語音辨識及資訊檢索技術於數位典藏多媒體文物之應用

陳鴻彬　陳柏琳　林順喜
國立台灣師範大學 資訊工程研究所
{james,berlin,linss }@csie.ntnu.edu.tw

摘要

當今博物館的數位典藏內容，可謂多采多姿、無所不包。這些典藏資料通常以影音、圖片及文字等形式儲存，不但分散於不同系統與平台，而且這些典藏文物並不容易取得與使用。有基於此，在本論文中，我們發展了一套無線網路環境下以語音為基礎的多媒體資訊檢索介面，其中包括了大詞彙連續語音辨識和多尺度索引等核心技術。我們將此多媒體資訊檢索介面與國立歷史博物館豐富的數位典藏內容作結合，研發成一套以PDA為平台的多媒體資訊檢索雛形系統。未來，希望能夠提供博物館的參觀民眾一個自然且有效率的數位典藏文物學習環境。

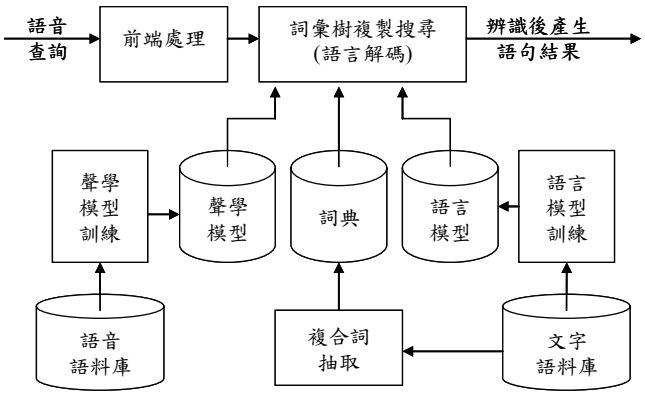
關鍵詞：大詞彙語音辯識、多尺度索引技術、多媒體、數位典藏。

一、 簡介

現今的數位典藏多媒體資料迅速成長，內容也益趨於多樣化，如相同的一件典藏文物，可能同時的擁有純文字的內涵描述文件和呈現其外在風貌的多媒體電子檔，而文物與文物間的檔案又交叉使用，更造成內容的複雜化[1]。為了要讓使用者從大量龐雜的數位典藏資料中快速且正確找尋出所需的內容，除要能為龐雜的數位典藏文物建立起完整的詮釋資料外，如何提供一個親切且便捷的資訊檢索環境已經變成了當務之急[2]。然而，目前大多數的數位典藏搜尋系統是以文字輸入的方式來搜尋數位典藏多媒體資料。同時，使用者僅能靜態地透過個人電腦等設備經由有線網路去找尋想要的資料。這樣的學習方式已經無法滿足使用者希望能在任何時候、任何地點、任何設備情況下查詢數位典藏內容的需求。

另一方面，眾所皆知的是語音是人與人之間長久以來最重要也最自然的溝通方式。隨著許多更輕薄短小的智慧型電子設備不斷地被發展出來，傳統以鍵盤為輸入的方式已不再方便。而且因為無線通訊和無線網路的普遍盛行，使用者不斷的要求在任何情況下都要能查詢資訊，這種渴望資訊隨手取得的心態日益強烈，語音也將會扮演更重要的角色，擔任起人們與各種不同智慧型電子設備間最主要的人機介面。因此，語音辨識技術的應用在未來勢必成為日常生活中不可或缺的一個環節[3]。

由於上述的觀察，我們發展了一個無線網路環境下以語音為基礎的多媒體資訊檢索介面，其中核心技術包括了自動關鍵詞抽取、大詞彙連續語音辨識、以及多尺度索引等[4]。我們將此多媒體資訊檢索介面與國立歷史博物館豐富的數位典藏內容作結合，研



圖一、大詞彙連續語音辨識。

發成一套以PDA(Personal Digital Assistant，簡稱PDA)為平台的多媒體資訊檢索雛形系統。未來，希望能夠提供博物館的參觀民眾一個自然且有效率的數位典藏文物學習環境。

二、 大詞彙連續語音辯識

大詞彙連續語音辯識(Large Vocabulary Continuous Speech Recognition, LVCSR)的流程如圖一所表示，主要包括前端處理(Front-end Processing)、聲學模型訓練(Acoustic Model Training)、複合詞抽取(Compound Word Extraction)、語言模型訓練(Language Model Training)和詞彙樹複製搜尋(Tree-Copy Search)等部分，各部分核心技術將於以下各小節中詳細介紹說明。

2.1 前端處理

前端處理包括PDA環境下語音的錄製以及聲學特徵參數的萃取。首先，為了能準確地在PDA環境下錄製使用者的語音片段以及去除其它不相關雜音片段，我們發展以能量為基礎的端點偵測技術。在實際的PDA使用狀況下，我們觀察到PDA本身在開啟麥克風錄音時產生的脈衝雜訊主要為訊號低頻成分，因此我們先將輸入的語音經過在頻譜上的預強處理(Preemphasis，為一高頻濾波器)用以降低分佈在低頻的PDA雜訊成分[5]，同時強調分佈在高頻的語音成分。然後我們根據先前動態取得的環境靜音(Silence)統計量，分別設定了語音前後端點的門閥值，藉以擷取出語音片段供作聲學特徵參數萃取使用。其次，聲學特徵參數萃取主要是在頻譜上作分析，求取出可作為辨識依據的重要頻譜特徵向量，而在本論文中我們使用了常用的梅爾倒頻譜(Mel-frequency Cepstrum)特徵(簡稱MFCC特徵)。在求取MFCC特徵時，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，為了降低不同PDA麥克風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)嘗試在梅爾倒頻譜上移除PDA麥克風通道效應的統計量。

2.2 聲學模型訓練

本論文中所使用的訓練語音語料是經由宏碁(Acer 公司)所生產的兩款PDA裝置(N10與N20)錄製，錄製期間從92年7月至93年5月，共8.5小時，約有成年男女各二十人參與錄製。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。我們最後總共使用了151個隱藏式馬可夫模型(Hidden Markov Models)來作為這些INITIAL-FINAL聲學模型的統計模型。在隱藏式馬可夫模型中，每個狀態則會依據其對應到之訓練語料的多寡，用2到128個高斯統計分佈來表示，不管男女性別都是使用同一套聲學模型[4]。

2.3 詞典和語言模型

詞典(Lexicon)與語言模型(Language Models)在大詞彙連續語音辨識中扮演了極為重要的角色，詞典決定了可以輸出的詞彙，語言模型則用來限制語句中詞彙與詞彙間相互接連的可能性。由於國立歷史博物館的數位典藏詮釋資料中經常有文物相關的專有名詞，如器物名、地名、年代名、人名等，並沒有涵蓋在一般的語音辨識詞典中，在語言模型訓練階段的文字語料處理時就會被斷詞成一連串單字詞，使得外詞彙(OOV, Out of Vocabulary)的問題常常發生，進而導致語音辨識率的降低。為了解決外詞彙問題，我們使用了歷史博物館提供的文物文字語料(約150M字)並輔以中央社的新聞文字語料(約70M字)，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$P_f(w_j|w_i)$，與後雙連(Backward Bigram)機率$P_b(w_i|w_j)$，並以前後

龍窯、
青銅劍、
鎮墓獸、
龍紋貴妃榻、
紅木靈芝如意、
長沙窯龜形玩具、
長沙窯黃釉褐彩貼花

圖二、部分新產生的複合詞。

雙連(Forward and Backward Bigrams)的機率幾何平均大小$FB(w_i, w_j)$當作複合詞選擇的標準：

$$FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}. \quad (1)$$

文字語料先經過一個使用含有1至4字詞約六萬八千個詞的詞典的斷詞程序處理過，然後利用上述的公式，經數次的疊代以及不同的基準閥值設定，產生了約一萬六千個2至10字詞的複合詞，使得最後的語音辨識詞典約含有八萬三千個1至10字詞。圖二顯示部分新產生的複合詞。

另外，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)，並以上述的文字語料為語言模型訓練資料。在本論文中的語言模型使用了 Kneser-Ney 模型平滑技術，在訓練時是採用 SRL 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[6]。

2.4 詞彙樹複製搜尋

我們所發展的大詞彙連續語音辨識系統是採用由左至右(Left-to-right)、音框同步(Frame-synchronous)的詞彙樹搜尋方式

[7]。在詞彙樹中每個 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 arcs 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)若擁有相同的語言模型歷史則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)。當在每個音框，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的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為標記，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，僅要建立一棵詞彙樹作為搜尋時路徑展開參考之用即可，而分別要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，因此必需以光束剪裁(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的語言模型機率，並以其中最大者當為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) [7]，再加上內部狀態節點本身搜尋時所累積的解碼分數(Decoding Socre)當成剪裁比較的依據。在本研究中，我們採用的是詞單連(Word Unigram) 語言模型前看，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，做為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片段)，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)。並且在這詞圖上使用更高階的語言模型，如詞三連、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，而在詞圖搜尋階段是使用詞三連語言模型。

三、 資訊檢索

在本節將介紹我們所發展的多尺度索引(Multi-scale Indexing)技術及以向量空間(Vector Space Model)檢索模型。

3.1 多尺度索引技術

我們首先為數位典藏文物的詮釋資料建立起索引，以供資訊檢索使用。由於音節(Syllable)是中文非常重要的組成成分，每個中文詞都是由一到數個音節所構成。而且，在過去的研究中也發現以音節層次的索引特徵在中文語音資訊檢索上有很不錯的效果[8]。因此，在本論文中我們同時採用了詞層次及音節層次的索引來表示詮釋資料。每一個詞句(不管是語音輸入或是詮釋資料)在這裡都會先被轉成對應的音節串。對於以詞為索引的方式，每一個單詞(Single Word)會被當成一個索引單位；對於以音節為索引的方式，每一個單音節(Single Syllable)以及任兩個相鄰的音節對(Syllable Pair)都會分別被當成一個索引單位。

3.2 資訊檢索模型

在本論文中，我們以向量空間模型作為資訊檢索模型。每篇文章 d 便可以被一組的特徵向量表示式所代表，其中每一個向量 d_j 表示式存放某一類索引特徵，如單詞、單音節、音節對。在每一個向量表示式中的每個元素 z_{jt} 則是代表者某個特定索引 t 在文章 d 的加權統計量，如下式：

$$z_{jt} = (1 + \ln(c_t)) \cdot \ln(N/N_t) \quad (2)$$

其中 c_t 是索引 t 在文章中出現的次數，$1+\ln(c_t)$ 則是進一步代表 c_t 在文章中的頻率，其中對數運算是為了壓縮頻率的分佈範圍。另一方面，$\ln(N/N_t)$ 是所謂的的反文件頻率(Inverse Document Frequency，簡IDF)，N 是所有文章數，N_t 是索引 t 出現的文章數。當索引出現的較多的文章時，它的重要性會將低，反之亦然。使用者的查詢輸入 q 也同樣是利用上述的方法表示。對於每一類索引特徵向量，我們可以利用查詢向量與文章向量間的餘弦值估測(Cosine Measure)來代表查詢與文章間的相關度分數：

$$R_j(\vec{q}_j, \vec{d}_j) = (\vec{q}_j \bullet \vec{d}_j) / (\|\vec{q}_j\| \cdot \|\vec{d}_j\|) \quad (3)$$

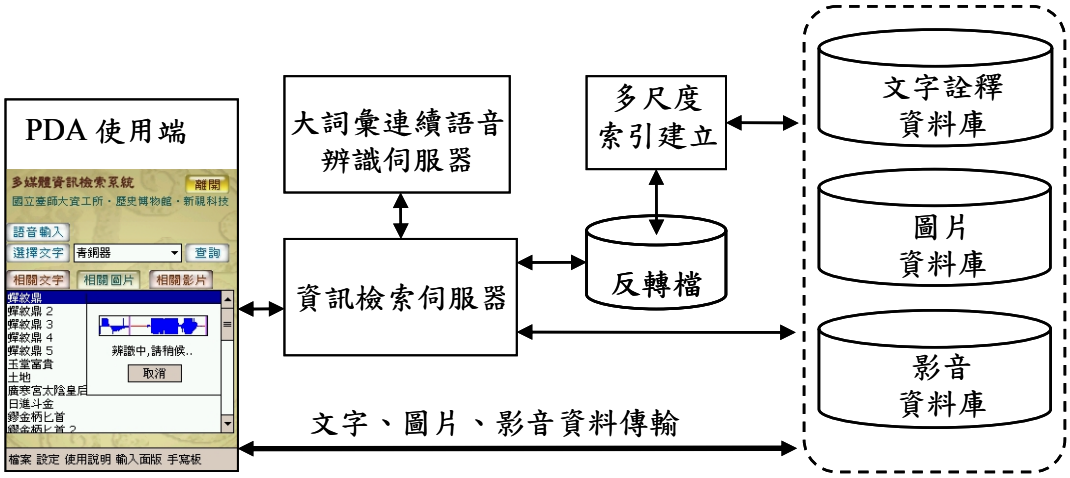
最後查詢與文章的總相關度分數，可由各類索引(單詞、單音節、音節對)所計算出的相關度分數作加權而得：

$$R(Q,D) = \sum_j w_j \cdot R_j(\vec{q}_j, \vec{d}_j), \quad (4)$$

其中 w_j 是不同類索引的加權值。

四、 多媒體資訊檢索系統

目前我們已經結合國立歷史博物館豐富的數位典藏文物，研發成一套以PDA為平台的多媒體資訊檢索雛形系統，希望提供使用者一個自然且有效率的檢索與瀏覽環境，圖三為多媒體資訊檢索系統的示意圖。它主要包括了大詞彙連續語音辨識、多尺度索引建立與資訊檢索等三部分。使用者可以透過PDA在無線網路環境下以語音輸入、文字輸入或樹狀點選(如圖四之(d)所示)的方式來檢索歷史博物館的數位典藏內容。當使用者以語音輸入檢索時，所講的語句先經過無線網路傳送到資訊檢索伺服器，再轉傳到



圖三、多媒體資訊檢索系統。

大詞彙連續語音辨識伺服器，產生代表語音辨識結果的詞串。然後資訊檢索伺服器以辨識結果為查詢去比對並取得存放在多媒體資料庫內的相關數位典藏文物電子檔，把檢索結果傳回PDA，最後在PDA系統介面顯示相關數位典藏文物的標題資訊。數位典藏文物也將依類型而做不同的呈現方式，如「相關文章」類的文章內容閱讀(如圖四之(a))、「相關圖片」類的圖片賞析與文章介紹(如圖四之(b))和「相關影片」類的影片欣賞(如圖四之(c))之數位典藏文物顯示。

五、 結論與未來展望

我們目前正在執行的國科會九十二年「適用於數位典藏多媒體內容之自動分類索引與複合式多媒體檢索技術」的數位典藏計畫裡[9]，已經成功地將大詞彙連續語音辨識和多尺度索引等核心技術整合於PDA平台的多媒體資訊檢索雛形系統中，將來的使用者不再只是使用個人電腦的數位學習方式，更可以經由PDA語音輸入方式，迅速、準確地查詢數位典藏文物，加強了未來行動學習的可能性，目前系統已正式於九十三年四月在國立歷史博物館提供來館參觀民眾測試使用。而展望今年(九十三年度)的「無線網路環境下複合式數位典藏文物導覽及電子商務系統之研發」數位典藏計畫，我們除了將持續發展適用於數位典藏內容之複合式多媒體檢索技術外，也將針對典藏單位所亟需的其它技術，如室內定位、多媒體無線傳輸與壓縮技術、電子商務等，作更進一步的研究發展。而在使用者介面研究上部份，將會發展口語對話機制，讓多媒體資訊檢索系統在未來能更具人性化。

六、 致謝

感謝本研究部分經費由國科會九十二年度數位典藏計畫NSC-92-2422-H-003-0370及九十三年度NSC-93-2422-H-003-003-計畫補助，並感謝歷史博物館[10]與新視科技股份有限公司[11]提供典藏文物資料與技術支援。

(a) 相關文章　(b)相關圖片　(c)相關影片　(d)樹狀點選式

圖四、系統展示畫面。

七、 參考文獻

[1] 鐘子帆、何建明,“數位典藏系統之多媒體檔案管理與呈現”,第一屆數位典藏技術研討會,July.2002

[2] 徐典裕,“以服務及應用為導向之博物館數位知識庫建構模式”,第一屆數位典藏技術研討會,July.2002

[3] B.H. Juang, S. Furui, *Automatic Recognition and Understanding of Spoken Language—A First Step Toward Natural Human–Machine Communication*, Proceedings of the IEEE, vol. 88, NO. 8, August 2000.

[4] B. Chen, J.W. Kuo, W.H. Tsai, *Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription*, IEEE Int. Conf. Acoustics, Speech, Signal processing, May 2004.

[5] X. Huang, A. Acero, H. Hon, *Spoken Language Processing*, Prentice Hall, 2001

[6] A. Stolcke, *SRI language Modeling Toolkit*, version 1.3.3, http://www.speech.sri.com/projects/srilm.

[7] X. L. Aubert, *An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,* Computer Speech and Language 16 , 2002.

[8] 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, Vol. 10, No. 5, July 2002.

[9] 國立台灣師範大學資訊工程研究所「無線網路環境下複合式數位典藏文物導覽及電子商務系統之研發」計畫網站：http://prjda.csie.ntnu.edu.tw。

[10] 國立歷史博物館：http://www.nmh.gov.tw。

[11] 新視科技：http://www.visionnext.com。