感測器，如血壓、心跳等，而這些感測器的特徵正好可用來幫助語音辨識，如裝置知道使用者的位置，就可以將當地景點、餐廳在語言模型中的機率增加，如裝置知道使用者血壓、心跳增高，可能代表使用者處於憤怒狀態，此時憤怒的詞彙在語言模型中的機率也可以被適當地增加。因此如果可以擅用這些感測器的結果，將能很好地幫助語音辨識。

本章中的個人化語音系統的應用皆實作於 Google 推出的 Google 眼鏡 (Google Glass) 之上 (如圖 6.1)，Google 眼鏡主體為其右上角之藍色部分，其前面包含了一塊透明的顯示器，Google 眼鏡會將顯示畫面投影到透明的顯示器上讓使用者閱讀，藍色的部分還包含了麥克風、喇叭、相機、觸控板（可偵測多種手勢）等裝置。

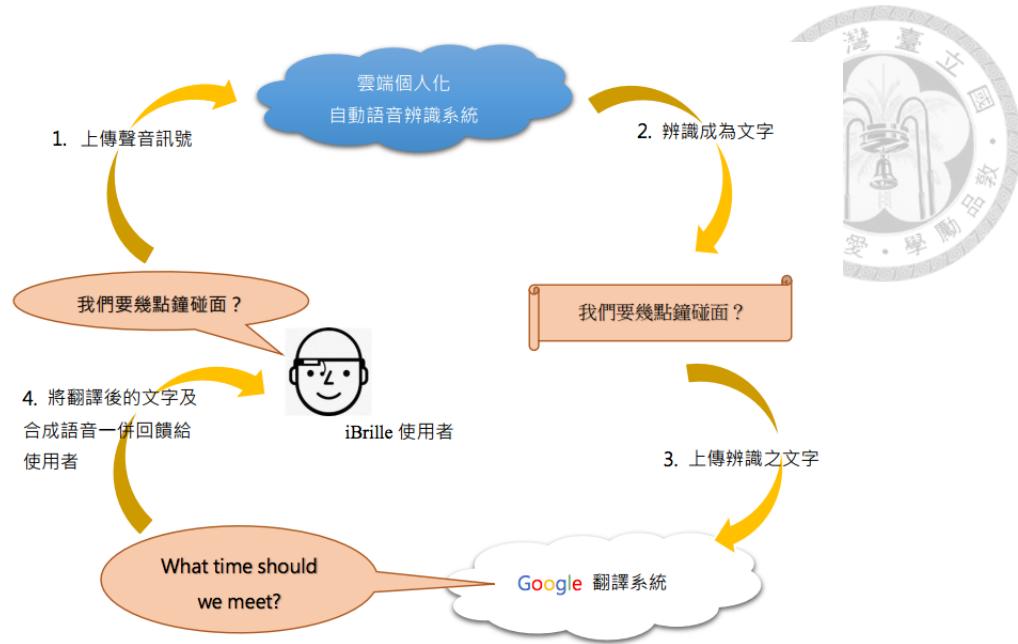


圖 6.2: 雲端個人化語言翻譯系統架構

6.2 個人化的語言翻譯系統簡介

以下將介紹本實驗室開發的個人化語言翻譯系統，其開發動機是由於穿戴型裝置日益流行，如果能讓使用者在旅遊或是與外國人對話的期間，一旦有不會講的話就能直接用自己的母語對裝置說話，並由裝置進行辨識與翻譯後告訴使用者該如何用另一個語言講這段話。

因此我們設計了一套個人化語言翻譯系統，利用智慧型行動裝置如手機、甚至是近年來逐漸受到重視的智慧型穿載裝置，像Google眼鏡等，讓使用者能隨時隨地地進行為自己量身打造的體驗：使用者以自己的母語講出想學的內容，行動裝置會將訊號送至雲端的個人化語音辨識系統，將語音辨識成文字，再將辨識後的文字送至Google提供的翻譯服務將文字翻譯成使用者想學的目標語言，再回傳到行動裝置讓使用者同時看到翻譯後的文字與聽見範例發音，系統示意如圖 6.2

本系統可分為兩部分：

個人化自動語音辨識系統 (Personalized Automatic Speech Recognition System)

目前使用的雲端辨識系統可分為幾個區塊，包含聲學模型、語言模型、詞典，依序介紹如下：



- 聲學模型：

聲學模型架構使用的是三連音(Triphone)隱藏式馬可夫模型(Hidden Markov Model, HMM)，每一個馬可夫模態(HMM State)則是由深層類神經網路(Deep Neural Network)的輸出層中的一個神經元(Neuron)來表示，也就是近年流行的上下文相關深層類神經網路隱藏式馬可夫模型 (Context-dependent Deep Neural Network Hidden Markov Model, CD-DNN-HMM)。類神經網路端所輸入的語音特徵則是疊加了九個音框長度的梅爾濾波器組特徵(Mel-Filter Bank Features)，維度是621維。網路中共有四層隱藏層(Hidden Layer)，每一層中有2048個隱藏神經元(Hidden Neuron)，輸出層則是有6647個節點。用來訓練聲學模型的語料則是結合了雙語料，其中聲碩麥克風(ASTMIC)做為中文語料以及台灣區英語(English Across Taiwan)當中非母語語者(Non-native Speaker)部分做為英文語料。兩個語料都是收集自台灣區語者的錄音，並且都是多語者(Multi-speaker)語料，且性別各半均衡。

- 語言模型：

語言模型採用傳統的N連詞(N-gram)模型，並且為三連詞模型。訓練語料包含了Yahoo! 新聞(Yahoo! News)、Gigaword、公視新聞 (PTV)，並且在訓練完成後，與以聲碩麥克風與台灣區英語的訓練集語料訓練的語言模型作內插調適。

- 詞典：

此系統詞典中文部分包含所有常見中文單字，以及經由PAT-Tree所產生的字詞。英文部分，則是包含卡內基美儂大學所公開之英文辭典當中，詞頻較高者。所有字詞的發音皆是音素(Phoneme)序列，包含中英雙語的所有音素。



翻譯系統

經辨識所得之文字，將送至 Google 的雲端翻譯系統，並由 Google 翻譯將其翻譯為使用者想學的目標語言，再將翻譯後的文字與合成發音一併傳回 Google 眼鏡上，將使用者想翻譯的語言呈現於屏幕上，並同步撥放發音以助翻譯。此一部分由於本實驗室尚未有完善的翻譯系統，不過一旦未來研發出辨識系統後，即可將 Google 翻譯系統替換為本實驗室自行開發之翻譯系統。

6.3 個人化的語音文件檢索系統簡介

以下將介紹本實驗室自行開發的個人化行動語音文件檢索系統，其開發動機為使用者在行動時也往往有搜尋資訊的需求，而在移動時使用鍵盤輸入對於使用者又極度地不方便，因此我們就決定開發了這套基於 Google 眼鏡的語音文件檢索系統，讓使用者能夠隨時隨地取得自己想要的資訊，使用情境如下：使用者對 Google 眼鏡輸入語音查詢詞，然後 Google 眼鏡再將訊號上傳到個人化語音辨識系統，辨識為文字後，系統再將文字的語音查詢詞上傳至 3.2 中提到的監督式語意文件檢索系統。檢索系統會再將檢索後的內容傳回給使用者，並呈現於 Google 眼鏡上供使用者閱讀。

本系統可分為兩部分：

個人化自動語音辨識系統

此部分同上一節個人化的語言翻譯系統中的個人化語音辨識系統。



監督式語意檢索系統

此處使用之系統同 3.2，使用的語料為2001年間從電台廣播中錄下的4小時新聞，並手動切成5034篇語音文件，每篇語音文件大約包含了1至3句的語句。用來辨識的語言模型是用1999年間收集的新聞文章(包含4000萬個詞彙)訓練而成，辭典中包含了62000個詞彙，聲學模型是用2000年間收集的8小時廣播新聞訓練而成的音節內(Intra-syllable)右方資訊相依(Right-context-dependent)聲韻母模型(Initial-Final models)。辨識後的唯一最佳序列的字元正確率為(Character Accuracy)為75.27%。

6.4 系統展示

此節將展示本系統之部分截圖與使用示例。

6.4.1 個人化的語音翻譯系統展示

本實驗室開發的個人化語音翻譯系統的名稱取為iBrille，取Brille在德語中為眼鏡，在西班牙語及法語中有閃耀的意思。圖 6.3 中為 iBrille 系統的首頁，一進入系統後會顯示「歡迎來到 iBrille 教學系統」，圖 6.4 中則為使用者對 iBrille 講出一段欲翻譯的話後，由個人化的語音翻譯系統辨識、並翻譯後顯示的畫面。

6.4.2 個人化的語音文件檢索系統展示

本實驗室開發的個人化語音文件檢索系統的名稱取為「新聞隨手查」，希望能讓

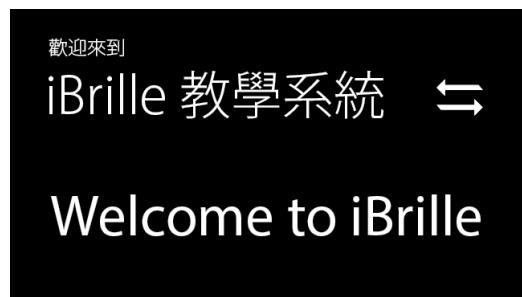


圖 6.3: iBrille 首頁



圖 6.4: iBrille 使用示範

使用者無論在隨時隨地都能很方便迅速地吸收到感興趣的資訊。圖 6.5 為新聞隨手查的首頁，而圖 6.6 則為使用者說出「颱風」這個查詢詞後顯示的畫面。

6.5 本章總結

行動裝置在這幾年內漸形重要，因此也使得在行動裝置上的語言輸入漸受重視，本章中利用了 Google 眼鏡實作了個人化的語音翻譯系統與個人化的語音文件檢索系統，將個人化的語音辨識系統與行動裝置上的應用程序結合起來，成為對使用者更為方便的應用程序。

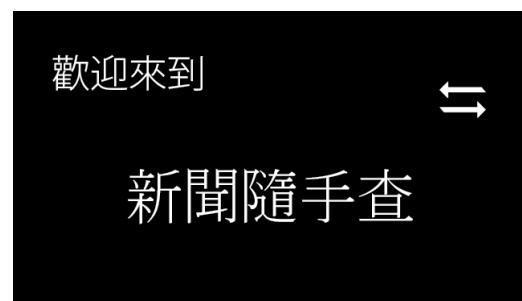


圖 6.5: 新聞隨手查首頁

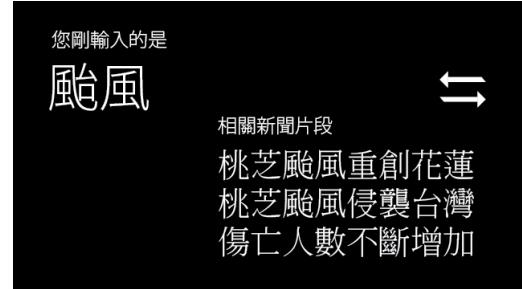


圖 6.6: 新聞隨手查使用示範

第七章 結論與展望

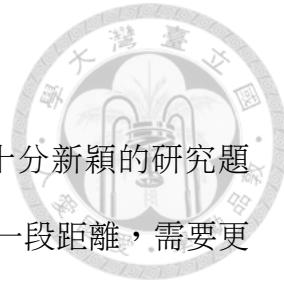


7.1 本論文主要的研究貢獻與未來展望

7.1.1 使用聲學組型加強語音文件檢索

本論文中主要貢獻在於利用聲學組型解決傳統語音檢索的難題，由於傳統的語音文件檢索是先辨識後在詞圖上檢索，但有許多聲學上的資訊在辨識之後就消失了，因此本論文試圖在檢索時加入聲學組型的資訊以提升檢索系統成效，主要貢獻條列如下：

1. 使用聲學組型加強監督式語音文件的語意檢索系統，因為傳統的語音文件檢索是先辨識後在詞圖上檢索，但如果此時有辭典外詞彙或是辨識錯誤的話，檢索結果就會很差了，因此本論文使用聲學組型解決此難題。
2. 使用聲學組型達成非監督式語音文件的語意檢索系統，傳統的語意檢索系統需要先將語音文件辨識成詞圖後才進行語意檢索，但是這樣需要已訓練得很好的聲學模型和語言模型，而這兩者的訓練通常是非常昂貴的，因此我們使用聲學組型解決這個問題。
3. 使用聲學組型雖然能加強以上兩種狀況，但是聲學組型有一個缺點：聲學組型在訓練時並不知道聲音和字之間的關聯，因此會將所有音同字不同的聲音都歸類到同一個聲學組型中，但如此一來在檢索時便無法區別不同的字義了，進而導致檢索的成效很差，所以我們試圖用遞迴式類神經網路語言模型產生的詞表示法來區分出同一個聲學組型中對應到不同字義的聲學組型。



未來的改進方向

由於使用聲學組型進行完全非監督式的語音文件檢索是一個十分新穎的研究題目，因此目前系統的平均準確率距離監督式語音文件檢索尚有一段距離，需要更進一步地研究，而可能的研究方向如下：

1. 將同音的聲學組型盡可能按照它對應到的字區分開來：由於目前的聲學組型是將同音的組型盡量分在一起，但這在檢索時會造成困難，因此如果能將聲學組型按照它對應到的字區分開來的話，將能使檢索的效能進步許多，第 5 章 中嘗試了一些做法，但還可以嘗試不同的分群演算法和更好的詞表示法。
2. 使用詞表示法改進檢索系統：目前的檢索系統仍然是使用傳統的詞表示法，因此無法得知詞與詞之間的相似關係，如果能將檢索系統的詞表示法改為第 5 章中的方法，應能在語意檢索上再進步許多。

7.1.2 實作雲端語音辨識與應用程式於 Google 眼鏡

行動裝置在這幾年間快速地崛起，連帶使得行動裝置上最自然的輸入方式-語音輸入受到許多人的重視，另一方面，行動裝置上的語音辨識系統會遭遇到許多與傳統語音辨識系統不同的挑戰與機會，諸如語速的變化、聲學/語言模型/辭典的個人化、行動裝置上的感測器等等。本論文將本實驗室開發的雲端個人化語音辨識系統實作於目前最流行的穿戴式裝置-Google 眼鏡之上，並於其上建立了兩個應用程序：

1. 雲端個人化語言學習系統：使用雲端語音辨識加上Google翻譯，使學生能夠隨時隨地學習自己需要的內容。
2. 雲端個人化新聞查詢系統：使用雲端語音辨識加上 3.2 中提到的監督式語意檢索系統，讓使用者隨時隨地都能查詢自己感興趣的新聞。

參 考 文 獻



- [1] Yaodong Zhang and James R Glass, “Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriograms,” in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.
- [2] Ciprian Chelba, Timothy J Hazen, and Murat Saracclar, “Retrieval and browsing of spoken content,” *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 39–49, 2008.
- [3] Lin-shan Lee and Berlin Chen, “Spoken document understanding and organization,” *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 42–60, 2005.
- [4] Murat Saracclar and Richard Sproat, “Lattice-based search for spoken utterance retrieval,” *Urbana*, vol. 51, pp. 61801.
- [5] ChengXiang Zhai, “Statistical language models for information retrieval,” *Synthesis Lectures on Human Language Technologies*, vol. 1, no. 1, pp. 1–141, 2008.
- [6] Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng, “Statistical lattice-based spoken document retrieval,” *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 1, pp. 2, 2010.
- [7] Chengxiang Zhai and John Lafferty, “A study of smoothing methods for language models applied to ad hoc information retrieval,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 334–342.

- [8] Chun-an Chan and Lin-shan Lee, “Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping.,” in *INTERSPEECH*, 2010, pp. 693–696.
- [9] Timothy J Hazen, Wade Shen, and Christopher White, “Query-by-example spoken term detection using phonetic posteriogram templates,” in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.
- [10] John S Garofolo, Cedric GP Auzanne, and Ellen M Voorhees, “The trec spoken document retrieval track: A success story.,” .
- [11] Ian Ruthven and Mounia Lalmas, “A survey on the use of relevance feedback for information access systems,” *The Knowledge Engineering Review*, vol. 18, no. 02, pp. 95–145, 2003.
- [12] Gerard Salton, Anita Wong, and Chung-Shu Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
- [13] Chengxiang Zhai and John Lafferty, “Model-based feedback in the language modeling approach to information retrieval,” in *Proceedings of the tenth international conference on Information and knowledge management*. ACM, 2001, pp. 403–410.
- [14] Stephen E Robertson and K Sparck Jones, “Relevance weighting of search terms,” *Journal of the American Society for Information science*, vol. 27, no. 3, pp. 129–146, 1976.

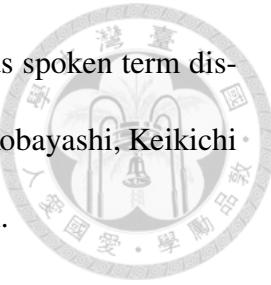


- [15] Simon Tong and Edward Chang, “Support vector machine active learning for image retrieval,” in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 107–118.
- [16] King-Shy Goh, Edward Y Chang, and Wei-Cheng Lai, “Multimodal concept-dependent active learning for image retrieval,” in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 564–571.
- [17] Jingrui He, Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, and Changshui Zhang, “Mean version space: a new active learning method for content-based image retrieval,” in *Proceedings of the 6th ACM SIGMM international workshop on Multi-media information retrieval*. ACM, 2004, pp. 15–22.
- [18] Diane Kelly and Jaime Teevan, “Implicit feedback for inferring user preference: a bibliography,” in *ACM SIGIR Forum*. ACM, 2003, vol. 37, pp. 18–28.
- [19] Oren Kurland, Lillian Lee, and Carmel Domshlak, “Better than the real thing?: iterative pseudo-query processing using cluster-based language models,” in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005, pp. 19–26.
- [20] Jinxi Xu and W Bruce Croft, “Query expansion using local and global document analysis,” in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 4–11.
- [21] Shipeng Yu, Deng Cai, Ji-Rong Wen, and Wei-Ying Ma, “Improving pseudo-relevance feedback in web information retrieval using web page segmentation,” in

Proceedings of the 12th international conference on World Wide Web. ACM, 2003,
pp. 11–18.



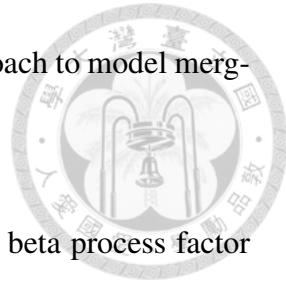
- [22] Tetsuya Sakai, Toshihiko Manabe, and Makoto Koyama, “Flexible pseudo-relevance feedback via selective sampling,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 4, no. 2, pp. 111–135, 2005.
- [23] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson, “Selecting good expansion terms for pseudo-relevance feedback,” in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 243–250.
- [24] Kyung Soon Lee, W Bruce Croft, and James Allan, “A cluster-based resampling method for pseudo-relevance feedback,” in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 235–242.
- [25] Yuanhua Lv and ChengXiang Zhai, “A comparative study of methods for estimating query language models with pseudo feedback,” in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 1895–1898.
- [26] Yuanhua Lv and ChengXiang Zhai, “Positional relevance model for pseudo-relevance feedback,” in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 579–586.



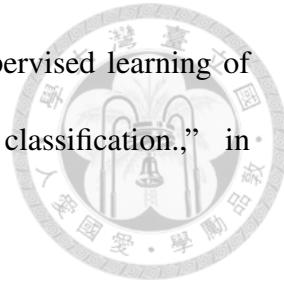
- [27] Aren Jansen, Kenneth Church, and Hynek Hermansky, “Towards spoken term discovery at scale with zero resources,” in *INTERSPEECH*, Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, Eds. 2010, pp. 1676–1679, ISCA.
- [28] Mikael Bodén, “A guide to recurrent neural networks and backpropagation,” *The Dallas project, SICS technical report*, 2002.
- [29] Tao Tao and ChengXiang Zhai, “Regularized estimation of mixture models for robust pseudo-relevance feedback,” in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 162–169.
- [30] Victor Lavrenko and W Bruce Croft, “Relevance based language models,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 120–127.
- [31] Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran, “Query-by-example spoken term detection for oov terms,” in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 404–409.
- [32] Atta Norouzian, Aren Jansen, Richard C Rose, and Samuel Thomas, “Exploiting discriminative point process models for spoken term detection.,” in *INTERSPEECH*, 2012.
- [33] Hung-yi Lee, Po-wei Chou, and Lin-shan Lee, “Open-vocabulary retrieval of spoken content with shorter/longer queries considering word/subword-based acoustic feature similarity.,” in *INTERSPEECH*, 2012.



- [34] Hung-Yi Lee, Yun-Nung Chen, and Lin-Shan Lee, “Improved speech summarization and spoken term detection with graphical analysis of utterance similarities,” *Proc. APSIPA*, 2011.
- [35] Hung-yi Lee, Chia-ping Chen, and Lin-shan Lee, “Integrating recognition and retrieval with relevance feedback for spoken term detection,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2095–2110, 2012.
- [36] Chia-ying Lee and James Glass, “A nonparametric bayesian approach to acoustic model discovery,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.
- [37] Aren Jansen and Kenneth Church, “Towards unsupervised training of speaker independent acoustic models.,” in *INTERSPEECH*, 2011, pp. 1693–1692.
- [38] Aren Jansen, Kenneth Church, and Hynek Hermansky, “Towards spoken term discovery at scale with zero resources.,” in *INTERSPEECH*, 2010, pp. 1676–1679.
- [39] Alex S Park and James R Glass, “Unsupervised pattern discovery in speech,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 186–197, 2008.
- [40] Veronique Stouten, Kris Demuynck, and Hugo Van Hamme, “Discovering phone patterns in spoken utterances by non-negative matrix factorization,” *Signal Processing Letters, IEEE*, vol. 15, pp. 131–134, 2008.



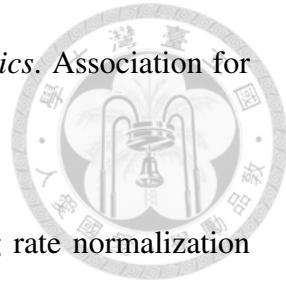
- [41] Lei Wang, Eng Siong Chng, and Haizhou Li, “An iterative approach to model merging for speech pattern discovery,” 2011.
- [42] Niklas Vanhainen and Giampiero Salvi, “Word discovery with beta process factor analysis.,” in *INTERSPEECH*, 2012.
- [43] Joris Driesen and H Van Hamme, “Fast word acquisition in an nmf-based learning framework,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5137–5140.
- [44] Yaodong Zhang and James R Glass, “Towards multi-speaker unsupervised speech pattern discovery,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4366–4369.
- [45] Man-Hung Siu, Herbert Gish, Arthur Chan, and William Belfield, “Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision.,” in *INTERSPEECH*, 2010, pp. 2838–2841.
- [46] Timothy J Hazen, Man-Hung Siu, Herbert Gish, Steve Lowe, and Arthur Chan, “Topic modeling for spoken documents using only phonetic information,” in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 395–400.
- [47] Herbert Gish, Man-hung Siu, Arthur Chan, and William Belfield, “Unsupervised training of an hmm-based speech recognizer for topic classification.,” in *INTERSPEECH*, 2009, pp. 1935–1938.



- [48] Sourish Chaudhuri, Mark Harvilla, and Bhiksha Raj, “Unsupervised learning of acoustic unit descriptors for audio content representation and classification.,” in *INTERSPEECH*, 2011, pp. 2265–2268.
- [49] Marijn Huijbregts, Mitchell McLaren, and David van Leeuwen, “Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4436–4439.
- [50] Chun-An Chan and Lin-Shan Lee, “Unsupervised hidden markov modeling of spoken queries for spoken term detection without speech recognition.,” in *INTERSPEECH*, 2011, pp. 2141–2144.
- [51] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li, “An acoustic segment modeling approach to query-by-example spoken term detection,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5157–5160.
- [52] Matthew Riley, Eric Heinen, and Joydeep Ghosh, “A text retrieval approach to content-based audio retrieval,” in *Int. Symp. on Music Information Retrieval (ISMIR)*, 2008, pp. 295–300.
- [53] Yang Liu, Wan-Lei Zhao, Chong-Wah Ngo, Chang-Sheng Xu, and Han-Qing Lu, “Coherent bag-of audio words model for efficient large-scale video copy detection,” in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 89–96.

- [54] Cheng-Tao Chung, Chun-an Chan, and Lin-shan Lee, “Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8081–8085.
- [55] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur, “Recurrent neural network based language model.,” in *INTERSPEECH*, 2010, pp. 1045–1048.
- [56] Tomas Mikolov, Stefan Kombrink, Lukas Burget, JH Cernocky, and Sanjeev Khudanpur, “Extensions of recurrent neural network language model,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5528–5531.
- [57] Tomas Mikolov and Geoffrey Zweig, “Context dependent recurrent neural network language model.,” in *SLT*, 2012, pp. 234–239.
- [58] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, “Linguistic regularities in continuous space word representations,” in *Proceedings of NAACL-HLT*, 2013, pp. 746–751.
- [59] Ronan Collobert and Jason Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.
- [60] Joseph Turian, Lev Ratinov, and Yoshua Bengio, “Word representations: a simple and general method for semi-supervised learning,” in *Proceedings of the 48th*

Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010, pp. 384–394.



- [61] Ching feng Yeh, Hung yi Lee, and Lin-Shan Lee, “Speaking rate normalization with lattice-based context-dependent phoneme duration modeling for personalized speech recognizers on mobile devices,” In Bimbot et al. [65], pp. 1741–1745.
- [62] Tsung-Hsien Wen, Aaron Heidel, Hung yi Lee, Yu Tsao, and Lin-Shan Lee, “Recurrent neural network based language model personalization by social network crowd-sourcing,” In Bimbot et al. [65], pp. 2703–2707.
- [63] Tsung-Hsien Wen, Hung-Yi Lee, Tai-Yuan Chen, and Lin-Shan Lee, “Personalized language modeling by crowd sourcing with social network data for voice access of cloud applications,” in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 188–193.
- [64] Jerome R Bellegarda, “Statistical language model adaptation: review and perspectives,” *Speech communication*, vol. 42, no. 1, pp. 93–108, 2004.
- [65] Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, Eds., *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013.