臺語多聲調音節合成單元資料庫暨文字轉語音雛形系統之發展

佘永吉、鍾高基、*吳宗憲

國立成功大學醫學工程研究所、*資訊工程研究所

Establish Taiwanese 7-Tones Syllable-based Synthesis Units Database for the Prototype Development of Text-To-Speech System

Yung-Ji Sher, Kao-Chi Chung, *Chung-Hsien Wu

*Institute of Biomedical Engineering, *Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan*

摘　要

臺語為具有入聲與鼻音等特性的多聲調語言，故建立聲調組合的音節資料庫，為計算語言學研究的基礎。本研究之文字轉語音雛形系統，係由漢字或臺語現代文輸入、單音字詞資料庫、字轉音處理模組、合成單元音檔資料庫及臺語語音輸出處理模組等子系統所構成。2557 個音節合成單元資料庫的建立，係參考漢字與羅馬拼音文獻及字典。合成單元音檔採用載字句錄音，合成器使用基週同步疊加法。系統效能評估則以臺語現代文語詞輸入，聽寫測驗十位受測者，紀錄可辨度與自然度。結果顯示可辨度平均為 83.5%，MOS 自然度量表平均為 3.188。

關鍵詞：臺語、合成單元資料庫、音韻學與數位訊號處理、語音合成

緒　論

閩南語或稱廈門話，源起於西元三世紀以前，隋唐時代中原的口語， 隨著族群陸續遷徙到閩南；十六世紀起，再隨著移民逐漸散佈到臺灣暨東南亞各國，如泰國、新加坡、菲律賓、馬來西亞、印尼等[1]。東南亞地區約有四千九百萬人的母語是閩南語，人口數在世界語族排名第二十一位，比某些歐洲語言如義大利語及荷蘭語還多[2]。

臺灣是多語言的社會，現有語言源自於原住民語、閩南語、 客語及北京語。依據學者之研究，語言依語系及起源地區形成不同特性，東方語系屬聲調語言(tone languages) ，原住民語屬南島語系，閩南語與客語屬漢藏(Sinitic)語系，北京語屬阿爾泰(Altaic)語系[1, 3]。在日常生活中，各族群經常使用自己的母語來交談，週邊國家的語言如英語及日語也常出現於某些特定的場合。本研究所稱的臺語，係指占母語人口數四分之三強的閩南語。閩南語(Amoy Chinese)源起於福建省南方之漳、泉州、及廈門，早期隨移民遷徙至臺灣後分別形成不同的腔調，如府城(今臺南)、鹿港(偏泉州)、艋舺(今臺北萬華)及宜蘭(偏漳州)腔等。因口音交流、歷史經驗、地理背景、族群融合與外來文化，臺語各不同腔調的口音逐漸融合，且形成有別於原鄉閩南之獨特語音(phonology)、語彙(morphology)、語意(semantics)及語法(syntax)[4]。臺語具有多聲調、入聲、詞前轉調、多鼻音(nasals)而缺乏捲舌音等音韻學上之特色，直接影響計算語言學之語音研究方法與實驗設計，例如基本合成單元(basic synthesis units)之選取、字轉音規則庫之建立、語音特徵參數(parametric extraction)之選擇與量化分析[5]、語音合成法則(synthesis algorithm)之決定等。目前臺語相關之計算語言學研究相當缺乏，此領域的語音資訊實屬研究開發的處女地。

臺灣語音溝通科技之研發肇始於 1980 年代的中文語音合成與辨識，現有研究成果應用於人性化溝通之電腦人機介面、電話語音系統等，均專注於北京語(Mandarin)系統的研發。現階段的臺語研究，較專注於臺語語言學特性的探討[6, 7, 8, 9]，而臺語計算語言學研究仍屬萌芽階段[10]。近年來臺語計算語言學論文發表的研究群，包括長庚大學呂仁園與清華大學江永進教授的大量詞彙辨識系統[2, 11, 12]、臺灣大學陳信希教授[13]、成功大學王駿發、吳宗憲、鍾高基教授的文字轉語音系統[14, 15]等。臺語計算語言學研究所遭遇之困難，主要是缺乏高效率的文字系統及參考文獻。臺語為典型的口語式語言(oral language)，尚無統一之文字，對於人文科技之發展與本土文化之傳承，深具負面影響。目前研究所採用的文字系統，多沿用以北京語發音、語彙及語法為主的漢字。漢字字集龐大複雜且總字數無法確定，不僅文字音、形分離，發音及轉調對應規則複雜且特殊情況甚多；至於採用罕見漢字之臺語漢字系統，在目前普遍使用的中文電腦內碼系統中並不完整，導致甚難提供科學化語音合成、辨識與理解系統之研發[16, 17]。

臺灣目前大力提倡母語教育，但首先需要針對書寫的文字作科學化的分析與研究，以發展有效率的母語教材。目前使用的臺語文字根據互異之基本音素(phonemes)單元，可分成三大類，包括：漢字(Chinese-substituted)、拼音字(Spelling)與音漢混合字。漢字系統之使用，最早追溯於 1566 年的”荔鏡記”文獻，教會羅馬字(Romanization，以下簡稱教羅)，源於 1832 年西方傳教士所建的”廈門白話字”， 是臺語拼音字的雛形。以“東”字為例，漢字”東”的臺語標音方式為”德紅切”，即採用前字的聲母與後字的韻母和聲調，來表示該字的發音。而教羅僅需標示”tang”，音形意合一[1, 18]。因此，臺語漢字音、形分離，缺乏教育及訓練的效率性。閩南語十五個子音最早記載於西元 1800 年黃謙的”彙音妙悟” 字典，教羅則以六個最基本的單母音 a, e, i, o•, o, u 及其組合來標示臺語的母音。傳統閩南語字典使用漢字標音，難以準確標示讀音，無法符合資訊化處理的原則。教羅針對臺語語言的特色，應用特殊符號或數字來標示聲調，以字尾變化配合特殊符號來區分入聲。其缺點乃未善用拼音字母，且屬於表音式音標而非拼音文字。

臺語為多聲調的語言，一般將臺語分成七或八個聲調[17]，不同聲調的音節即具有不同的語意，聲調記號的標示是臺語拼音文字化的重要課題。臺語的特色是具有許多的鼻音，以子音”k”與母音”oa”所組合的音節”koa”為例：前鼻音與後鼻音分別代表”汗”及”縣”，亦即相同子音與母音組合的音節，會因鼻音在前與鼻音在後之不同，而代表不同的意念。使用漢字無法直接表達讀音，而教羅則以”n”的上標與否表達鼻音位置之不同，例如”koa^{n}”及”koan”分別代表” 汗”及” 縣”，相同的字母組合無法明確的表達鼻音的差異[18]。臺語最大特色在於複雜的轉調規則，例如單音節位於單字詞與詞末時，發本調音；若位於多字詞的前幾個字時，則發轉調音，稱為詞前轉調的特性。若以漢字紀錄臺語，必須正確斷詞後才能轉調，斷詞是漢字文章的一大問題，因此漢字完全無法表達轉調的特色。且漢字的臺語讀音常為一字多音，難以取捨讀音，例如”香港的香很香”。教羅為音節(syllable)單元的記音式音標(phonemic transcription)，一律紀錄本調音，無法明確表達詞前轉調的特性，難以記憶背誦與教育訓練。

林繼雄教授於 1943 年提出以語詞(words)為單元之音、形、意三合一文字概念，稱為臺語現代文(Modern Literal Taiwanese, MLT)[19]。臺語現代文承襲教羅，

逐一改進缺點，共使用 27 個字母及其組合來記錄臺語，包括英文的 26 個字母，及特殊字母’ò’。首先將易混淆的母音’o•’及’o’，標示成’o’及’o̲’(或’ò’)；子音’c’依所接母音之不同分成’c’及’z’；前鼻音化音節則以’v’為標記的字母，改接於子音字母之後；選取教羅不常用或不使用之字母為聲調記號，透過不同的字母代表臺語的八個聲調，分別是高調音(raised tone)、上突音(pushed out tone)、下突音(depressed tone)、低促音(low stop tone)、迴旋音(bend tone)、迴升音(bend-up tone)、基調音(fundamental tone)及高促音(high stop tone)；再以字尾之不同字母變化標示入聲；以語詞為單位，語詞拼字乃直接拼寫轉調後的拼字組合等。臺語現代文符合臺語現代化與國際化的需求，期許經由減少方言腔調爭議之妥協拼字，改善現有臺文表達困難與不確定性(ambiguous)等缺點，進而達到臺語文字化、資訊化發展之終極目的。

根據語言學及音韻學原理[20]，所有語言的語音，均可由有限之單元(units)所組成。語音合成系統的合成單元包括：音素(phonemes)、單音節(syllables)、雙音(diphones)、三音(triphones)、多音(polyphones)，次音節(semi-syllables or demi-syllables)與不定長度(non-uniform)等，應依需求選取適當的合成單元。西方語系屬句調語言(intonation language)，音節僅指子音與母音的組合，音節內僅區分重音與次重音等。

近年來文字轉語音系統，多使用連續語音資料庫來建立合成單元，其流程包括基週偵測及平滑化、語音單元的平滑化、頻譜參數的求取、合成單元挑選、以及人工檢驗。基週偵測及平滑化指對於語音資料庫中的每一個語音單元，使用聲母/韻母分割的演算法以及自相關法來自動地求取基週。基週軌跡的量化是以離散的雷建德多項式[5]，轉換成四維的基週向量(a_0, a_1, a_2, a_3)來表示基週軌跡，並將基週軌跡作平滑化的處理。由於某些語音單元的發音太短促、音高太高或太低等，故必須過濾這種語音單元，稱為語音單元的平滑化。相鄰合成單元間頻譜軌跡的連接越平滑，產生的語音會比較平順悅耳，因此在語音合成時，以較接近共振頻率(Formant Frequency)的線頻譜對頻率(Line Spectrum Pair Frequency，簡稱LSF)來作為頻譜參數[21]，選取最小的音節間頻譜失真的單元。合成單元挑選則使用兩種失真度的量測來決定每一種發音的基本合成單元：第一種為「音節失真

度」，乃對於同一種發音的語音單元，計算分析其間的距離；第二種是「音節間失真度」，用來計算分析不同發音單元之間的頻譜失真度。目前合成單元之篩選，係先由電腦自動化輔助過程，再經由人工檢查，剔除音質不好的合成單元，最後挑選一個適當的合成單元。

研究目的及重要性

本研究之目的為應用語音溝通的數位訊號處理科技以及臺語現代文，探討與建立臺語語音合成單元基本資料庫並發展文字轉語音合成系統雛形。本論文的特定目標(Specific Aims)為：

1. 探討分析臺語音韻學中的音素、音節、聲調、相關拼音法則與轉調規則的特性，建立資訊化處理時最佳紀錄及分類方式，以發展臺語多聲調單音節(syllable-based) 合成單元基礎資料庫；

2. 依據語言學的原則與方法：選擇載字句，錄製多聲調載字句電子音檔，並利用數位訊號處理方法切割、評估與篩選，以建立合成單元語音資料庫；

3. 依據臺語聲調與轉調特性，發展字轉音處理模組，以建立漢字與臺語現代文轉語音合成系統之雛形。

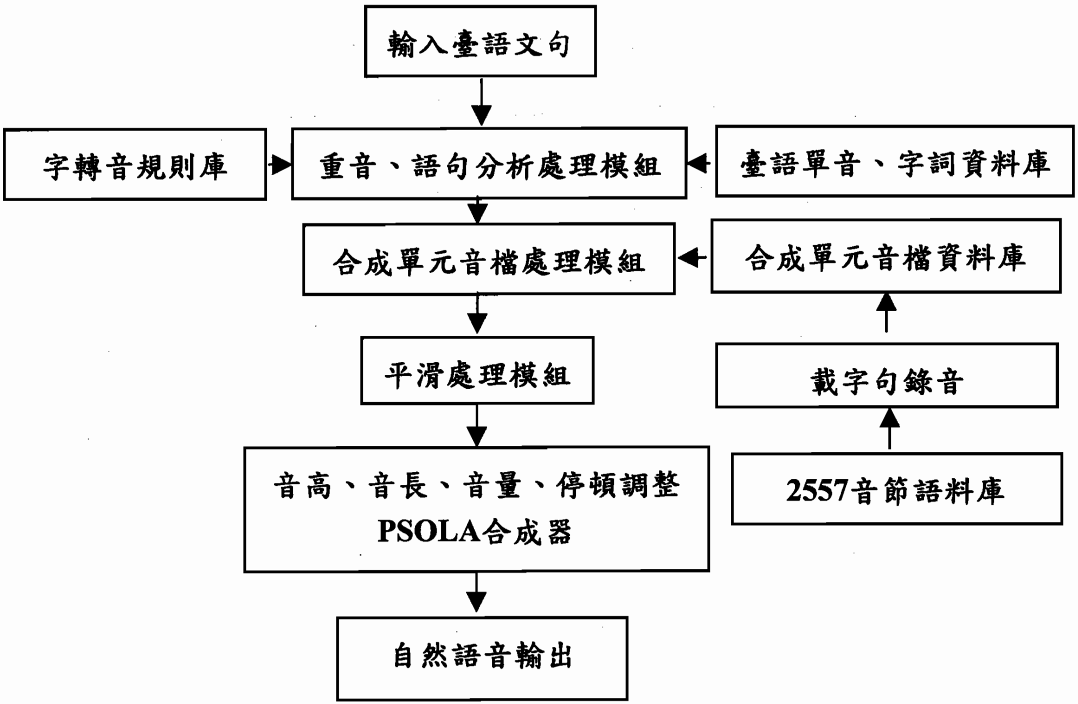
本研究的重要性乃針對臺語文字化、現代化、資訊化及國際化的前瞻性發展需求，提供臺語資訊化發展及計算語言學相關研發的基礎，同時廣泛的應用到一般成人與兒童早期的母語教育、電腦輔助語言教育與訓練、聽語障礙者特殊教育及語音溝通輔助復健科技之發展。

方法

本研究考量臺語的聲調特性與內含的音韻訊息，因此採用子音、母音與聲調組成的音節作為基本合成單元，並根據林繼雄教授的臺語現代文，將鼻音化的子音與不同聲調的母音，均歸類為相異之基本音素單元，以明確區分不同聲調的特性。故本系統的臺語基本音素包括 32 個子音、33 個單聲調母音與其八個聲調的變化組合[15]。

實驗設備/特徵參數萃取： 錄音系統為具方向性且抗背景雜音之電容式麥克風，架設在 Pentium PC 上，取樣頻率為 22.05 kHz，解析度為 16 bits。以短時間視窗(short-time windowing)之時域及頻域分析及萃取特徵參數，每一個視窗長度約為 40 ms，且至少包含七個基週。並依據能量及過零率曲線之分析，區分音節與語音段。

實驗方法與步驟： 本研究之整體架構如圖一所示，系統發展分成三個階段：第一階段為文獻探討及考量不同的音節單元組合，發展單音、字詞資料庫、字轉音規則庫與分析處理模組，並建立以單音節為基礎之臺語七聲調基本合成單元資料庫；第二階段進行載字句之錄製、切割、評估與篩選，建立七聲調音節合成單元電子音檔資料庫；第三階段為七聲調臺語語音合成系統之雛形研發及系統效能評估。具體的步驟與方法詳述如下：



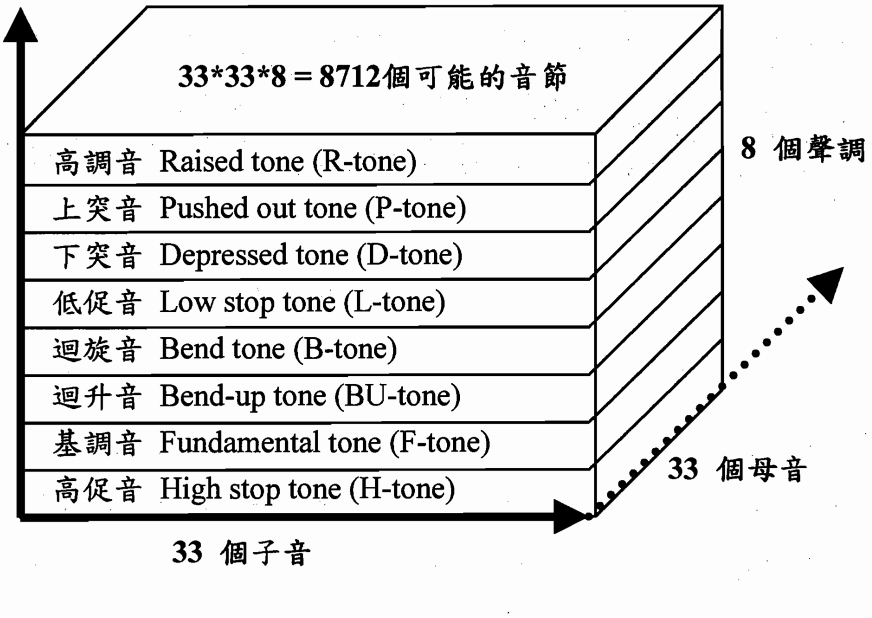
圖一：本研究之架構圖

1. 發展及建立多聲調音節合成單元資料庫：

1-1 首先建立臺語字典與辭典電子檔，並針對臺語的特性、轉調規則、文白讀音與不同腔調，發展與建立字轉音規則庫及分析處理模組。經由文獻探討歸納出子音音素共計 33 個，包括：1 個音節內只存在母音而無子音之情況、18 個一般子音及 14 個後接'v'的前鼻音化子音。其中 b, j, l, m, n 五個子音的鼻音不存在或本身即為鼻音，鼻音韻母則在母音字母前加上字母'v'。

1-2 本研究中的臺語母音音素，不特意區別元音、介音與韻尾，只將其組合成的單母音、雙母音及複母音，均視為互異之母音。故母音的總數為 264 個，包含 33 個單聲調母音及其八聲調之組合。雖然母音總數較多，然其主要的優點為不同聲調的音節可經由不同的拼字方式予以明確區別。

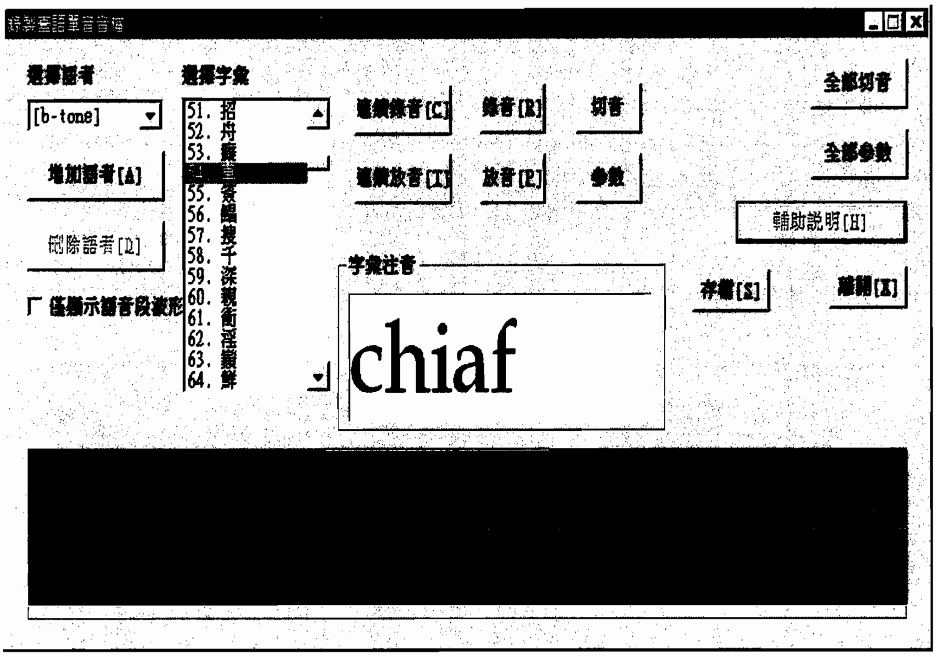
1-3 依據上述基本音素的訂定，共形成 8712 (33*33*8)個音節(如圖二)，組合成的音節不一定具有語意。根據拼音臺文漢字輸入法、臺語電子字典、辭典等相關文獻的探討，將 8712 個音節中，無對應漢字的音節剔除，以建立單音節基本合成單元資料庫[15, 22, 23]。


33*33*8 = 8712個可能的音節
高調音 Raised tone (R-tone)
上突音 Pushed out tone (P-tone)
下突音 Depressed tone (D-tone)
低促音 Low stop tone (L-tone)
迴旋音 Bend tone (B-tone)
迴升音 Bend-up tone (BU-tone)
基調音 Fundamental tone (F-tone)
高促音 High stop tone (H-tone)
8 個聲調
33 個母音
33 個子音

圖二：臺語單音節的組合

2. 建立多聲調合成單元電子音檔語音資料庫

首先錄製上述多聲調音節的載字句。載字句係指一段連續的多音節語句，涵蓋欲錄製的音節，卻未必具有特定語意。載字句選擇的標準，是以該載字句前後單字與所欲錄製語音音節的界線是否清晰為原則。本研究所選用的載字句為”Goafthak ‘S’ ciuxhor” (我讀’S’就好)，’S’表示欲錄製的語音音節，置於載字句之間，以控制音高、音長與音量之變異，並保留較佳的音韻訊息。錄音程式可以同時顯示所欲錄製的單音節漢字與所對應之臺語現代文(如圖三)。錄製載字句連續語音過程中，每秒速度約為三個單音節，音量大於 65dB，音高、音量、速度、及音長儘量保持一致。載字句的錄音錄製過程非一次完成，因此每次錄音時均播放先前錄音之音檔，以維持音高的一致性。



圖三：錄音程式之視窗介面

載字句電子音檔錄製完成後，重新播放以評估其音高、音長、音質、音量與可辨度；根據波形的時域觀察，並經由能量曲線(energy contour)及過零率曲線(zero-crossing-rate contour)分析之輔助篩選，判斷音節界線並切割所需之音段，切割後每個單音節語音音段約為 300 毫秒。若可辨度不佳，或音高、音量、速度及音長相差太大，則重新錄製與分析。

3. 臺語多聲調語音合成系統之雛形研發及系統效能評估

本系統包括文字輸入、單音字詞資料庫、字轉音規則庫與分析處理模組、合成單元音檔處理模組、及語音合成輸出等子系統。系統提供臺語現代文或漢字輸入之功能：若以臺文拼字輸入，則由拼字檢查(spelling check)程式檢查拼字；若為漢字輸入時，則經由漢字轉臺文之詞翻譯系統轉換成臺文拼字，再以字轉音程式規則庫、語句分析處理模組等，調整詞調及句調，對應到合適的合成單元電子音檔，經由平滑化處理及PSOLA合成器，最後輸出合成單元資料庫中經過調整處理的自然音節語音。

3-1 漢字轉臺文之詞翻譯系統： 現有電腦大五碼(big-5)中的所有漢字，均已鍵入於本系統所建立的臺語單音資料庫中，亦即所有現存於電腦系統的通行漢字，均可找到對應之臺語拼字。故當本系統以漢字輸入時，首先經由電子辭典進行斷詞(segmentation)，再根據臺語語詞的漢字及拼音文字對照資料庫，利用語詞為單位轉換到對應之臺語現代文拼字；無法在辭典中找到的語詞均視為單字詞，並以臺文漢字字典逐字轉換成臺文拼字。

3-2 字轉音規則庫： 直接輸入或經詞翻譯後的臺語現代文拼音字，均具有八聲調變化，但是當臺語連續語音整句合成輸出時，有豐富且複雜的轉調規則。基於臺語有文、白讀音、與各地腔調等不同的歧異，例如臺南(偏漳州)與臺北(偏泉州)的迴旋音與入聲，分別適用不同的轉調規則，且漢字的臺語讀音常有一字多音的現象，本研究嘗試根據學者的歸納與整理，發展字轉音的規則程式，以配合字、辭典的音標選取，減少錯誤讀音的發生率。

3-3 重音、語句分析處理模組： 在整句臺語的連續讀音時，必須考量加重音、輕讀或綴詞等之讀音轉調，故應分析處理以提高合成語音的品質。經過語句的分析處理，可以作為詞調與句調調整的依據，以提供合成單元音檔處理模組、音韻調整模組、與平滑處理模組調整參數的基本資料。

3-4 其他處理模組及PSOLA合成器： 本研究目前仍持續收集臺語語料之電子檔，以提供PSOLA合成器參數調整規則的歸納或參數調整訓練的參考。

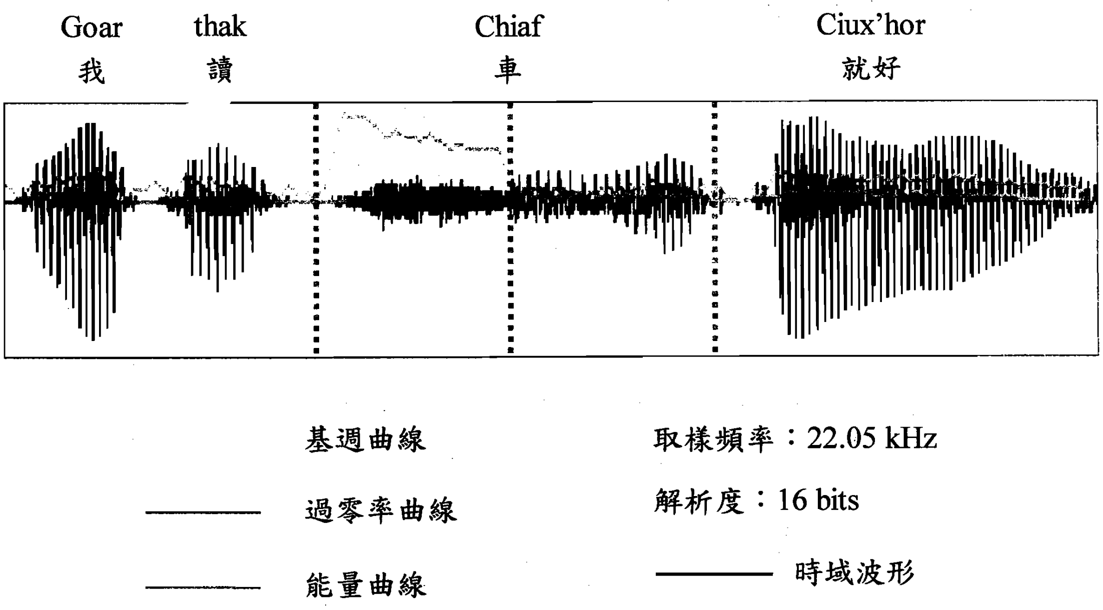
3-5 本系統之表現評估： 包括可辨度(intelligibility)與自然度(naturalness)，受測者共計十人，分成一般人士組五人，實驗室研究生五人。可辨度之評估，係選取日常生活最常使用的20個二字詞、20個三字詞、20個四字詞成語、與臺語教材文句10句。其中二字詞與三字詞多為白話音，四字詞則為文言音，文句則屬生活對話(如附錄一)，不同子音開頭之語詞均加以選取，由系統合成臺語語音，受測者以聽寫方式寫下所聽到的語詞，計算受測者書寫語詞之正確率，作為可辨度的評估結果。

自然度則採用平均鑑定分數(mean opinion scores, MOS)，區分為五級，即優良(excellent)、良好(good)、尚可(fair)、差(poor)、極差(unsatisfactory)，由受測者在聆聽系統合成語音，分別依所感覺的自然度給予5至1分的評分。

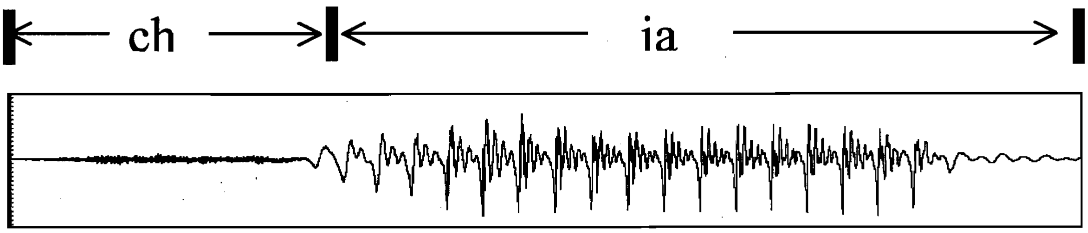
結　　果

建立2557個七聲調的音節資料庫 本研究建立的2557個七聲調音節資料庫，包括：394個基調音的音節(如附錄二)、418個高調音的音節、426個上突音的音節、394個下突音的音節、292個低促音的音節、387個迴旋音的音節單元、及246個高促音的音節單元。傳統臺文僅有七個聲調，臺語現代文將迴升音視為一個獨立的聲調，然文獻中甚難確定迴升音的音節總數，此情況頗類似國語的輕聲。本研究僅建立傳統七個聲調的音節資料庫，至於臺語的聲調區分，有待進一步的探討與量化。

建立七聲調合成單元語音資料庫 對應上述之音節資料庫，共錄製2557個載字句的電子音檔，錄音時的速度、音量、音高及音長，以及訊號處理時擷取的長度(length of window)，均採用中文系統目前研發的經驗參數值。此未必符合臺語之語音特性，故將來更需要進一步的探討。以臺語的" chiaf'(車)為例，載字句"Goarthak chiaf ciuxhor"(我讀車就好)的時域波形及過零率、能量、基週曲線如圖四所示。自左端算起第三個較大的波形為本實驗所切割的音節合成單元"chiaf'(車)，其前後音節"讀"(~thak)和"就"(ciux~)的母、子音和"chiaf'形成可區分的連音波形，音節界線清晰而易於切割。圖五為"chiaf'(車) 切割後的時域波形，前段振幅較小為子音'ch'，後段類週期訊號為母音'ia'，'f'為高調音標示記號。

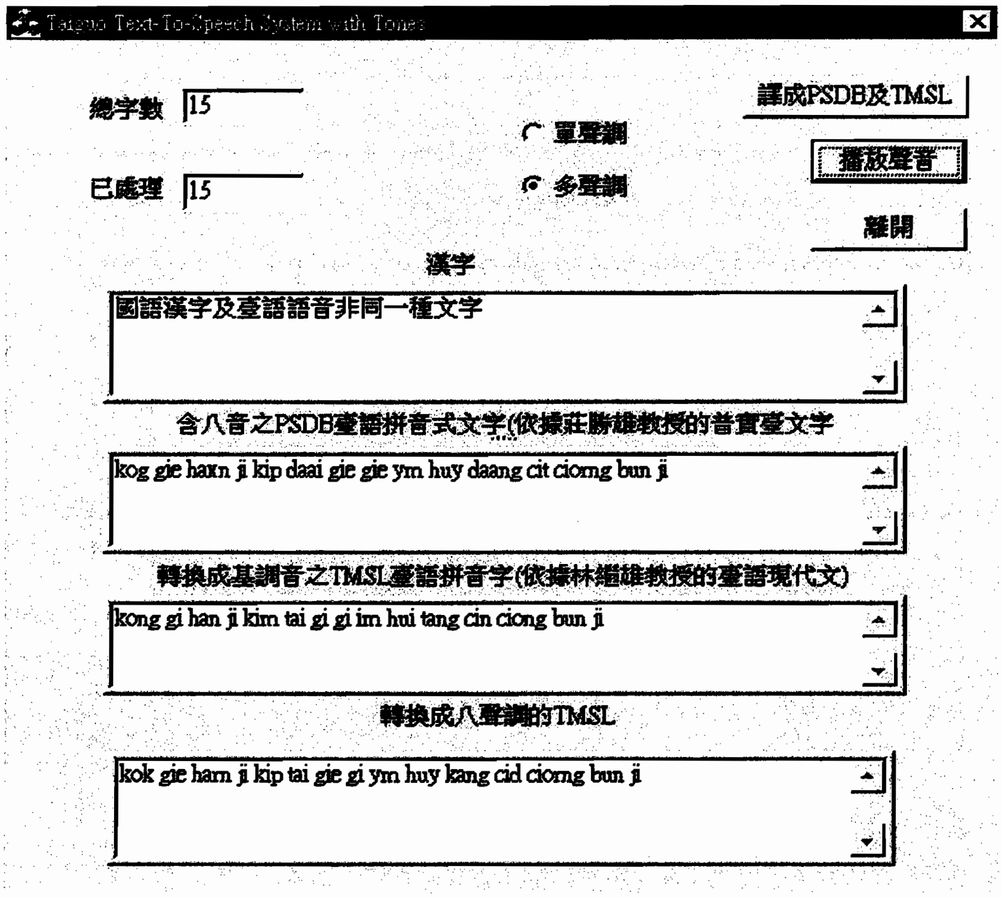


圖四：載字句 "Goafthak chiaf ciuxho̱r." (我讀車就好) 語音訊號之特徵



圖五：合成單元"chiaf"(車)的時域波形

<u>七聲調語音合成之雛形系統</u>　　雛形系統之程式係在中英文電腦上編譯，為程式編寫方便，必須英文字母化，故本研究將臺語母音'ò'以'o̱'標示。臺語並無統一之文字，文獻中記載的不同文字系統均具參考價值，因此本研究針對最為廣泛使用之漢字、教會羅馬字、臺語現代文與普實臺文四種文字系統，建立文字及音標方案之音素對照表，提供不同臺語文字系統的轉換。教會羅馬拼音及普實臺文文字系統皆經由上述音素對照表自動轉換譯成臺語現代文，而漢字系統則必須經由詞翻譯系統轉換成臺語現代文，以合成語音輸出，如圖六所示。普實臺文文字系統目前具有電子字辭典、拼字檢查程式、及臺文輸入轉漢字之功能。



總字數 15

已處理 15

○ 單聲調
● 多聲調

漢字

國語漢字及臺語語音非同一種文字

含八音之PSDB臺語拼音式文字(依據莊勝雄教授的普實臺文字

kog gie haxn ji kip daai gie gie ym huy daang cit ciomg bun ji

轉換成基調音之TMSL臺語拼音字(依據林繼雄教授的臺語現代文)

kong gi han ji kim tai gi gi im hui tang cin ciong bun ji

轉換成八聲調的TMSL

kok gie ham ji kip tai gie gi ym huy kang cid ciomg bun ji

圖六：臺語文字轉語音系統雛形

本研究目前所發展的臺語現代文單字詞資料庫包括 20,000 個漢字與拼字之對照表，及多字詞資料庫包括 40,000 個語詞漢字與拼字之對照表(表一、二)；且針對上述單字詞及多字詞資料庫所對應的臺語文言音、白話音、漳州音、泉州音等逐步增訂音標及建立字轉音規則。拼音檢查程式乃根據上述發展之資料庫，檢查比對輸入之臺文拼字是否正確。為配合臺灣官方漢字使用之現況，發展漢字轉臺文的詞翻譯系統，以提供使用漢字作為文字輸入的使用者，仍可透過系統輸入漢字以合成臺語語音(如圖六)。本雛形系統包括：漢字轉臺文的詞翻譯系統、字轉音規則庫與分析處理模組、及語音處理模組等。

本系統可辨度測試結果如表三、表四所示；自然度測試結果如表五、表六所示。

表一：臺語單字詞資料庫

編號	漢字	臺文
6194	棠	toong
6200	棟	toxng
6231	棉	mii
……	…….	…….

表二：臺語多字詞資料庫

編號	漢字	臺文	音節
2355	今日	kinafjit	kin'af'jit
3967	螢火蟲	Huefkimkof	huef'kim'kof
…..	……	……	……

表三：可辨度實驗結果

	測試詞句	數量	可辨度
一般人士組(5 人)	二字詞	20	88.1%
	三字詞	20	83.3%
	四字詞成語	20	79.6%
	文句	10	72.2%
	平均		80.8%

表四：可辨度實驗結果

	測試詞句	數量	可辨度
實驗室研究生(5 人)	二字詞	20	84.5%
	三字詞	20	88.3%
	四字詞成語	20	92.0%
	文句	10	81.9%
	平均		86.68%

表五：自然度實驗結果

	測試詞句	數量	自然度(MOS)
一般人士組(5 人)	二字詞	20	3.6
	三字詞	20	3.5
	四字詞成語	20	3
	文句	10	2.5
	平均		3.150

表六：自然度實驗結果

	測試詞句	數量	可辨度
實驗室研究生(5 人)	二字詞	20	3.3
	三字詞	20	3.5
	四字詞成語	20	3.3
	文句	10	2.8
	平均		3.225

討　　論

此2557個音節資料庫係經由剔除不具語意之合成單元及篩選出具有對應漢字之音節，有效減少合成單元資料庫的數量，而不影響整體的完整性。

本系統目前輸出的臺語語音，係經由載字句切割出音節單元所合成。由於載字句未必具有語意，也不一定符合文法規則及語用情境，故所錄製的合成連續語音，其音韻品質較差。二字詞、三字詞與四字詞之可辨度均優於整句文句，係因臺語除詞末音節發本調音外，其餘音節皆必須依規則轉調，稱為詞前轉調。本系統採用之臺語現代文以語詞單位拼寫，其語詞拼字乃直接拼寫詞前音節的轉調音，及詞末音節的本調音，故系統不必處理詞前轉調的問題。但是整句文句除詞調之轉調規則外，仍必須考慮句調之轉調，目前之系統無法有效調整句調，必須進一步錄製平衡句以萃取句調參數。臺灣目前相當缺乏臺語連續語音之音檔處理資料，先天性的限制有關臺語合成法則及合成器參數之研究。因此建立平衡詞句資料庫，實為發展高品質臺語語音合成系統之當前最重要的課題。

臺灣教育部頒訂常用的繁體標準漢字約4,800餘字，均可對應到本系統的電子音檔，並透過語音合成系統輸出臺語語音。臺文中某些常使用的罕見漢字，在現行電腦的通行內碼中並不完整。長期受到百越族方言、日語、英語等之影響，現代臺語中有些常使用的語音沒有對應的漢字[19]，因此必須經由日常生活中之對話、臺語廣播與電視新聞的語音資料、及拼音式文字的文獻中持續的研發探討，以增進本資料庫的完整性。

國語與臺語實質上具有不同之音韻、語意、文法、語用等特徵[16]，例如國語之"颱風"與臺語之"風颱"為語詞的倒裝，故有些國語漢字與臺語語音轉換時，必須考慮機器翻譯(machine translation)的使用。臺語語音合成系統之研發初期，首先應以臺文文字轉臺語語音為基礎，待雛形系統完成後，再考量國語漢字轉臺文之機器翻譯系統等，以減少在初步系統研發過程中的複雜度，且可符合目前臺灣官方使用國語漢字之現況。

本研究發現許多國語漢字無法直接翻譯成符合臺語口語的語詞，主要原因乃目前的系統係以語詞為基礎的直接對照翻譯。本系統若以國語漢字輸入，經詞

庫斷詞後，直接對照漢字之臺語文言音拼字，由於尚未建立臺語口語之拼音文字對照，故系統輸出為臺語之文言音，而非普遍使用之臺語口語詞；若以臺語漢字輸入，由於尚未建立臺語漢字詞庫，故無法以臺語詞庫斷詞，所輸入的漢字將被系統判斷為單字詞，直接對照單字詞庫中之臺語拼音，因此系統輸出為漢字的單字詞拼音，而非以語詞為單位之臺語口語詞。解決之道為：第一種方法乃針對國語漢字詞庫建立對照的臺文口語詞拼音辭典，並依據臺語之口語語詞加入綴字，例如漢字語詞”今日”或”今天”，依臺語口語習慣，最好對應成” kin'af'jit”而非” kim'jit/kim'tiefn”，國語漢字語詞對照臺文口語拼字時，必須依據許多文獻考證之結果，以減少對應之爭議；第二種方法為訓練使用者直接以臺語拼音文字作為系統之文字輸入。

臺語漳州與泉州腔之轉調規則不同，若欲產生不同腔調之臺語連續讀音，仍須建立適當之機讀轉調規則。臺語連續語音中，有加重音、輕讀音、介係詞轉調等特殊規則，必須仰賴系統所建立的規則庫，以轉換成正確之音調。語句分析處理模組負責剖析輕讀字、介係詞、重音等之位置，並依規則轉調後，再由合成單元音檔處理模組對應正確之合成單元。

音韻調整處理模組負責句調之調整，本系統目前的研發進度侷限於肯定句、否定句與疑問句三種句型之音高、音長、音量及停頓參數調整。為提高系統語音合成之品質，仍需持續的針對更多之句型探討與選擇適當的調整參數。

臺語現代文係以多音節語詞為拼音化的基本單元，其優點是符合使用者的語言學習經驗，可以有效減少發音與斷詞的不確定性。本研究應用臺語現代文的語言學理論及音韻規則，發展與建立語音合成系統，期待透過教育與訓練，並經由視覺與聽覺的回饋，發展具親和性及效益性的電腦人機介面應用。

結論與未來展望

本研究應用臺語現代文、音韻學學理及數位訊號處理科技，發展並建立2557個單音節合成單元語料庫，經由載字詞音檔錄製2557個音節合成單元電子音檔資料庫，配合單音、字詞資料庫、字轉音規則庫、文句分析與音韻調整模組之程式發展，研發完成臺語文字轉語音(text-to-speech)系統的雛形。

本研究未來的發展方向：(1) 首先蒐集文獻資料並運用統計方法，發展臺語平衡句語料庫；(2) 利用自相關法(auto-correlation) 或離散的雷建德多項式(discrete Legendre polynomial)求出基週 (pitch)、能量(energy)與音段(duration)曲線(contour)的特徵參數；(3) 由頻域(frequency domain)計算共振波(formant)參數、倒頻譜參數(MFFC: mel frequency cepstrum coefficient)及差值倒頻譜參數(Delta MFFC)等特徵值，共同組成每一個單音節語音之特徵參數矩陣；(4) 利用統計之多變量分析(multi-variable)區分不同之音素、音高、聲調與音韻訊息，以選取最佳化合成單元及調整音韻訊息參數，改善目前臺語語音合成系統的音韻品質。

使用母語具有許多無法取代之優點，而發展一套合乎資訊化、國際化的臺語文字系統，為當前臺灣步入資訊科技時代最重要的課題。發展音、形合一的拼音式現代臺灣語言，必能提高教育、訓練與學習的效率，提供科學研究的有利條件，提昇臺灣人文與科技的品質，對人類文明提供更具體之貢獻。

誌謝

承蒙林繼雄教授悉心的指導與提供臺語現代文詳盡的文獻資料，本論文得以完成。感謝莊勝雄教授研究群提供的文獻與字、辭典檔案、臺文輸入法與拼音檢查工具程式。感謝成大資訊所學弟林順傑、郭啟祥、醫工所邱毓賢、劉俊明先生等協助系統之架設。

參考文獻

1. 施炳華，”附錄：臺語文常識—臺語尋根之旅，”府城國中雙語教材第一冊，pp. 35-41。

2. Y. C. Chiang, et. al., “A New Hybrid Duration Hidden Markov Model with Application to Large Vocabulary Taiwanese (Min-nan) Word Recognition,” 1st International Symposium on Chinese Spoken Language Processing (ISCSLP98), 1998.

3. 許極敦，臺灣語概論，臺灣語文研究發展基金會，臺北市，初版，1990，pp. 33-54。

4. 許極敦，臺語文字化的方向，自立晚報，臺北市，1992，pp. 3-55。

5. S. H. Chen and Y. R. Wang, "Vector Quantization of Pitch Information in Mandarin Speech," IEEE Transactions on Communications, 38(9), 1990, pp. 1317-1320.

6. F. H. L. Jian, "Boundaries of Perception of Long Tones in Taiwanese Speech," ICSLP1998.

7. S. H. Peng, "Production and Perception of Taiwanese Tones in Different Tonal and Prosodic Contexts," Journal of Phonetics, 25, 1997, pp. 371-400.

8. F. H. Jian, "Perception of Long and Short Tones in Taiwanese Speech," J. Acoust. Soc. Am. (1997), 102(5), pp. 3095.

9. S. H. Chen, "Phonetograms of Normal Taiwanese Young Adults," A Thesis of Doctor of Philosophy in University of Wisconsin-Madison, 1996.

10. F. H. L. Jian, "Classification of Taiwanese Tones Based on Pitch and Energy Movements," ICSLP1998.

11. R. Y. Lyu, Y. J. Chiang, R. Z. Fang, W. P. Hsien, "A large-vocabulary Taiwanese (Min-nan) speech recognition system based on inter-syllabic initial-final modeling and lexicon-tree search," ROCLING XI Conference 1998, pp. 139-149.

12. R. Y. Lyu, Y. J. Chiang, W. P. Hsieh, "A Large-Vocabulary Taiwanese (Min-Nan) Multi-Syllabic Word Recognition System based upon Right-Context-Dependent Phones with State Clustering by Acoustic Decision Tree," ICSLP1998.

13. 林川傑 & 陳信希，"中文到閩南語之線上翻譯及閩南語之語音合成，"1999 語文處理技術研討會集刊。

14. 黃保章，"國語文句翻臺語語音系統之研究，"成功大學電機工程研究所碩士論文，1999。

15. Y. J. Sher, K. C. Chung, C. H. Wu, "Taiwanese Syllable-based Synthesis Units Database," Chinese Journal of Medical and Biological Engineering, 19(1), 1999,

pp. 47-58.

16. 鄭良偉，走向標準化的臺灣話文，自立晚報文化出版部，1989，pp. 69-100。

17. 洪惟仁，臺灣河佬話語聲調研究，自立晚報，臺北市，1985，pp. 1-47。

18. 林繼雄，林華英，陳煜楠，由漢字注音到臺語拼字文，，育德文教基金會，初版，1997，pp. 7-15, 72-133。

19. 林繼雄，臺語現代文，大夏出版社，初版，1990，pp. 1-19。

20. J. P. Gee, An Introduction to Human Language, New Jersey, Prentice-Hall Inc., 1993, pp. 65-134.

21. S. Furui, “Digital Speech Processing, Synthesis, and Recognition,” Marcel Dekker, 1989, pp. 5-43.

22. 林繼雄，Taiwanese Dictionary of Words with Modern Spelling，1988。

23. Maryknoll, English Amoy Dictionary, Taichung, 1995.

附錄一：可辨度及自然度評估詞彙集

二字詞部份 (共 20 個)

編號	臺文縮寫	漢字	臺語現代文
1	AK	愛睡(睏)	Aekhuxn
2	EK	浴室	Egkefng
3	HH	高興(歡喜)	Hvoahie
4	KI	喜歡	Kah’ix
5	KT	客廳	Khehthviaf
6	SZ	刷牙	S0efzhuix
7	PJ	小便	Parngji0
8	PK	房間	Pangkefng
9	PaS	大便	Parngsae
10	PiS	廁所	Piexnsor
11	PV	醫院	Pvixvi
12	SiB	失望	Sitbong

13	SọB	洗臉	S0efbin
14	SL	想(思)念	Suliam
15	SeK	逛街	Seqkef
16	SiK	生氣	Siuxkhix
17	SP	散步	Sarnpo
18	SS	傷心	Siongsym
19	ZS	穿衣	Zhexngsvaf
20	TI	討厭	Th0’iax

三字詞部份 (共 20 個)

編號	臺文縮寫	漢字	臺語現代文
1	SSK	洗澡(身軀)	S0efsefngkhw
2	BSK	不舒服	B0sofngkhoaix
3	GBH	我不會	Goar buexhiao
4	GKK	我感覺	Goar kafmkag
5	GMA	我不要	Goar m aix
6	KPZ	看報紙	Khvoax p0rzoar
7	KTS	看電視	Khvoax tiexnsi
8	ZIS	找醫生	Zh0exisefng
9	TOL	臺灣人	Taioaan-laang
10	HZT	蕃薯湯	Hanzuu-thngf
11	TJP	糖尿病	Thngji0xpvi
12	CSL	一世人(一輩子)	Cidsielaang
13	TTK	土地公	Thoftixkofng
14	HZS	現此時	Hiexnzhwsii
15	HTP	虎頭蜂	Hofthauphafng
16	IET	游泳池	Iu’efngtii
17	PTK	怠惰骨	Pintvoaxkud
18	PTH	布袋戲	Port0exhix
19	SLK	少年家	Siaolienkef
20	ZHT	臭火乾	Zhaoh0eftaf

四字詞部份 (共 20 個)

編號	漢字	臺語現代文
1	人山人海	Jinsafn-jinhae
2	三頭六臂	Samthi0o-liogpix
3	千變萬化	Chienpiexn-baxnhoax
4	山珍海味	Santyn-haybi
5	不三不四	Putsafm-putsux

6	不義之財	Putgi-cy-zaai
7	心甘情願	Sym-kafm-zeeng-goan
8	甘拜下風	Kampaix-haxhofng
9	安分守己	Anhun-siwkie
10	作惡多端	Zok’og-t0toafn
11	門當戶對	Buntofng-hoxtuix
12	莫名其妙	Bogbeeng-kibiau
13	耀武揚威	Iauxbuo-iong’uy
14	繡花枕頭	Siuohoaf-cymthi0o
15	談天說地	Tamthiefn-soatte
16	滿面春風	Boafnbien-zhunhofng
17	寬宏大量	Khoanhoong-taixliong
18	福如東海	Hog-juu-tonghae
19	飲水思源	Ymsuie-sugoaan
20	頂天立地	Tefngthiefn-libte

文句部份 (共 10 句)

編號	漢字	臺語現代文
1	放棄自己所有的一切	Parngsag kaki sofu ee itzhex.
2	他一時糊塗偷拿錢	Y cidsii hotoo thautheq cvii.
3	你會曉講英語沒	Lie exhiao korng Enggie b0e?
4	廢除壞的風俗習慣	Hoeatii phvae ee hongsiok sibkoaxn.
5	媽媽教我如何掃地板	Mamaf kax goar afnnar saux thokhaf.
6	厝裡有一點仔事情要辦	Zhuxnih u cidtiafmar taixcix boeq pan.
7	大家都是台灣人	Taixkef lorng Si Taioaan-laang.
8	那內面就是冷凍的物件	Hit laixbin ciux si lefngtoxng ee miqkvia.
9	太陽從東邊升起	Jidthaau tuy tangpeeng khielaai.
10	我要養一隻小隻的鳥仔	Goar b0eq chi cid ciaq s0eaciaq ee ciawar.

附錄二：臺語 2557 個七聲調的音節資料庫—僅節錄基調音 394 個

Taiwanese Fundamental-Tone Syllables

Taiwanese F-tone.xls

	a	ai	am	an	ang	au	e	eng	i	ia	iam	iang	iau	ien	im	in	iong	io̲	iu	m	ng	o	oa	oai	oan	oe	om	ong	o̲	o̲e	u	ui	un
無	也		餡	限	旺	後	下	用	義	夜	鹽	癢	要	緣		孕	用		又	毋	暈	芋	樺		援	話		王		能	有	位	韻
b	密	覓		萬	網	貿	賣	命	味				妙	面		臉		廟	謬			幕			墁	妹		望	磨	未	霧		聞
c									舐	謝	暫			弄		盡	狀		就														
ch									飼			銜					匠	照	樹														
chv																			象														
cv									舐										癢														
g	訝	礙	偘	諺	愣	睾	藝	迎	誼	嚳	驗	攘	蕎	研	擒	恨	邛	蕎				五	外		願	藝		呆	餓		遇	危	
gv									硬													訛											
h	夏	害	憾	限	巷	校	系	幸	耳	蟻			邀	現	噤	恨		後	復	盍	遠	雨	畫	壞	患	會		鳳	號	會	婦	惠	份
hv	跨								耳	艾												喔	岸										
j									二	揖		讓	尿		任	認	讓	尿					若			銳					裕		潤
k	咬				共	厚	低	間	居	陡		彊	罵	健	妗	近	共	轎	咎			苟	樺	乖	縣	易		狂	膏	街	舅	跪	郡
kh								虹	柿	豎	儉								糗				譁					誆	裸	刮	懼		
khv																			強														
kv										件									強			軒	汗	乖									
l	撈	賴	混	難	弄	漏	麗	令	字		唸	亮	料	練		磷	量	尿	餾			路	賴		亂	耒		浪	勞	耒	呂	類	論
m	罵	昧				貌	罵		麵	命										梅	問		曼	妹		妹			帽			媒	
n	若	耐				鬧			呢	岭			鳥					讓	讓		蛋		唾						耨			卵	
p	罷	敗		辦	棒	暴	父	並	備					辯		臏		鰾			飯	部			叛	倍		傍	埠		孵	吠	笨
ph	皰				縫	抱	被	併	砒			棒				蘋						簿	旆		伴	被		膨	抱	被	噴		
phv	泡	揹							鼻														伴										
pv							病		病														拂										
s		侍		墠			謚	盛	是	射	贍	最	召	善	甚	腎	上		受				逝		罵				唆	垂	士	祟	順
sv									豉	盛									想														
t	大	代	淡	但	動	豆	袋	定	箸		墊	丈	調	電	朕	陣	重	稻	稻		斷	肚	大		段	第		動	稻	地	箸	隊	遁
th		態			弄	毒		挺	雉		填		柱	填		滕					燙		導			頽	垂	堆	它			累	填
thv	飾	態							縫	挺																						燙	
tv	大						鄭		漲	定									丈				彈										
v	餡	揹							院	飄									樣				換	椏									
z	載	財	站	贊		找	坐	淨													旋	助	誓	扭	饌	罪		狀	造	多	自	誰	陣
zh	吵	祀	潛			嘈	坐	穿															娶			找				尋	娶		
zhv																							鱔										
zv	掬																						濺										

33*33=1089

教羅第七聲

基調音

國語文句翻台語語音系統之研究
A Study for Mandarin Text to Taiwanese speech System

十王駿發　十黃保章　*林順傑
{wangjf, hwangbj ,linsj }@server2.iie.ncku.edu.tw

十國立成功大學電機工程學系
*國立成功大學資訊工程學系

摘要

本論文建立了一套能說台語的中文文句翻台語語音系統。使用者只要輸入中文詞句，便可輕易地得到對應的台語發音。文中主要針對台語語音性質在語音系統實作時所產生的:(1) 漢字的一字多音、(2) 文句的斷詞、(3) 聲調運作的問題、(4) 合成單元處理等四個問題進行探討。

1. 簡介

就文字轉語音合成系統而言，由於國語為普遍性的溝通語言，研究單位也多偏向此領域來進行研究，並且有不錯的研究成果發表。但是在方言的語音合成系統方面，卻鮮少論及。由於母語在現今傳播媒體的使用頻率與日俱增，母語的語音合成系統，在運用上也變得有其需要性。有鑑於此，乃嚐試開發會說台語的文字轉語音合成系統。

國語文句斷詞處理完在漢字對應於台語發音所產生的一字多音問題，本文是先對各音做優先音排序，若仍有取音混淆情形則將某音以詞庫窮舉的方式來解決。

台語聲調的變化是與國語差異最大的部份。本論文先決定了一套語音合成系統上所適用的台語聲調種類與變調規則；然後再提出一套規則來處理句子中，何字該變調，何字不該變調的問題。最後，對於本系統所用的合成單元是直接取自本論文提出的「固定音高旋律性錄音法」所錄得的語音資料。

本論文的章節架構如下：第二節是研究現況與研究背景；第三節則介紹我們資料庫的建立；第四節說明研究方法-斷詞、台語的取音及聲調變化的處理情形[21]；實驗的結果則在第五節描述；最後，第六節則是結論與討論。

2.研究現況與研究背景

2-1 研究現況

台灣語音溝通科技之研究肇始於1980年代的中文語音合成與辨識，現有研究成果應用於人性化溝通之電腦人機介面、電話語音系統等，均專注於北京語(Mandarin)系統的開發。現階段的台語研究，較專注於台語語言學特性的探討，而台語計算語言學研究仍屬萌芽階段。近年來台語計算語言學論文發表的研究群，包括長庚大學呂仁園與清華大學江永進教授的大量詞彙辨識系統[15][16][17]，台灣大學陳信希教授[18]、成功大學吳宗憲、鍾高基教授的文字轉語音系統[19][20]等。

2-2 拼音符號介紹

台語和國語皆屬漢語語系分支，是單音節語言，由所謂聲母、韻母和聲調三部份所組成。本文修改部分【國台雙語辭典】的音標方案，以利於論文中的台語語音標示，其音標形式詳見表2-1。聲母有17種、韻母48種，各欄名稱解釋如下：「國」是指注音符號，「台」、「楊」是指國台雙語詞典所採行的音標符號，「黃」是指本論文的音標方案，「教羅」台灣教會所採行的音標方案，「例字」是發此音的漢字。

2-3 一字多音的處理

由於時間悠久、空間遼闊及其他因素造成台語的一字多音的現象。本系統的漢字對應台語音個數，一個漢字對應台語音節的數量以九個音為最多，計有「撓那落」三個字。傳統台語的漢字發音教學方式，在由字發音時，多半只教文白音的分別念法，並未指定規則來擇取。今若以系統實做考量，這種分類方式是很籠統的，況且單一漢字的台語音節數目也不止兩種而已。要決定取該字的哪一個音節來發音，便成為猶疑難決的問題。

例如：香港的香很香。(hiong2-kang2 ye7 vhio1 ziong8 pang1)三個『香』皆發不一樣的音，而字典中「香」有五種讀音(vhiu1 白話音,vhio1 白話音,hiong1 文言音,hiang1 文言音,pang1 俗音)。

國台聲母（CONSONANT）

國	ㄅ		ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ		ㄎ	ㄏ	ㄐ		ㄑ	ㄒ	ㄓ	ㄔ	ㄕ	ㄖ	ㄗ		ㄘ	ㄙ
台	ㄅ	ㆠ	ㄆ	ㄇ		ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㆣ	ㄎ	ㄏ	ㄐ	ㆢ	ㄑ	ㄒ					ㄗ	ㆡ	ㄘ	ㄙ
黃	b	bh	p	m		d	t	n	l	g	gh	k	h	z(i)	zh	ch(i)	s(i)					z	zh(u)	ch	s
楊	b	v	p	m		d	t	n	l	g	q	k	h	z(i)	j	c(i)	s(i)					z	j(u)	c	s
教羅	p	b	ph	m		t	th	n	l	k	g	kh	h	ts c ch	j	tsh ch chh	s					ts c ch	j	tsh ch chh	s
例字	褒	帽	波	冒		刀	桃	奴	囉	哥	鵝	科	號	之	字	癡	施					資	裕	此	思

國台韻母（VOWEL）

國	ㄩ	ㄩㄝ	ㄩㄢ	ㄩㄣ	ㄩㄥ	ㄚ			ㄜ	ㄝ		ㄞ		ㄟ	ㄠ		ㄡ		ㄢ	ㄣ	ㄤ	ㄥ			
台						ㄚ	ㆩ		ㄜ	ㄝ	ㆥ	ㄞ	ㆮ		ㄠ	ㆯ		ㆰ	ㄢ	ㄣ	ㄤ	ㄥ	ㆬ	ㆭ	
黃						a	va		er	e	ve	ai	vai		au	vau		am	an	(n)	ang	*	m	ng	
楊						a	a^{n}		∂	e	e^{n}	ai	ai^{n}		au	au^{n}		am	an	(n)	ang	*	m	ng	
教羅						a	a^{n}		o	e	e^{n}	ai	ai^{n}		au	au^{n}		am	an	(n)	ang	*	m	ng	
例字						膠 阿	監 餡		高 婀	家 挨	更 嬰	皆 哀	閒 口哀		交 甌	爻 藕		甘 庵	干 安		江 尪		姆 姆	鋼 秧	

國	ㄧ		ㄧㄚ			ㄧㄢ	ㄧㄤ	特別		ㄧㄣ	ㄧㄥ	ㄧㄛ			ㄧㄝ	ㄧㄞ	ㄧㄠ		ㄧㄡ	
台	ㄧ	ㆪ	ㄧㄚ	ㄧㆩ	ㄧㆰ	ㄧㄢ	ㄧㄤ	ㄩㄥ	ㄧㆬ	ㄧㄣ	ㄧㄥ	ㄧㆦ	ㄧㆧ	ㄧㄜ			ㄧㄠ	ㄧㆯ	ㄧㄨ	ㄧㆫ
黃	i	vi	ia	via	iam	ian	iang	iong	im	in	ing	io	vio	ir			iau	viau	iu	viu
楊	i	i^{n}	ia	ia^{n}	iam	ian	iang	iong	im	in	i/eng	io	io^{n}	i∂			iau	iau^{n}	iu	iu^{n}
教羅	i	i^{n}	ia	ia^{n}	iam	ian	iang	iong	im	in	eng	無	io^{n}	io			iau	iau^{n}	iu	iu^{n}
例字	居 衣	梔 嬰	迦 夜 7	驚 營	兼 閹	堅 煙	姜 央	恭 雍	金 音	巾 因	經 英		鴦 (漳	茄 腰			嬌 妖	口梟 喵	ㄐ 憂	薑 鴦

國	ㄨ		ㄨㄚ		ㄨㄛ	ㄨㄞ		ㄨㄟ	ㄨㄢ					ㄨㄣ	ㄨㄤ				ㄨㄥ
台	ㄨ	ㆫ	ㄨㄚ	ㄨㆩ		ㄨㄞ	ㄨㆮ		ㄨㄢ	ㄨㄝ	ㄨㆥ	ㄨㄧ	ㄨㆪ	ㄨㄣ	ㄨㄤ	ㆦ	ㆧ	ㆱ	ㄨㄥ
黃	u	(vu)	ua	vua		uai	vuai		uan	ue	ue^{n}	ui	ui^{n}	un	uang	o	vo	om	ong
楊	u	u^{n}	ua	ua^{n}		uai	uai^{n}		uan	ue	ue^{n}	ui	ui^{n}	un	uang	o	$_{o}n$	om	ong
教羅	u	u^{n}	oa	oa^{n}		oai	oai^{n}		oan	oe	oe^{n}	ui	ui^{n}	un	oang	o^{o}	$_{o}{}^{o}n$	om	ong
例字	龜 污		瓜 哇	官 鞍		乖 歪	糜 歪		觀 彎	檜 煨		規 威	口關 偉 n	君 溫	光 嚾黨	沽 菇	姑	箴	公 翁

表 2-1 台語音標符號相關表

2-4 聲調特性介紹：

眾所周知，國語聲調計有陰平(1)，陽平(2)，上(3)，去(4)，輕聲，五種，本文姑且稱為本調聲調。當上聲音節接上聲音節時，前音節會變成陽平聲以便接上聲；另外還有「一」、「不」的聲調變化情形。我們稱此現象為國語聲調的變調。

國語聲調產生變調的原因乃是基於發音困難而不得不變。但台語則不然，台語聲調產生變調的情形不僅較國語多，而且台語的變調運用，並不是單單基於發音困難與否而產生，變與不變二者往往有辨義或語態上的不同作用。

譬如：

“後(yau7)日(zin8)” ，

台語念(yau3-zin8)是”改天”的意思，若念成(yau7-zin3*)則是”後天”的意思。

“驚(vgia1)死(si2)”，

台語念(vgia7-si2)是”怕死”的意思，台語念(vgia1-si3) 則是”嚇死”的意思。

2-4-1 聲調種類：本調與變調

自古以來台語聲調種類的歸納分類並未脫離「八音」之說，意即台語有八種不同聲調的意思，名稱如表 2-2，括弧內為其對應的聲調代號。但是現在的臺灣除了鹿港發音外，已無陽上(6)，只剩七種聲調。

	平	上	去	入
陰	陰平(1)	陰上(2)	陰去(3)	陰入(4)
陽	陽平(5)	陽上(6)	陽去(7)	陽入(8)

表 2-2 台語聲調的傳統分類

觀察台語音調本調的頻率走勢，取現在的台南口音來描繪實際的本調變調頻率走勢(如圖 2-1(b))，比較王育德博士(台南人)在 1954 年夏天於東京大學理工研究所曾做實驗所畫出的台語音調走勢[3]如圖 2-1(a)，可知台南的聲調的變動情形很小。

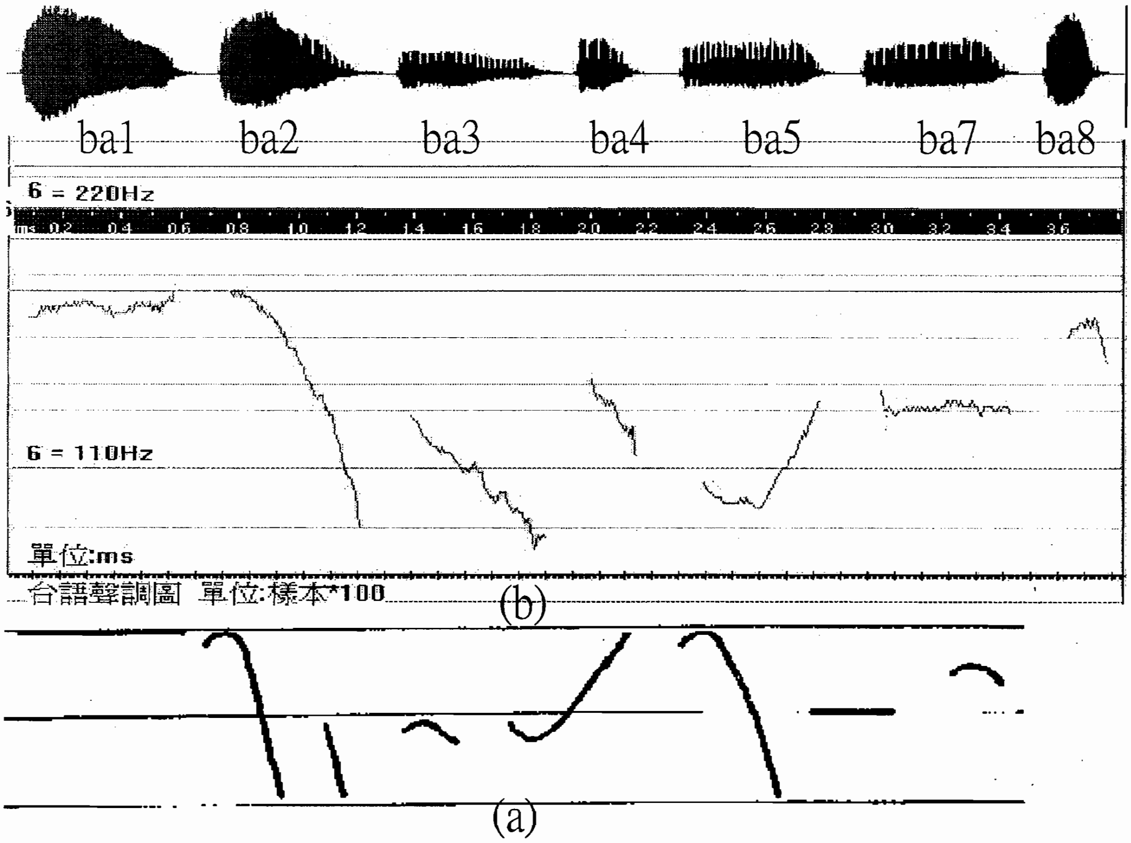


圖 2-1 台南口音本調頻率走勢圖

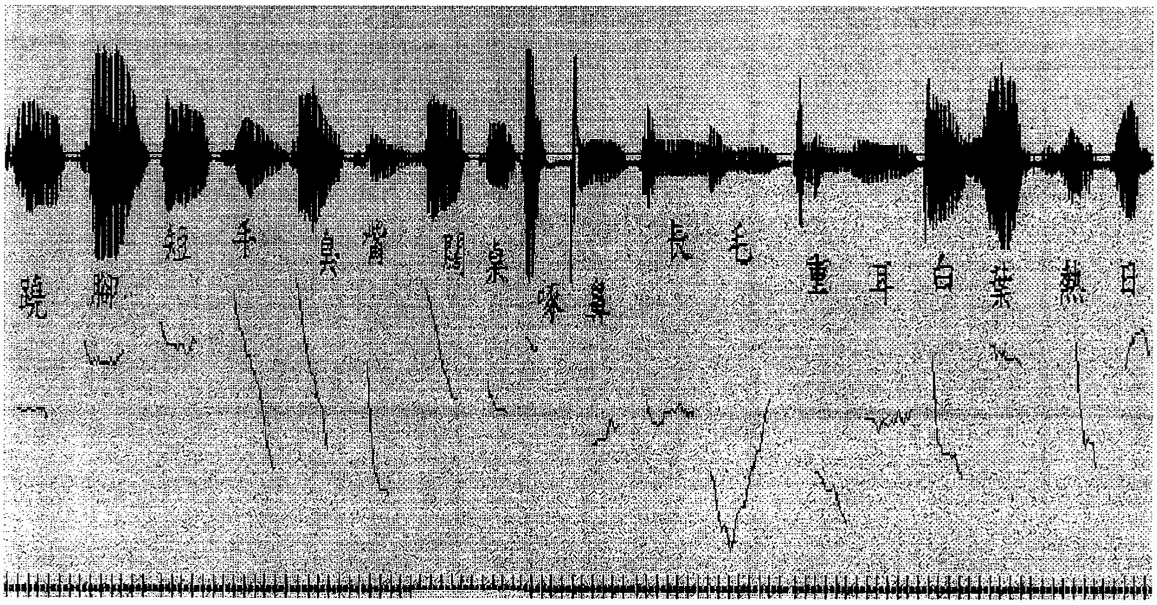


圖 2-2 台南口音變調頻率走勢圖

2-4-2 一般變調規則

再觀察二字詞的發音情況，它的頻率走勢如圖 2-2 所示。參考現今台語教學著作所描述台語變調規則計有三種版本[1][12][13][14]（見圖 2-3(a)、(b) 、(c)）。分析其中差異出現在入聲調變化的問題上。以台南音，台北音而言，經實驗證實可以用圖 4-2 的規則來處理台語發聲系統的聲調轉換。此規則與圖 2-2 的頻率走勢相驗，是無問題的，並且經本系統證實是可用的。

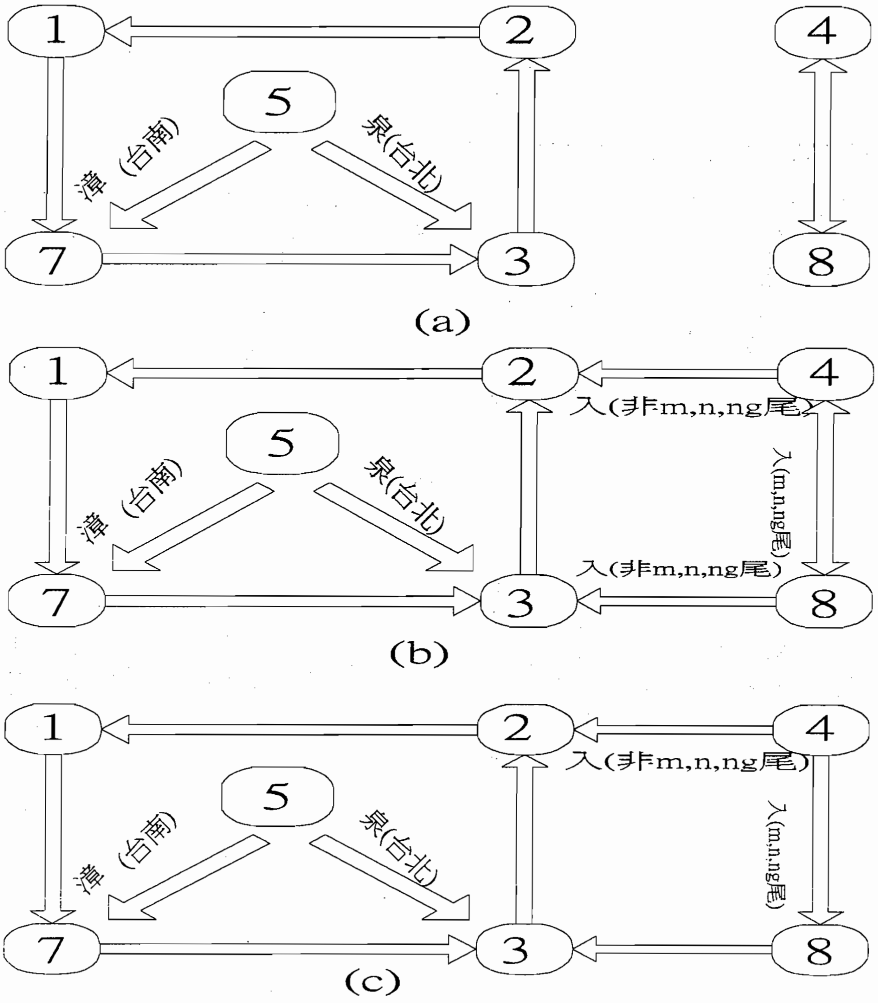


圖 2-3 台語一般變調規則(a),(b),(c)

3.資料庫的建立

3-1 系統架構

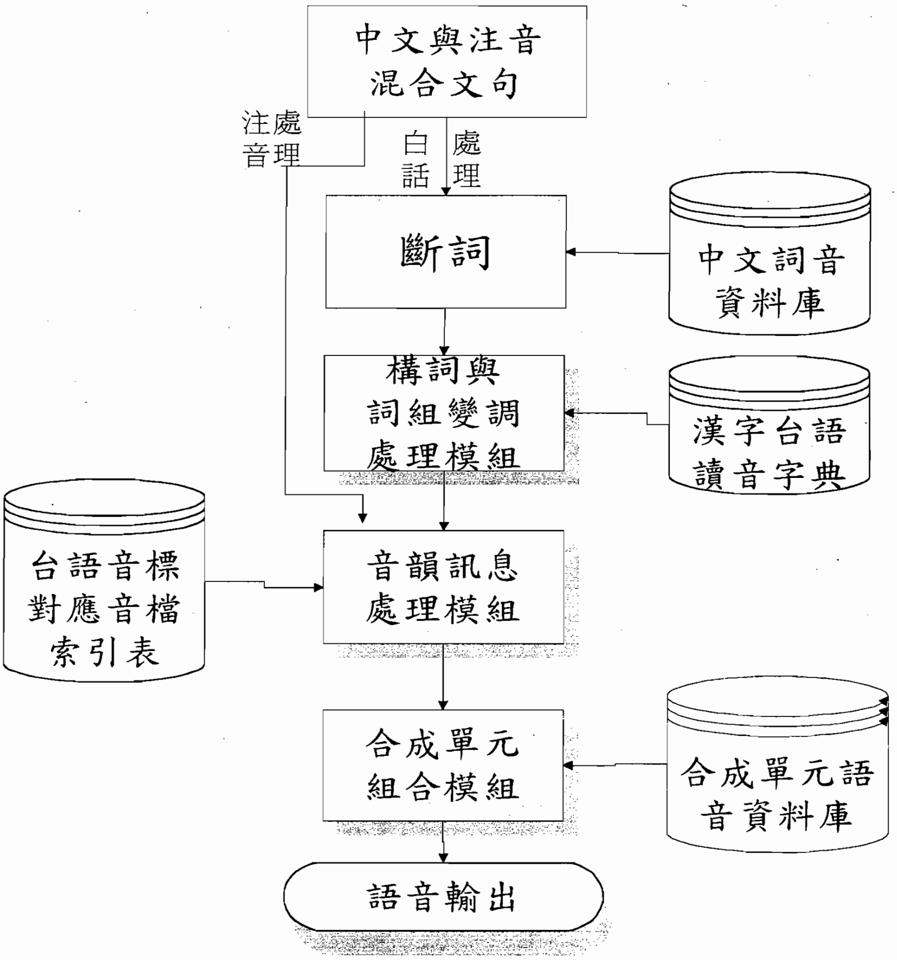


圖 3-1. 系統架構圖

整個系統架構如圖 3-1 所示，由四個資料庫及四個處理模組構成。本節先介紹資料庫，下一節再介紹四個處理模組方塊運作情形

3-2 漢字台語讀音字典與中文詞音資料庫

漢字台語讀音字典決定斷詞後漢字對應的台語注音字，以及作為單字發音查詢使用。本系統漢字對台語音字典的建立，主要是以楊青矗先生所編寫的【國台雙語辭典】為範本而建立[2]。本系統收錄漢字對應台語音的字計有 8168 字，其中對應注音數目如下表 3-1：

總音數	1	2	3	4	5	6	7	8	9
總字數	4418	2504	866	247	93	27	7	6	3

表 4-1 漢字對應注音總數表

而中文詞音資料庫共建立二字詞有18730條，三字詞有4207條，四字詞有3983條，共有26920條的常用詞彙，已包含用窮舉方式所解決的單一漢字取台語音混淆情形的詞條，作為斷詞結果的主要依據。

3-3 台語音標對應音檔索引表

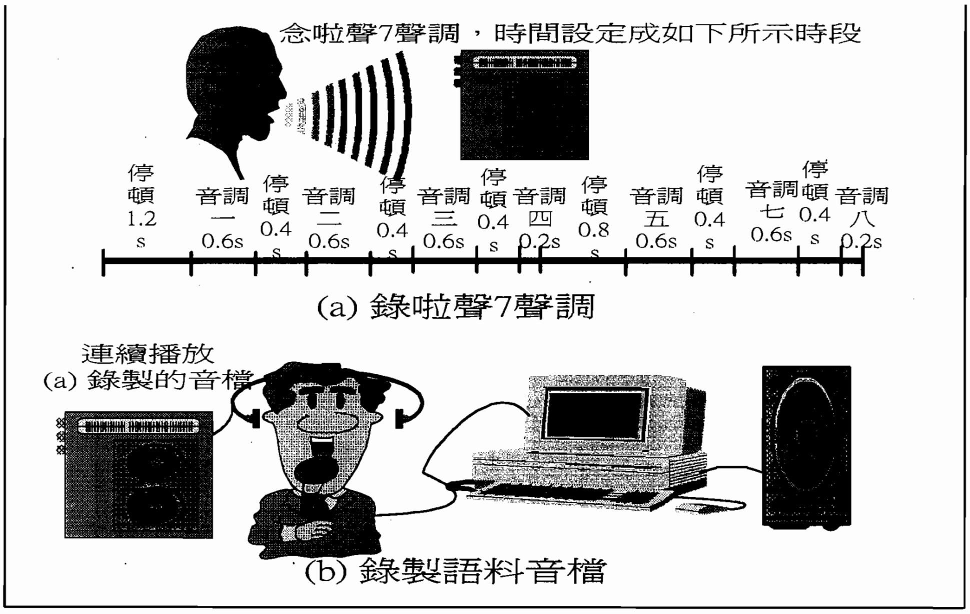
此表由552個音節各含7個聲調的組合，共3864個注音音符組成，包含了本系統所使用的所有音節符號，用以對應到該音節的聲音檔索引名稱，對於各種形式的注音(拼音)音符間的對應關係，可以對此表做擴充，以增加系統的運用彈性。

當輸入編輯器接受所輸入注音字或漢字後，皆可透過此表轉換成聲音檔的索引名稱，以便供給變調規則運用和語音樣本擷取時所需的資訊。

3-4 合成單元語音資料庫

觀察人講話的語音特性，在不同時空之下，同樣一句話，其音高特性會有差異，也就是說語音音高特性是相對的，不是如樂音旋律音高值是絕對的。

對系統的語音合成所需取得音節聲音樣本而言，各同調音節語音樣本卻極需具有相同的音高特性，並且不同聲調音節的語音樣本間的音高特性更需控制在一定範圍內，以便產生良好的合成語音音質輸出。合成單元語音資料是以本論文所提出的「固定音高旋律性錄音法」錄音方法錄製而成，共3864個音節，並且紀錄各音節的起點，結束點，以及詞中、詞尾的端點。



(a) 錄啦聲7聲調

(b) 錄製語料音檔

圖 4-2 固定音高旋律性錄音法步驟一與步驟二

4. 研究方法-斷詞、聲調變化及台語的取音

4-1 斷詞

本論文採用蒐集大量的詞彙利用統計分析建立一個詞庫加以斷詞[4]，然而有限的詞庫亦無法包含所有的詞彙，因此有可能發生在詞庫找不到匹配的詞而造成錯誤，但是若要獲致滿意的結果，最終雖仍需要靠語法及語意的分析才能解決。

一般語言模型的機率可以下式表示：

$$P(W)=\prod_{i=1}^{n} P(W_i\,|\,W_1W_2\ldots W_{i-1}) \qquad \text{(式 4-1)}$$

其中 $P(W_i\,|\,W_1,W_2,\ldots,W_{i-1})$ 代表詞 Wi 出現在字串 W1,W2,…,Wi-1 之後的機率，但是此機率在實際的情況下是無法估算出來的，因為假如我們的語言含有 L 個詞，就可能須要記錄 $L^{(i-1)}$ 筆資料，這是相當大的資料量，所以我們簡化這機率公式，假設 Wi 出現的機率只和它的前幾個詞有關而已，和在之前的詞無關。例如 Wi 只和它之前的一個詞有關時，即為 bigram 關係，以 4-2 式來表示：

$$P(W)=\prod_{i=1}^{n} P(W_i\,|\,W_{i-1}) \qquad \text{(式 4-2)}$$

其中前一詞和下一詞之間相連的機率 $P(W_i\,|\,W_{i-1})$ 的機率估算由下式來計算：

$$P(W_i\,|\,W_{i-1}) = f(W_i\,|\,W_{i-1}) = \frac{C(W_{i-1}W_i)}{C(W_{i-1})} \qquad \text{(式 4-3)}$$

上式中的分子分母分別是 bigram 函數和 unigram 函數。結果如圖：

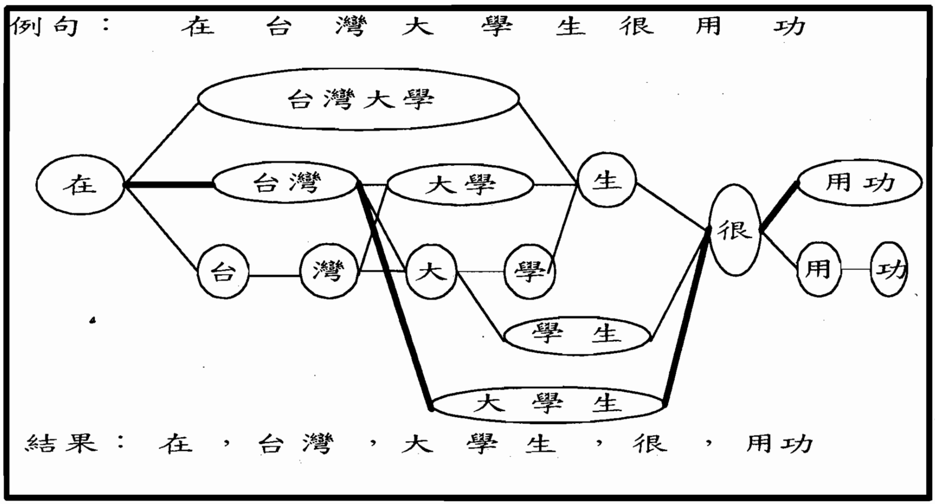

例句： 在 台 灣 大 學 生 很 用 功
台灣大學
在
台灣
大學
生
用功
很
台
灣
大
學
用
功
學生
大學生
結果： 在 ， 台灣 ， 大學生 ， 很 ， 用功

圖 4-1 斷詞狀態轉移圖

4-2 構詞、選音與詞組邊界變調處理

台語文句的音調變化的準確與否，決定了聽覺的流利度感覺與辨義功能。就詞彙本身而言，一般詞本身的基本變調規則為：最後一字不變調，前面全變調。

譬如：

『安』念 yan1 .『安平』念 yan7-beng5 .『安平港』念 yan7-beng7-gang2 .

『碼』念 beng2 .『碼頭』念 beng1-tau5 .

但是有少數詞彙及大量成語皆有特例性質，不適用上述規則，需以詞庫建立處理。如：心*悶(sim1-bhun7) ，縱虎*歸山(ziong2-ho2 gui7-san1)

上述的變調規則為系統斷詞結果與自然人說話所認定的斷詞點相同時所採用的規則，然而有些詞並無法在系統詞庫中列舉出來，如數字，人名，地名等，所以我們需以構詞方式處理來補強斷詞結果。下一小節將介紹系統處理數字的方法。

另外還得解決詞組邊界的聲調處理。意義以下列例子說明：

看右面句子：在台灣大學生很用功

斷句成: 在,台灣,大學生,很,用功 .

台語注音音節表示為：

zai7,dai7-wuan5,dai3-hang3*-seng1,ziong4,yiong3-gong1.

由上說明即知：所謂詞組邊界的聲調處理，意即是要決定『在，灣，生，很，功』等字是否要變調。圖 4-2 為本系統所採用的變調規則，其中 3*是當音節尾音有 m, n, ng 音，且聲調值是 8 的變調後聲調值。

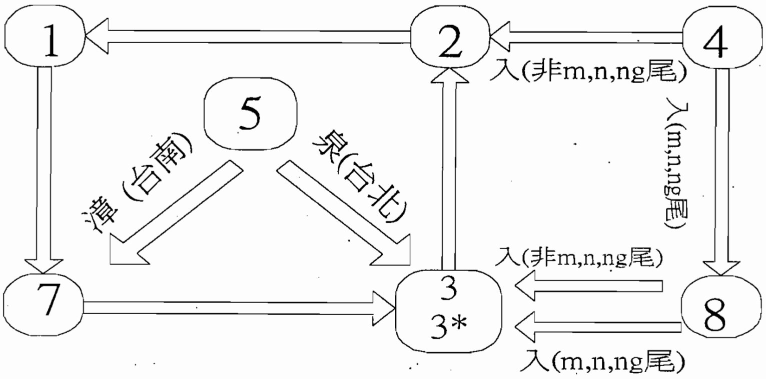


圖 4-2 本系統採用的台語一般變調規則

4-2-1 數詞的構詞、選音與變調規則處理

本文雖以「詞庫窮舉法」的方式解決台語系統一字多音選音混淆的問題，但是仍有難以決定誰是第一順位的問題存在，這在數字念法上特別明顯。處理說明於下。

漢字(1)	漢字(2)	文言(1)	白話(2)
一	壹	yin4	zinj
二	貳	zhi7	lng
三	參	sam1	vsac
四	肆	su3	siq
五	伍	vgho2	gho
六	陸	liong8	langj
七	柒	chin4	
八	捌	ban4	be4
九	玖	giu2	gau2
十	拾	sim8	zam8
	零		leng5
百	佰	beng4	ba4
千	仟	chian1	cheng1
萬		bhan7	
億		yeng4	
兆		diau7	

表 4-1 數詞注音表

表 4-1 是數詞注音表，其文言音與白話音在口語中常並列使用，前者常用在成語或術語中，後者則用在數量詞用語中。文言音因為詞庫畢竟無法全部包含，故選音排名為優先地位。當我們以構詞方式定出數量詞用語，需再改變注音符號以達成正確選音的需求。再來還需探討數量詞連結時，變調處理的問題。

基本上，我們將數詞分成 A，B 二類，如下表 4-2：

A 類	一，二，三，四，五，六，七，八，九，十
B 類	百，千，萬，億，兆，零

表 4-2 數詞分類表

則詞或字結合點的變調處理規則為:

只有 A 的組合則念文言音。如：一九九九年。

- A+B 的組合則念白話音。如：一百五十年，但是二十、十二，的 ”二” 須念文言音。
- [N*（A+B）or N*（A+B+B），N>1]的組合時，雖是念白話音，但是千，萬，億，兆，固定不變調。如：二千一百三十五億人 ，「千」念本調音。
- 組合詞是 (數詞+量詞) 或 (數詞+名詞)，則數詞尾字需變調，該組合詞視為獨立詞。
- 組合詞是 (數詞+量詞+名詞) 則數詞、量詞尾字均需變調，該組合詞視為獨立詞。

4-2-2 詞組邊界的聲調處理

- 中文詞庫的詞屬性分成變動詞、獨立詞、連接詞、的、數詞、量詞、可形容詞化名詞、感嘆詞(語助詞)。
- 所有詞最後只剩 變動詞 與 獨立詞。

  變動詞接任何詞則變動詞尾音節需變調。 獨立詞則否。
- 「連接詞」、「的」視為句子段落，之前若有斷詞斷點，該分界不需變調。
- 「的」本身成為變動詞。
- 有「連接詞」、「的」的句子，該詞即視為句子段落。也就是說，之前若有斷詞斷點，該分界不需變調。而「的」本身之後若有斷詞斷點，則本身需做變調處理。「連接詞」分成二類，一類需要變調處理，如「既然」等，一類不需要變調處理，如「但是」等。
- 將初步斷詞為一字詞或二字詞或三字詞或四字詞者，分成變動詞與獨立詞，變動詞與下一詞相連時，尾音節需變調，獨立詞則否。
- 句子尾字與獨立詞尾字念本調。

- 句子尾字若是感嘆詞或語助詞如「啊，呀，而已」等語助詞則其前一字需念本調，本身毋需做變調處理。

以例子說明：

「安平港碼頭」念 yan7-beng7-gang2　beng1-tau5.

因為安平港，碼頭皆為獨立詞，所以接合處音節 gang2 無須變調。

再看下面句子：

「在台灣大學生很用功」，斷句成: 在,台灣,大學生,很,用功 .

「在，很，用功』等字是變動詞，所以接合處音節須變調，但「功」在句尾，所以不變調。而「台灣，大學生」皆為獨立詞，所以接合處音節無須變調。故結果變成 zai3,dai7-wuan5,dai3-hang3*-seng1,ziong8,yiong3-gong1.

經過上述詞組式變調處理後，文句大致可以得到還不錯的結果。

4-3 聲音輸出處理

語音合成器基本上可分為兩大類：第一類是利用全極點(all-pole)的語音發聲模型，將發聲的機制分為聲帶激發訊號(glottal excitation)及一個口腔模型(acoustic tube model)。應用此一模式所發展出來的語音合成技術多源自語音壓縮技術，如線性預估編碼 (LPC vocoder)[5][6]，單、多脈衝激發(multi-pulse excitation)[7][8][9]等。

第二類的合成器則是所謂波型合成器，此類合成器在時域上直接調整語音基週頻譜特性而不需藉由簡化的發聲數學模型來處理合成語音的音韻。雖然儲存的空間需求較大，但所合成出來的語音品質較第一類合成器的效果要好，而且此類合成器的運算量較少。目前的波型合成器所使用的方法以時域基週同步疊加法(Time Domain Pitch synchronous Overlap-Add,TD-PSOLA)[10][11]最廣為採用。本系統並未運用上述技術，而是以音節直接加以合成，整個作法說明如下：

經由斷詞及構詞以及詞組變調旗號處理之後，我們得到該文句各字的注音字、變調旗號、音長旗號、停頓間隔旗號，送入音韻訊息處理模組。

音韻訊息處理模組便根據文句分析之後所得到的上述參數，將注音字轉換成發音代碼，並以一般變調規則進一步轉換成實際的發音代碼，之後再藉由音長旗號與停頓間隔旗號決定各音節在合成時所需的語音樣本長度及所需的停頓間隔。

對於台語的入聲音節，由於本身是短促音，口語化時與相鄰音節相接時必有一停頓現象，可以本身音節長度做為靜音長度，以增加語氣的流利度。

經由上述處理，合成單元組合模組便可藉由適當的音節斷點與靜音長度的組合，自合成單元語音資料庫取得語音樣本，便將語音結果輸出。

5. 實驗

5-1 實驗環境

本系統是以 Pentium 133 個人電腦加上 16 位元聲霸卡在 Win95 作業系統下以 Microsoft Visual C++ 5.0 開發完成，而實驗環境也是在如上述的環境下進行。各項目簡述如下：

- 測試人員：分一般人士組及實驗室同仁組兩組，各組 5 人。
- 測試樣本：選取二、三及四字詞包含七聲調變化調與本調各 20 條。報紙短句 10 句。
- 測試方式：

可辨度：

在可辨度的評估方面，在系統合成測試樣本後，由受測者將所聽到的語音以聽寫方式寫出。最後統計正確文句與受測者所寫文句之間的差異，以作為評估可辨度的方式。

自然度：

我們採用平均鑑定分數(Mean Opinion Scores, MOS)做為評估的標準。這種評估方式將合成語音的自然度分為優良(excellent)、良好(good)、尚可(fair)、差(poor)及極差(unsatisfactory)五個等級，分別給予 5 至 1 不等的分數。測試人員在聽過合成的語音後，以所感覺到的自然度評分。

5-2 實驗結果

實驗結果如下所示：

一、可辨度

測試種類	數量	一般人士組可辨度	實驗室同仁組可辨度
二字詞	20	95.0%（38/40）	95.0%（38/40）
三字詞	20	90.0%（54/60）	93.3%（56/60）
四字詞	20	87.5%（70/80）	93.8%（75/80）
短句	10	60.7%	60.2%
平均		83.3%	85.58%

表 5-1. 可辨度實驗結果

由實驗結果的數據顯示，本系統合成語音的平均可辨度為 84.4%，而且測試樣本字數愈多可辨度愈低。推究其原因，是因為字數長度愈長，變調處理錯誤機會也越高，可見變調問題的處理對台語系統的發音的重要性。

二、自然度

測試種類	數量	一般人士組自然度	實驗室同仁組自然度
二字詞	20	4.0	4.0
三字詞	20	4	3.5
四字詞	20	3.5	3.7
短句	10	3	3.1
平均		3. 625	3.575

表 5-2. 自然度實驗結果

自然度評估的結果，平均鑑定分數為 3.6125，介於尚可與良好之間，表示本系統所輸出的語音雖未在合成技術上多所著墨，但是採用的錄音方法，亦可達成某一程度的水準。另外，變調在台語流利度的感受占著極大的因素，也由實驗數據可以得知。

6. 結論與討論

在本文中，我們實作了一套能說台語的國語文句翻台語語音的系統。在面對漢字一字對台語多音的取音抉擇問題上，試以詞庫窮舉的方式來解決。

本論文亦建立了一套語音合成上可用的台語一般變調規則，並對詞與詞連接時，所衍生前一詞的詞尾是否變調的問題處理提出作法。而以「固定音高旋律性錄音法」所錄

得語音資料，可以有效降低聲音樣本因時間差而產生的語音頻譜差異過大的現象。這對於連音部分自然度的摩擬，很有幫助。此種錄音方式亦有利於單音節多聲調的系統做初步的開發。譬如國語或客語發音系統。

由於變調處理在台語發音上，是為必要的成分，所以應該變調之處就必須變調，不該變調處就不要變調，所以斷詞正確與否對台語發音有關鍵性的影響。本文所用的斷詞演算法可增加斷詞的正確性。同時，也建立構詞組詞規則來增加變調處理的正確性。

系統雖完成初步成果，但離自然流利的口語仍很遠。以下提出了未來應該改進的幾個方向：

- 語法的分析對提昇斷詞、變調組位置及字轉音選哪音的正確性有很大的幫助。尤其是斷詞結果必與轉調相關聯，也就是對於合成語音的自然度有很大的影響。因此，一個包含語法分析器的文句分析模組是大家該努力的方向。
- 漢語的特色是單音節多聲調，當可嚐試建立一套共同技術系統，來直接開發各種漢語方言的發音系統。

7. 參考文獻

[1] 鄭良偉，鄭謝淑娟合編，「台灣福建話的語音結構及標音法」，學生書局台北，民 76 年 8 月四刷

[2] 楊青矗，「國台雙語辭典」，敦理出版社，台北市，1996.2 五版

[3] 王育德，「台灣話講座」，台北， pp. 34，自立出版社， 1993.5

[4] 陳世達，「應用音中仙中文聽寫機之全球資訊網語音輸入查詢系統」，成大電機研究所碩士論文, 民國 87 年 6 月

[5] 歐陽明、李琳山，「一套中文的文句翻語音系統」，台大電機研究所碩士論文, 1985 年 6 月

[6] L. R. Rabiner and R. W. Schafer, “Digital Processing of Speech Signals”, p398-p403, 1987.

[7] 劉繼謐等，「以線性預測編碼為合成器的中文文句翻語音系統」，電信研究季刊，第 19 卷第 3 期，民國 78 年 9 月

[8] S. Singhal and S. Atal, “Amplitude Optimization and Pitch Prediction in Multipulse Coders”, IEEE Trans. Acoust, Speech, Signal Processing,

ASSP-37, pp.317-327, March 1989

[9] 朱國華等，「一套以多脈衝激發語音編碼器為架構之即時中文文句翻語音系統」，電信研究季刊，第 21 卷第 4 期，民國 80 年 12 月

[10] J. Charpentier and M.G. Stella, “Diphone Synthesis Using an Overlap-Add Technique For Speech Waveforms Concatenation”, Intern. conf. on ASSP, ICASSP-86, pp. 2015-2019, 1986

[11]Christian Hamon, Eric Moulines, Francis Charpentie, “A diphone synthesis based on time-domain prosodic modifications of speech”, ICASSP, pp.238-241,1989

[11] 施炳華，「台語教學法」，啟人書局，台南市，民 82

[12] 林繼雄，「台語教學法」，大夏出版社，台南市，民 79

[13] 方南強，「大家來說台灣母語（閩南語篇）」，時報出版公司，1996.5 初版七刷

[15] Chiang Yuang-Chin, et. al., “A New Hybird Duration Hidden Markov Model with Application to Large Vocabulary Taiwanese(Min-nan) Word Recgonition,” 1st International Symposium on Chinese Spoken Language Processing(ISCSLP98),Dec, 1998, Singapore .

[16] Ren-yuan Lyu,Yuang-jin Chiang, Ren-zhou Fang, Wen-ping Hsien, “A Large-Vocabulary Taiwanese (Min-nan) Speech Recognition System based on Inter-Syllabic Initial-Final Modeling and Lexicon-Tree Search” , ROCLING XI Conference 1998, p. 139-149

[17] Ren-yuan Lyu, Yuang-jin Chiang, Wen-ping Hsieh, “A Large-Vocabulary Taiwanese (Min-nan) Multi-Syllabic Word Recognition System based upon Right-Context-Dependent Phones with State Clustering by Acoustic Decision Tree”, ICSLP1998 .

[18] 林川傑&陳信希，中文到閩南語之線上翻譯及閩南語之語音合成，1999 語文處理技術研討會集刊。

[19] Y.J.Sher, K.C. Chung, C. H. Wu, ”Taiwanese Syllable-based Synthesis Units Database”, Accepted by Chinese Journal of Medical and Biological Engineering (CJMBE-233),1999,volume 19,no 1, page47-58.

[20] Y.J.Sher, K.C. Chung, C. H. Wu, “Establish Taiwanese 7-Tones Syllable-based Synthesis Units Database for the Prototype Development of Text-To-Speech System” , Accepted by ROCLING XII .

[21] 鄭良偉，「台語的語音和詞法」，遠流出版社，台北，民 86 年出版。

Semantic classification for Patterns Containing Non-Text Symbols in Mandarin Text

Feng-Long Hwang[12], Ming-Shing Yu[1],Ming-Jer Wu[1], Shyh-Yang Hwang[1]

[1]TTS Lab., Department of Applied Mathematics, National Chung-Hsing University
Taichung, Taiwan, R.O.C. 402, Tel: +886-4-2860133 ext: 609
[2]National Lien-Ho Institute of Technology, Miauli, R.O.C,Tel:+886-37-332997
flhwang@amath.nchu.edu.tw, {msyu,mjwu,syhwang}@dragon.nchu.edu.tw

ABSTRACT

In this paper, we address the semantic classification of non-text symbols in Mandarin text using multiple decision classifiers. Some non-text symbols (e.g., "/" and ":") appear frequently within the Mandarin texts (such as newspaper, magazine and files in Internet). Usually, these symbols in sentence may have more than one possible oral expression. In contrast to *2-gram*, *3-gram* and *n-gram* language models, the paper proposes the multiple layer decision classifiers, which can resolve the category ambiguities of oral expression for patterns containing one or several non-text symbols in Mandarin texts efficiently. There are two principal phases in our proposed approach: training phase and classification phase. Currently, classification phase contains two decision classifiers. We can predict the correct category of the non-text symbols then translate the non-text symbols into correct oral expression further. The empirical precision rates for inside and outside test are *97.8% and 93.0%* respectively.

1 Introduction

The goal of Text-To-Speech (TTS) system is to translate the text input into correct Mandarin speech. There are three principal phases in a TTS system: 1) text analysis, 2) prosody generation and 3) speech synthesis phase. The task of text analysis is to analysis the syntax and semantic information of text and to generate the phonetic transcription and part-of-speech (POS). The prosody generating is to generate the prosodic feature of text, such as duration, speech energy and pitch. The phase of speech synthesis, which should transforms the prosodic feature and synthesis units in the acoustic inventory according to the prosody of speech, is to generate the output of Mandarin speech with clear intelligibility and great comprehensibility. The acoustic inventory may contain about 400 synthesis units with monotone or 1345 synthesis units with 4 tones (tone 1, 2, 3 and 4) in Mandarin speech.

Within the process for translating text to speech output, one situation is frequently encountered: because of existence of homograph words and non-text symbols, there are several possible different oral expressions based on its contextual information and non-text symbols in sentence. There are some non-text symbols (e.g., "／" and "：") within the Mandarin texts (such as newspaper, magazine and files in Internet). For example, the pattern of “2/3” can be translated into “February three ” or “two third”; and the pattern of “9:15” may be translated into “nine versus fifteen” or “fifteen minutes past nine”. The pattern of "3/5" in (A) is categorized into *date* category(三月五日) while pattern of "3/5" in (B) into *fraction* category(五分之三) (A') and (B') are the oral expression with respect to (A) and (B). Some major types of homographs are listed in [Yarowsky,1997].

(A) 3／5，電算中心出版使用手冊。
March 5th , Computer Center publish the users' manual.

(A') 三月五日，電算中心出版使用手冊。
Suan1 yue4 wu3 r4, dian4 suan4 jung1 shin1 chu1 bian3 shi3 yuan4 shou3 che4.

(B) 產品價格比台灣的價格便宜3／5左右。
Products' price is less about three-fifth than that in Taiwan.

(B') 產品價格比台灣的價格便宜五分之三左右。
Chan2 pin3 jia4 ge2 bi3 tai2 wan1 de jia4 ge2 pian2 yi2 wu3 fen1 jr1 suan1 tzuo3 you4.

The Academic Sinica Balanced Corpus version 3.0 (ASBC) [黃居仁等,1995] includes 317 text files distributed in different topics, occupying 118MB memory and 5.22 millions of words totally. In ASBC, sentences have been segmented into several words (詞, or so-called lexicons) based on corpus of Academia Sinica Chinese Electronic Dictionary (ASCED), and each word in the sentence is tagged with its related part-of-speech (POS). There are several kinds of non-text symbols (such as /, %, :, X ,…., and so on). Each non-text symbol may have different meanings subject to the syntax and semantics, such situation (like sentence A & B above) is so-called oral ambiguity. Different semantic category for each non-text symbol should be translated into its related oral expression. On the other hand, there is a one-to-many possible correspondence between a non-text lexical symbol and its possible semantic meanings. Whether the real meanings of non-text symbols can be expanded into its oral expression or not will affect seriously the correct output of Mandarin speech in TTS system. Based on the linguistic knowledge and usage of prosody in TTS systems, the possible semantic categories of non-text symbol slash “／” are classified and shown on Appendix A.

The so-called **non-text symbols** are defined as follow: the symbols that are not the Mandarin characters and have several different semantic meanings and oral expressions within a sentence. Such symbols including some punctuation (such as ":", “ • ” , and so on) will be found in text frequently.

The paper is organized as follow: in section 2, we first present previous works and then addresse the overall structure of proposed approach. Section 3 focuses on the multiple decision classifiers. Section 4 displays the empirical the testing results of evaluation. Finally, we will present the conclusions and future works.

2 The Proposed Approach

2.1 previous works

There are several methods that resolve the classification problems of linguistic and semantic ambiguity for natural language processing :

1) N-garm taggers: [Merialdo,1990] may be used to tag each in a sentence with its part-of-speech (POS), thereby resolving those pronunciation ambiguities.

2) Bayesian classifiers: Bayesian have been used for a number of sense disambiguation. An implementation was proposed in [Golding ,1995].

3) Decision tree: [Brown ,1991] can be effectively at handling complex conditional dependencies and nonindependence, but often encounter severe difficulties with very large parameter space.

4) Hybrid methods : [Yarowsky,1997] combines the strengths of each of preceding paradigms. It is based on the formal model of decision tree.

5) Multiple Decision classifiers: [Rodova,1997] take interest in speaker identification.

2.2 The Proposed System Structure

The system structure is shown as Figure 1. It contains two principal phases:1) **training phase** and 2) **classification phase**. In the training phase, the feature corpus will be trained using several parameters of linguistic knowledge of the pattern containing non-text symbols. In the classification phase, the patterns containing non-text symbols in sentence will be classified using the multiple decision classifiers, in which the output of predicted category will be sent into the translating phase to translate the pattern to correct oral expression. The output text can be processed for linguistics analysis further, which could promote the overall performance of TTS system. In contrast to *2-gram*, *3-gram* and *n-gram* Language models (LMs), this paper proposes an approach of multiple decision classifiers which can resolve the category ambiguity of oral expression for non-text symbols efficiently. In multiple decision classifiers, currently we have generated two classifiers: the first decision classifier is constructed as decision tree under the linguistic knowledge and plays as a binary function. Within first classifier some impossible categories will be excluded and all remaining categories are the promising categories. The second classifier employes statistical method, in which all the words (lexicons) in sentence play as voter under voting criterion and vote for each category with statistical parameters.

These multiple decision classifiers are combined together with *multiply* operation. Like the political mechanism, all voters will give their suffrage to each category with a statistical score. Finally the category with maximum voting score can be predicted as the goal category for non-text symbol. Basically, the decision tree classifier is generated according to linguists' experience and knowledge. The remained categories are all the possible categories that the non-text symbol may belong to.

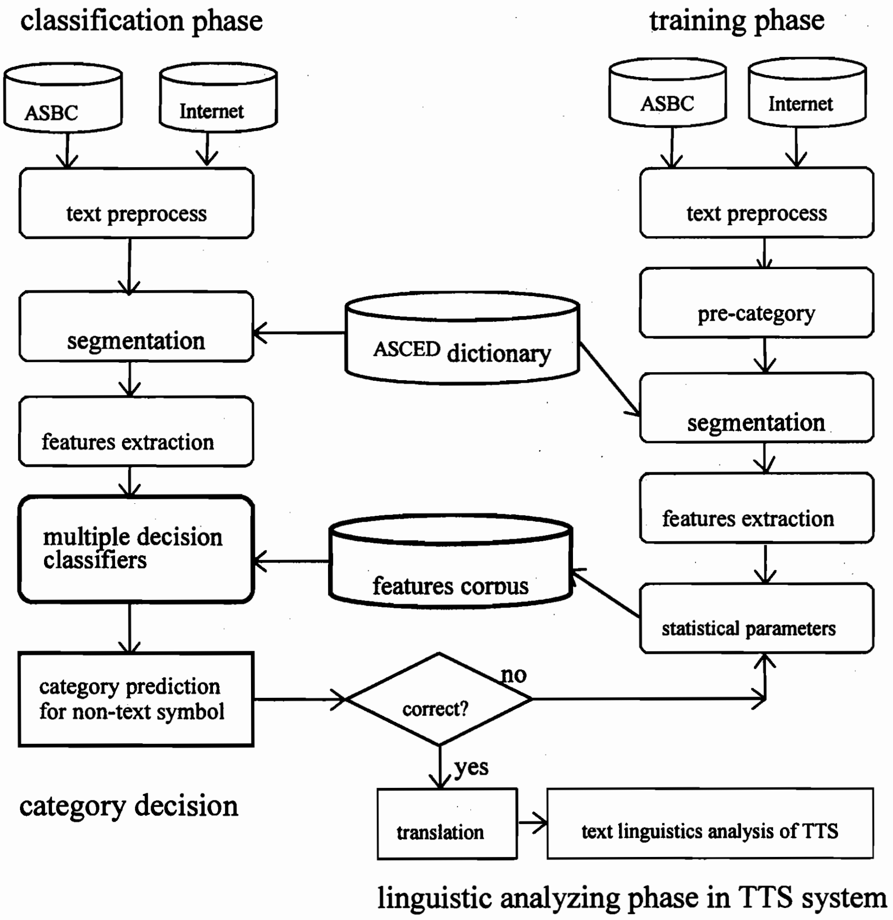


Figure 1:The principal phases of statistical decision classifier with voting criterion.

2.3 Training Phase

i) The text preprocess

The Academic Sinica Balance Corpus (ASBC) contains 317 text files and 4.55M characters in Chinese Mandarin [黃居仁等,1995]. Each sentence in original ASBC is tagged with part-of-speech (POS) and segmented into several words, the tags and white separation (space) between words will be removed during processes. In the text preprocesses, we further collect and download the more text from HTML source and BBS posted papers, and then remove all the HTML tags (such as <HTML>, <P>, <A href=" ....", and so on) and other unnecessary symbols in these files.

ii) The pre-category of each non-text symbol

The text source for training phase can be extracted from ASBC and Internet HTML and BBS files semi-automatically. First, we category the source for each non-text symbol, the extracted sentences will be distributed into one or several categories related the symbol based on the lexical and semantics knowledge. The eight possible categories for non-text symbol "／" are listed in Appendix A.

iii) Segmentation

Word segmentation paradigm is based on the Academia Sinica Chinese Electronic Dictionary (ASCED), which contains near 80,000 words. The words in ASCED are composed of one to 10 characters. Our principal rules of segmentation are subject to maximal length of word first and then to least number of words in a segmented pattern based on the **dynamic programming method** (Viterbi searching). The priority scheme is that segmented pattern which contains the maximal length of word will be chosen. If two patterns have same maximum length, we compare further the total number of words in the pattern; then the pattern that is composed of least number of words will be chosen. The same segmentation's priority will be used within the training phase and testing phase.

iv) Constructing corpus for statistical parameters

After the segmentation for *CHa* and *CHb*, the feature of each word will be used as the statistical parameters, all of which will be recorded in the training corpus statistically. Each record contains the four feature evidences explained above.

2.4 Classification Phase

i)The text preprocess

Text preprocess in this phase process the same task as that in training phase.

ii) Segmentation

segmentation task in classification phase uses same criterions as that used in training phase shown in precious section also. A sentence with non-text symbols will be divided into substring *CHa* and *CHb.* For each word, the probability of each category can be calculated and summed up based on the parameters found in feature corpus respectively.

iii) The features extraction

Feature extraction in this phase does same task as that in training phase.

iv) Multiple decision classifiers

The goal of multiple decision classifiers is to predict the correct category, to which the non-text symbols belong. The structure details will be described in next section.

Within the classification phase, some categories output in sentence could be mispredicted. To make the multiple decision classifier more robust, these sentences can be sent back into statistical parameter process in training phase and adapts dynamically the parameters of feature corpus to raise the precision rate. The feedback usually can solve the unseen events (words) in training text, the situation of unseen words often appears in natural language processing.

3 The Multiple decision classifiers

3.1 The Structure of Multiple decision classifier

In contrast to *2-gram*, *3-gram* and *n-gram* Language Models, this paper proposes an approach of multiple decision classifiers, which can resolve the category ambiguity of oral expression for non-text symbols efficiently. Within the classification phase, we have constructed two classifiers: the first decision classifier is generated and shown as decision tree based on the linguistic knowledge. Some impossible categories will be excluded while the remaining categories are all the promising categories. The second classifier employes a corpus statistics-oriented technique to estimate the final category with maximum score. All the words (lexicons) in sentence play as voter under the voting criterion and vote for each category with statistical parameters score.

These multiple decision classifiers are combined together with *multiply* operation. Like the political mechanism, all voters will give their suffrage to each category with a statistical probability score. Finally, the category with maximum statistical parameters score can be predicted as the goal category for non-text symbol. The overall system structure of multiple decision classifiers is shown as Figure 2.

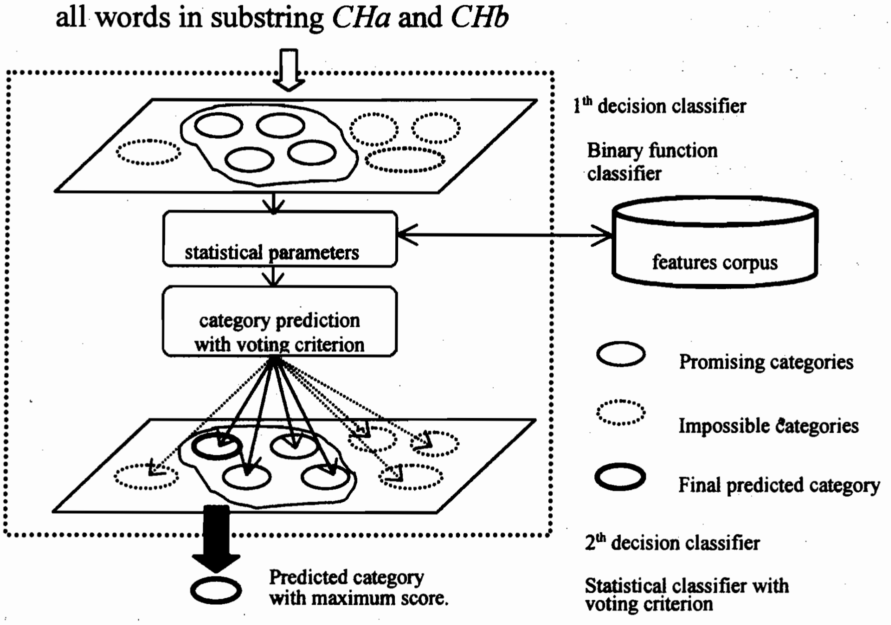

all words in substring *CHa* and *CHb*
1th decision classifier
Binary function classifier
statistical parameters
features corpus
category prediction with voting criterion
Promising categories
Impossible Categories
Final predicted category
2th decision classifier
Statistical classifier with voting criterion
Predicted category with maximum score.

Figure 2: Multiple decision classifiers contain two classifiers, which are merged together with *multiply* operation.

The function of multiple decision classifiers can be described as follow:

Suppose that C denotes the sentence with non-text symbols, $\Phi_1$and Φ_2 denote the 1st and 2nd classifier respectively. *set* is the set containing all promising categories induced by 1st classifier. Φ denotes the multiple decision classifiers, which is composed of the 1st decision tree classifier and 2nd statistical decision classifier merging with *multiply* operation. *TS*(•) will compute the total score for all categories based on the voting criterion and statistical parameters schemes.

$$\Phi_1(C) = set, \tag{1}$$

$$\Phi_2(C) = \mathrm{TS}(\Omega_j), \qquad \Omega_j \in set \tag{2}$$

$$\Phi(C) = \Omega_{j^*}, \qquad \exists!\Omega_{j^*} \in set \text{ and } \mathrm{TS}(\Omega_{j^*}) = \arg\max_{j=1,2,\ldots J} \mathrm{TS}(\Omega_j) \tag{3}$$

where j is the number of category for non-text symbols.

3.2 The Binary Function Classifier based on Decision Tree

The decision tree classifier plays as a binary logical function, which is to induce all promising categories for the non-text symbol based on Mandarin linguistic knowledge. The classifier will assign probability value 1 to the promising categories. On the other hand, some categories will be excluded and assigned a probability value 0. For example, the pattern of "3/4"may belongs to several possible categories: *date* (March 4th), *fraction* (three fourth) and *tempo*(three slash four pulses) , these categories will be assigned a value 1. But the pattern of "14/2" and "SUN4/75" could not belong to the category *date* and *tempo,* all these categories will not be the possible category for non-text symbol and be assigned a probability value 0.

A successive answers to questions: $Q_1, Q_2, \cdot\ \cdot\ \cdot, Q_n$, which are the questions about the syntax and semantic meaning for left and right neighbor (tokens or words) of non-text symbol in sentence, will decide which path should trace into based on the linguistic knowledge. Finally, one leaf node in decision tree will be reached and a *set* of categories will be contained. Within the *set*, all the categories will be assigned a probability value 1 while all other categories will be assigned a value 0. The key point for constructing an effective decision tree is how to exploit the linguistic knowledge and the skill of making decision tree. All possible categories should be keep inside the *set*, otherwise the precision rate will be reduced. In our proposal, the probability value for each category cab be described as follow:

$$\underset{i=1,2}{P_i}(\Omega_j) = \begin{cases} 0 & \text{if } \mathrm{i} \in 1 \text{ and } \Omega_j \notin set. \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

where i=1,2, …,I. i is labeled as the ith decision classifier. I is the number of total decision classifiers (currently, we have developed two decision classifiers, so I=2). If we have *J* categories $\Omega_1\Omega_2\cdot\ \cdot\ \cdot\Omega_J$ and Ω_j denotes the category j for non-text symbols. $P_i(\Omega_j)$ is the probability value of category j for the ith classifier . *set* is induced from the decision tree classifier and contains all promising categories. These promising categories will be passed into 2nd decision classifier further, one of which will be the final predicted category. First classifier plays as a binary function in our approach. So, Equation (4) can be explained further as follow: if i=1 and $\Omega_j \notin set$, then $P_1(\Omega_j) = 0$. Otherwise, $P_i(\Omega_j) = 1$.

Basically, the decision tree classifier is generated according to linguists' experience and theories. The remained categories are all the possible categories that the non-text symbol may belong to. Thus, the voting approach can predict the only one among possible categories. It is so apparent that processing of adopting decision tree can improve the precision rate.

3.3 The Statistical Decision Classifier with voting criterion

The segmentation task of testing phase adopts same criterions as that in training phase shown in section 2.3. A sentence will be divided into substring *CHa* and *CHb.* For each word, the probability of each category can be calculated and summed up based on the evidence (parameters found in feature corpus) respectively. It is called the **voting criterion**.

Based on the *voting criterion,* each word in CHa and CHb have a statistical probability value, which looks like the voting suffrage, to every category of the non-text symbol. Like the political voting mechanism, the only category, which gets the tickets in majority (maximum score in our approach) will become to be the predicted category. In our voting criterions, three scoring schemes are proposed: which are the *preference scoring* and the *winner-take-all* criterion. These voting criterions will be implemented and compared with each others to find which one can achieves the best empirical results.

3.3.1 Voting criterion with preference scoring

The prediction processing is based on the occurrence of each word inside training corpus for each category. Usually, the sentence C is composed of three parts: substring CHa ,non-text symbol N and substring CHb. C, CHa and CHb could be expressed as:

$$\begin{aligned} &C = CH_a + N + CH_b \\ &CH_a = w_{a_1} w_{a_2} \cdot \cdot \cdot w_{a_j} \cdot \cdot \cdot \mathrm{w}_{\mathrm{a_m}} \\ &CH_b = w_{b_1} w_{b_2} \cdot \cdot \cdot w_{b_j} \cdot \cdot \cdot \mathrm{w}_{\mathrm{b_m}} \end{aligned} \tag{5}$$

where a_m and b_n are the total number of words in CHa and CHb respectively. It is apparent that CHa and CHb contain one or several different non-text symbols. Also, CHa and CHb may be an empty substring.

For each word in CHa and CHb, the word appearance probability appearing in category j of non-text symbol can be computed based on three different statistical parameters scheme: which are word-based, category-based and corpus-based . In this work, the word appearance probability can be considered as the probability the word may appear in certain category for non-text symbol. The appearance probability can be regarded as a score for each word in CHa and CHb to vote for each category of non-text symbol further.

There are three statistical probability schemes, on which the value can be considered as the probability for each word to appear in each category.

(1) word-based statistical probability

For all words in CHa and CHb, the appearance probability score S_a and S_b of each word voting for category j $(\Omega_j)$of non-text symbol can be computed as:

$$S_a(w_{ak_1} \mid \Omega_j) = \frac{C_a(w_{ak_1} \mid \Omega_j)}{TN_a(w_{ak_1})}, \quad S_b(w_{bk_2} \mid \Omega_j) = \frac{C_b(w_{bk_2} \mid \Omega_j)}{TN_b(w_{bk_2})} \tag{6}$$

where $1 \le k_1 \le m$ and $1 \le k_2 \le n$, w_{ak_1} and w_{bk_2} are labeled as the k_1th and k_2th word in CHa and CHb. $C_a(w_{ak_1} \mid \Omega_j)$ and $\mathrm{C}_b(w_{bk_2} \mid \Omega_j)$ are the occurrence of w_{ak_1} and w_{bk_2} for category j of non-text symbol. $TN_a(w_{ak_1})$ and $TN_b(w_{bk_2})$ stand for the total frequency of w_{ak_1} and w_{bk_2} within features corpus with respect to the location proceeding and following non-text symbol, which can be computed as follow:

$$TN_a(w_{ak_1}) = \sum_{j=1}^{J} C_a(w_{ak_1} \mid \Omega_j), \quad TN_b(w_{bk_2}) = \sum_{j=1}^{J} C_b(w_{bk_2} \mid \Omega_j) \tag{7}$$

$$\sum_{j=1}^{J} S_a(w_{ak_1} \mid \Omega_j) = 1 \ , \qquad \sum_{j=1}^{J} S_b(w_{bk_2} \mid \Omega_j) = 1 \tag{8}$$

Based on the definition above, $S_a(w_{ak_1} \mid \Omega_j)$ and $S_b(w_{bk_2} \mid \Omega_j)$ can be considered as the probability value in which the w_{ak_1} and w_{bk_2} will appear in the category j . As the result, our voting criterions are based on this probability value.

In the paper, $S_a(w_{ak_1} \mid \Omega_j)$ and $S_b(w_{bk_2} \mid \Omega_j)$ stand for the suffrage for each word (voter) to vote for certain category j (Ω_j) .

(2) category-based statistical probability

With respect to Equation (6), the denumerator will be computed based on the total occurrence for the all words which appear in category j (Ω_j) . Equation (8) can't hold in this scheme.

(3) corpus-based statistical probability

With respect to Equation (6), the denumerator will be computed based on the total occurrence for the all words which appear in feature corpus. Equation (8) can't hold in this scheme.

For the 2nd decision classifier, the total score TS_a and TS_b for all words in substring *CHa* and *CHb* to vote for categories j of non-text symbol can be computed.

The overall total score TS of 1st and 2nd decision classifier for category j is computed with the *multiply* operation:

$$\underset{j=1,2,\ldots J}{TS(\Omega_j)} = P_1(\Omega_j) * \ P_2(\Omega_j) * TS(\Omega_j) \ , \ \ \Omega_j \in set \tag{9}$$

where $P_2(\Omega_j)$ denotes the probability value of category j (Ω_{j^*}) in the 2nd classifier. In our approach, $P_2(\Omega_j) = 1, \mathrm{j} = 1,2,....,\mathrm{J}$. *set* is composed of all the promising categories induced by 1st decision tree classifier.

$$TS(\Omega_{j^*}) = \underset{j=1,2,\ldots,J}{\arg\max}\,(TS(\Omega_j)) \tag{10}$$

where $TS(\Omega_{j^*})$ will return the maximum score subject to category j* (Ω_{j^*}) based on 1st decision classifier and 2nd statistical decision classifier. $TS(\Omega_{j^*})$ will be used in Equation (3) for the multiple decision classifiers to predict the final category j*(Ω_{j^*}) .

3.3.2 Voting criterion with winner-take-all scoring

In construct to the preference scoring criterion above, the Voting with *winner-take-all* adopts a different scoring rule. For each word in *CHa* and *CHb*, $S_a(w_{ak_1} \mid \Omega_j)$ and $S_b(w_{bk_2} \mid \Omega_j)$ will have the total parameter score 1 of category j* for word w_{ak_1} and w_{bk_2} and assigned a score value 1. S_a (similar to S_b) in Equation (6) should be changed as follow:

$$S_a(w_{ak_2} \mid \Omega_{j^*}) = \begin{cases} 1 & \exists ! \Omega_{j^*} \in \Omega \text{ and } S_a(w_{ak_1} \mid \Omega_{j^*}) = \underset{j=1,2,\ldots J}{\arg\max}(S_a(w_{ak_1} \mid \Omega_j)) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Within the classification phase, some sentences could be mispredicted. To make the statistical decision classifier more robust, these sentences can feedback into category process in training phase and adopt parameters in features corpus. The feedback usually cab solve the unseen events in source, the situation appear often in natural language.

3.4 Unknown word

There are a lot of words in natural language, usually more several ten thousands. New lexicons or tokens will be generated in near future. Within natural language processing (NLP), it is so hard to collect all the words. In our paper, the so-called unknown words can be considered that the words do not appear in our corpus (feature corpus), which have been generated in the training phase. It is so apparent that the distribution and total number of collected word will affect the statistical parameters seriously, especially on the statistical models. Another situation is the data sparsity. The smoothing techniques can resolve the situation.

Based on the ASBC and ASCED corpus, the ASBC source is divided into four groups. we compute the total frequency and number of words in these four groups to derive the relation, in which we can predict the probability unknown words. The fitting regression curve can be employed to estimate the probability for unknown words. $Y(X) = aX^2 + bX + c$. We can find the derivative of $Y(X_j)$. Within classification phase, Value X_j represent the total frequency of collected words in feature corpus for category j $(\Omega_{j}.)$. The first order derivative of $Y(X_j)$ can be considered that the probability of unknown word in category j $(\Omega_{j}.)$. Such probability will be used as voting score for unknown words to vote for category j.

3.5 Translate Oral Expansion

The output of multiple decision classifiers is the unique predicted category. Based on the category, the non-text symbol can be translated into its oral expression of text in which the category has been predicted by testing phase. Sentence (C) contains a non-text symbol "/", which is predicted as the *date* category and the pattern of "4/10" in (C) will be translated into the oral expression "四月十日" in sentence (C'). The output text of this phase will be processed further with text linguistic analysis in TTS system.

(C) 這本雜誌已於上週六（4∕10）出版。

This magazine was published last Saturday (April tenth).

(C') 這本雜誌已於上週六（四月十日）出版。

Je4 ben3 tza2 jr4 yi3 yi2 sang4 jou1 liou4 sz4 yue4 sz4 r1 chu1 bian3.

4 Implementation and Evaluation

Our approach has been implemented on a platform of personal computer (PC) with Intel Pentium III. The language package for system development is in C++ environment. Two decision tree classifiers have been generated. We evaluate the results of inside test and outside test for 2nd *statistical classifier* with two different *voting criterions*, then we combined it with *decision tree classifier* to compare the performance of precision rate. The precision rate is defined as:

Precision rate(PRs) = # of correct prediction categories ⁄ total # of non-text symbol (12)

4.1 Evaluation only for statistical decision classifier

The results for 2nd classifier with different voting score criterion and statistical parameters are listed in Table 1. Total number of non-text symbol "/" for inside and outside test are 564 and 202 respectively.

4.2 Evaluation for merging two decision classifiers together.

Under the multiple decision classifier structure, the 1st and 2nd decision classifier are merged together to improve the overall precision rate. exploiting the 1st classifier to exclude some impossible categories first, the results are attractive and listed in Table 1 also. As shown, the final results of inside test and outside test is 97.3% and 92.9%, which are obtained by merging the 1st and 2nd classifier with voting criterion of preference score and category-based statistical parameters in 2nd classifier.

Table 1: The overall precision rate of inside test and outside test of 2nd statistical decision classifier for symbol "/"

Precision rate(%)	multiple decision classifier, merging or not?	2nd decision classifier, word-based statistical scheme			
		voting with preference score		winner-take-all score	
		inside test	outside test	inside test	outside test
word-based statistical scheme	without 1st classifier	95.4	86.3	85.9	77.3
	with 1st classifier	96.2	91.2	90.5	85.7
category-based statistical scheme	without 1st classifier	96.0	92.8	92.9	84.8
	with 1st classifier	97.3*	92.9*	96.1	89.4
corpus-based statistical scheme	without 1st classifier	95.5	86.1	89.2	81.1
	with 1st classifier	96.3	89.9	90.1	85.5

Table 2 is the results for non-text symbol " : ", based on the preference score voting criterion and word-base statistical parameters. The average rate of inside testing and outside testing are 97.8% and 93.0%. Notation of *N* in Table 2 stand for non-text symbol. The total word occurrence for non-text symbol " : " is 14406.

Table 2: The overall precision rate of inside test and outside test for non-text symbol " : ", the 1st and 2nd decision classifier merging.

multiple decision classifier	voting with preference score , word-base statistical parameters													
	inside testing							outside testing						
category	1	2	3	4	5	6	7	1	2	3	4	5	6	7
PRs rate(%)	99	100	98	95	100	100	97	86	100	100	88	100	78	97
Total no. of *N*	272	105	126	21	85	35	351	68	31	30	8	22	9	83

5 Conclusion and Future Works

In the paper, we have developed an effective approach, which can classify the semantic category of patterns containing non-text symbols and resolve the category ambiguity in Mandarin text. In contract to the *2-gram* and *n-gram* Language Models, our approach just need smaller size of corpus and still can hold the semantic and linguistic knowledge for statistical parameters and features. Currently, we have developed two decision classifiers: one is based on the decision tree to induce promising categories the other is on the statistical decision classifier with two voting criterion with word-based, category-based and corpus-based statistical parameter schemes. Final precision rate of inside and outside test achieves the performance of 97.8% and 93.0% respectively.

In addition to the non-text symbols "／" addressed in the paper, there are some other symbols, such as *, %, [] and so on, in which the oral ambiguity problems will be incurred and should be resolved. The topics which should be researched further in the future include:

1) Patterns of special and frequent cases for non-text symbols in text.
2) The extraction training parameters and learning algorithms.
3) The POS of word and smoothing techniques for unknown words.
4) Expand the current two classifiers into more classifiers to resolve complicated linguistic classification problem.

References

- 黃居仁等, *中央研究院平衡語料庫簡介*, Proceeding of ROCLLING VII, pp. 81-99, 1995.
- P. Brown, S. Della Pietra, V. Della Pietra and R. Mercer. *Word Sense Disambiguation Using Statistical Methods*. In Proceeding of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, pp. 264- 270, 1991.
- A. R. Golding, *A Bayesian hybrid method for Context-Sensitive Spelling Correction*, In Proceedings of the third workshop on Very Large Corpora, pp. 39-53, Boston, USA, 1995.
- B. Merialdo, *Tagging Text with a Probabilistic Model*, In Proceeding of the IBM Natural Language ITL, Paris, France, pp. 161-172, 1990.
- V. Rodova and J. Psutka, An Approach to Speaker Identification Using Multiple Classifiers, ICASSP, 1997,pp. 1135-1139.
- David Yarowsky, *Homograph Disambiguation in Text-to Speech Synthesis*, Book of Progress in Speech Synthesis, 1997, chapter 12,pp 157-172.

Appendix A: 8 categories and its related oral expression for non-text symbol slash "／".

category	some lexical patterns with non-text symbol "／"	oral expression in Mandarin
1. date	３／４	三月四日
2. fraction	３／４	四分之三
3. tempo	３／４	四分之三拍
4. path, directory	／ｄｅｖ／ｎｕｌｌ	根目錄ｄｅｖ斜線ｎｕｌｌ
5, computer words	Ｉ／Ｏ	silence(or 斜線)
6. production version	ＶＡＸ／ＶＭＳ	silence(longer pause or 斜線)
7. frequent words in	ＴＣＰ／ＩＰ	silence(or 斜線)
8. others	中／日／韓文著錄	silence(longer pause)

動詞詞構與語法功能互動初探

張麗麗、陳克健

中央研究院資訊科學研究所

Email:lili@iis.sinica.edu.tw, kchen@iis.sinica.edu.tw

摘要

漢語的詞類劃分存在一個特殊現象：詞類和語法功能之間並非一對一的關係。一個詞的詞類通常表示該詞的主要語法功能，但並不表示該詞就不能扮演其他的語法功能[1]。同屬一個詞類的詞也不表示它們所能扮演的語法功能都相同。也就是說，即使知道一個詞的詞類，也並不一定就能準確掌握該詞在個別語句中所扮演的語法功能。這一點對於語句剖析的確造成一定的困擾。這個現象尤以動詞最為紛亂，因此本文將提出一些協助判定的標準以利掌握動詞在個別語句中的語法功能。由於雙音節動詞的語法功能分佈比起單音節動詞來得複雜多變，所以本文的研究對象以雙音節動詞為主。在文中我們將先探討動詞次分類，也就是動作動詞和狀態動詞大致的語法功能分佈情況。然後探討詞彙內部結構和語法功能分佈之間的關係，這是本論文的焦點所在。我們認為動詞的內部結構能夠幫助判定其語法功能分佈的大體趨勢，並從「中研院平衡語料庫」(CKIP 1995) 抽取資料統計，說明這方面的現象。文末我們將從詞彙語義的角度就詞構對動詞語法功能分佈的影響提出一些解釋。

1. 前言

在現代漢語中，常常可以看到同一個詞扮演許多不同的功能，尤其以動詞

[1] 在有詞形變化的語言中，一個詞扮演不同功能時，多半都有不同的詞尾變化。例如在英語中，happy 是個形容詞，能作述語和定語；happiness 是名詞，只作主賓語；happily 是副詞，只作狀語。也就是說，英語中的每個詞類通常只代表單一的語法功能，因此憑藉詞類就能做出正確的語法功能判定，比漢語單純得多了。在漢語中，「快樂」無論用作述語、定語、主賓語或狀語都沒有任何詞型上的變化，但是一般作法都是將「快樂」歸類為狀態動詞，因此從詞類並不能輕易地判斷它在個別用法中的語法功能。

的變化最大。以「痛苦」為例，它是一個狀態動詞，除了作述語外，它還可以作補語、狀語、定語，甚至主賓語[2]，如例(1)至例(5)。

(1) 述語：可是我很痛苦。
(2) 補語：在清華，不練到此等功力，你會活得很痛苦。
(3) 狀語：一遍一遍，對著其間的空白痛苦嘶吼。
(4) 定語：解除痛苦強度的第一個步驟是面對那個痛苦源。
(5) 主賓語：縱使已經醫藥罔效，還是盡量要減輕他的痛苦。

但是並不是所有的狀態動詞都可以扮演這麼多種功能，例如「難過」也是狀態動詞，而且和「痛苦」意義相近，可是它卻很少作補語、定語和主賓語。甚至於現代漢語中還有很多動詞就只能用作述語，像是「打破、喚醒」。為什麼「痛苦」可以扮演這麼多的語法功能，「難過」多半只用作述語或狀語，而「打破、喚醒」卻只能用作述語？究竟一個動詞所能扮演的語法功能是由什麼所決定的？這就是本文所要探討的主題。

這樣的研究對於自然語句的自動剖析將有一定的幫助。因為唯有先判斷出句中每個詞的語法功能，才能判斷出這些詞之間的關係。我們常常見到同一個句子中含有兩個以上的動詞，但是通常只有一個作述語，其他的動詞則扮演別的語法功能。在這樣的情況中，如果我們知道每個動詞所能扮演的功能，對於語句的剖析將有一定的幫助。例如，下面的例句含六個詞，其中竟有五個是動詞，但其實只有「體驗到」和「躲開」作述語，「深刻」是作狀語，「糾纏」和「快樂」則作主賓語。在這個句子中，除了「的」能幫助我們判斷出「快樂」是中心語，前有修飾語外，沒有其他的功能詞能幫助判斷此句的結構。因此，如果系統能夠先判斷出「體驗到」和「躲開」只能用作述語，「深刻」、「糾纏」、「快樂」可以扮演很多語法功能，再來分析這個句子就容易得多了。

(6)	深刻	體驗到	躲開	糾纏	的	快樂。	
	動詞	動詞	動詞	動詞		動詞	→詞類
	狀語	述語	述語	主賓語		主賓語	→功能
	[狀	中心語	[[述	賓	]$_{\text{關係子句}}$	中心語]$_{\text{賓}}$]$_{\text{句}}$	→結構

在本文中，我們將只探討雙音節動詞，因為單音節動詞和雙音節動詞的功能判定所涉及的層次和所需標準並不相同。一方面，單音節動詞往往還兼有其他意義和詞類，在語法功能的判定前，還要先篩選出適當的意義和詞類，比起雙音節詞還要複雜；另一方面，純為動詞的單音節詞，其語法功能的判斷反而相當單純的。根據張國賓(1990a, 1990b)的觀察，僅就動作動詞來看，雙音節動詞

[2] 在本文中，為避免詞類和語法功能的混淆，一律以「動詞、名詞、副詞…」作為詞類的名稱，以「述語、主賓語、定語、狀語、補語」作為語法功能的名稱。

的語法功能分佈比起單音節動詞要來得複雜多變。單音節動作動詞搭配一定的詞類時，通常只扮演一個特定的語法功能，但是雙音節動作動詞卻有多種可能。例如，當單音節動詞前接名詞時，只可能作述語，像是「雞叫、雪花飄」；當雙音節動詞前接名詞時，除了作述語，像是「公雞啼叫、雪花飄揚」，也可能作主賓語，像是「貨物運輸、小說評論」。

2. 動詞次類和語法功能分佈

最重要的動詞次分類就是動作動詞和狀態動詞的區分。這個區分對於動詞語法功能的判定具有一定的幫助。基本說來，動作動詞不用作狀語或補語，只有狀態動詞才能作另一個動詞的修飾語，扮演狀語或補語的功能。

狀態動詞：除了述語用法外，可能扮演主賓語、定語、狀語、補語。

動作動詞：除了述語用法外，可能扮演主賓語、定語（、補語[3]）。

有了以上的瞭解，我們可以更有效地掌握自然語句的結構。由於只有狀態動詞可以扮演狀語和補語的功能，因此，當一個句子中連續出現兩個動詞，一個動作動詞和一個狀態動詞，那麼不論狀態動詞接在動作動詞的前面或後面，我們都可以猜測那可能是修飾語和中心語的關係[4]。

狀態動詞-動作動詞：可能是狀語修飾述語

(7) 漂亮 擊出 一 支 全壘打。

動作動詞-狀態動詞：可能述語後接補語

(8) 把 桌面 擦拭 乾淨。

但是在定語和主賓語的功能分佈上，就無法藉著動作/狀態這樣的次分類來推判。因為無論動作或狀態動詞都可能作定語和主賓語。

[3] 動作動詞中只有少數幾個動詞可用作補語，像是「去、來、走、跑、開、進、出、上、下、進去、出來」…等。在「中研院平衡語料庫」中，這些詞作補語時，是和前面的動詞合為一詞，像是「拿去、趕走」。

[4] 基於動作動詞和狀態動詞功能分佈的差異，因此在「中研院平衡語料庫」中我們對兩種動詞採取不同的詞類標記原則。因為狀態動詞用作狀語和補語是相當普遍的現象，因此當它作狀語或補語時仍然維持動詞詞類標記。由於動作動詞用作狀語是少見且不規律的，因此當動作動詞用作狀語時，會標示為副詞。在該語料庫中，共有 199 個這樣的詞。我們發現絕大部分這樣的詞其動詞和狀語的意義差別很大，應該視為不同的詞，的確應該分別標示詞類。以「比較」為例：

(1) 比較兩個時代的不同。→作述語，標示為動作動詞。

(2) 他比較謹慎。→作狀語，標示為副詞。

動作動詞作定語：閱卷老師、調查結果、求證過程、撰寫速度
狀態動詞作定語：漂亮老師、精彩結果、慘烈過程、緩慢速度
動作動詞作主賓語：警方的調查、大樓的管理、這趟飛行、一系列展示
狀態動詞作主賓語：他的痛苦、強烈不滿、幼童的安全、他的虛偽

但是並不是所有的動詞都具有上述兩種功能，有的不能作主賓語，有的不能作定語，有的就只有述語用法，例如：

除了述語外，只能作定語：加油（加油站）、代課（代課老師）
除了述語外，只能作主賓語：不悅（老闆的不悅）、自責（他的自責）
只能作述語：打破、喚醒

究竟有沒有什麼其他的方法可以推測個別動詞的語法功能分佈呢？在下面一節我們將提出一個輔助判定主賓語和定語功能的準則，那就是動詞的內部結構。

3. 動詞內部結構和語法功能分佈

在我們今年兩篇文章（Chang et al, 1999a 和 1999b)中，我們觀察到雙音節動詞的內部結構和動詞的語法功能具有一定的關係。在 Chang et al (1999a)這篇文章中，我們發現情緒動詞可以分成兩組[5]：第一組都是「非並列動詞」，主要用作述語（88.51%）[6]。「並列動詞」都在第二組，除了述語功能（30.70%）外，還常常用作主賓語（44.36%）和定語（14.20%）。這是我們首度注意到詞構，特別是並列動詞，和語法功能間的關係。因此我們將研究範圍從情緒動詞拓展到所有的動詞，在 Chang et al (1999b)這篇文章中，我們提出統計數字支持上篇研究的結論，證明並列動詞具有兩個明顯的傾向：作主賓語傾向高、作定語時修飾的名詞種類多。

[5] 我們只挑選在「中研院平衡語料庫」中出現 40 次以上的情緒動詞，第一組動詞計有：高興、開心、痛快、難過、痛心、傷心、後悔、生氣、害怕、擔心、擔憂、憂心；第二組動詞計有：快樂、愉快、喜悅、歡樂、歡喜、快活、痛苦、沈重、沮喪、悲傷、遺憾、憤怒、氣憤、恐懼、畏懼、煩惱、苦惱。除了語法功能的分佈不同外，這兩組動詞還有一些語法行為上的差異：第一組動詞修飾的中心語種類少、常後接時態標記「了」、及物性高、可以搭配祈使句和價值判斷句；第二組動詞修飾的中心語種類多、不常後接時態標記「了」、及物性低、不搭配祈使句和價值判斷句。

[6] 這個百分比是從情緒動詞的十四個代表詞的所有用法中統計出來的。我們觀察的動詞分屬七種不同的情緒類型，分別是快樂類、難過類、傷心類、後悔類、生氣類、害怕類、擔心類。每類的兩個最高頻動詞正好分屬兩組，所以共得十四個代表詞。

在本文中，我們則要廣泛探討動詞的所有內部結構，以通盤瞭解動詞內部結構和語法功能之間的關係，特別是主賓語和定語這兩個功能和詞構的關係，因為這兩個功能無法由一般所用的動詞次分類特徵來判定。

動詞的內部結構相當多，計有並列、述賓、偏正（又稱狀中結構）、述補、主述，甚至一些後接詞尾或介詞的複合動詞，如下所示：

a. 並列：調查、痛苦
b. 述賓：編班、排名
c. 偏正：不幸、反彈
d. 述補：挑出、拿開
e. 主述：心痛、地震
f. ~詞尾：美化、淨化
g. ~介詞：便於、提到

不過，現代漢語中主要還是以前四種結構為主，這可以從表一看出。「中研院平衡語料庫」中共有 30536 個雙音節動詞，我們從其中隨機抽樣五百個動詞，並計算每種詞構的數量，列於表一。從得出的結果可以看出，在現代漢語的動詞裡，並列動詞數量最多，佔了三分之一強；其次是述賓動詞，佔了四分之一；然後是偏正和述補動詞，各是 17.6%和 15.6%。其他詞構的總數佔不到 7%，其中佔 1.6%的「名詞」指的是由名詞轉化為動詞的詞。因此在以下的討論中，我們只把焦點放在前四大類詞構，即並列、述賓、偏正、述補。

	並列	述賓	偏正	述補	主述	名詞	其他	總數
數量	173	127	88	78	11	8	15	500
百分比	34.6%	25.4%	17.6%	15.6%	2.2%	1.6%	3%	100%

表一：「中研院平衡語料庫」30536 個雙音節動詞中隨機抽樣五百個動詞的詞構分佈比例

因為狀語和補語功能可以利用動詞的次分類來判斷，而且和動詞的詞構並沒有太密切的關係。所以在下面的討論中，我們將討論這四大類詞構和主賓語及定語這兩個語法功能的關係。從下列兩個小節的統計數字中，我們看到這四大類詞構和主賓語及定語的大致關係如下：

一、述補動詞不扮演定語和主賓語的功能；
二、相對地，並列動詞的這兩種功能都非常活躍；
三、述賓和偏正動詞都是在一些特殊情況下才用作主賓語；
四、述賓動詞作定語的傾向比偏正動詞強；
五、有一群述賓和偏正動詞除了述語用法外，只能用作定語。

3.1 動詞內部結構和主賓語功能的關係

「中研院平衡語料庫」中共有 7425 個雙音節動詞扮演過主賓語（即名詞中心語）的功能。我們從其中隨機抽出 500 個，並統計各種詞構的數量列於表二[7]。

	並列	述賓	偏正	述補	主述	名詞	其他	總數
數量	263	103	87	14	5	7	23	500
百分比	**52.6%**	20.6%	17.4%	**2.8%**	1.0%	1.4%	4.6%	100%

表二：「中研院平衡語料庫」中 7425 個作主賓語動詞中隨機抽樣五百個動詞的詞構分佈比例

比較表一和表二，我們看到並列動詞的比例上升了：從 34.6%上升到 52.6%，述補動詞明顯下降，從 15.6%降到僅有 2.8%；述賓動詞微幅下降，從 25.4%降到 20.6%；偏正結構的比例並無太大的變化（17.6%比 17.4%）。可見在所有複合動詞中，並列動詞用作主賓語的傾向最高，而述補動詞幾乎不能當主賓語。

除此之外，最常用作主賓語的三類動詞，即並列動詞、述賓動詞和偏正動詞，在這個功能上還有細微的用法差異。述賓動詞和偏正動詞都是在一些特定的情況中才會被用作主賓語。述賓動詞絕大多數都是動作動詞，動作述賓動詞之所以可以用作主賓語，主要是以下兩種現象所致，否則它不太會被用作主賓語。

一、述賓動詞前有名詞性修飾語，多半是該動詞意義上的賓語[8]。例如：任務編組、家庭抽樣、出版品分級、十大惡性腫瘤的排名。

二、述賓動詞又被分析為偏正名詞，由動詞後接賓語的結構轉為名詞前接謂語性修飾語的結構。例如：這個編組、十幾個分區、三個抽樣、電話收費。

狀態動詞較少述賓結構，少數的例子，像是「無知、無理、缺德」，在作主賓語時，多半出現在「的」之後，像是「他的無知」。

偏正動詞用作主賓語以幾種狀語性領頭詞所組成的複合動詞最為常見，主要是「不、自、反、重、互、相」組成的動詞。其他的偏正動詞用作主賓語的情況並不多見。下面的例子中，以「不」開頭的例子都是狀態動詞，以「自」、「反」、「重」、「互」或「相」開頭的例子都是動作動詞。

[7] 在「中研院平衡語料庫」中，動詞用作主賓語或動作動詞用作定語會標上[+nom]這樣的標記。該語料庫中帶有[+nom]的動詞共有 9601 個，不過有 2176 個只作定語，所以有 7425 個動詞扮演過主賓語功能。

[8] 在這樣的結構中，述賓動詞前面的名詞其實扮演兩種語義功能：一、作該動作的賓語性修飾語，例如：城市排名（對城市的排名）；二、作為該動詞執行的標準，例如：成績排名（依照成績做出的排名）。不過以第一種語意關係居多。

不~V：不滿、不便、不幸、不安、不耐、不悅、不當、不忍、不捨、不解…
自~V：自殺、自救、自創、自焚、自衛、自責、自制、自戕、自述、自棄…
反~V：反駁、反思、反省、反射、反饋、反彈、反串、反問、反常、反擊…
重~V：重建、重整、重修、重生、重排、重劃、重組、重現、重測、重讀…
互~V：互信、互動、互諒、互愛、互通、互助、互惠、互補、互賴、互斥…
相~V：相處、相容、相聚、相認、相逢、相遇、相左、相通、相殘、相剋…

基於述賓、偏正動詞上述的限制，他們所搭配的結構也就有所差異。動詞用作名詞中心語時可以出現在下列五種結構中[9]，並列動詞和這五種結構的搭配都非常普遍，數量都很大。述賓和偏正動詞雖然大略說來也可以出現在這五種結構中，但是述賓動詞以 c 種和 d 種結構的用法最為常見，而偏正動詞則多半出現在 a 種結構中。

並列動詞（“*”此符號表示為常用結構）：

*a. 關係子句的中心語：警方對此案的調查
*b. 狀態動詞後：正式調查、慎重的調查
*c. 數量詞後：一項調查、種種調查、此次調查、幾番調查
*d. 複合名詞的中心語：田野調查、民意調查
*e. 搭配名詞性賓語的動詞之後[10]：接受調查、等候調查、展開調查

述賓動詞（“*”此符號表示為常用結構）：

a. 關係子句的中心語：十大惡性腫瘤的排名、他的無知
b. 狀態動詞後：混合編組
*c. 數量詞後：這個編組、十幾個分區、三個抽樣
*d. 複合名詞的中心語：任務編組、家庭抽樣、出版品分級、電話收費
e. 搭配名詞性賓語的動詞之後：完成編組

偏正動詞（“*”此符號表示為常用結構）：

*a. 關係子句的中心語：你對他的不滿、三毛的自殺、他對此案的反駁、信念的重建、人與人之間的互信、異性的相處
b. 狀態動詞後：強烈不滿
c. 數量詞後：種種不滿、一些反駁
d. 複合名詞的中心語：利他性自殺
e. 搭配名詞性賓語的動詞之後：引起不滿、提出反駁

[9] 這五種結構並不包含動詞用作主語的情況。在「中研院平衡語料庫」中，動詞單獨出現在主語位置，像是「審判獨立與否」中的「審判」，是不帶[+nom]的屬性。也就是說在該語料庫中動詞用作主語並不視為名物化，我們也採用同樣的原則。

[10] 當名物化動詞作賓語時，所搭配的動詞相當有限，多半限於：做、展開、開始、引起、造成、導致、提出、接受、拒絕、完成、進行…等。

另外，我們從兼有動詞名詞兩個詞類的詞中，也看到平行的現象，顯示出詞構和主賓語功能之間具有相同的關係。在「中研院平衡語料庫」中，有的詞被標示了動詞和名詞兩個詞類。這些詞大部分是由動詞轉用作名詞，少部分是由名詞轉作動詞。這樣的詞共有 2448 個。其中雙音節詞有 1662 個，動詞和名詞用法都超過十次以上的有 477 個，參見表三。我們依據這 477 個動詞來統計每種詞構的數量，得出的結果在表四。從表四可以明顯看出，它們的詞構分佈比例比起表二的統計還要鮮明，兼有名詞詞類的動詞中以並列動詞最多，佔了 63.10％，其他各類詞構的比例就更低了。而且從表三的詞頻統計我們知道這些詞都是高頻詞，平均詞頻達 454 次，這表示表四的統計結果相當具代表性。

兼具動詞和名詞的詞	Type	Token	平均詞頻
總數	2448	391778 名+278689 動=670467	273.88
雙字詞	1662	175824 名+127179 動=303003	182.31
出現十次以上的雙字詞	477	125582 名+90927 動=216509	453.90

表三：「中研院平衡語料庫」中兼具動詞和名詞詞類的詞

	並列	述賓	偏正	述補	主述	名詞	其他	總數
專指事件名詞	262	41	26	3	2	51	5	390
普通名詞(和事件名詞)	37	**17**	9	0	0	8	0	71
數量詞	2	0	12	0	0	1	0	15
專有名詞	0	0	0	0	0	1	0	1
總數	301	58	47	3	2	61	5	477
百分比	**63.10%**	12.16%	9.85%	0.63%	0.42%	12.79%	1.05%	100%

表四：「中研院平衡語料庫」兼具動詞和名詞用法各十次以上的雙字詞之詞構分佈比例

在表四中我們特別針對這些詞作名詞時的意義將其分成四類[11]。此表告訴我們絕大多數動詞是轉作事件名詞，少數可轉作普通名詞。特別的是，述賓動詞轉為普通名詞的比例相當高，58 個中有 17 個。這是因為述賓動詞往往被分析為偏正式名詞的緣故，像是「移民、定義」。這一點觀察和我們上述對述賓動詞

[11] 這些兼有動詞和名詞詞類的詞其名詞類型可以分成四類，這四類分別是事件名詞、普通名詞、數量詞和專有名詞：

一、專指事件名詞：這類詞大部分的名詞用法是用作事件名詞，像是「研究, 發展, 管理, 影響, 決定…」。

二、普通名詞（和事件名詞）：用作普通名詞，可以指人，像是「領導, 編輯, 移民」，也可以指物，像是「報導, 投資, 作業」。這些詞往往也用來指事件名詞，例如：

普通名詞用法：他是一名優秀的領導。

事件名詞用法：名眾不服他的領導。

三、數量詞：在該語料庫中數量詞屬於名詞的次分類，有一部份數量詞也兼有動詞分類，像是「很多, 不少, 更多, 太多, 一般, 很少, 較多, 重重, 過多…」。

四、專有名詞：在該語料庫中，恰好有一個專有名詞也用作動詞：「青青」。

的描述是不謀而合。

(9) a. 他想移民到澳洲。（述賓結構的動詞）
　　b. 在澳洲華人移民相當多。（偏正結構的名詞）

(10) a. 試問真正的快樂是什麼，你能定義嗎？（述賓結構的動詞）
　　 b. 他替「快樂」下了三個定義。（偏正結構的名詞）

從以上的討論我們知道動詞內部結構和主賓語功能有以下幾個明顯的關係：

一、 述補動詞不用作主賓語；

二、 並列動詞最常用作主賓語；

三、 述賓動詞作主賓語時，動作動詞前面多半是名詞性修飾語或數量詞，狀態動詞前面多半出現「的」；

四、 偏正動詞則以「不、自、反、重、互、相」所組成的複合動詞用作主賓語的情況較為常見，前面多半出現「的」。

根據以上的訊息，對語句剖析而言，凡是述補動詞，就不必考慮它扮演主賓語的可能；並列動詞和以「不、自、反、重、互、相」開頭的偏正動詞出現在上述五種結構中就可以判定為主賓語。動作述賓動詞前接名詞或定量詞，或狀態述賓動詞前接「的」，也可能是作主賓語。

3.2 動詞內部結構和定語功能的關係

在討論動詞內部結構和定語功能的關係時，我們只討論動作動詞，不討論狀態動詞，因為狀態動詞作定語是相當普遍的，其內部結構並不會影響狀態動詞在定語用法上的分配。因此本節所有統計數字皆只以動作動詞為對象，不包括狀態動詞。在本文的討論中，對定語的定義較為嚴格，只討論不後接「的」的定語用法，因為後接「的」的定語對電腦剖析而言是很容易辨別出來的。「中研院平衡語料庫」中共有 4689 個動作動詞扮演過定語的功能，我們從其中隨機抽出 500 個動詞，並統計各種詞構的數量，列於表五。

	並列	述賓	偏正	述補	主述	名詞	其他	總數
數量	182	182	91	18	6	7	14	500
百分比	36.4%	**36.4%**	18.2%	**3.6%**	1.2%	1.4%	2.8%	100%

表五：作定語的動作動詞中隨機抽樣五百個動詞的各個詞構分佈比例

比較表一和表五，我們看到述賓動詞的比例上升了：從 25.4%上升到 36.4%；述補動詞明顯下降，從 15.6%降到 3.6%；並列和偏正結構的比例並無太大的變化（34.6%比 36.4%，17.4%比 18.2%）。可見在所有複合動詞中，述賓動詞用作定語的傾向最高，述補動詞則幾乎不用作定語。

表五告訴我們定語用法一樣是以並列、述賓和偏正動詞為主，不過我們觀察到並列動詞和述賓/偏正動詞在定語功能上有兩點明顯的差異。第一點關係到定語功能的重要性和比例上的不同。基本說來，能作定語的並列動詞也都可以用作主賓語。但是我們卻發現許多述賓和偏正動詞能夠作定語，卻完全不用作主賓語，而且作定語的比例相當高。也就是說，對這些述賓和偏正動詞而言，定語功能是它們的首要或僅次於述語的功能。以述賓動詞「代課」和偏正動詞「主治」為例，他們用作述語和定語的例子和次數如下：

代課：述語(6 次)：宋主任雖曾擔任督學並在學校代課過，但未教過國中生。
定語(17 次)：代課老師、代課教師、代課教員

主治：述語(1 次)：後者是主治居家旅行小毛病的藥油。
定語(16 次)：主治醫師、主治大夫

第二點差異在於述賓和偏正動詞修飾的名詞種類比較固定，而並列動詞所修飾的名詞種類則較為廣泛。例如述賓動詞中，「代課」一詞的定語用法絕大部分都是「代課老師、代課教師」，「剪紙」一詞的定語用法大部分都是「剪紙藝術」。偏正動詞也是如此，「主治」大部分的定語用法都是「主治醫師、主治大夫」，「郵遞」大部分的定語用法都是「郵遞區號」。我們可以很清楚地從表六看出這個傾向。該表中的第一欄相當於表五，是各詞構作定語的分佈比例。當我們篩選出修飾名詞種類最多的動詞時，述賓動詞和偏正動詞的比例都大幅下降，而並列動詞則大幅上升。作定語的 4689 個動作動詞中有 262 個動詞修飾十到二十種名詞，有 165 個修飾二十種以上的名詞。述賓動詞作定語的常態比例是 36.4%，修飾十到二十種名詞的動詞中下降到 24.43%，在修飾二十種以上名詞的動詞中又繼續下降到 20.61%。偏正動詞作定語的常態比例是 18.2%，在修飾十到二十種名詞的動詞中下降到 11.07%，在修飾二十種以上名詞的動詞中又繼續下降到 7.27%。也就是說，修飾名詞種類愈多，這兩類動詞的比例就越少，這表示述賓動詞和偏正動詞多半分佈在修飾少數名詞的範圍內。和述賓及偏正動詞相對的是並列動詞，從表六我們也可以看出修飾名詞的種類愈多，並列動詞的比例越高。在「中研院平衡語料庫」中修飾名詞超過二十種以上的動作動詞就有高達 67.88%是並列動詞，幾乎是它作定語的常態分佈比例（36.4%）的兩倍。這表示並列動詞作定語較靈活，所修飾的名詞種類較廣泛。

	並列	述賓	偏正	述補	主述	名詞	其他	總數
所有作定語的動作動詞中抽樣五百個	182	182	91	18	6	7	14	500
	36.4%	**36.4%**	**18.2%**	3.6%	1.2%	1.4%	2.8%	100%
作定語且修飾十到二十種名詞的動詞	153	64	29	10	1	4	1	262
	58.40%	**24.43%**	**11.07%**	3.82%	0.38%	1.53%	0.38%	100%
作定語且修飾二十種以上名詞的動詞	112	34	12	5	0	0	2	165
	67.88%	**20.61%**	**7.27%**	3.03%	0%	0%	1.21%	100%

表六：定語功能中詞構分佈比例的比較

以下的述賓動詞和偏正動詞都具有上述兩個特色[12]：作定語的比例相當高，修飾的名詞中心語較為固定。

述賓：代課、升等、剪綵、造勢、頒獎、摸彩、剪紙、選機、防疫、煉油、控股、育兒、抑癌、戡亂、點字、通勤、在野、任課、緝私、救災、旅日、離線、救國、瘦身、觀光、辦公、施政、放款、造船、划船、排水、美容、灌籃、控球、反毒、考古、搶孤、殖民、掃毒、驗尿、照明、修業、徵文、排版、計分、致病、追價、交誼、犯罪、祭祖、跳水、製片、抗日、求才、靠行、滑水、就業、購物、入會、導覽、拼音、斷詞、撤僑、候機、計價、降水、轉診、避暑、問政、入學、避孕、啟蒙、營利、登山、繪圖、轉帳、立法...

偏正：主治、電傳、電鍍、公賣、郵遞、預審、自學、私營、零售、安養、虛擬、約聘、遊牧、附加、協談、自治、聯歡、內銷、專用、對流、草創、主打、疏運、試用、自救、新任、槍擊、義賣、反射、聯誼、公演、接種、復健、自衛、集訓...

從以上的討論我們知道動詞內部結構和定語功能有以下幾個關係：

一、述補動詞不用作定語；

二、述賓動詞作定語的傾向最強；

三、有一部份述賓動詞和偏正動詞常用作定語，甚至超過述語用法；

四、述賓動詞和偏正動詞所修飾的名詞中心語比較固定，並列動詞所修飾的名詞中心語則較為廣泛。

因此，在剖析自然語句中，對於述補動詞，我們不必考慮它扮演定語的可能性。相反地，對於那些定語傾向強的述賓和偏正動詞可以優先猜測它是用作定語。此外，當一個述賓動詞後接名詞時，我們可以優先猜測它是用作定語，而不是作述語。一來，正如前述，述賓動詞作定語的傾向強；二來，述賓動詞由於其內部結構中已經含有賓語，所以多半不再後接賓語。相對地，如果是並列動詞或偏正動詞後頭緊接一個名詞，它可能是述語，後接賓語；也可能是定語，後接名詞中心語，如下所示。不過，這兩種結構的區分涉及後面名詞的語義類型，已經超過本文討論範圍，故不再申論[13]。

[12] 這些詞雖然最常用作定語，但它們也都具有述語的用法，在詞類劃分上仍屬於動詞。在這裡，不以多數語法功能為詞類，似乎違反了詞類劃分原則。但是由於動詞兼有定語用法是常見現象，而且動詞是基本的詞類，這些詞所表示的概念也的確是個動作，因此還是被劃分為動詞。在詞庫和一般語言學上的分類，漢語中只能作定語的詞是劃分為「非謂形容詞」，這個詞類的詞不兼有述語功能。

[13] 張國賓(1989)的文章中提到當雙音節動詞與專有名詞組合時，只能是述賓結構；與一般名詞和集體名詞組合時，較多是述賓結構；與抽象名詞組合時，則多為偏正結構。此外，陳克健和洪偉美(1995)的文章中也提出了在電腦剖析上區別這兩種結構的方法。

並列動詞：　述語-賓語：研究水果、管理校務
　　　　　　定語-名詞中心語：研究成果、管理業務
偏正動詞：　述語-賓語：復育螢火蟲、郵購顏料
　　　　　　定語-名詞中心語：復育計畫、郵購公司

4. 討論

在本文中，我們探討了動作/狀態以及動詞詞構和動詞功能分佈的關係。簡言之，詞構可以輔助判斷動作動詞的主賓語和定語功能以及狀態動詞的主賓語功能，如下表以雙線所框出的範圍。

		述語	狀語	補語	主賓語	定語
動作	並列	標準用法	不可	不可	強	可
	述賓	標準用法	不可	不可	可（名詞~、數量詞~）	強
	偏正	標準用法	不可	不可	可（自/反/重/互/相~）	可
	述補	標準用法	不可	不可	不可	不可
狀態	並列	標準用法	標準用法	標準用法	強	標準用法
	述賓	標準用法	標準用法	標準用法	可（的~）	標準用法
	偏正	標準用法	標準用法	標準用法	可（不~）	標準用法

表七：動詞功能分佈和動作/狀態以及詞構的關係[14]

由上表得知，動作動詞是否可以用作主賓語和定語，或是狀態動詞是否可以用作主賓語可由詞構得出一個基本的判斷，並佐以領頭詞或搭配詞類等補充條件。我們認為這些涉及語法功能分佈的規律並非任意武斷的，是可以解釋的。在以下的討論中我們將試圖從詞彙語意來解釋為什麼動詞詞構與其功能分佈有著如上的關係，並探討各個詞構的名物化強度。

4.1 詞義和語法功能分佈

依照我們的看法，這兩個牽涉動詞功能分佈的因素，即動作/狀態的區分以及

[14] 如果只將動詞區分成動作和狀態兩類的話，述補動詞則應歸為動作動詞，因為它具有動作動詞的諸多語法特點，像是可以搭配時態標記「了」或是狀語「常常、故意」，不可以搭配程度副詞「很、非常」。如果嚴格區分事態，述補動詞應該和動作動詞、狀態動詞區分開來。依據 Vendler (1957)的分類，事態可分成四種：activity、state、accomplishment 和 achievement。述補動詞應該算是 accomplishment，和動作(activity)或狀態(state)不同。事態的區分還可以更精細，詞庫小組區就分出十三種不同的事態，請參見 Huang et al (1999)以及張麗麗等(1999)。

詞構，都和動詞的語義密切相關。動作動詞和狀態動詞的區分傳統認為是語法上的分類，但是越來越多的研究顯示這其實是意義上的分類。從語意來看，動作動詞和狀態動詞的的差別在於它們的「事件形態」(event types)。動作動詞的事件形態有開端、有過程、有終點，所以我們可以說「開會了」、「在開會」、「開完會」，分別指涉「開會」這個動作其中一個階段；狀態動詞的事件形態沒有過程，不能指涉開端或終點，所以不說「疲倦了」、「在疲倦」、「疲倦完」[15]。大部分的狀態都有強弱之別，所以可以加上程度副詞來修飾，像是「很疲倦」。上述這些語法行為差異是傳統用來區分動作動詞和狀態動詞的語法依據，但其實是由於事件形態的差異所造成的。我們認為這個差異也是導致二者功能分佈差別的主因。狀語或補語這兩個語法功能在語意上通常是用來描述主賓語的狀態，或整個事件進行的狀態[16]，而非指涉另一個獨立的事件，所以只有狀態動詞適用。

動詞的內部結構同樣也涉及了詞彙語義。以下我們將針對並列結構、述補結構、述賓結構以及偏正結構分別探討。我們將側重詞構對於詞彙意義的影響，並說明每種詞構的語義特色是如何決定詞彙功能的分佈。

在 Chang et al (1999a, 1999b) 中，我們強調並列動詞和所有其他動詞的詞構形成一個鮮明的對比，正可以解釋為何並列動詞當作主賓語的傾向強以及並列動詞所修飾的名詞種類多。並列動詞是由兩個概念相近的動詞所組成，在語意上是將每個單字動詞各代表的事件結構融合起來。其他詞構則是由一個主要動詞和一個事件成分結合，二者之間有兩種主要的明確的語義關係：一種是動詞和論元的關係，包含述賓結構和主述結構；另一種是動詞和修飾語的關係，包含偏正結構和述補結構。因此，在概念上「非並列動詞」表示一個更完整、更精確的事件結構。也就是說，並列動詞和非並列動詞最重要的差別在於：並列動詞沒有明確的內部結構關係，而非並列動詞卻具有明確的內部結構關係。這種結構上的對比也造成了詞義上的差距：並列動詞所表達的詞義傾向概念化，而非並列動詞所表達的詞義傾向精確化。當我們將一個事件用作名詞，即主賓語，我們是將該動詞當作一個指稱，用來指涉該事件的整體，所以傾向選擇沒有具體內部關係、詞義概念化的動詞。這也就是為什麼並列動詞作主賓語的傾向最強。當我們要說明一件事物的性質時，我們傾向選擇該性質的代表詞。並列動詞是將兩個近義詞的意義融合起來，表達更廣泛的概念，因此最容易成為代表詞。這也就是為什麼並列動詞容易用來修飾各式各樣的名詞。

述補動詞不作狀語和補語，也幾乎不作主賓語和定語，這是它和其他詞構最

[15] 少數狀態動詞可以指涉狀態的改變，所以可以說「漂亮了起來」、「不再漂亮」。

[16] 例如，在「他憤慨指出社會的不公」中，「憤慨」描述了主事者的情緒狀態；「把他罵得慘兮兮」中，「慘兮兮」描述了受事者的狀態；「緩慢移動腳步」中，「緩慢」描述事件進行的狀態。

大的差別。我們認為這也是導因於述補動詞的結構意義。在非並列動詞中，述補動詞所傳達的事件訊息最為複雜，不但說明事件的過程，也明確指出該事件的結果。過程指涉一個動作，而結果指涉一個狀態，所以述補動詞是動作和狀態兩種事件的組合，既表示時間先後關係，也表示因果關係。正因它不表示單純的狀態，所以不用作狀語或補語。也因為它所傳達的概念複雜，而非單純的概念，所以不適合用作主賓語或定語。當我們將動詞作為一個指涉的事件（即主賓語用法），或是用來說明某個名詞的性質時（即定語用法），都是取動詞單一性質的概念，因此表達複雜事件概念的述補動詞就不適合作為主賓語或定語。相對地，述補動詞的動詞性非常強，除了它絕大部分的語法功能都是扮演述語外，這也可以從它所搭配的句型看出。例如在現代漢語中就只有述補動詞可以直接搭配「處置性」（disposal）強的「把」字句（Li and Thompson 1981），而不需要再加上任何修飾語[17]。

述賓動詞的定語傾向強，也有其語意基礎。述賓動詞相當於一個完整的謂語結構，在事件結構的概念上它比起其它的詞構都來得完整，因為從一個事件的成分來看，賓語是事件結構中一個很重要的成分。語言學家認為當動詞帶上賓語就會使得該句所表示的事件具有完整的終點（telicity）。明確說出事件施事的對象，也就是將該事件結構作更進一步、更精確的勾勒。因此，述賓動詞在區分不同事件類型上具有最佳功能，用作定語時就等於對名詞中心語作最精確的描繪和區分。這也就是為什麼述賓動詞非常容易被當作定語。

在定語功能上，述賓和並列動詞各有特色。述賓動詞作定語的傾向最強，但是並列動詞修飾的名詞種類最多。這樣的差異也是受其結構意義的影響。述賓動詞表達的謂語概念最完整，所以是最佳的定語，許多述賓動詞往往以定語為主要功能。但是由於其事件結構完整，能夠修飾的名詞反而有限。而並列動詞因為表達的概念最廣泛，所以能夠修飾各式各樣的名詞。縱使如詞，定語功能絕非其最重要的功能。以「代課」和「替代」為例，前者是述賓動詞，在「中研院平衡語料庫」中其定語功能佔總次數的 70%，但是修飾的名詞就僅限於「代課」這個事件結構上的的主語，像是「代課老師、代課教師、代課教員」。而「替代」是並列動詞，在同一語料庫中其定語功能只佔總次數的 27%，但是所修飾的名詞種類各色各樣都有，只要是和「替代」這個事件結構相關的任何語意成分都可以被修飾，例子見下。另以「排版」和「排列」為例，「排版」作定語的比例有 68%，偏重修飾該事件結構中的工具，像是「排版系統、排版軟體、排版工具」；「排列」作定語的比例僅僅 1%，縱使如此，它也和其他並列動詞一樣，

[17] 語言學家一致認為「把」字句具有「處置」的特性，只能搭配將事件的過程和結果都說明清楚的動詞或動詞組。例如，「搬開」是述補動詞，「搬運」是並列動詞，我們可以說「把貨物搬開」，但不說「把貨物搬運」，一定要再加上一些成分，像是「把貨物搬運到倉庫」，或是「把貨物搬運出去。」

可以修飾和其事件結構相關的任何語意成分，例子見下。

代課(16/23=**70%**)：代課老師、代課教師、代課教員

替代(22/81=**27%**)：替代方案、替代方式、替代構想、替代產品、替代場地、替代道路、替代效果、替代詞、替代開發案

排版(19/28=**68%**)：排版系統、排版軟體、排版工具、排版功能

排列(13/125=**1%**)：排列方式、排列方向、排列次序、排列位置、排列形狀、排列順序、排列遊戲

基本說來，述賓動詞和偏正動詞作主賓語的傾向要比並列動詞弱得多，因為它們所表達的事件概念都過於精細，不適合用來指涉一個單純的事件概念。雖然有一些述賓和偏正動詞的確可以作主賓語，但這些例子都有其特殊的語意轉變，正好證明述賓和偏正動詞作主賓語的傾向偏弱。在3.1小節中我們提到述賓作主賓語時多半前有名詞性修飾語，是該動詞意義上的賓語，像是「任務編組、出版品分級」。在這個結構中述賓動詞所含的結構意義已經大幅下降，因為該動詞又外含一個賓語，表示該動詞內含的賓語不再具有影響動詞及物性的功用。也就是說，該動詞的內部結構對動詞語法行為的影響已經減低。這個現象告訴我們，當述賓動詞的結構意義下降，才有可能作主賓語。因此這一點反倒證明了結構意義強的述賓結構是不用作主賓語的。

偏正動詞作主賓語多半是以「不、自、反、重、互、相」開頭的動詞。從語意來看，這些偏正動詞和一般偏正動詞有所不同。一般的偏正動詞說明該動作的等級或方式，像是「痛哭、輕唱、前仰、亂吐」，而以「不、自、反、重、互、相」開頭的偏正動詞則說明了動詞基本性質的差異。我們認為是這個差異導致前者較少作主賓語，而後者較常作主賓語。以「不」為例，雖然「不安」是個偏正式狀態動詞，但是它不是「安」這個概念的修飾，而是和「安」的意義相對。因此，在概念上它表示的是「基本」的狀態的概念，所以適合用來作一個事件概念（名物化）的指涉，作主賓語傾向也就提高。縱使如此，這些偏正動詞作主賓語的用法還是十分有限，不像並列動詞那麼靈活。它們作主賓語時絕大多數都是出現在關係子句的結構中，像是「你對他的不滿、三毛的自殺、他對此案的反駁、信念的重建、人與人之間的互信、異性的相處」，而在另外四種主賓語結構中的分佈都不高。

以上對狀語的區分我們也找到平行的現象。現代漢語中，不以單個動詞，而以動詞組作為名詞組中心語的情況並不多見，但是在我們所找倒的有限例句中狀語成分也都和「不、自、反、重、互、相」的概念相近，例子如下：

(11) 讓他漸漸感覺到彼此之間的不協調，...

(12) 並指出突破這困境的必要條件是：人的自我剖析、批評、否定、解放和再肯定，...

(13) 啟發年輕一代對傳統的重新認識。

(14) 也可減輕了同音詞的互相干擾。

4.2 詞構和名物化強度

我們的觀察除了和自然語句的自動剖析相關外，也涉及語言學上動詞名物化的現象。主賓語和定語這兩種語法格式都和動詞名物化有關。基本說來，名詞組中心語，即主賓語，是最典型的名物化格式，而定語則是比較弱的名物化格式[18]。依據本文的觀察我們可以看出不同結構的動詞具有不同強度的名物化傾向：並列動詞強過述賓動詞，述賓動詞又強過偏正動詞，偏正動詞則又強過述補動詞。我們得出這樣的結論主要是依據本文的三項觀察：一、主賓語用法中，並列動詞最為活躍；二、定語用法中，述賓動詞強過偏正動詞；三、述補動詞不用作主賓語或定語。因此，我們得到不同詞構的名物化強度如下：

名物化強度：並列 > 述賓 > 偏正 > 述補

這個排序正好和表一中各詞構分佈比例的排序相當，也就是說，現代漢語動詞中比例越高的詞構其名物化強度也越高。此外，能用在強名物化格式中的動詞必能用在弱名物化格式中，這是單向進行的，反之不然。我們觀察到能作主賓語的動詞幾乎都能作定語，但是能作定語的動詞卻不一定就能作主賓語。在3.2 小節中，我們列舉了一些述賓和偏正動詞，除了述語用法外，它們只能用作定語，卻不能用作主賓語。這樣的例子在「中研院平衡語料庫」中相當豐富，在現代日常生活中也隨處可見。

我們所觀察到的名物化強度和張國賓所提的模式相似性相當高。張國賓(1989)提到「雙音節動詞由於正處在向名詞一端飄移的進程中，所以雙音節動詞之間上存在著某些功能差異，這似與雙音節動詞內部的構成方式有關，經初步考察，我們發現現代漢語雙音節動詞的功能差異情況大致如下圖：

動/名雙功能詞←聯合式、支配式、陳述式、補充式、附加式→純粹動詞」
(並列)　(述賓)　　　(述補)　(偏正)

兩個模式唯一的差別在於：我們認為述補動詞名物化傾向最低，但是張國賓卻認為附加式動詞（即偏正動詞，或稱為狀中結構）才是最純粹的動詞。由於張的文章中只列舉此模式，沒有任何說明或證據，我們無從根據他所提的論證來反駁。不過我們認為本文的表二和表五中述補動詞作主賓語和定語極低的分

[18] 朱德熙(1990)認為作定語的動詞也算是名物化，他說「這類格式中的動詞（如：建築材料、研究方法）已經名物化了，它是以名詞的資格充任定語的，因此凡是不大能被名物化的雙音節動詞也就不能直接作定語」。

佈比例就已經充分證明了述補動詞才是最純粹的動詞。此外，從句型的搭配也顯示述補動詞的動詞性最強。我們在本節稍早提到只有述補動詞可以直接和處置性強的把字句搭配（請參見註釋 17），而無須添加任何修飾成分。

5. 結論

在本文中，我們嘗試提出一些判定動詞語法功能的準則，希望能提升電腦剖析中文語句的能力。我們所提出的準則有二：從宏觀角度來看，動作動詞和狀態動詞各有其大致的語法功能分佈趨勢；從微觀的角度來看，動詞的內部結構可以更進一步幫助判斷個別動詞的功能分佈。具體的結論有：狀態動詞比動作動詞多了狀語和補語兩個語法功能、並列動詞作主賓語的傾向強、述賓動詞作定語的傾向強、述補動詞只作述語。此外，述賓動詞和偏正動詞用作主賓語時，各有一些辨識條件。述賓動詞作主賓語多半前接名詞或定量詞，偏正動詞作主賓語多數是以「不、自、反、重、互、相」為詞首的動詞。再者，有一部份述賓動詞和偏正動詞其定語用法遠遠超過述語用法。依照這幾個原則，在剖析語句中，我們便能以較有效率的程序來推測動詞的語法功能，減少錯誤的分析，加快分析的速度。當然剖析系統要能運用這些判定的準則並不是非常容易，必須在具有動詞詞構訊息的前提下才能操作這些原則。同時我們也瞭解，詞構也只能指出動詞功能分佈的大體趨勢，並不是絕對的標準。縱使如此，要想解決這方面的問題，詞構仍然是不容忽視的訊息。

在語言學界，已經有越來越多的學者認為詞彙的用法基本上是由其意義所決定的。（Levin 1993, Pustejovsky 1995）我們接受了這樣的看法，也希望能夠透過詞彙意義來具體有效地掌握現代漢語的整體面貌，並對自然語言處理提出一些幫助。我們認為文中所提影響語法功能的兩方面因素也都是語義的因素，等於是初步探討了詞彙意義和語法功能的關係。同時，藉助本文的觀察，我們也提出對動詞名物化的一些看法，並修正了前人所提的模式。在這些方面，這篇文章都只能算是一個初步的嘗試，作法和結果都還很粗淺。我們將以此為基礎，對詞彙意義和詞構作多方面深入的探討，希望能夠更進一步瞭解詞構對詞彙意義和用法的影響，並朝著詞彙意義和用法這個大方向作更多的研究。

參考書目：

朱德熙，1987，“定語和狀語”，漢語知識講話，上海教育出版社。

陳克健、洪偉美，1995，“中文裡「動-名」述賓結構與「動-名」偏正結構的分析”，第八屆計算語言學研討會論文集，第 1-13 頁。

張國賓，1989，““動＋名”結構中單雙音節動作動詞功能差異初探”，中國語文，1989 年第三期，第 186 至 190 頁。

張國賓，1990a，“單雙音節動作動詞搭配功能差異研究”，上海師範大學學報，1990 年第一期，第 141 至 145 頁。

張國賓，1990b，“V 單短語與 V 雙短語探異”，懷北煤師院學報，社會科學版，1990 年第四期，第 117 至 123 頁。

張麗麗、陳克健、黃居仁，1999，“漢語動詞詞彙語意分析：表達模式與方法”，Working Papers on Chinese Verbal Semantics (Vol.I), ed. by Kathleen Ahrens, Chu-Ren Huang, and Mei-chih Tsai.. Taipei: Corpus Research Group and Chinese Knowledge Processing Group.

Chang, Li-li, Keh-jiann Chen and Chu-Ren Huang. 1999a. “Alternation Across Semantic Fields: A Study of Mandarin Verbs of Emotion.” Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation, pp39-50, Taipei.

Chang, Li-li, Keh-jiann Chen and Chu-Ren Huang. 1999b. “The Semantic Properties of Mandarin VV compounds.” To be presented at the eighth International Association of Chinese Linguistics (IACL-8), Melbourne.

CKIP. 1995. *A Description to the Sinica Corpus.* Technical Report 95-02. Academia Sinica. Taipei.

Huang, Chu-Ren, Liu Mei-chun, and Mei-chih Tsai. 1998. “From Lexical Meaning to Event Structure Attributes: Across Semantic Classes of Mandarin Verbs.” The 6th International Conference on Chinese Linguistics/The 10th North American Conference on Chinese Linguistics. June 26-28. Stanford.

Huang, Chu-Ren. 1998. “Classifying Event Structure Attributes: A Verbal Semantic Perspective from Chinese.” Chinese Workshop. The 1998 International Lexical-Functional Grammar Conference. June 30-July3. Brisbane, Australia.

Huang, Chu-Ren and Kathleen Ahrens. 1999. “The Module-Attribute Representation of Verbal Semantics.” Working Papers on Chinese Verbal Semantics (Vol.I), ed. by Kathleen Ahrens, Chu-Ren Huang, and Mei-chih Tsai. Taipei: Corpus Research Group and Chinese Knowledge Processing Group.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Li, Charles & Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar.* California: University of California Press.

Liu, Mei-chun. 1997. “Lexical Meaning and Discourse Patterning – the three

Mandarin cases of 'build'." Paper presented at the Third Conference on Conceptual Structure, Discourse, and Language. Boulder, Colorado.

Liu Mei-chun, Chu-Ren Huang, and Charles C.L. Lee. 1998. "When Endpoint Meets Endpoint: A Corpus-based Semantic Study of Throwing Verbs." The 6th International Conference on Chinese Linguistics/The 10th North American Conference on Chinese Linguistics. June 26-28. Stanford.

Liu, Mei-Chun, Chu-Ren Huang and Ching-Yi Lee. 1999. "Lexical Information and Beyond: Constructional Inferences in Semantic Representation." Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation, pp27-38, Taipei.

Pustejovsky, James, S. Bergler, and P. Anick. 1993. "Lexical Semantic Techniques for Corpus Analysis." Computational Linguistics, 19.2, pp331-358.

Pustejovsky, James. 1995. *The Generative Lexicon.* Cambridge: MIT Press.

Smith, C. 1991. *The Parameter of Aspect.* Dordrecht: Kluwer.

Tenny, C. 1992. "The Aspectual Interface Hypothesis." In I. Sag and A. Szabolcsi Eds. *Lexical Matter*. Standford: CSLI.

Tsai, Mei-chi, Chu-Ren Huang, Keh-jiann Chen, and Kathleen Ahrens. 1998. "Towards a Representation of Verbal Semantic: An Approach Based on Near-Synonyms." International Journal of Computational Linguistics & Chinese Language Processing, pp62-74.

Vendler, Zeno 1957. "Verbs and Times." Philosophical Review 56, 143-160. Also in Z. Vendler. 1967. *Linguistics in Philosophy.* Cornell University Press. Ithaca, NY, 97-121.

Semantic Representation of Verbal Information – A Case from Mandarin Verbs of Judging

Mei-Chun Liu, *National Chiao Tung University*
Chu-Ren Huang, *Academia Sinica*
Jia-Ying Lee, *National Chiao Tung University*

All correspondences to:
Mei-chun Liu
Dept. of Foreign Languages and Literatures,
Natl. Chiao Tung University, Hsinchu, Taiwan ROC
E-mail: mliu@cc.nctu.edu.tw

Abstract

This paper aims to introduce a recently-developed framework for lexical semantic representation of Mandarin verbal information, using verbs of judging as an illustration. The framework (MARVS) takes each verbal sense as conveying one unique eventive structure and seeks to represent all syntactically relevant information with modular and attributive characterization. By exploring the semantic-syntactic interdependencies pertaining to verbs of judging, the study is able to identify the meaning components that are crucial for syntactic distinction and ultimately represents the semantic information in a systematic and principled way with MARVS.

1. Introduction

Semantic representation has always been a central issue in Natural Language Processing (NLP). At the core of our semantic knowledge is the complex information encoded by verbs. The question as to how to fine-tune and distinguish the meaning lexicalized in each individual verb remains to be solved and presents a challenging task for semantic representation of Mandarin.

1.1 Semantic Representation and Verb Meanings

In order to represent verbal information, efforts of research have been made to identify the semantic factors that are syntactically crucial and to work out some general principles governing the mapping between lexical semantics and syntax. Traditionally, the main concern on verbal information is limited to their subcategorization frames and semantic restrictions. Most formal theories of linguistics assume that verbs are the structural head of the sentence and hence the concern is how many and what kind of argument(s) each verb can take. Clear distinctions of verb meanings are treated only as general tendencies in selectional preferences, and the semantic details of individual verbs are largely neglected. However, as pointed out in Liu, Huang and Chang (1999), recent development in lexical research has shifted the focus to investigating the grammatically-relevant semantic properties of verbs. Researchers believe that the full range of syntactic realization of a verb depends largely on the meaning of the verb, and attempts have been made to define and establish patterns of interdependencies between verb meanings and syntactic behavior (cf., Levin 1997, Pustejovsky 1995, Levin 1993, Atkins and Levin 1991, Atkins et al. 1988, etc.). In particular, Levin (1993) presents a comprehensive attempt and categorizes English verbs into semantically distinct classes on the basis of their argument alternation patterns. Pustejovsky (1995) proposed a generative framework of lexical information with a multi-layered representational scheme that includes Argument Structure, Event Structure, Qualia Structure, and Inheritance Structure. His goal is to fully represent the interaction of word meaning and compositional constraint.

In practice, Levin et al (1997) has suggested that careful consideration of the range of argument expression options exhibited by members of *various classes of verbs* may help reveal the syntactically-relevant meaning components. Based on corpus patterns of verb behavior, their case study on English verbs of sound (Levin et al 1997) has successfully factored out the grammatically crucial elements of verb meaning,

1.2 Lexical Semantic Studies of Mandarin Verbs to Date

Lexical studies on Mandarin verbal semantics have just started in recent years. Collaboration between Academia Sinica and National Chiao-Tung University has rendered some preliminary results based on a series of corpus-based studies (e.g. Chang et al 1999; Liu et al 1999; Liu, Huang, and Chang 1999; Liu et al 1998, Huang et al 1998, Tsai et al 1996, etc.). These studies can all be characterized as exploring the meaning contrast among verbs of the same semantic field by way of comparing their syntactic behavior observed in the Sinica Corpus. The earlier works focused mainly on differentiating near-synonym sets, with the goal to fine-tune the interaction between semantic features and syntactic realization. The scope was then expanded to a whole class of verbs. For example, Chang et al (1999) investigated all subgroups of 'emotion' verbs and pointed to the morphological make-up (VV vs. non-VV compounds) as the key to their syntactic variation. Liu, Huang and Chang (1999) explored verbs of surface contact and found that this group of verbs may take either the location or the substance to be the object (termed Locus-Locatum Alternation) and can be further divided into three sub-groups in terms of directional/ locational change of the substance. Taking the effect of *construction* (association of structural pattern and meaning) into consideration, Liu, Huang and Lee (1999) spelled out the importance of constructional inferences beyond lexical specification, using verb of rushing (趕) as an example.

As Liu , Huang and Lee (1999) pointed out, Mandarin lexical semantic studies are advancing but remain still in a pioneering and primitive stage. More comprehensive investigation is needed to identify the set of crucial semantic attributes as well as compositional principles that have syntactic consequences. This present study can then be viewed as one more effort in building a sound and solid foundation for further exploration of the wonder and wealth of lexical semantics of Mandarin verbs.

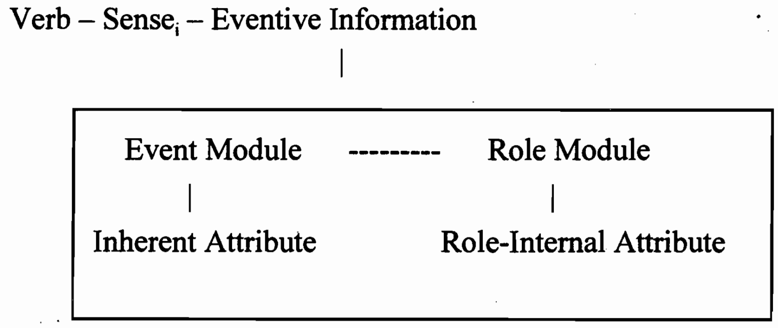
2. A Framework for Representing Mandarin Verb Semantics (MARVS)

The studies mentioned above all lead to one important question: What would be a principled way of representing verbal distinctions in Mandarin? In Huang and Ahrens (1999), a lexically based model called Module-Attribute Representation of Verbal Semantics (MARVS) was proposed as a first step toward developing a comprehensive framework for detecting and representing Mandarin verb meanings.

2.1 Basic Constructs

The model takes each verbal sense as one *event structure* conveying distinct *eventive information* which consists of two modules: Event Module with event compositional information and Role Module with salient participant role information. Within each module, detailed specifications are represented as attributes: Inherent Attributes are features concerning the semantics of the event itself and Role-internal Attributes are features further specifying a participant role. The model can be schematized as follows:

(1) Module-Attribute Representation of Verbal Semantics (MARVS):



Verb – Sense$_i$ – Eventive Information

Event Module --------- Role Module

Inherent Attribute

Role-Internal Attribute

The model is built upon three theoretical premises. First, all grammatical information is encoded in the lexicon. Grammar is information-based and lexicon-driven. Second, verbs express eventive information. The identification of verbal senses is then dependent on the identification of event types and event structures. Third, the classification of information is twofold: structural vs. attributive. There are therefore two ways to break down verbal semantic information to atomic units. Structural components are viewed as modules while attributive information are treated as features.

More specifically, Event Modules are the basic building blocks of the event contour. There are five event modules:

- Boundary : an event module that can be identified with a temporal point and must be regarded as a whole (including Complete Event);
- Punctuality: an event module that represents an single occurrence of an activity that cannot be measured by duration.

- Process: an event module that represents an activity that has a time course; i.e. it can be measured in terms of temporal duration.
- State: a homogeneous event module in which the concept of temporal duration is irrelevant; i.e. it is neither punctural nor has a time course.
- Stage: an event module consisting of iterative sub-events.

The five modules can be symbolized as follows:

(2) Symbol Representaion of Event Modules

a.	Boundary	•
b.	Punctuality	/
c.	Process	/////
d.	State	____
e.	Stage	^^^^

The five basic building blocks may be combined to render three event composition types attested in Mandarin: Nucleus Event, Simplex Events, and Composite Events (for details of the these event types, please see Huang and Ahrens 1999). The next section provides a simple illustration of the framework.

2.2 An Illustration with Verbs of Construction

There are three verbs in Mandarin which can all be translated as 'build' – 建、蓋、造, but their meanings are actually distinct if we observe carefully the typical object they take:

(3) Objects for Verbs of Building:
a. 地主在河川地 蓋／建／*造 房子。
b. 政府在山上造／建／*蓋 水庫
c. 計劃與波音合作造／*建／*蓋 飛機。

It is clear that 蓋 only occurs with objects denoting 'building', 建 takes an architecture as its object, while 造 requires the object to have some kind of internal design. Their difference in the semantic requirement of the object (or the incremental theme) also explains why only 造 can be used in the following sentence:

(4) 工程師造／*建／*蓋 不出房子。

Since 工程師 'engineers' are not designers, they are not able to create any houses.

Besides, the three verbs also differ in aspectual composition. Only 建 can be used in the sentence below, pointing to the fact that 建 may allow a focus on the endpoint or completion of the activity:

(5) 房子建／*蓋／*造 了三年了還沒人住。

In sum, although the three verbs share the same Role Module (all taking an incremental theme), they can be differentiated in terms of Event Module and Role-internal Attribute, as specified below:

(6) MARVS Representation of 建、蓋、造

```
建    • ///// • (Bounded Porcess)    <Agent, Incremental Theme>
                                                 |
                                          [architecture]
蓋    • ///// (Inchoative Process)     <Agent, Incremental Theme>
                                                 |
                                            [building]
造    • ///// (Inchoative Process)     <Agent, Incremental Theme>
                                                 |
                                             [design]
```

To show in more details how this framework can be used for differentiating as well as representing Mandarin verbal semantics, we investigate another group of verbs – verbs of judging – in the following sections.

3 Mandarin Verbs of Judging

Verbs of judging, as a semantic group, can be defined as verbs that describe a person's judgmental attitude towards another person (or institute) on a certain, presumably factual ground. These verbs may be purely mental (eg. 滿意、不滿) or accompanied with speaking act (eg. 稱讚、責罵). To narrow the scope of our study, we first look at verbs of negative judgement. Its class members include: 不滿、埋怨、批評、指責、斥責、責備、責難、責罵、責怪、駁斥、痛斥、怒斥、罵、咒罵、叫罵、破口大罵, etc.

At first sight, we noticed that these verbs are quite heterogeneous in terms of verbal kinetics, or the Stative vs. Active distinction:

(7) Distinction in Verbal Kinetics
Highly stative: 不滿
Highly active: 斥責、罵

It is also observable that the active verbs in this group can also be characterized as verbs of speaking in that they denote a verbal act outwardly reflecting the negative judgement. One immediate question follows: does the distinction in kinetics bear any significant consequences in their syntactic behavior? To answer the question, we looked carefully at their uses in the corpus and found that they have quite different distributions in the following aspects.

3.2 Grammatical Roles

These verbs differ in terms of the major grammatical functions they may be used for. Although they all occur as verbs, their distributions among other grammatical functions vary. Among all the verbs, 不滿 displays the widest range of grammatical roles: it may be used as adjectival modifier, as in (8a); adverbial modifier, as in (8b), nominal object or complement, as in (8c), and verbal predicate, as in (8d):

(8) Grammoatical Roles:
a. Adjectival modifer: 人民的不滿情緒
b. Adverbial Modifier: 陳水扁強烈不滿地指出…
c. Nominal Complement: 大陸漁民仍表示不滿
d. Verbal Predicate: 部份黨員不滿提名作業不符黨內民主

In the table below, we listed the distributional differences for six of the verbs in this group:

(9) Distribution among Major Grammatical Roles:

	不滿	批評	指責	斥責	責怪	罵
Total #	178	833	200	93	86	272
Adjectival	4%(8)	3%(24)	0	0	0	<1%(2)
Adverbial	2%(4)	0	0	0	0	0
Nominal	52%(92)	25%(208)	18%(34)	13%(12)	2% (2)	<1%(2)
Verbal	42%(80)	72%(601)	88%(166)	85%(81)	98%(86)	99%(268)

It is clear from the table that the mental verb-不滿, as the most stative verb in the group, is most flexible in its grammatical realization, while verbs with speech act, such as 指責, do not function as modifiers at all and their use as nominal complement is also significantly lower[1]. This syntactic difference can in part be attributed to their inherent properties in event denoting: Although they all involve some kind of judgmental evaluation, verbs like 不滿 are Attitude-denoting, focusing more on internal state change and thus more 'attributive', while speech act verbs like 指責、責罵 are Action-denoting, focusing more on the verbal act being performed as a result of the negative judgement. Verbs such as 批評、埋怨 are, on the other hand, either Attitude-denoting or Action-denoting since they may allow non-actional, attributive use:

(10) Attitude-Denoting Use with 批評、埋怨:
a. Adjectival: 面對自己的 批評／埋怨 心態
b. Adverbial: 埋怨地看著他

3.3 Argument Expression

When used as verbal predicates, most of the verbs display a similar range of argument expression. They can take a single NP-Goal, as in (11a), or a clausal complement denoting Goal with Cause, as in (11b):

(11) a. Goal: 埋怨／批評／指責 政府（or 政府的無能)
b. Goal-Cause: 埋怨／批評／指責 政府 毫無行政效率 (or 執法不力)

Aside from this similarity, a clear difference is found with some Action-denoting verbs as they can also be used as quotation verbs with or without '說', where the content of speaking is taken as a salient argument:

(12) a. 以台語斥責說：車子是怎麼開的。
遭中共人員斥責：這裡是大陸，不是香港。

Among the Action-denoting verbs, 罵 (and related members as 叫罵、謾罵) singles itself out as it does not allow any *inanimate* Goal, as shown in (13a), and its

[1] The adjectival and nominal uses with 罵 are highly idiomatic and restricted, as show in the examples:
Ajectival: 罵話語彙
Nominal: 挨了一頓罵 (derived from 挨罵, which itself should be treated as a verb entry.)

occurrence with direct quotation is much higher than other speech verbs, as exemplified in (13b):

(13) a. 一再罵 政府／*政府的無能
b. 上船就罵：他媽的，要給你死
大聲叫罵：國民黨走狗。

It is obvious that 罵 differs from other Action-denoting verbs in its specification of the Goal-argument (if there is one) and its tendency of taking the content of speaking as its sole argument. Here, as in English, a Manner of Speaking verb (i.e. 罵) can be used as a Content of Speaking verb (e.g. 說) to introduce direct quotations.

3.3 Passive Construction

It is widely known that Mandarin passive construction is semantically negative, i.e., associated with negative evaluation. Therefore, we looked at the co-occurrence of these negative judgement verbs with the passive marker 被 or 遭. What we found was that 不滿, as a highly stative and attitude-denoting verb, is incompatible with passive construction. In the corpus, 不滿 never occurs with passive markers such as 被 or 遭, as shown below:

(14) Occurrence with Passive Markers

	不滿	批評	斥責	責怪	責備
Total #	178	833	93	86	49
被	0	6%(46)	3%(3)	2%(2)	10%(5)
遭(到/受)	0	8%(65)	13%(12)	1%(1)	8%(4)

This finding is not surprising given that stative verbs in general cannot be passivized, as an universal trend in most languages.

3.4 Degree vs. Manner Modifier

Another interesting observation related to the Attitude-denoting vs. Action-denoting distinction is that the two types of verbs display different patterns of adverbial modification. Attitude-denoting verb 不滿 occurs only with *degree* modifier such as 強烈、十分、極度, etc., while the Action-denoting verbs occur predominantly with manner modifier, such as 大聲、嚴厲, etc., as made clear in the table below:

(15) Different Types of Adverbial Modification

	不滿	批評	斥責	指責	責備
總筆數	178	833	93	200	49
Degree	29%(51)	3%(22)	0	<1% (1)	2%(1)
Manner	0	6%(50)	12%(11)	7%(12)	6%(3)

And again, verbs capable of either attitude-denoting or action-denoting (eg. 批評、埋怨) display more evenly between both types of modification, as exemplified below:

(16) a. Degree: 強烈批評民政局
更加埋怨對方
b. Manner: 嚴詞批評中共的對台政策
大聲埋怨對方

4 MARVS Representation of Verbs of Judging

Adopting the representational scheme MARVS, as introduced in section 2, we can identify the meaning differences among verbs of judging in terms of the following Module-Attribute characterization, using 不滿、埋怨、指責、罵 as four representative verbs:

- With regard to Event Module, 不滿 differs from other verbs in that it denotes a state rather than a process. More specifically, 不滿 encodes an effect state or inchoative state (schematized as · ____), which allows an event focus on either the effect or the durative state. Other verbs behave more like inchoative process (symbolized as · /////). The difference between 埋怨 and 指責／罵 can then be captured with a further specification on Inherent Attribute: 埋怨 allows attitude-denoting, which enables it to be used as an adjectival or adverbial modifier.

- With regard to Role Module, 不滿 and 埋怨 both take a Goal or Goal-Cause as their argument, while 指責 may in addition take the Content (direct quotation) as a salient argument. In contrast, although 罵 may also take a Goal-NP, it differs from the others in that it does not occur with Cause-argument; instead, it takes a Content-argument, as either a direct quotation or a clausal complement. Furthermore, 罵 enforces a Role-internal restriction on the semantics of the Goal: it has to be animate.

(17) MARVS Representation of Four Types of Negative Judging Verbs

	不滿	埋怨	指責	罵
Event Module	· ____	· /////	· /////	· /////
| Inherent Attribute	Attitude-denoting	Attitude-denoting Action-denoting	Action-denoting	Action-denoting Speech Act
Role Module	[Goal – (Cause)]	[Goal – (Cause)]	[Goal – (Cause)] [Content]	[Goal-(Content)] [Content]
| Role-Internal Attribute				Goal: +Animate

These four verbs are typical of four sub-groups of judgement verbs. Among them, the 指責-group seems to be the largest. It is also tentatively noted that the four-way distinction may apply to positive judgement verbs as well, with corresponding members such as 滿意、讚許、稱讚、誇. A follow-up study is needed to confirm the speculation.

5 Conclusion

This study has shown that based on corpus observation and analysis, the group of negative judging verbs can be further divided into four sub-groups, each with distinct syntactic behavior that stems from their unique properties in lexical meaning. The representational framework based on Module-Attribute taxonomy (MARVS) was adopted for systematic sense differentiation. The model helps to delimit and identify the meaning components that are syntactically crucial and provides a principled way to represent these features as well-defined eventive information.

Given that the processing of Mandarin depends largely on semantic information, a representational framework that is semantically-constrained is indeed needed. Focusing on verbal semantics, the present work can be seen as a preliminary effort towards developing a comprehensive model for knowledge representation as well as future application.

● References

Atkins, B.T. and Levin, B. 1991. Admitting impediments. In Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, ed. by U. Zernik. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, E.V. and Clark, H.H. 1979. When nouns surface as verbs. Language 55: 767-811.

Croft, William. 1991 Syntactic categories and grammatical relations, University of Chicago Press,

Chicago,IL

------. 1990. Possible verbs and the structure of events. In Meanings and Prototypes: Studies in

Linguistic Categorization, ed. by S. Tsohatzidis. London/New York: Routledge.

Dowty, David. 1991. Thematic proto-roles and argument selection. Language 67:3.547-619.

Filmore, C.J. 1968. The case for case. In Universals in Linguistic Theory, ed. by E. Bach and R. T.

Harms, 1-88. Holt, Rinehart and Winston New York, NY.

Grandy, Richard E. 1992. Semantic fields, prototypes, and the lexicon. In Frames, Fields, and Contrasts: new Essays in Semantic and Lexical Organization, ed. by Lehrer and Kittay, 10-122. Hillsdale: Lawrence Erlbaum.

Huang, Chu-ren, and Ahrens, Kathleen. 1999. Module-Attribute Representation of Verb Semantics.

Working Papers on Mandarin Verb Semantics. Taipei: Academia Sinica.

Huang, Chu-Ren, Mei-chun Liu, and Mei-chih Tsai. 1998. From lexical meaning to event structure attributes: across semantic classes of Mandarin verbs. Paper presented at the 6th International Conference on Chinese Linguistics/The 10th North American Conference on Chinese Linguistics. Stanford U.

Jackendoff, R. S. 1983. Semantics and Cognition. Cambridge: MIT Press.

-------. 1990. Semantic Structures. Cambridge: MIT Press.

-------. 1996. Conceptual semantics and cognitive linguistics. Cognitive Linguistics 7: 93-129.

-------. 1997. Twistin' the night away. Language 73 (3): 534-559.

Juffs, A. (1993). The syntax ans semantics of locative verbs in Chinese. ESCOL *92,* 137-148.

Levin, Beth. 1993. Verb Classes and Alternation. Chicago: U of Chicago Press.

Levin, B., G. Song and B.T. S. Atkins. 1997. Making sense of corpus data: A case study of verbs of sound. International Journal of Corpus Linguistics 2(1): 23-64.

Levin, B. and M. Rappaport Hovav. 1991a. The lexical semantics of verbs of motion: The perspective from unaccusativity. In Thematic Structure: Its role in grammar, ed. by I. Roca. Berlin: Walter de Gruyter.

-------. 1991b. Wiping the slate clean: A lexical semantic exploration. Cognition 41: 123-151.

Levin, Beth, and M. Rappaport. Hovav. To appear. From lexical semantics to argument realization.
In Handbook of Morphosyntax and Argument Structure, ed. by H. Borer.

Liu, Meichun. 1996. A pilot study on Chinese verb classes and alternations. NSC Project Report
(NSC85-2418-H-009-003).

-------. 1997. Conceptual basis and categorial structure: a study of Mandarin V-R compounds as a
radial category. In Chinese Languages and Linguistics IV: 425-51. Taipei: Academia Sinica.

-------. 1999. Lexical meaning and discourse patterning - the three Mandarin cases of '*build*'. In
Cognition and Function in Language, ed. by Barbara Fox, Dan Jurafsky, Laura Michaelis,
181-199. Stanford: CSLI.

Liu, Meichun, Chu-Ren Huang, Charles Lee, and Ching-Yi Lee. 1998. When endpoint meets endpoint: a corpus-based lexical semantic study of Mandarin verbs of '*throwing*'. Paper presented in IACL-7/NACCL-10, Stanford University.

Liu, Meichun, Chu-Ren Huang, and Ching-Yi Lee. 1999. Lexical information and beyond: constructional inferences in semantic representation. Proceedings of the 13th Pacific Asia
Conference on Language, Information and Computation, Taipei.

Liu, Meichn, Chu-Ren Huang, and Chih-Ling Chang. 1999. The locus-locatum alternation: a lexical
semantic study of Mandarin verbs of surface contact. Paper presented at the 11th North American
Conference on Chinese Linguistics, Harvard University.

Lyons, John. 1977. Semantics. Cambridge: Cambridge University Press.

Pustejovsky, James. 1991. The syntax of event structure. Cognition 41:47-81.

-------. 1995. The Generative Lexicon. Cambridge: The MIT Press.

Pustejovsky, James, S. Bergler, and P. Anick. 1993. Lexical semantic techniques for corpus
Analysis. Computational linguistics 19 (2):331-358.

Rappaport, M., & Levin, B. 1988. What to do with theta-roles. In Thematic Relations, ed. by W.
Wikins, 7-36. New York: Academic Press.

Smith, Carlota. 1991. The Parameter of Aspect. Dodrecht: Kluwer Academic Publishers.

Teng, Shou-hsin, ed. 1994. Chinese Synonyms Usage Dictionary. Taipei: Crane Publishing Co.

Tenny, Carol. 1992. The aspectual interface hypothesis. In Lexical Matters, ed. By Ivan A. Sag and Anna Szabolcsi. Stanford: CSLI.

-------. 1994. Aspectual Roles and the Syntax-Semantics Interface. Dordrecht: Kluwer.

Tsai, Mei-Chih, Chu-Ren Huang, Keh-Jiann Chen. 1996. 由近意詞辨義標準看語意、句法之互
動。Paper presented in IsCLL V, Taipei: National Cheng-Chi University. In the Proceedings of
the Fifth International Symposium on Chinese Languages and Linguistics, 167-180.

Tsai, Mei-Chih, Chu-Ren Huang, Keh-Jiann Chen, and Kathleen Ahrens. 1998. Towards representation of verbal semantics -- an approach based on near synonyms. Computational Linguistics and Chinese Language Processing 3 (1): 61-74.

Vendler, Zeno. 1967. Linguistics in Philosophy. Ithaca: Cornell University Press.

階層式文件自動分類之特徵選取研究

柯淑津
東吳大學資訊科學系
ksj@volans.cis.scu.edu.tw

陳振南
銘傳大學資訊管理系
jnchen@mcu.edu.tw

摘要

文件分類（Text Categorization）是指針對一組事先設定好的類別集，透過特徵選取的作法，將自然語言文件標上適當的主題類別。文件分類的應用範圍非常廣泛，包括：電子郵件與新聞過濾、資訊檢索、自動索引、以及詞彙語意解析等等。

有關文件分類的研究，常由文件內容中抽取重要的特徵（feature）來代表這個文件，而特徵抽取的來源包羅萬象，可以簡單地從文件作者、出版機構著手，或是由蘊含豐富資訊的語言結構來作為抽取文件特徵的依據。先前的研究通常只由歸屬同類別的文件選出特徵集，很少將類別間是否具相關性納入考慮，而且當選完特徵後通常不再加以變動。這樣的作法對於線性分類或許是可行的，若是應用於階層式分類便顯得不恰當。

本研究提出一個適用於階層式文件自動分類系統的特徵選取方法，經初步選完特徵集後，再依各特徵與相近類別間所具的分類意義做適當的調度。我們以『財經記事』的新聞資料進行分類實驗，結果驗證系統的強健性。另外，也得到下列幾個結論：（1）少的特徵數目有利於分類的進行，（2）階層式分類優於線性分類，（3）適當的特徵選取將更凸顯階層式分類的效能。

1. 簡介

自動文件分類（Text Categorization）是指針對一組事先設定好的類別集，透過自動化的作法，將自然語言文件標上適當的主題類別。文件分類的應用範圍非常廣泛，譬如電子郵件與新聞過濾(E-mail and News filtering)、資訊檢索(Information Retrieval)、自動索引（Automatic Indexing）、詞彙語意解析（Word Sense Disambiguation）等等。

在網際網路的蓬勃發展下，資訊的傳播沒有國度、時間的限制。Internet 持續地累積多樣化的資訊，已經形成一個巨大、分散的多媒體。然而這些大批的文件資料若是未能做妥善的整理，對個人或者資訊服務系統而言都將造成資訊氾濫。文件分類的技術正是解決這個難題的利器。現行著名的網路資訊服務系統，如：Yahoo (http://www.yahoo.com/)以及 Virtual Library (http://vlib.stanford.edu/) 便提供這類的服務。他們將蒐集到的網路文件組織成適當的結構，像是依照文件的地理區域、出版時間、出版機構、或是內容主題等，對文件加以分類（林頌堅，1998）。透過這些分類資料，使用者可以根據其需求，自系統中快速地選取相關資訊。

一般而言文件分類的建構方式可分為兩種：（1）由機器自動學習，（2）以人工對文件進行主題標示。由機器自動學習的作法中可歸納為兩類，督導式（Supervised）及非督導式（Unsupervised）學習。督導式的學習是由使用者或一些專家先對部分文件進行分類，然後再將分類好的文件作為自動分類系統的訓練資料。通常這種督導式學習的分類效果較非督導式為佳（Lewis, 1996）。至於，以人工對文件進行主題標示的工作，則常需仰賴所謂的分類專家，如圖書館員或是專業領域中的專家學者，以其專業知識對文件進行分類。這樣的作法，雖然可以得到較準確的結果，卻要付出相當的時間與人力。面對現在資訊爆炸的時代，我們急需一個良好的自動處理技術與工具來進行文件分類工作。

本研究提出一個適用於階層式文件自動分類系統的特徵選取方法，經初步選完特徵集後，再依各特徵與相近類別間所具的分類意義做適當的調度。我們以『財經記事』的新聞資料進行分類實驗，結果驗證系統的強健性。

本文的其他部分結構如下：在第二節中介紹先前有關文件分類的研究，第三節是我們對於資料所做的一些觀察，第四節提出適用於階層式文件分類的演算法，接著是實驗描述與結果討論，最後，提出結論與探討未來的研究方向。

2. 先前有關文件分類的研究

先前有關文件分類的研究，常由文件內容中抽取重要的特徵（feature）來代表這個文件，而特徵抽取的來源包羅萬象，可以簡單地從文件作者、出版機構著手(Blosseville, et al., 1992; May, 1997），或是由蘊含豐富資訊的語言結構：語彙(Lexical)、語法

(syntactic)、或是語意(semantic)等資訊來作為文件特徵抽取的依據。

文件的語彙資訊是語言結構中最容易抽取的特徵內容，常見的有：字、詞、片語等單位，有些研究利用語詞出現在文件中的頻率值(tf, term frequency) (Frakes and Baeza-Yates, 1992; Witten, Moffat and Bell, 1994)做為文件的特徵(Salton 和 McGill, 1983)，較常見的是除了頻率值外，再加上語詞本身的重要性這個考量，即是以各語詞的 tf x idf (inverse document frequency, Witten, Moffat and Bell, 1994)值所組成的向量做為文件的特徵(Salton 和 Buckley, 1988)。另外，有些研究人員認為所有的語詞都併入特徵值的處理並不恰當，他們建議以統計方法 χ^2 檢定來選取重要的語詞當作文件的特徵(Watanabe et a1., 1996; Ng, Goh and Low, 1997)。

在文獻中以語詞所含的語意代替語詞本身來設定文件特徵的作法有下列幾個，Liddy 提出的 DR-LINK 系統（Liddy et al., 1993）利用朗文機讀字典將文件中的每個語詞轉換成主題碼(SFC-Subject Field Code)，若有歧義情形發生時再依句子中 SFC 的分佈狀況等資訊，設定出合適的 SFC 碼（Liddy, Paik and Yu, 1994）。最後，以經正規化後的 SFC 向量來表示文件特徵。另外，Schütze 等人提出的隱含語意索引(LSI - Latent Semantic Indexing)，將文件看成為空間上的一個特徵向量，藉著分析語詞共生模式(word co-occurrence pattern)，再利用奇異值分析(SVD – Singular Value Decomposition)的技巧，將高維度的向量轉化成為一個具較低維度的向量(Schütze, Hull and Pedersen, 1995)，他們的實驗證實了這個方法的有效性，尤其是在減少計算量的部分。

在文件分類的處理中以語詞為單位來粹取文件特徵的研究方法，很明顯地存在著下述幾個缺點：同義字問題、一詞多義問題、參數數量問題、以及多字詞問題等等。另外，Yang 和 Chute 他們觀察到以同義詞典為主的分類方法，往往因為一般的同義詞典所涵蓋的字不足以應付各種不同領域需求(Yang and Chute, 1994)，因此，利用同義詞典將語詞轉化成為主題(Subject)的有關研究，相較於直接用語詞當作文件特徵的作法，並無法得到較佳的精確度。他們認為存在於文件中的自由文體與同義詞典中的控制詞彙間的詞彙漏洞(vocabulary gap)，可以利用人類知識來加以彌補。一者是利用先前由人工對應過的訓練資料，或者由相關性回饋(relevant feedback)的技巧來收集資訊。

過去有關分類的研究，證實分類結果的精確度與召回率常隨著類別個數增加而降低（Apte, Damerau and Weiss, 1994; Yang, 1996）。而階層式分類由樹根開始，在每一節點只需考慮往其子節點細分，因此，在處理過程的每一階段，所需面對的類別個數較少，這是階層式分類往往有較佳效果的原因之一。另外，隨著樹的階層數（level）遞增，所處理的文件常設定在愈來愈窄的特殊範疇（specific domain），此時詞彙的歧義程度較易規範（D'Alessio et al., 1998），這是階層式分類的另一個優點。

3. 階層分類的特徵詞彙調整

經觀察『財經記事』新聞資料後，我們發現存在著這個現象：樹狀結構中距離相近的類別共用特徵詞彙。相似的類別往往會共用特徵詞彙，這種情形尤其常出現在階層式分類系統中，被歸屬在同一大類別下的幾個細層類別，往往具有相當高的相似度。因此，它們會共用特徵詞彙，這種現象若不加以處理，將導致細層類別不易區分。如表一所示，『央行』、『交易』、以及『利率』等詞彙，同時以高頻率出現在金融篇的幾個中類別裏。

當文件被歸到階層式分類系統中的某一節點後，往下進行細層分類時，這些共用詞彙應被視為該節點的停用字（stopword）。各細層類別需靠它們之間相異的特徵詞彙來彼此競爭，因此，這些共用詞彙應自其特徵集中移除。

表一　共用詞彙分佈在細層類別的頻率值

詞彙	語料中出現頻率	中類別	中類別出現頻率
台幣	346	金融	65
		外匯	95
外匯	260	銀行	62
		外匯	65
央行	797	金融	392
		銀行	142
		外匯	77
交易	773	金融	126
		外匯	22
		股票	221
利率	844	金融	168
		銀行	337
		外匯	19

4. 階層式文件分類演算法

線性分類系統將各個類別之間的關係當成彼此完全獨立，而這種假設在階層式分類系統是不恰當的。因為在階層架構中擁有同一父節點的兄弟節點間的關係顯然較其他節點更為密切。因此傳統適用於線性分類的特徵選取方式，直接搬移到階層式分類系統，並不完全可行。

4-1 特徵詞彙選取

起始特徵詞彙選取

階層式分類系統的特徵選取處理，我們區分為葉節點(leaf node)與非葉節點(non-leaf node)兩個部分。對於葉節點的部分，我們利用訓練資料透過詞頻統計的方式，為每個分類選取適當的特徵詞彙集，並且依據每個特徵 f 與類別 c 間的關係強度，設定權重值 W(f, c)。本研究採用 tf-idf 來計算權重值（如公式１、２所示）。這些特徵分為正項特徵與負項特徵兩類，其中，正項特徵詞彙以高頻率出現於歸屬類別 c 的文件中，他們擁有正值權重。而負項特徵詞彙在類別 c 的文件中並不出現，而且以高頻率出現在與 c 擁有相同父節點的其他兄弟節點類別中，因此他們在類別 c 的權重值為負數。

至於，非葉節點 p 的特徵選取，我們以其所有子節點 c 特徵詞彙的聯集當為 p 的候選特徵集。節點 p 與候選特徵集合中的特徵詞彙 f 間的權重值 W(f, c) 為 p 的所有子節點 c 與 f 的權重值 W (f, c)的總和，如公式３所示。

$$W(f, c) = tf_{f,c} \times idf_f, \qquad \text{(當 } c \text{ 屬葉節點時)} \qquad (1)$$

$$idf_f = \log(\frac{T}{df_f} + 1), \qquad (2)$$

$$W(f, c) = \sum_{a \in c\text{的子節點}} W(f, a) \qquad \text{(當 } c \text{ 屬非葉節點時)} \qquad (3)$$

$tf_{f,c}$ ： 詞彙 f 出現在類別 c 中的頻率值，

df_f ： 詞彙 f 出現的類別數，

T ： 所有類別總數。

特徵詞彙調整

透過觀察我們瞭解階層式分類的特徵集，需依特徵所具的分類意義而做調整。對於分類意義的量化，我們以 Lin 在 1997 年的論文中所提出的分支比率－BR(Branch Ratio) 來決定（Lin, 1997），如公式 4 所示。當 BR 值愈小，代表該特徵之高權重普遍來自各個子節點，因此，此特徵對所有子節點而言不具分類意義。所以需將此特徵自其子節點的特徵集合中移除。反之，BR 值愈大時，表示此特徵值主要來自某個特定子節點，因此，我們必須保留此特徵不加以更動。

$$\mathrm{BR}(f,p)=\frac{\underset{c\in p\text{的子節點}}{\mathrm{MAX}}\ \mathrm{W}(f,c)}{\underset{c\in p\text{的子節點}}{\mathrm{SUM}}\ \mathrm{W}(f,c)} \qquad (4)$$

4-2 相關函數計算

對於文件與類別的相關程度，我們仍以 tf-idf 的方式來計算，如公式 5 所示。其中，將文件 *d* 所含詞彙 *f* 與類別 *c* 計算權重 W(*f*, *c*)，再加總所有的 W(*f*, *c*)，即得它們的相關強度 R(*c*, *d*)。

$$\mathrm{R}(c,d)=\sum_{f\in d}\mathrm{W}(f,c)\times idf_f \qquad (5)$$

至於當階層樹中各個類別的特徵詞彙集訓練完成後，面對一個待分類的文件 *d*，我們要如何決定 *d* 的歸屬類別呢？首先，我們由階層樹的樹根開始，往下計算位居第二階層的各個節點類別 *c* 與 *d* 的相關度 R(*c*, *d*)。此時，若是所有的 R(*c*, *d*)皆無法高過預先設定的門檻值 θ，則停止往下層分類的工作，並將文件 *d* 分派給父節點。若存在節點類別 *c* 使得 R(*c*, *d*)高過 θ，則將文件 *d* 分派給擁有最大 R(*c*, *d*)值的節點類別 *c*。然後照相同方式，繼續往下層節點進行分類。

5. 實驗設計及結果說明

5-1 實驗資料

本研究以『財經記事』中所含的財經新聞標題（見附錄）進行實驗，其內共含 132606 則新聞標題，其來源為民國八十一年間中國時報、工商日報、聯合報、民生報等各

報社之新聞標題。這些標題事先經人工標示採三層式分類，共分為金融、產業等九大類別，大類別下細分為 39 個中類別，中類別下又分小類別。本研究僅以大、中兩層類別進行實驗，其編碼方式以及類別內容等詳見附錄。

經審視資料後，發現有少數標題完全重複，另外，大部分的標題僅設定一個類別，少部分標題給定多種分類。在整個實驗中，為方便效能評估我們將重複標題去除，並且僅採用單一類別之標題，最後剩下 119845 則新聞標題。我們將這些標題以隨機方式取出 20% 作為測試語料，其餘的標題當訓練語料。

5-2 實驗設計

為驗證本研究提出的階層分類特徵選取方法之效能，特設計一連串實驗，以『財經記事』的新聞資料進行分類。其中，第一組實驗採線性方式分類，直接以第二層的中類別進行自動文件分類。而第二組實驗則為階層式分類，先將文件由根節點開始，分至合適的大類別，再由此大類別往下分派給所屬的中類別。每組實驗我們皆以不同的特徵個數, *k*, 來設定特徵詞彙集，並比較 *k* 值大小所造成的效果差異。

5-3 實驗結果

在效能部分，本研究採資訊檢索領域中最常用的精確率（Precision Rate）及召回率（Recall Rate）進行評估。在線性分類的部分，當選取 25 個特徵詞彙進行分類實驗時，我們可以得到 59.1%的精確率以及 77.8%的召回率。而隨著特徵詞彙數量的降低，我們發現實驗的效果愈來愈見提升（請詳見表二）。當每個類別只選取 10 個特徵詞彙時，可達到 61.0%的精確率及 81.0%的召回率。

階層式分類的結果列於表三，同樣地，我們可以發現小的特徵數，可以達到較佳的效果。當特徵詞彙的選取由 25 個降至 10 個時，召回率可以由 95.4%提升至 98.1%，同時，精確率也由 63.8%提升至 66.0%。這種現象是因為小的特徵集所含詞彙的歧義性較少，這結果同時也驗證了 Lewis 在 1992 年對路透社新聞進行分類的研究經驗（Lewis, 1992）。另外，值得注意的是在效果方面，階層式分類優於線性分類。而且，這種現象對不同大小的特徵集皆有一致的結果，這似乎驗證本研究所提出的特徵選取方法具強健性。

6. 結論與未來研究方向

本文提出一個適用於階層式文件自動分類系統的特徵選取方法，並且經由實驗證實它的強健性。另外，也得到下列幾個結論：（1）少的特徵數目有利於分類的進行，（2）階層式分類優於線性分類，（3）適當的特徵選取將更能凸顯出階層式分類的效果。

未來我們將嘗試以詞彙概念代替詞彙本身作為特徵的選取單位，探討這種作法對文件分類系統所帶來的衝擊。另外，將特徵出現在文件的位置併入計算權重值的考量，這些研究方向對分類效果應該會有更上一層的空間。

表二　線性分類的實驗結果

特徵詞彙數	召回率	精確率
10	81.0 %	61.0 %
15	80.0 %	59.4 %
20	79.8 %	59.0 %
25	77.8 %	59.1 %

表三　階層式分類的實驗結果

特徵詞彙數	召回率	精確率
10	98.1 %	66.0 %
15	96.5 %	64.4 %
20	96.4 %	63.7 %
25	95.4 %	63.8 %

致謝

本實驗獲得行政院國科會贊助（計畫編號：NSC 88-2213-E-031-003），特此致謝。

參考文獻

1. Apte, C., F. Damerau and S. M. Weiss, “Automated Learning of Decision Rules for Text Categorization,” ACM Transactions on Information Systems, 12 (3), 1994, pp. 233-251.

2. Blosseville, M., G. Hebrail, M. Monteil, and N. Penot, “Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together,” In Proceedings of the 15th Annual International ACM

SIGIR Conference on Research and Development in Information Retrieval (SIGIR-92), 1992, pp. 51-58.

3. D'Alessio, S., K. Murray, R. Schiaffino, I. College and A. Kershenbaum, "The Effect of Topological Structure on Hierarchical Text Categorization," In Proceedings of the Sixth Workshop on Very Large Corpora, 1998, pp. 66-75.

4. Frakes, W. B., and R. Baeza-Yates, Information Retrieval Data Structures & Algorithms, Edited by Frakes, W. B., and R. Baeza-Yates, Prentice Hall, New Jersey, 1992.

5. Lewis, D. D., "Feature Selection and Feature Extraction for Text Categorization," In Proceedings of Speech and Natural Language, 1992, pp. 212-217.

6. Lewis, D. D., "Challenges in Machine Learning for Text Classification," In Proceedings of the Ninth Annual Conference on Computational Learning Theory, 1996, pp. 1.

7. Liddy, E. D., W. Paik, and E. S. Yu, "Text Categorization for Multiple Users Based on Semantic Features from a Machine-Readable Dictionary," ACM Transactions on Information Systems, 12 (3), 1994, pp. 278-295.

8. Liddy, E. D., W. Paik, E. S. Yu and K. A. McVearry, "An Overview of DR-LINK and its Approach to Document Filtering," In Proceedings of the Human Language Technology Workshop, Princeton, N.J., 1993.

9. Lin Chin-Yew, Robust Automated Topic Identification, PhD Dissertation, University of Southern California, 1997.

10. May, A. D., "Automatic Classification of E-mail Message by Message Type, Journal of the American Society for Information Science," 48 (1), 1997, pp. 32-39.

11. Ng, H. T., W. B. Goh, K. L. Low, "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization," In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-97), 1997, pp. 67-73.

12. Oard, D. W., "Adaptive Filtering of Multilingual Document Streams," In Fifth RIAO Conference on Computer Assisted Information Searching on the Internet, 1997.

13. Salton, G. and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, NY, USA, 1983.

14. Schütze, H., D. A. Hull and J. O. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem," In Proceedings of the 18th

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-95), 1995, pp. 229-237.

15. Watanabe, Y., M. Murata, M. Takeuchi, and M. Nagao, “Document Classification Using Domain Specific Kanji Characters Extracted by Method,” In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), 1996, pp. 794-799.

16. Witten, I. H., A. Moffat and T. C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, International Thomson Publishing Company Press, New York, 1994.

17. Yang, Y. and C. G. Chute, “An Example-Based Mapping Method for Text Categorization and Retrieval,” ACM Transactions on Information Systems, 12 (3), 1994, pp. 252-277.

18. Yang, Y., “Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval,” In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-94), 1994, pp. 13-22.

19. Yang, Y., “An Evaluation of Statistical Approaches to MEDLINE Indexing,” In Proceedings of the AMIA, 1996, pp. 358-362.

20. 林頌堅，“自動化文件分類在資訊服務上的應用，21世紀資訊科學與技術的展望，1998, pp. 255-280.

附錄　財經記事資料介紹

財經記事為卓越出版社出版的財經新聞資料，每則共分為三部分：第一部份為分類碼、日期及報社等，第二部份為編者，通常是記者；第三部份為新聞標題。附表一為取自財經記事的部分範例。財經記事採階層式分類，共分為大類別、中類別、及小類別三層。其中，大類別共分為金融、產業等九類，附表二為大類別的類別種類及文件分佈情形。大類別之下又細分為 39 個中類別，其分類情形如附表三。由附表三，我們可以看出平均每則新聞標題約含 60 個位元組，即約 30 個中文字長。另外，由最長及最短兩個欄位，我們可以發現文件的長度變化很大，長度短於 5 個中文字的標題通常是不完整的文件。原本財經記事的分類，在中類別下以序號細分為小類別，但本研究僅以大、中兩個類型進行實驗。附表四及附表五分別為『總體篇』下的『經濟』及『稅賦』兩個中類別往下細分的情形。

附表一 財經記事部分範例

[1]HX.7104;810106;經;16(M);1;	[2]任文姍,江東峰	[3]日本區域性股市與店頭市場抬頭
[1]HF.01;810106;經;26(M);1;	[2]陳嫣妮	[3]經部商業司及多處機構全力推動,81 年商店自動化帶來市場新契機
[1]HB.9005;810106;聯;09(M);1;	[2]編輯部	[3]積極尋求與獨立國協會員國發展關係,中共與烏克蘭及塔吉克建交
[1]HDl.20;810106;聯;06(M);1;	[2]吳媛華	[3]3 分之 1 女性專任太太,鼓勵已婚婦女回到就業市場,勞委會擬妥草案報院
[1]HGb.01;810112;中晚;04(M);1;	[2]初馨怡	[3]新銀第 1 年可吸收存款 1000 億,佔資金市場 35 分之 1 金融環境將質變,產生消費者的銀行
[1]HB.9010;810112;中晚;03(S); 1;	[2]社評	[3]大陸企業來台投資
[1]J.60;810112;自早;02(M);1;	[2]社論	[3]妥善因應對韓關係的可能變動
[1]J.2007;810112;自晚;03(S); 1;	[2]彭琳淞	[3]兩岸條例延而難立
[1]J.01;810112;自晚;03(M);1;	[2]黃煌雄	[3]給黨團成員的一封公開信
[1]HJd.11;810112;工;01(M);1;	[2]陳志峻	[3]第 2 波擴大適用加值稅小店舖選定,鐘錶眼鏡等 9 行業須開立發票,決自 4 月 1 日起,展開強力輔導
[1]HX.12;810117;貿;05(L);1;	[2]編輯部	[3]我輸加鞋類退居第 3,業者應多警惕(加拿大)(附表 1:1990 年加拿大鞋類進口主要供源)
[1]J.10;810118;聯晚;02(M);1;	[2]社論	[3]民主化的嚴峻考驗:評立法院正副院長選舉
[1]HN.20;810118;聯晚;15(M);1;	[2]方紫苑	[3]我思・我捐系列(5):慈善有夢夢當圓,聯合勸募,加油!

附表二 財經記事大分類表

類別名稱	文件篇數	類別名稱	文件篇數	類別名稱	文件篇數
公營事業篇	2017	金融篇	12404	貿易篇	3308
其他	38432	國際篇	15742	農業篇	1777
服務業篇	12669	產業篇	16960	總體篇	29293

附表三 財經記事中分類表

大分類	中分類					
類別名稱	代碼	類別名稱	文件數	文件長度（位元組）		
				平均	最長	最短
公營事業篇	HDn	公營事業	2017	59	250	6
其他	HBp	人口、移民	223	55	162	8
	HDm	公共建設	3154	65	212	8
	HEs	郵政,電信	465	58	164	8
	HEm	大眾傳播	79	46	236	10
	HJ	財政	908	60	218	10
	HN	社會	2280	54	240	8
	J	政府,政治	15086	57	248	0
	L	教育	2368	59	251	10
	RA	醫療衛生	2383	61	201	8
	T	科技	758	61	157	6
	TD	環境	2927	60	229	6
	HFm	企業管理	3280	41	246	8
	W	公司檔案	956	51	215	7
	Hp	人物檔案	2564	54	234	8
	HPt	人事動態	770	51	218	8
	WG	集團企業	231	56	158	8
服務業篇	HF	服務業	8165	54	247	6
	HE	交通運輸業	2766	63	237	8
	HEt	觀光旅遊	1738	53	232	8
金融篇	HG	金融	3166	59	235	4
	HGb	銀行	2874	60	214	10
	HGe	外匯	564	56	157	4
	HGs	股票	4758	62	251	6
	HGt	租賃	1042	57	186	8
國際篇	HX	國際政經	15742	50	250	4
產業篇	HDo	各項產業	16960	57	240	4
貿易篇	HFt	貿易	3308	64	248	4
農業篇	S	農業	1250	60	206	8
	SD	林,牧,漁,礦	527	58	169	10
總體篇	HB	經濟	17404	60	251	4
	HBc	消費	1296	47	182	4
	HD	土地	1536	64	240	8
	HDi	工業	1966	62	238	8
	HDl	勞工	2858	59	201	10
	HJd	稅賦	2949	58	241	7
	HJt	關稅	384	62	176	14
	HFc	商標,智慧財產	900	57	181	8

附表四　總體篇經濟（HB）之細層分類

分類碼	類別名稱	分類碼	類別名稱
HB.01	綜合動態(利益輸送)	HB.20	經濟研究機構
HB.02	管理法令政策	HB.30	經濟辭彙
HB.03	統計數據,指標	HB.40	經濟建設計劃(六年國建)
HB.0301	經濟成長	HB.50	技術合作
HB.0302	國民生產毛額	HB.60	整廠輸出
HB.0303	國民所得	HB.70	自由化國際化
HB.0304	國民儲蓄	HB.80	經濟犯罪
HB.0305	物價	HB.85	地下經濟,走私,洗錢
HB.04	景氣循環	HB.8501	地下工廠(攤販)
HB.05	經濟史	HB.90	大陸政經
HB.06	生產力(經濟生產變化)	HB.9001	綜合動態
HB.10	投資,海外合作發展基金	HB.9002	貿易
HB.1001	對外投資	HB.9003	產業(含企業集團)
HB.1002	僑外投資(華僑)	HB.9004	商業(含金融,服務業)
HB.1003	投資環境介紹	HB.9005	政治
HB.1005	創業頭資(V.C)	HB.9010	兩岸經濟交流(貿易)
HB.1007	企業購併(事件性)	HB.9020	其他交流
HB.15	中小企業(自創品牌)	HB.9030	民運

附表五　總體篇稅賦（HJd）之細層分類

分類碼	類別名稱	分類碼	類別名稱
HJd.01	綜合動態	HJd.30	地方稅
HJd.02	管理法令政策	HJd.3001	土地稅
HJd.05	賦稅改革委員會	HJd.3003	房屋稅
HJd.10	稅捐稽徵(含機關)	HJd.3005	契稅
HJd.11	統一發票	HJd.3007	加值型營業稅
HJd.13	逃漏稅	HJd.3009	牌照稅
HJd.20	國稅	HJd.3011	印花稅
HJd.2001	個人所得稅(綜合所得稅)	HJd.40	規費(行政費用)
HJd.2003	營利事業所得稅	HJd.4001	商港建設費
HJd.2005	遺產,贈與稅	HJd.4003	工程受益費
HJd.2007	貨物稅	HJd.4005	都市建設捐
HJd.2009	證券交易稅,證券交易所得		

Automatically Controlled-Vocabulary Indexing for Text Retrieval

Kuang-hua Chen and Chien-tin Wu
Department of Library and Information Science
National Taiwan University
1, SEC. 4, Roosevelt RD.
Taipei, TAIWAN, 10617, R.O.C.
E-mail: khchen@ccms.ntu.edu.tw; jtwu@steelman.lis.ntu.edu.tw

Abstract

The IR society has made efforts in free-term indexing for a long time. By contrast, few efforts are made in controlled-vocabulary indexing. A new model for controlled-vocabulary indexing is proposed in this paper. This proposed model, TF×OSDF×CSIDF, distinguishes subject-specific words from common words and domain-specific words in documents. 60,400 MEDLINE records are used as training data and testing data and 100 MeSH subject headings are used as the testing controlled vocabularies. The preliminary experiments show good results. The precision and the recall concurrently exceed 90% using abstracts as training materials. The precision reaches 90% and the recall still keeps at 70% using title only. The problem of indexer's consistency could be alleviated using the proposed model to automatically generate index terms.

1. Introduction

The quality of indexing not only depends on professional knowledge and experience of librarians or subject specialists, but also is restricted by time and cost. Lack of indexing experts and subject specialists, the information exploration has confronted libraries with manpower problem. In addition, the issue of indexer consistency still cannot be resolved effectively. In 1950, researchers started to employ machine to enhance indexing process. In recent years, the Internet has made the subject access become the mainstream of information seeking behavior and prompted researches of automatic indexing, classification and abstracting. The researchers of automatic indexing always take complete substitution of human indexing as the ultimate goal. Although there is a long way to go, many researchers claim that the performance of automatic indexing is the same as that of manual indexing at least. (Cleverdon and Mills, 1963; Cleverdon, 1976)

Most researches of automatic indexing focus on the free-term indexing. (Salton, 1988; Ponte and Bruce Croft, 1998) By contrast, the researchers do not pay much attention to the automatic indexing for controlled vocabularies. The free-term indexing is to identify keywords or key phrases, which represent subjects of document and use them as index terms directly. Basically, these keywords and key phrases couldn't represent true "concept" of user's information need. As to controlled-vocabulary indexing, indexer has to translate subject concepts into controlled vocabularies. From this viewpoint, controlled-vocabulary indexing may be regarded as concept indexing. Besides, the free-term indexing usually increases recall rather than precision. This has pushes researchers to study controlled-vocabulary indexing for information retrieval again.

This paper proposes a new model, uses titles and abstracts of documents which have been indexed manually as the training materials, and makes controlled-vocabulary indexing automatic easily. Section 2 discusses the idea and proposes the new model. Section 3 describes the design of experiments and carries out a series of experiments. Section 4 discusses the experimental results in detail. Section 5 is the short conclusions.

2. The Idea and the Proposed Model

The idea behind the proposed model is based on the content-bearing words. It is assumed that there should be some kind of relationships among controlled vocabularies and content-bearing words. If some content-bearing words are found in a document, the related controlled vocabulary should be assigned to the document.

The training process will construct a function between document and subject headings (a set of controlled vocabularies). After this function is determined, documents could be transferred into correspondent feature values, and then calculate indexing scores of documents for certain subject headings. Indexing score implies the possibility that documents are indexed in some subject headings.

The previous researches on automatic indexing have been associated with the exploitation of statistical techniques. Luhn (1997) considered that the justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and he elaborates on an aspect of a subject. Salton (1989) suggested that the general aim of statistical measures be to reject both very high and very low frequency words from the texts being indexed.

In tradition, the considered types of frequency are term frequency (TF) and document frequency (DF). DF is often transferred into inverse document frequency (IDF) while adopted. TF indicates occurrence of word in a document, while DF refers to distinct occurrence of word in a document collection. If there are N documents in a document collection, IDF is represented as *log*(*N*/*DF*). (Sparck Jones, 1972) Low IDF will decrease weight of word and make word be rejected from the candidate list of index terms. Although IDF has adopted widely in various indexing models and has been verified as an effective measure for weighting words, it also filters out some kind of words with great benefit in subject identification. Take Figure 1 as an example.

A+B+C = Words with high DF and low IDF
A = Common Words
B = Domain-specific Words
C = Subject-specific Words

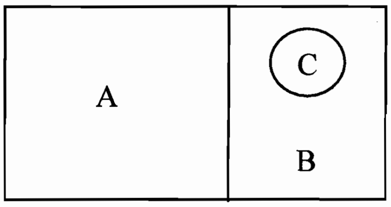


Figure 1. Words with Low IDF

In Figure 1, the largest rectangle indicates the words with low IDF. A, B and C areas within rectangle refer to common words, domain-specific words and subject-specific words, respectively. The words of A and B cannot offer useful information in subject identification actually. For example, in documents discussing education, there are high-frequency words, such as “education”, “school”, “teacher”, and “student”, which are not discriminative enough. Unlike words of A and B, the words of C benefit subject identification greatly. Take documents concerning about AIDS as instance. Occurrence of AIDS in such kind of documents should be very high. If IDF is adopted as unique measurement, then it will assign low weight to AIDS which reflects subject of these documents, and then AIDS will be rejected finally.

In general, there should exist some subject-specific words in the documents with the same subject, which are highly close to the subject and with high occurrence. As mentioned above, although subject-specific words are very useful, the IDF measurement cannot distinguish them from common words and domain-specific words in a document collection with the same subject. Therefore, we could not use IDF directly. That is to say, the indexing model has to be enhanced with a capacity to separate subject-specific words from high DF words and increase weight of subject-specific words while evaluating significance of words. Increasing no additional training documents, a new model is proposed for solving problem mentioned above. In this new model, the training documents, which originally belong to distinct subject headings, will be combined into a document. Please refer to Figure 2. For convenience, the following description takes original training documents as "original set", and the combined documents as "combined set". The new model will weight words using the multiplication of TF in documents, DF in original set (OSDF) and IDF in combined set (CSIDF).

The distributional tendency of common words, domain-specific words and subject-specific words are shown in Table 1. Common words and domain-specific words will be of lower CSIDF. CSIDF of a word will not be changed for different subject headings. Therefore, common words and domain-specific words will be of low weight when they found in documents of different subjects. Unlike CSIDF, OSDF of a word varies by different subject headings, and some subject-specific words with high weight will be figured out for certain subject headings.

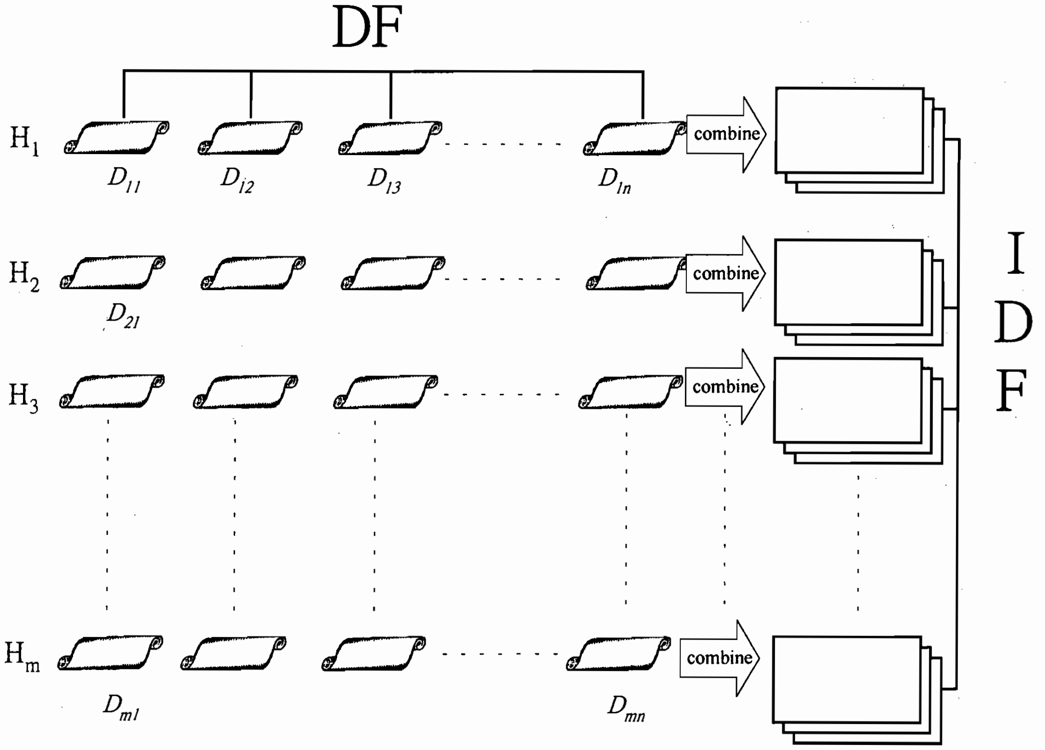


Figure 2. Diagram for Calculation of Term Weight

Table 1. Distributional Tendency of Words

	OSDF	CSIDF
Common Words	High	Low
Domain-specific Words	High	Low
Subject-specific Words	High	High

3. Experiments

3.1 Learning Process

We choose MEDLINE (MEDlars onLINE) as the source of training documents (MEDLINE, 1998), and MeSH (Medical Subject Headings) as the controlled vocabularies. (Medical Subject Headings, 1998) Since the documents were collected from late of 1997, subject headings for training are extracted from 1997 MeSH Tree, not 1998 MeSH Tree. A sample training text is shown in Figure 3.

Title
A method to test blood flow limitation of peritoneal-blood solute transport.
Local Messages
Undefined
Abstract
Current transperitoneal transport models assume that effective blood flow to the microcirculation does not limit solute exchange with dialysate in the cavity. Despite evidence that gas transfer across the peritoneum (assumed to equal the effective blood flow) occurs at rates that exceed maximum urea transfer rates by a fact or of two to three, the assumption has been strongly challenged. To address this problem at the tissue level, a technique to determine the effect of local blood flow on small-solute transport was developed in this study. Diffusion chambers were affixed to the serosal side of the anterior abdominal wall of rats, and solutions containing radiolabeled urea or mannitol were placed in the chambers. During each experiment, the local blood flow beneath the chamber was monitored with laser Doppler flowmetry and the disappearance of the tracer versus time was simultaneously measured under three conditions of blood flow: control, 30% of control, and zero blood flow. The results demonstrated no significant differences for either solute between control and the condition in which blood flow was reduced by 70%. However, there was a significant reduction in the rate of mass transfer with no blood flow. It was concluded that blood flow at > or = 30% of control values does not limit solute transfer across the abdominal wall peritoneum during dialysis.

Figure 3. Sample Text of MEDLINE Record

One hundred of subject headings of 1997 MeSH Tree were selected for training and testing. The selection of subject headings is a crucial step. In order to test the model in terms of average performance, the extracted subject headings should equally distribute in the MeSH Tree. This could avoid subject headings concentrating on certain fields.

Besides distribution, both depth and width of the subject headings in MeSH Tree have to be taken into account. On the one hand, the amount of records in database could be used to judge the width of a subject heading. Therefore, the subject headings associated with 1,000 to 2,000 records from 1991 to 1997 in MEDLINE were chosen. On the other hand, the distance from root or leaf node to a subject heading could reflect relative depth. The subject headings with the following criteria are chosen.

- Distance to root equal or more than 1 layer and less than 4 layers
- Distance to leaf are less than 6 layers

The average distances to root and to leaf are 1.58 and 1.62, respectively.

600 records for each subject heading under consideration are collected. 400 out of 600 records are for positive training; the others are for positive testing. There are totally 400 records collected for negative testing, which are not indexed by any subject headings under consideration. Table 2 lists detailed statistics of records in the three sets. The collected records only contain titles and abstracts. The volume of whole experimental records is 84 MB: positive training set is 56 MB; positive testing set is 27.8 MB; negative testing set is 654 KB. The collected records are English documents and the average number of words in abstract and title are 107.6 and 11.3, respectively, after filtering out the stop words.

Assume there are m subject headings, H_1, H_2, H_3, …, H_m and l distinct words, W_1, W_2, W_3,…, W_l. We will take H_j as instance to illustrate the following experimental processes.

Table 2. Amount of Experimental Records

	Training Set	Testing Set	Total
Positive	40,000	20,000	60,000
Negative	--	400	400
Total	40,000	20,400	60,400

The OSDF and CSIDF of words in the training set are calculated. Formula of CSIDF is shown below,

$$CSIDF(W) = \log_2\left(\frac{P - O(W)}{O(W)}\right)$$

whereas P represents the amount of documents in combined set, $O(W)$ is the number of documents which contain word W. CSIDF is negative when W appears more than half of documents in combined set. After this step, relationship between H_j and a set of words R_j is constructed. Words with high weight in R_j have the high possibility to be subject-specific words. R_j can be regarded as the weighted vector shown as follows,

Subject Heading	Weighted Vector
H_1	$R_1=\{w_{11}, w_{12}, \ldots, w_{1k}, \ldots w_{1l}\}$
H_2	$R_2=\{w_{21}, w_{22}, \ldots, w_{2k}, \ldots w_{2l}\}$
…	…
H_j	$R_j=\{w_{j1}, w_{j2}, \ldots, w_{jk}, \ldots w_{jl}\}$
…	…
H_m	$R_m=\{w_{m1}, w_{m2}, \ldots, w_{mk}, \ldots w_{ml}\}$

whereas w_{jk} is the weight (OSDF×CSIDF) in R_j for the word W_{jk}.

3.2 Evaluating Process

After R_j has been constructed, we will calculate indexing score (*IS*) of a document for each subject heading according to the weighted vectors. The *IS* is shown as follows.

$$IS = \frac{\sum(OSDF \times CSIDF) \times (TF)}{\text{number of words in the document}}$$

The normalization is used to avoid favoring the lengthy documents. When an unseen document appears, the process of automatic indexing will preprocess it first, and then compute the indexing score based on each subject heading. The lower *IS* indicates that this document should not be indexed by H_j; the higher *IS* indicates that it is likely to assign H_j as one index term for this document.

An indexing threshold T_j, is determined to distinguish documents with high *IS* from those with low *IS*. If *IS* is larger than T_j, the subject heading H_j will be assigned to the document. Otherwise, it will not. Because, it is not easy to determine a threshold, ten thresholds from 0.1 to 1.0 were used in our experiments. Finally, one best threshold will be chosen.

The precision and recall are used for performance evaluation. Note that the precision and recall are from viewpoint of subject headings rather than from documents. The precision and recall of individual subject heading (H_1-H_{100}) will be merged to compute the average precision and average recall for the final evaluation.

4. Experimental Results

There are four sets of data in experiments: data for abstract part both in training set and testing set, data for title part both in training set and testing set. The detailed results of abstract part could be referred to Table 3; those of title part can be referred to Table 4.

4.1 The Precision vs the Recall

Let's consider the performance for the abstract part of training set. When the threshold is between 0.3 and 0.6, both the precision and recall are higher than 90%. When the threshold equals to 0.43, both the precision and recall are higher than 94% according to the interpolation. Consider the performance for the abstract part of testing set. While the threshold equals to 0.43, both the precision and recall are higher than 90%.

Consider the performance for the title part of training set. When the threshold equals to 0.4, the precision is higher than 90% and recall is higher than 70%. When the threshold equals to 0.1, both the precision and recall are higher than 76%. Consider the performance for the title part of testing set. When the threshold is close to 0.3, the precision reaches 90% and recall reaches 70%. When the threshold is 0.1, both the precision and recall are higher than 78%.

Table 3. Precision & Recall of Abstract Part (TF×OSDF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	77.31%	99.62%	87.06%	76.78%	96.63%	85.57%
0.2	87.17%	98.83%	92.63%	86.47%	92.90%	89.57%
0.3	91.68%	97.36%	94.43%	91.01%	89.49%	90.24%
0.4	93.92%	95.32%	94.61%	93.32%	86.19%	89.61%
0.5	95.49%	92.88%	94.17%	95.00%	83.39%	88.82%
0.6	96.33%	90.24%	93.19%	95.91%	80.62%	87.60%
0.7	96.92%	87.31%	91.86%	96.56%	77.87%	86.21%
0.8	97.32%	84.51%	90.46%	97.01%	75.51%	84.92%
0.9	97.62%	81.69%	88.95%	97.35%	73.15%	83.53%
1.0	97.83%	79.09%	87.47%	97.59%	71.09%	82.26%

Table 4. Precision & Recall of Title Part (TF×OSDF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	80.52%	76.48%	78.45%	80.87%	78.24%	79.53%
0.2	86.85%	74.86%	80.41%	86.78%	74.38%	80.10%
0.3	89.94%	72.86%	80.50%	89.67%	70.76%	79.10%
0.4	91.92%	70.73%	79.94%	91.54%	67.34%	77.60%
0.5	92.98%	68.50%	78.88%	92.57%	64.48%	76.01%
0.6	93.60%	66.09%	77.48%	93.18%	61.76%	74.28%
0.7	94.10%	63.78%	76.03%	93.71%	59.58%	72.85%
0.8	94.51%	61.47%	74.49%	94.14%	57.67%	71.52%
0.9	94.88%	59.30%	72.98%	94.56%	55.67%	70.08%
1.0	95.19%	57.12%	71.40%	94.92%	53.92%	68.77%

4.2 The Abstract vs the Title

Basically, words in titles are much fewer than those in abstracts. Therefore, the performance of title part is supposed to be unstable. The experimental results show the consistency with this prediction. In comparison with abstract part, the recall of title part is 80% of that of abstract part. Although the recall agrees with the original supposition, the precision is much better than the predicted one. In fact, the precision of title part could reach 95%. Generally speaking, the title contains lots of useful information, which is very effective in subject identification, and worthy of using in the construction of indexing models.

Take testing subject heading 001 as instance. Table 5 is the statistics of average indexing score of testing documents in abstract and title parts. In 200 documents indexed by subject heading 001, the indexing scores of title part are divergent. The higher standard deviation and range reveal unbalanced distribution of indexing scores and imply the higher possibility of error. Although the recall in title part decreases, the precision does not drop too much. The stable precision indicates that once the title provides information, it will be useful.

Table 5. Indexing Score of Documents in Testing Set

Heading 001		Range	Min.	Max.	Mean	Std. Dev.
Positive Document	Abstract	16.70	0.18	16.72	3.53	3.45
	Title	38.96	0.00	38.96	8.27	8.35

4.3 The Proposed Model vs the Traditional Model

We carry out the same experiments for the traditional model as a baseline model. These experiments are divided into two parts: one is for TF×OSIDF; the other is for TF×CSIDF.

The experimental results of TF×OSIDF show that the precision and the recall are zero when threshold is 0.1. Obviously, subject-specific words play the crucial role in subject identification. As to TF×CSIDF, although the training documents belong to different subjects, the weight of subject-specific word is not enhanced without the aid of OSDF. Despite of the better performance than TF×OSIDF, the performance of indexing model using TF×CSIDF is still inferior to our model. Table 6 and Table 7 show the performance of TF×CSIDF in details.

In comparison with the traditional model in terms of recall (please refer to Table 3 and 4), our model not only shows less diversity in the training set and the testing set, but also performs stably. Because our model identifies the importance of subject-specific words, we shorten the gap between training set and testing set. By contrast, the recall of traditional model in testing set drops quickly. The largest difference between traditional model and our model in recall is higher than 94%; the least difference is also higher than 60%. In terms of precision, some results of the traditional model are better than those of our model. However, it sacrifices the recall to the precision.

Table 6. Recall and Precision of Abstract Part (Traditional Model TF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	86.67%	97.80%	91.90%	84.86%	84.30%	84.58%
0.2	93.33%	89.01%	91.12%	90.51%	60.65%	72.63%
0.3	95.08%	73.26%	82.76%	91.18%	39.20%	54.83%
0.4	96.10%	54.45%	69.51%	91.54%	23.92%	37.93%
0.5	97.09%	38.04%	54.66%	92.62%	14.30%	24.77%
0.6	98.53%	24.84%	39.68%	95.55%	7.94%	14.66%
0.7	99.81%	15.46%	26.77%	99.34%	4.52%	8.65%
0.8	99.89%	9.47%	17.30%	99.60%	2.46%	4.80%
0.9	100.00%	5.70%	10.79%	100.00%	1.36%	2.68%
1.0	100.00%	3.45%	6.67%	100.00%	0.71%	1.41%

Table 7. Recall and Precision of Title Part (Traditional Model TF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	87.81%	90.38%	89.08%	84.94%	70.80%	77.23%
0.2	92.34%	84.16%	88.06%	89.35%	58.54%	70.74%
0.3	94.43%	76.79%	84.70%	91.31%	47.57%	62.55%
0.4	96.46%	68.30%	79.97%	93.71%	37.38%	53.44%
0.5	98.00%	59.40%	73.97%	95.99%	28.93%	44.46%
0.6	99.24%	50.77%	67.17%	98.27%	22.18%	36.19%
0.7	99.69%	42.44%	59.53%	99.24%	17.08%	29.14%
0.8	99.91%	34.94%	51.77%	99.77%	12.90%	22.85%
0.9	99.96%	28.33%	44.15%	99.90%	9.84%	17.92%
1.0	100.00%	22.64%	36.92%	100.00%	7.36%	13.71%

5. Conclusions

A new indexing model is proposed for controlled-vocabulary indexing in this paper. Increasing no additional training documents, the new model uses various frequencies through combination and separation of the same training documents, and distinguishes subject-specific words from common words and domain-specific words. The preliminary experiments show good results using 100 MeSH subject headings and 60,400 abstracts and titles. The precision and recall concurrently exceed 90% using abstracts as training materials. As to title, the precision reaches 90% and the recall still keeps at 70%.

The future works should consider phrase terms, enhance the indexing procedure, and test the performance for full texts. Firstly, phrases bear more semantic information than single words. Therefore, the performance of indexing model will be improved using phrase terms. Secondly, it's not efficient for a system to compute index features of all controlled vocabularies in the present design. Clustering could be employed to deal with the problem. Thirdly, there are more and more online full-text databases in recent years. We could use full texts as training materials rather than abstracts and titles.

References

Cleverdon, C. W. (1976), "The Cranfield Tests on Indexing Language Devices," *Aslib Proceedings*, vol. 19, no. 6, pp. 173-194.

Cleverdon, C. W. and J. Mills (1963), "The Testing of Index Language Devices," *Aslib Proceedings*, vol.15, no. 4, pp.106-130.

Luhn, H. P. (1997), "The Automatic Derivation of Information Retrieval Encodements from Machine-Readable Texts," *Readings in Information Retrieval*, Morgan Kaufmann Publishers, Inc., San Francisco, pp. 21-24.

Ponte, J.M. and W. Bruce Croft (1998), "A Language Modelling Approach to Information Retrieval," *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275-281.

Salton, Gerard (1988), "Syntactic Approaches to Automatic Book Indexing," *Proceeding of the 26th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, NJ, pp.120-138.

Salton, Gerard (1989), *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, Inc., New York, p. 275.

Sparck Jones, K. (1972), "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation*, vol. 28, no. 1, pp.11-21.

A New Syllable-Based Approach for Retrieving Mandarin Spoken Documents Using Short Speech Queries

Hsin-min Wang

Institute of Information Science, Academia Sinica

Taipei, Taiwan 115, Republic of China

E-mail: whm@iis.sinica.edu.tw

Abstract

Intelligent and efficient information retrieval techniques allowing easy access to huge amount and various types of information become highly desired and have been extensively studied in recent years. Considering the fast growth of audio resources and the characteristic monosyllabic structure of the Chinese language, a syllable-based framework of retrieving Mandarin spoken documents using speech queries has been investigated. This paper presents a new syllable-based approach that is based on matching the whole syllable lattice directly instead of using the syllable or syllable pair information extracted from the syllable lattice. The experimental results show that the retrieval performance can be significantly improved.

1. Introduction

The network technology and the Internet are creating a completely new information era. With the rapidly growing audio and multi-media information on the Internet, a variety of exponentially increasing spoken documents such as the broadcast radio, television programs, video tapes, digital libraries, courses and lectures and so on, are now being accumulated and made available via the Internet. But most of them are simply stored there, kind of difficult to be further reused for lack of efficient retrieval technology. Development of the technology to retrieve speech information thus becomes essential and gets more and more important. Recently, with the advances in speech recognition technology, proper integration of information retrieval and speech recognition has been considered by many researchers (Bai et al., 1996, 1999; CMU Informedia, Glavitsch and Schäuble, 1992; James, 1995; Jones et al., 1996; Ng and Zue, 1997; Wechsler, 1998). In any case, retrieval of speech information using speech queries directly is apparently the most natural, convenient and attractive, although the technology involved will be the most difficult as well. This is because in such cases both the information to be retrieved and the input queries are in form of voice instead of texts, thus with unknown variabilities on both sides. For Chinese language, because the language is not alphabetic and the input of Chinese characters into computers has been a very difficult and unsolved problem even today, voice retrieval of speech information will be much more important and attractive for Mandarin Chinese than that for other languages.

Unlike the text information, the speech information can't be retrieved at all by directly comparing the input speech queries with the spoken documents. Therefore both the speech queries and the spoken documents must be transcribed into some kind of content features, such as phone strings or lattices, texts, keywords or concepts and so on using speech recognition techniques, based on which the similarity between the speech queries and the spoken documents can then be measured. Thus, there can be at least the keyword-based, the large-vocabulary-based, and the subword-based approaches. For the keyword-based approach (James, 1995; Jones et al., 1996), one can define a set of keywords for the spoken documents in advance, and whenever some keywords are extracted from the speech queries, the spoken documents with those or similar keywords can then be retrieved. This approach is efficient and cost-effective, and is very useful for retrieval of static databases with static queries, where the search words don't change frequently. However, usually it is not easy to define a set of adequate keywords for all the spoken documents to be retrieved unless we know the contents of all of them in advance. For the large-vocabulary-based approach, both the spoken documents and the speech queries are fully recognized into texts, thus many well-developed text retrieval techniques can be directly applied (CMU Informedia). However, for such an approach, the out-of-vocabulary problem is an important issue, since a large vocabulary speech recognizer needs a predefined lexicon for linguistic decoding, and some special words important for retrieval purposes, such as proper nouns (e.g. personal names or organization names), exotic words, and domain specific terms (e.g. special terms for business news or sports news), may be simply outside of this predefined lexicon. This leads to the concept of making comparison on the level of subword units instead, or the subword-based approach (Bai et al., 1996, 1999; Glavitsch and Schäuble, 1992; James, 1995; Ng and Zue, 1997; Wechsler, 1998). Because it is much more easier to obtain all necessary subword units to cover all possible pronunciations of a given language, the out-of-vocabulary problem existing in the ever-growing speech information may be somehow handled by directly measuring the similarity between the spoken documents and the speech queries on the subword level instead of on the word level. Because in such approaches the subword units are never decoded into words, therefore the retrieval is never limited by any lexicon either. Such a subword-based approach also has the advantages of bypassing the complicated lexicon matching and linguistic decoding processes, in addition to avoiding the out-of-vocabulary problem.

Considering the monosyllabic structure of the Chinese language, the syllable-based approach has been found to be an attractive special case of the subword-based approaches for retrieving Chinese text (Lin et al., 1995) and speech (Bai et al., 1996) information using speech queries. In this approach, the subword unit selected is the syllable due to various considerations on the characteristics of the Chinese language. The similarity between the spoken documents and the speech queries is measured on the syllable level based on the vector-space models widely used in many traditional text information retrieval systems. The feature vector of each document or query contains the presence information, frequency counts, and acoustic recognition scores of all syllables and adjacent syllable pairs in the syllable lattice obtained by speech recognition. However, the spoken document retrieval performance is obviously not satisfactory as compared to the upper-bound performance derived from text-based retrieval of transcripts of the spoken materials. As previously reported in Bai's work (Bai et al., 1999), for simple key phrase queries, the non-interpolated average precision rates (Harman, 1995) are 0.97 and 0.54 for text retrieval and speech retrieval respectively. While, for quasi-natural-language queries, the non-interpolated average precision rate for speech retrieval is 0.43, which is even worse. Of course the serious performance degradation is due to speech recognition errors and the increased ambiguity comes from the syllable lattice itself. Although both the single syllable information and the syllable-pair information extracted from the syllable lattice maybe somehow robust to speech recognition errors, they are not precise enough and many wrong syllables and syllable pairs are also included in the feature vectors. So good retrieving approaches should be able to make use of the increased correct syllables contained in the syllable lattice to achieve better results. Accordingly, in this paper, we propose a new syllable-based approach that is based on matching the whole syllable lattice directly instead of using the syllable and syllable pair information extracted from the syllable lattice. This approach has been evaluated based on the task of Mandarin spoken document retrieval using short key phrase queries. The experimental results show that the retrieval performance can be significantly improved to 0.73, based on exactly the same task and the same speech recognition front-end previously used in Bai's evaluation.

In the following, the speech recognition process is first introduced in Section 2. Section 3 briefly reviews the previous syllable-based approach, and Section 4 presents the new syllable-based approach. Finally, all experimental results are discussed in Section 5, and the concluding remarks are made in Section 6.

2. Syllable Lattice Construction

In Mandarin Chinese, there exists a total of 1,345 phonologically allowed tonal syllables, and these tonal syllables can be reduced to 416 base syllables and 5 tones. Base syllable recognition is thus believed to be the first key problem for large vocabulary Mandarin speech recognition as well as spoken document retrieval considered here. However, although the base syllable is a very natural recognition unit for Mandarin Chinese due to the monosyllabic structure of Chinese language, it suffers from inefficient utilization of the training data in the training phase and high computation requirement in the recognition phase. Thus, context-dependent Initial/Final's (Wang et al., 1997) are widely used acoustic units for Mandarin speech recognition specially considering the monosyllabic nature in Mandarin Chinese and the Initial/Final structure in Mandarin base syllables. Initial is the initial consonant of the base syllable and Final is the vowel (or diphthong) part but including optional medial or nasal ending. Each Initial or Final is then represented by a left-to-right continuous HMM. To allow anyone to use the system naturally without training, the retrieval system is operated under the speaker-independent mode. That is, the speaker-independent context-dependent Initial/Final HMM's are used to recognize the syllables and construct the syllable lattices. These models are trained by a training speech database including 5.3 hours of speech for phonetically balanced sentences and isolated words produced roughly by 80 male and 40 female speakers. Also, to deal with the silence segments in the spoken documents or speech queries, a single state HMM is used to represent the silence.

Based on the acoustic models mentioned above, the speech recognition processes for the spoken documents are described as follows: First, the speech recognizer performs the Viterbi search (Rabiner and Juang, 1993) on the whole spoken documents and outputs the best syllable sequence and the corresponding syllable boundaries. Then, based on the state likelihood scores calculated in the Viterbi search and the syllable boundaries of the best syllable sequence, the speech recognizer performs the second Viterbi search on each utterance segment which may include a syllable and outputs several most possible syllable candidates with their acoustic recognition scores. After the above speech recognition processes, a syllable lattice can be easily constructed.

The acoustic recognition score, $\log p(O \mid s)$, for a certain syllable candidate s in the syllable lattice and the feature vector sequence O for a certain speech utterance segment is first normalized with respect to the duration of the observed speech segment, and then transformed into a range between 0 and 1 by a Sigmoid function $\zeta(x)$,

$$\zeta(x) = \frac{1}{1 + \exp\left(-\alpha \cdot (x - \beta)\right)} \tag{1}$$

where α and β are used to control the slope and the range of the sigmoid function. Here, a simple utterance verification scheme is used to filter out the syllable candidates with less possibilities. Initially, 20 syllable candidates are obtained for each syllable segment after speech recognition, while only those with the acoustic recognition scores larger than a threshold can be left after utterance verification. The depth of the syllable lattice thus can be adjusted by simply changing the threshold value, and a more compact syllable lattice can be obtained.

3. The Previous Approach

This section will briefly review the overall system architecture, the feature vector, and the retrieving process of our previous syllable-based approach for retrieval of Mandarin spoken documents (Bai et al., 1996, 1999). This approach is primarily based on the vector-space models widely used in many traditional text information retrieval systems.

3.1 Overall System Architecture

The overall architecture of the previous syllable-based approach for Mandarin spoken document retrieval is shown in Figure 1. The whole system can be divided into three parts. The first part in the upper dotted square of Figure 1 is the off-line processing subsystem. All processes in this part should be performed off-line in advance. The second part in the middle dotted square is the initialization subsystem. All processes should be performed in the system initialization stage. The third part in the lower dotted square is the on-line retrieval subsystem, in which all processes must be performed on-line in real-time. The detailed operation of each part will be described separately below.

In the off-line processing subsystem, for each collected spoken document, speech recognition with utterance verification techniques is first applied to generate a syllable lattice, including the acoustic recognition scores for all syllable candidates, and the syllable lattice is then added to the syllable lattice database D_l. In this way, the most time consuming speech recognition process is performed off-line in advance, and all information necessary for retrieval is stored in the syllable lattice database D_l.

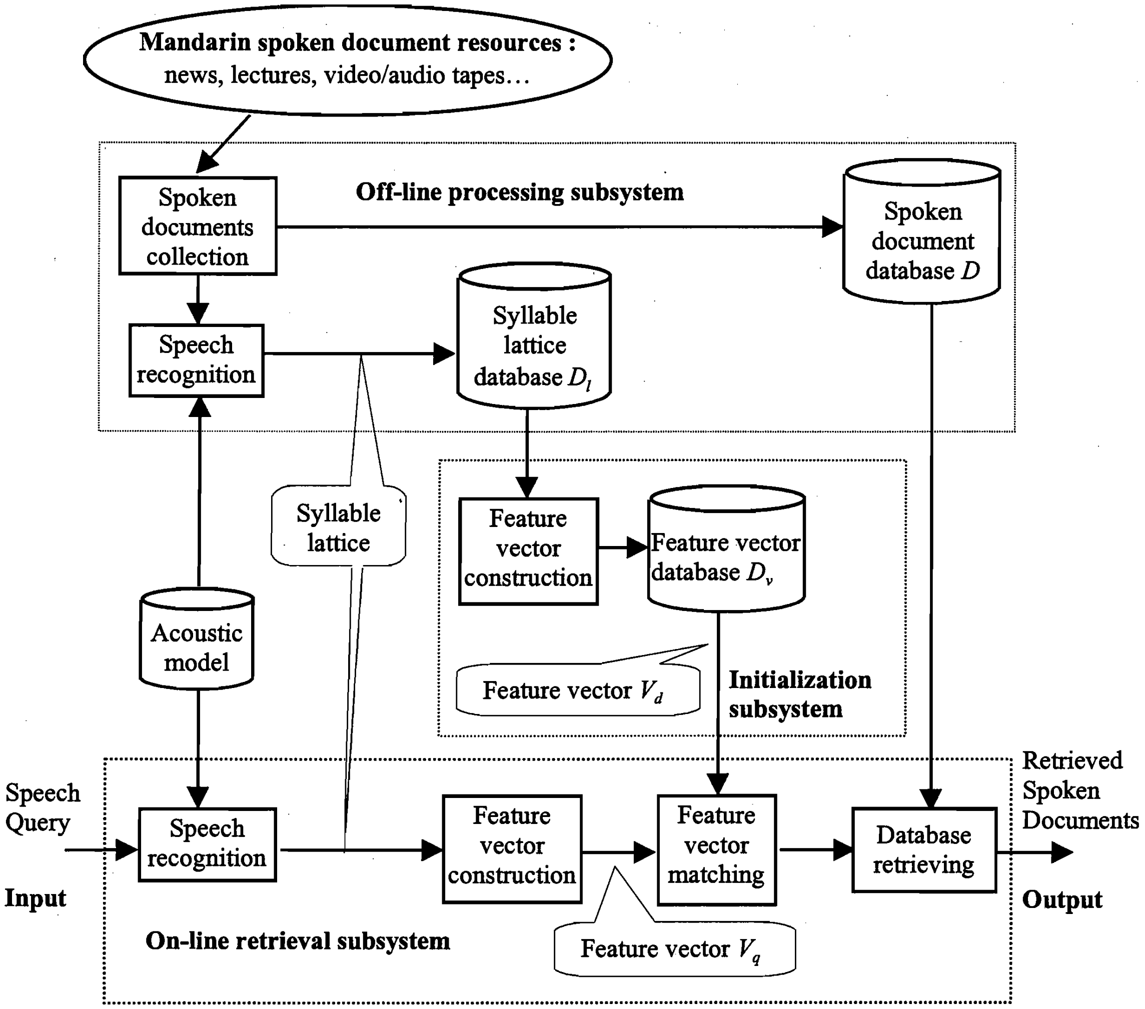

Mandarin spoken document resources :
news, lectures, video/audio tapes...
Spoken documents collection
Off-line processing subsystem
Spoken document database D
Speech recognition
Syllable lattice database D_l
Syllable lattice
Feature vector construction
Feature vector database D_v
Acoustic model
Feature vector V_d
Initialization subsystem
Speech Query
Input
Speech recognition
Feature vector construction
Feature vector matching
Database retrieving
Retrieved Spoken Documents
Output
On-line retrieval subsystem
Feature vector V_q

Figure 1: The overall architecture of the previous syllable-based approach for retrieving Mandarin spoken documents using speech queries.

The initialization subsystem is to obtain the feature vectors to be used for retrieval from the syllable lattice database D_l. The feature vector of each document contains the presence information, frequency counts, and acoustic scores of all syllables and adjacent syllable pairs in the syllable lattice. After the feature vectors have been constructed for all syllable lattices in the syllable lattice database D_l, the feature vector database D_v is established, which will be the target database to be physically retrieved. The whole process is also performed off-line in advance.

In the on-line retrieval subsystem, when a speech query is entered, speech recognition will first generate a syllable lattice for the speech query, and then the corresponding feature vector V_q will be constructed based on this syllable lattice via exactly the same processing procedures as those for spoken documents. Given the feature vector database D_v and the query feature vector V_q, the retrieving module then evaluates the similarity measure between V_q and all feature vectors of the database, and selects a set of documents with the highest similarity measures as the retrieving output.

3.2 Feature Vectors

For each spoken document *d* in the database *D*, through searching the syllable lattice, all acoustic scores of single syllables and adjacent syllable pairs in the syllable lattice can be extracted to form the feature vector V_d,

$$V_d = (as(s_1),...,as(s_i),...,as(s_{416}),as(s_1,s_1),...,as(s_i,s_j),...,as(s_{416},s_{416})) \tag{2}$$

where $as(s_i)$ is the acoustic score of the syllable s_i, and $as(s_i,s_j)$ is the acoustic score of the syllable pair (s_i,s_j). The feature vector constructing procedures were performed off-line on all documents in the database *D* to form a feature vector database D_v, which will be the target database to be physically retrieved. While regarding a query, the same feature vector constructing procedures must be performed on-line to construct the feature vector V_q right after the input query is entered.

3.3 Retrieving Process

Given the feature vector database D_v and a query *q*, the retrieving problem is actually a searching process to retrieve the document *d** in the target database D_v which is most related to the query. This searching process thus can be formulated as follows:

$$d^* \equiv \arg\max_{d \in D_v} Sim(d,q) \tag{3}$$

where *Sim(d,q)* is a similarity measure between a document *d* and the query *q*, and the Cosine measure (Salton, 1983) can be used to estimate the similarity:

$$Sim(d,q) = \cos(V_d,V_q) = \frac{V_d \cdot V_q}{|V_d||V_q|} \tag{4}$$

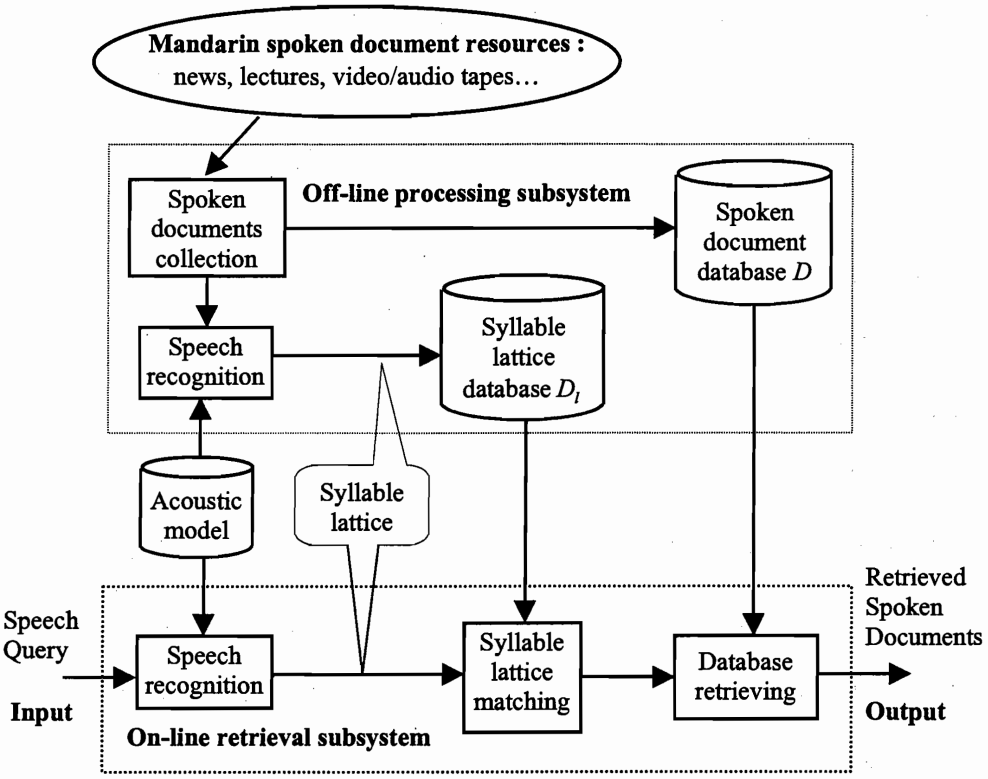

Mandarin spoken document resources :
news, lectures, video/audio tapes…
Spoken documents collection
Off-line processing subsystem
Spoken document database D
Speech recognition
Syllable lattice database D_l
Acoustic model
Syllable lattice
Speech Query
Input
Speech recognition
Syllable lattice matching
Database retrieving
Retrieved Spoken Documents
Output
On-line retrieval subsystem

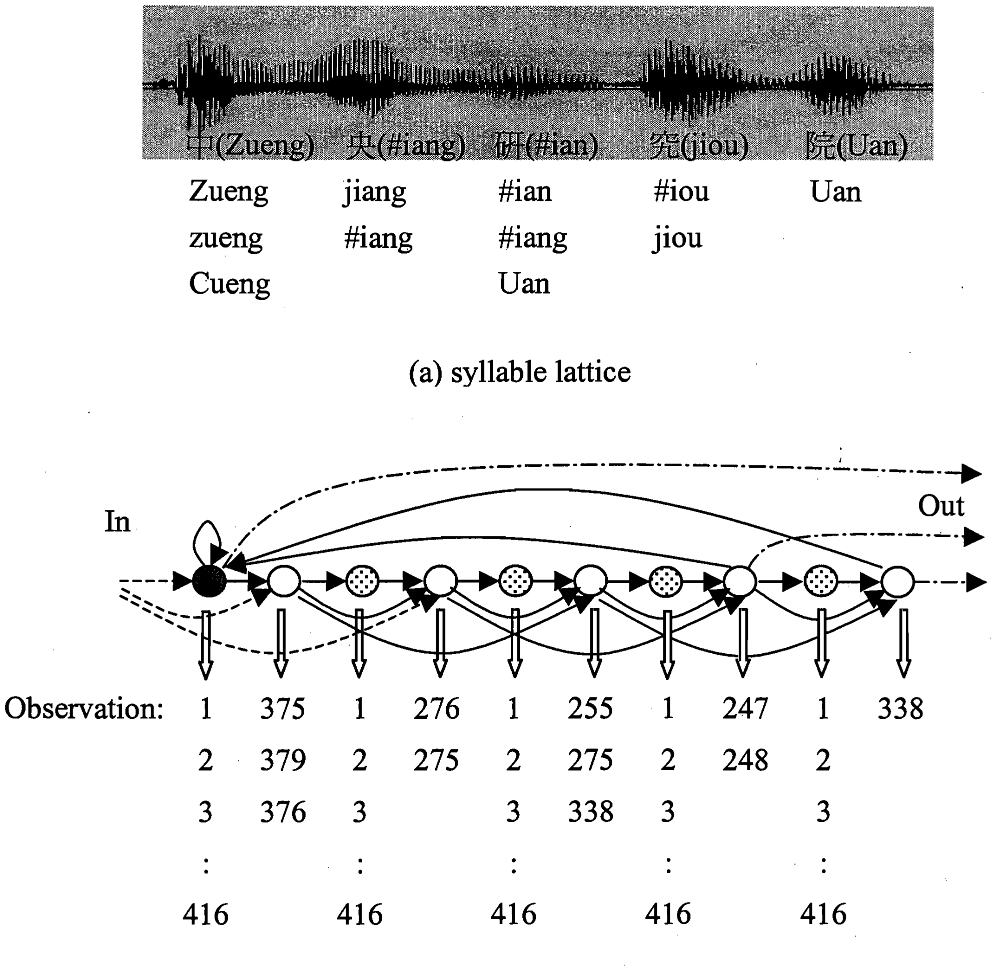
Figure 2: The overall architecture of the new syllable-based approach for retrieving Mandarin spoken documents using speech queries.

4. The New Approach

This section will briefly introduce the overall system architecture and the syllable lattice matching process of the proposed new syllable-based approach.

4.1 Overall System Architecture

The overall architecture of the new syllable-based approach for Mandarin spoken document retrieval is shown in Figure 2. The whole system is now divided into two parts. The first part in the upper dotted square of Figure 2 is still the off-line processing subsystem, which is exactly the same as the first part of the previous approach as shown in Figure 1. The second part in the lower dotted square is the on-line retrieval subsystem, in which all processes must be performed on-line in real-time. Note that here the similarity measure is based on directly matching the syllable lattice, the feature vector construction module in the previous approach is therefore no more necessary.


中(Zueng) 央(#iang) 研(#ian) 究(jiou) 院(Uan)
Zueng jiang #ian #iou Uan
zueng #iang #iang jiou
Cueng Uan
In
Out
Observation: 1 375 1 276 1 255 1 247 1 338
2 379 2 275 2 275 2 248 2
3 376 3 3 338 3 3
: : : : :
416 416 416 416 416

(a) syllable lattice

(b) DHMM representation of a syllable lattice

Figure 3: An example syllable lattice of a key-phrase speech query "中央研究院(Academia Sinica)", and the corresponding DHMM representation.

4.2 Retrieving Process

Given the syllable lattice database D_l and a query q, the retrieving problem is still a searching process to retrieve the document d^* in the target database D_l which is most related to the query as defined in equation (3). However, here we need a new method to evaluate the similarity measure between a document d and the query q.

As shown in Figure 3, the syllable lattice of a key phrase speech query can be represented as a discrete Hidden Markov Model (DHMM), $\lambda_q = (A, B, \pi)$ (Rabiner and Juang, 1993), where $A = \{a_{ij}\}$ is the state-transition probability distribution, $B = \{b_j(k)\}$ is the observation symbol probability distribution, and $\pi = \{\pi_i\}$ is the initial state distribution. The state number, N, equals to twice of the length (i.e., the syllable number) of the speech query. The first state (the dark one, as shown in Figure 3) is the filler state that is used for decoding surrounding non-key-phrase part of the spoken document, thus its observations include all syllables and they all share the uniform observation probabilities, i.e., $b_1(k) = 1/416, 1 \leq k \leq 416$. The dotted states are also filler states and are used for handling the possible insertion errors in the spoken documents and deletion errors in the speech queries, thus their observations also include all syllables and they all share the uniform observation probabilities. On the other hand, the other states are the key phrase states, which represent the corresponding syllable segments of the key phrase query respectively, thus the observations of each state only include the syllable candidates, and their observation probabilities can be the acoustic recognition scores. That is, $b_{j\times 2}(k) = as(s_k)$, if s_k is one of the candidates of the j-th syllable of the speech query, otherwise, $b_{j\times 2}(k) = 0$. To handle the possible deletion errors in the spoken documents and insertion errors in the speech queries, the DHMM topology allows the search process to skip one key phrase state each time. As a result, the distributions π and A can be easily derived according to the topology of the DHMM adopted here, as shown in Figure 3, e.g. $\pi_i = 1/3, i = 1,2,4$ while $\pi_i = 0, i = 3, \text{or } 4 < i \leq N$ because the entrance states include the first, second, and fourth states, and we would have $a_{ij} = 0$ for some (i,j) pairs. Furthermore, the exit states include the first state and the last two key phrase states only.

The syllable lattice of a spoken document can be thought as an unknown sequence with multiple observations at each time index. Then, each spoken document is an unknown utterance and the speech query is the keyword model, while the retrieving processes should identify all the segments in the spoken document that are similar to the keyword and generate the accumulated scores of all the matched spoken segments as the similarity measure between the spoken document and the speech query.

First of all, we can use the Viterbi search algorithm to find the best state sequence. The complete procedure is stated as follows (Rabiner and Juang, 1993):

1.Initialization

$$\delta_1(i) = \pi_i b_i(o_1), \qquad 1 \leq i \leq N \tag{5a}$$

$$\psi_1(i) = 0, \qquad 1 \leq i \leq N \tag{5b}$$

2.Recursion

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t), \qquad 2 \le t \le T, 1 \le j \le N \qquad (6a)$$

$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}], \qquad 2 \le t \le T, 1 \le j \le N \qquad (6b)$$

3.Termination

$$P^* = \max_{1\le i\le N}[\delta_T(i)], \qquad (7a)$$

$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)], \qquad (7b)$$

4.State sequence backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1 \qquad (8)$$

where $O = (o_1 o_2 ... o_T)$ is the observation sequence, $\delta_t(i)$ is the best score (highest probability) along a single path, at time *t*, which accounts for the first *t* observations and ends in state *i*, and $Q^* = (q_1^* q_2^* ... q_T^*)$ is the best state sequence. In this approach, the estimation of $b_j(o_t)$ can be formulated as follows:

$$b_j(o_t) = \sum_{k=1}^{K} b_j(o_{tk}) \times as(o_{tk}) \qquad (9)$$

where *K* is the number of syllable candidates contained in the syllable lattice of the spoken document at time index *t*, o_{tk} is the *k*-th syllable candidate contained in the syllable lattice of the spoken document at time index *t*, while $as(o_{tk})$ is the acoustic recognition score of the syllable candidate o_{tk}.

Then, based on the best state sequence, we can identify the matched spoken segments and estimate the similarity measure between a spoken document *d* and the speech query *q* using the following equation:

$$Sim(d,q) = \sum_{i=1}^{MSN} matched\_score(i) \qquad (10)$$

where *MSN* is the number of matched spoken segments and $matched\_score(i)$ is defined as follows:

$$match\_score(i) = \sum_{t=t_i}^{t_i+D_i-1} b_{q_t^*}(o_t) \qquad (11)$$

where t_i is the beginning time of the i-th matched spoken segment, and D_i is the duration of the i-th matched spoken segment. As a result, the documents with higher $Sim(d,q)$ will be selected and ranked as the retrieving results.

5. Experiments and Discussions

5.1 Speech Database Used in the Experiments

The example speech database to be retrieved in the following experiments consists of 500 Mandarin spoken documents for Chinese news. They were produced by 5 different male speakers. The text materials are news articles published in Taiwan area in 1997. On average, each spoken document contains about 100 characters (i.e., 100 syllables), while the individual length of the articles ranges from 44 to 269 characters. A set of 80 simple key phrase queries produced by 4 different male speakers were used for tests. Each of these queries contains only a key phrase for some news items. A typical example key phrase is "亞太經合會", which is a frequently used abbreviation of "亞洲太平洋經濟合作會議(Asia Pacific Economic Cooperation, APEC)". These key phrases were selected manually from the headlines of the original text materials. Each query contains 4.9 characters (or syllables) on average. For assessment of the retrieval performance, the relevant news articles for each query were selected manually by searching through the original text materials. Each query has on average 5.9 relevant documents among the 500 documents in the database, with the exact number ranging from 1 to 20.

Gender-independent speaker-independent context-dependent Initial/Final HMM's as mentioned in Section 2 were used to recognize the syllables and construct the syllable lattices for both the spoken documents and the speech queries. The top 1 syllable recognition rates for the spoken documents and the speech queries are 54.70% and 59.07%, respectively.

5.2 Experimental Results

5.2.1 Comparison between the two approaches

The first experiment was tested to make comparison between the previous syllable-based approach and the new syllable-based approach proposed in this paper. The non-interpolated

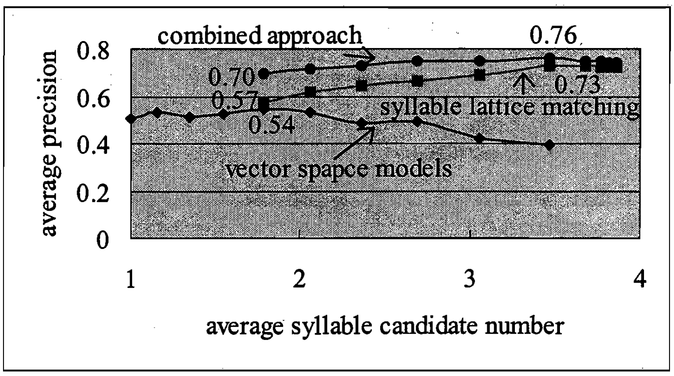


Figure 4: Results for retrieving Mandarin spoken documents using key phrase speech queries based on vector space models, syllable lattice matching, and the combined approach.

average precision rates with respect to the average number of syllable candidates are plotted in Figure 4. For the previous vector-space-based approach, it can be found that in general the performance becomes worse and worse when the number of syllable candidates increases, and the best precision rate achieved is 0.54 when the average number of syllable candidates is only 1.79. This result seems counter-intuitive at the first sight. When the number of syllable candidates is increased from 1 to n, the number of possible syllable pairs is increased from 1 to n×n. Although one of them may be correct and provide information regarding the desired documents, the other n×n-1 syllable pairs all include wrong syllables, and therefore inevitably increase the degree of ambiguity. Although the acoustic recognition scores $as(s_i, s_j)$ in the feature vectors can provide some degree of discrimination against less reliable syllable candidates, but the extra correct syllables included by the increase of the number of syllable candidates very often also have relatively lower acoustic scores. Therefore the information regarding the desired documents carried by these extra correct syllables may be easily swamped by syllable pairs constructed with wrong syllables with relatively higher acoustic recognition scores. These explain why the performance degrades with increased number of syllable candidates. So good retrieving approaches should be able to make use of the increased correct syllables to achieve better results. For the proposed lattice-matching-based approach, it can be found from Figure 4 that in general the performance becomes better when the number of syllable candidates increases, and the best average precision rate achieved is 0.73 when the average number of syllable candidates is 3.47. The new approach produced 0.19 (0.73-0.54) improvements in non-interpolated average precision, while the average number of syllable candidates in this case is almost twice (3.47/1.79=1.94) as the average number of syllable candidates used in the best case of the previous approach. It can also be

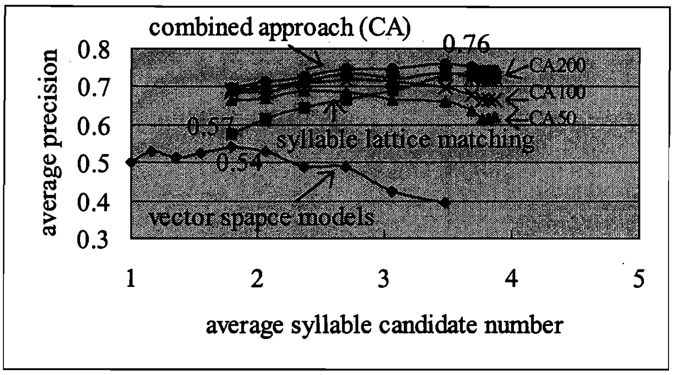


Figure 5: Results for retrieving Mandarin spoken documents using key phrase speech queries based on the two-stage approach (CA200, CA100, and CA50).

found that, for the proposed lattice-matching-based approach, the curve keeps relatively flat as the average number of syllable candidates further increases. The experimental results show that the new approach is better than the previous approach in making use of the syllable lattice, and thus the retrieval performance is significantly improved.

5.2.2 Results for the combined approach

Based on the above descriptions of the methodologies of the two syllable-based approaches and the experimental results, one may wonder how many further improvements can be obtained if we combine the above two approaches together. Since both the vector-space-based and lattice-matching-based approaches are based on exactly the same speech recognition and syllable lattice construction front-end. They can thus be very easily combined together as a combined approach. That is, the similarity measures obtained using equations (4) and (10) can be summed together to give a new similarity measure between a spoken document and a query. The top curve in Figure 4 shows that, in any case, the results for the combined approach are better than the results for either the vector-space-based approach or the lattice-matching-based approach. The best average precision rate achieved is 0.76, while the best average precision rates are 0.73 and 0.54 for the lattice-matching-based approach and the vector-space-based approach respectively.

Furthermore, it should be noted that, the computation requirement of the lattice-matching-based approach is much higher than that of the vector-space-based approach. In fact, in our experiments, the search time for the lattice-matching-based approach was about 10 times of that for the vector-space-based approach. It is thus a good idea to modify the combined approach to a two-stage search strategy for shortening the total search time. In the first stage, the vector-space-based approach can be applied to filter out the non-relevant documents and select a set of potential documents. Then, in the second stage, the lattice-matching-based approach is applied to these potential documents only. Finally, the potential documents are re-ranked based on the summation of two similarity measures obtained by two approaches and the final results can be obtained. Figure 5 shows the results for this two-stage approach, in which the three curves marked by "CA200", "CA100", and "CA50" represent the cases that the lattice-matching-based approach was applied to 200, 100, and 50 potential documents selected by the vector-space-based approach respectively. Because the retrieval performance of the first-stage vector-space-based approach is relatively poor, more potential documents are therefore necessary to cover the desired documents. This is why, as shown in Figure 5, the performance gets worse and worse with less potential documents applied in the second-stage lattice-matching-based approach. However, a very important result from Figure 5 is that the retrieval performance for the CA50 case (0.69) is still much better than that for using the vector-space-based approach only (0.54). But, in this case, the search time is only about 2 $(1+10\times 50/500)$ times of that for using the vector-space-based approach only.

5.3 Discussions

Currently, key phrase queries are widely used in text-based retrieval such as Internet search engines. In fact, key phrase queries are simple, convenient, and efficient. On the other hand, natural language queries inevitably cause more ambiguities and thus degrade the retrieval performance significantly. This is why this paper focuses on retrieval of Mandarin spoken documents using short key phrase speech queries. The experimental results indicate that the proposed approach can significantly improve the retrieval performance as compared to the previous approach. However, the previous approach can be directly applied to spoken document retrieval using natural language speech queries though the retrieval performance is even worse (Bai et al., 1996, 1999), but whether this new approach can be applied to spoken document retrieval using natural language speech queries is yet to be further investigated.

6. Conclusion

In this paper, we propose a new syllable-based approach for retrieving Mandarin spoken documents using short speech queries. This approach that is primarily based-on matching the whole syllable lattice directly can better make use of the syllable lattice obtained by speech recognition as compared to the previous syllable-based approach that using syllable and syllable-pair information extracted from the syllable lattice based on the vector space model.

The experimental results show that the retrieval performance can be significantly improved.

Acknowledgements

This work was partially supported by the Republic of China National Science Council under the grant No. NSC 88-2213-E-001-019. The author would like to thank Mr. Berlin Chen for providing the speech recognition front-end.

References

Bai, B. R., Chien, L. F., and Lee, L. S. (1996), "Very-large-vocabulary Mandarin voice message file retrieval using speech queries", *Proc. International Conference on Spoken Language Processing*, pp. 1950-1953.

Bai, B. R., Chen, B., and Wang, H. M. (1999), "Syllable-based Chinese text/spoken document retrieval using text/speech queries", *Proc. International Conference on Multimodal Interface*, pp. II46-II51.

CMU Informedia Digital Video Library project http://informedia.cs.cmu.edu/.

Glavitsch, U. and Schäuble, P. (1992), "A system for retrieving speech documents", *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, pp. 168-176.

Harman, D. (1995), "Overview of the Fourth Text Retrieval Conference (TREC-4)", Available at http://trec.nist.gov/pubs/trec4/t4_proceedings.html .

James, D. A. (1995), The application of classical information retrieval techniques to spoken documents, Ph.D. Dissertation, University of Cambridge, UK.

Jones, K. S., Jones, G. J. F., Foote, J. T., and Young, S. J. (1996), "Experiments on spoken document retrieval", *Information Processing & Management*, Vol. 32, No. 4, pp. 399-417.

Lin, S. C., Chien, L. F., Chen, K. J. and Lee, L. S. (1995), "Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains", *Proc. European Conference on Speech Communication and Technology*, pp. 1203-1206.

Ng, K. and Zue, V. (1997), "Subword unit representations for spoken document retrieval", *Proc. European Conference on Speech Communication and Technology*, pp. 1607-1610.

Rabiner, L. and Juang, B. H. (1993), *Fundamentals of Speech Recognition* (Prentice-Hall International, Inc.).

Salton, G. (1983), *Introduction to Modern Information Retrieval* (McGraw-Hill, NY).

Wang, H. M. et al. (1997), "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data", *IEEE Trans. Speech and Audio Processing*, Vol. 5, No. 2, pp. 195-200.

Wechsler, M. (1998), Spoken document retrieval based on phoneme recognition, Ph.D. Dissertation, Swiss Federal Institute of Technology (ETH), Zurich.

Recent Results on Domain-Specific Term Extraction from Online Chinese Text Resources

Lee-Feng Chien[1], Chun-Liang Chen[2], Wen-Hsiang Lu[3] and Yuan-Lu Chang[1]

1. Institute of Information Science, Academia Sinica
2. Dept. of CS and IE, National Taiwan University
3. Dept. of Computer Science and Information Engineering,
National Chiao Tung University, Taipei, Taiwan, R.O.C.

Abstract

This paper is to introduce recent results of an ongoing research called *Live Dictionary Construction,* which is investigating a number of efficient techniques for IR systems to automatically acquire Chinese terminological knowledge including domain-specific terms and similar terms from online text resources. Such research effort is pursued to be able to build a dynamic dictionary with IR systems, in which most of the necessary dictionary information can be dynamically extracted and adapted with the change of the indexed online resources. According to the obtained experimental results so far, it is promising that a live dictionary can be established and automatically grow.

1. Introduction

Automatic extraction of domain-specific terminological knowledge, such as keyterms and similar terms from online text collections is significant but very challenging for developing more effective information retrieval and also natural language processing systems. In this paper, we intend to introduce recent results of an ongoing research called *Live Dictionary Construction,* which is investigating a number of efficient techniques including *corpus classification*, *term extraction*, named entity extraction, similar term extraction to automatically acquire Chinese terminological knowledge. The research is pursued to build a *live dictionary* with IR and also NLP systems, in which most of the necessary dictionary information can be dynamically extracted and adapted with collection change.

Whether the employed dictionary is rigid and suitable for the database domain is very crucial in designing an effective IR system. It is clear that a well-prepared dictionary can help to identify representative keyterms in document indexing, find relevant terms in query expansion and perform exact term translation in cross-language information retrieval[Lewis'96, Wan'97]. Unfortunately, online resources most increase very fast. To most of the existing IR systems, it is cost-ineffective and even unrealistic to manually construct a domain-specific dictionary for each searching database. To avoid too many unknown searching terms and term translations appearing in database retrieval, the construction of a live dictionary which can grow with the update of the database like Altavista's LiveTopic is believed an alternative solution.

Our ongoing research is known as one of a few works towards the systematic construction of live dictionary for IR applications. The approach proposed for this purpose is based on proper integration of linguistic knowledge acquisition and IR technologies. This approach has achieved several technical breakthroughs. Like the technique designed for domain-specific term extraction, it has been proven performing well in extracting new terms incrementally. Compared with conventional research on knowledge acquisition[Zernik'91], the proposed approach has carefully considered the incremental characteristics of online information service. The developed techniques are all capable of handling large and dynamic texts and also easily to be integrated with IR systems. According to the obtained experimental results so far, it is optimistic that a live dictionary can be established.

2. Previous Work - PAT-tree-based Term Extraction

Keyterm extraction is frequently used in document classification and many other information retrieval applications. Since in Chinese language there is no "blanks" between words serving as word boundaries in printed and written sentences and the words are actually not well-defined, keyterm extraction has been a much more difficult and challenging problem in Chinese language processing as compared to western languages. An efficient approach for keyword extraction from Chinese texts has been developed previously, in which the difficult problem of large numbers of out-of-vocabulary words outside of any given lexicon and the sophisticated problem of word segmentation from sentences can both be avoided, and keywords or concatenated keywords (key terms) of arbitrary length which are very useful in information retrieval can be successfully extracted [Chien'97, Chang'99]. This approach is statistics-based and efficient in extracting major "significant lexical patterns (SLP)" from the Chinese texts.

3. Overview of the Proposed Approach

The proposed approach is formed as an abstract diagram shown in Fig. 1, where an IR system is designed as a composition of a searching engine and a live dictionary subsystem. The purpose of the live dictionary subsystem is trying to dynamically produce domain-specific term lists, term associations, and low-frequency named entities for the use of the searching engine. Such a subsystem contains several working modules, i.e., corpus storage and classification module, term extraction module, similar term extraction module, and named entity extraction module.

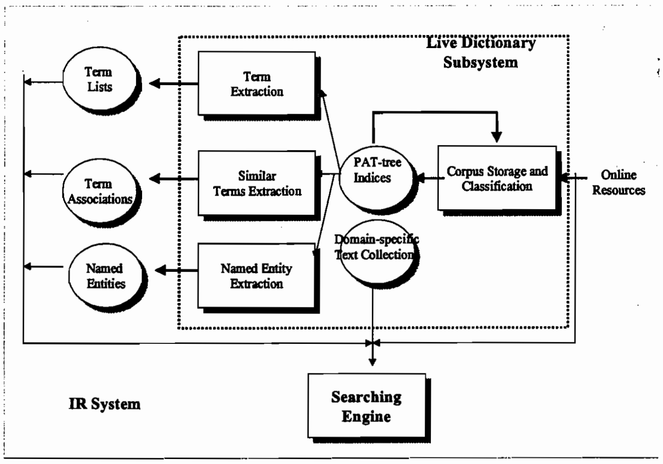


Fig. 1 An abstract diagram showing the proposed approach for live dictionary construction.

The execution of the corpus storage and classification module is the first step to construct a live dictionary. To allow domain-specific dictionary information can be effectively extracted, each input online document needs to be classified into corresponding collection(s) and serves as a training corpus for subsequent information extraction. Considering the demand on both document retrieval and corpus utilization, a method which employs *PAT trees* as the working structure for corpus storage and classification is presented. Each classifying document will be generated a PAT tree which records the occurrences of all of the composed character strings as the feature vector of the document, and then compares with the corresponding PAT-tree indices of each text collection in the system, by means of vector-space-based similarity estimation. The classified document will be then appended into the belonging collection (s) and update the corresponding PAT tree(s). The updated PAT tree(s) can record the up-to-date information of the text collection(s), on which rigid linguistic information can be incrementally extracted.

Once an input document has been classified and indexed, the term extraction module will be performed. It will extract new keyterms from the document by estimating the completeness and significance of the composed strings with the corresponding PAT trees. The underlying technique for term extraction is an extension of the previous work [Chien'97, Chien'99]. The extended technique here emphasizes the incremental ability in new term extraction. Besides, in the similar term extraction module, it will find similar terms from the extracted keyterms. The basic concept for this processing is to extract keyterms with near context[Smadja'93]. To deal with the extraction of low-frequency keyterms especially on named entities such as personal names and organization names, a named entity extraction module is then developing. Since it is hard to extract low-frequency named entities simply based on the previously-developed PAT-tree-based approach, the proposed technique is tried to compare the contextual similarity between each named entity candidate and a set of extracted high reliable named entities.

This paper will only focus on the introduction of incremental term extraction and similar term extraction and brief description of named entity extraction. Further description about the corpus classification and named entity extraction can be referenced in [Chen'99, Chang'99].

4. Incremental Term Extraction

To many online NLP and IR systems such as voice browsers, web-based machine translation systems, Internet searching engines et al., it is cost-ineffective and even unrealistic to manually construct a domain-specific dictionary for each service domain. To capture up-to-date information and reduce unknown vocabulary, incremental extraction of domain-specific terms from online text resources are necessary.

This section is to define the considering problem and give an overview of the proposed tecnique for incremental extraction of domain-specific terms. A *domain-specific term* is defined as a string that consists of more than one successive characters in Chinese (or words in English) which has certain occurrences and is specific to a text collection with a distinct subject domain. Such a string has a complete meaning and lexical boundaries in semantics; it might be a word, compound word, phrase or linguistic template.

4-1 Overview of The Proposed Method

Definition 1: The Incremental Term Extraction Problem

Given a new document *D,* a set of incremental and domain-specific text collections $C_{1\sim n}$ and corresponding term lexicons $T_{1\sim n}$, the goal of this problem is to determine the most promising collection C_i for *D*, extract *new terms X* from *D* and add to T_i, where $X = \{\ x|\ x$ occurs in *D; x* can be a domain-specific term of C_i but missed in T_i at present}.

The above problem is defined to deal with the extraction of domain-specific terms with the increase of an online resource. The online resource will be divided into different text collections with specific subject domains in advance. Once a term is found specific and becoming important in a certain text collection, it is pursued can be extracted as soon as possible. Such a domain-specific term often indicates the occurrences of a certain event. If it can be identified immediately, some kinds of real-time reaction and information services like event detection of online news service can be implemented.

In fact, considering the reliability of term extraction, the extracted terms should have a certain occurrence and is expected will be used in a period of time, although some of them may not be used in a long term. So as to, a term which is a keyword in a single document but rarely occurs in other documents is not considered as a domain-specific term. Many low-frequency proper names are not taken into consideration in this way.

The proposed technique is known as one of a few works considering such an incremental extraction problem. To deal with the problem, several difficulties need to be faced with. *The first difficulty is to identify new and meaningful terms with document inputs as soon as possible.* It is known that to extract meaningful terms in an automatic way is still a challenging problem in western languages, but it is more critical in Chinese and oriental language processing because of difficulties in word segmentation and unknown word identification[Wu'95]. Our idea to this difficulty is to develop an efficient algorithm which is able to monitor the frequency change and usage freedom of each candidate term in the text collections, with the input of the new documents. *The second difficulty is how to estimate the significance of the candidate terms.* In our solution each new document should be classified into corresponding text collection(s) and its composed candidate terms will be checked by observing their distributions in different collections in the system. The candidates which are "non-specific" will be removed. Moreover, *the third difficulty is the efficiency in handling large and dynamic texts.* Since real-time processing is required in many applications, the utilized techniques have to be efficient in execution. To reduce the difficulty, the PAT-tree-based working structure is adopted again.

The term extraction module, as shown in Fig. 2, consists of two elementary sub-modules: completeness analysis and significance analysis. The outputs obtained with the proposed technique will contain the classified text collections, the PAT-tree indices and the domain-specific term lexicons from online text resources.Because the words in Chinese are not well-defined anyway, in this technique all the character strings of any length in the texts are first taken as candidates of keyterms.

(1). ***Completeness Analysis***

The first step is to extract new complete terms from each examining document. Like the completeness analysis step of the previous approach, this step is mainly to check if the strings of candidate terms are complete in lexical boundaries. But in difference, the strings need to be checked here are only that occurred in the new document *D* which have certain occurrences in the corresponding text collection T_i but not found in the term lexicon K_i at present. For each string *X* in *D*, it will judge if *X* is complete in semantic by its distribution and context in the updated PAT tree I_i. *X* is defined as complete in semantic iff its *association norm* of the composed sub-strings is strong enough and has no *left and right context dependency.* The estimations defined below are the same with the previous work. Such a design really considers the characteristics of Chinese.

Definition 2: The association norm estimation

The association norm estimation *MI(X)* for each string *X* is defined below:

$$MI(X) = \frac{f(X)}{f(X_s) + f(X_e) - f(X)}$$

Where *MI(X)* is the mutual information of a target string *X*, X_s is the longest starting sub-string of *X*, i.e., the sub-string which is exactly *X* except that the last character of *X* is deleted, X_e is the longest ending sub-string of *X*, i.e., the sub-string which is exactly *X* except that the first character of *X* is deleted, and *f(X)*, $f(X_s)$, $f(X_e)$, are the frequency counts of *X*, X_s, and X_e, in the text collection respectively. Such a definition is based on the efficiency of calculation in real-time applications. Character stings with the above mutual information below a threshold are considered to be incomplete.

Definition 3: Left Context Dependency (LCD)

Each string *X* has left context dependency if $|L| < t1$ or $\mathrm{MAX}_\alpha\, f(\alpha X)/f(X) > t2$, where *t1, t2* are threshold values, *f(.)* is frequency, *L* is the set of left adjacent strings of *X*, $\alpha \in L$ and $|L|$ means the number of unique left adjacent strings.

Definition 4: Right Context Dependency (RCD)

Each string *X* has right context dependency if $|R| <$ *t1* or MAX_{β} $f(X\beta)/f(X) >$ *t2*, where *t1, t2* are threshold values, *f(.)* is frequency, *R* is the set of right adjacent strings of *X*, $\beta \in R$ and $|R|$ means the number of unique right adjacent strings. The stings with either left or right context dependency are considered to be incomplete.

In fact, the above metrics are actually used to check if *X* contains highly-associated composed strings and also has complete lexical boundaries, by judging the usage freedom of *X* according to its contextual information. The basic assumption is that if *X* has few unique left or right adjacent strings, or if it frequently occurs together with certain adjacent strings, it might be incomplete in semantics.

The above estimations are easy to be implemented using the PAT-tree indices [Gonnet'92]. To know if a candidate string in *D* is complete or not, it just needs to check its association norm of the composed sub-strings as well as left and right context dependency. All of the operations can be easily done with PAT-tree access.

(2). *Significance Analysis*

The second step is to find out domain-specific new terms. Like in the previous approach, the significance analysis step is to extract specific and significant candidate strings as the domain-specific terms. Using the following procedures, all of the remaining candidates strings will be checked using a common-word lexicon, a set of lexical rules and the analysis of the significance estimation function *S*(*Y*) shown below[Schutze'98]. If a candidate string appears either in the common-word lexicon or can be formed using the lexical rules, it is treated as a non-significant candidate and is removed. The remaining candidates will be further checked by observing their frequencies and distributions between the corresponding and different PAT trees in the system. The candidates which are also frequently appear in the different PAT trees are treated as non-specific and are removed too. The strings which satisfy the estimation (larger than a threshold value) will be selected as the new domain-specific terms.

Definition 5: The Significance Estimation Function

$S(Y) = (f_i(Y)/f(T_i))/ \ (f_g(Y)/f(T_g)$, where Y is a candidate term, f_i (Y) is the frequency of Y in collection T_i, $f(T_i)$ is total number of strings in collection T_i, $f_g(Y)$ is the frequency of Y in the general collection, and $f(T_g)$ is total number of strings in the general collection.

The above estimation compares the relative frequency in the text collection of interest with the relative frequency in a reference collection. The necessary parameters are all easy to be computed with the PAT-tree indices. Among them, $f_i(Y)$ and $f(T_i)$ can be obtained directly in PAT tree I_i. As to $f_g(Y)$ and $f(T_g)$ can be obtained by summing up all of the domain-specific PAT trees in the system.

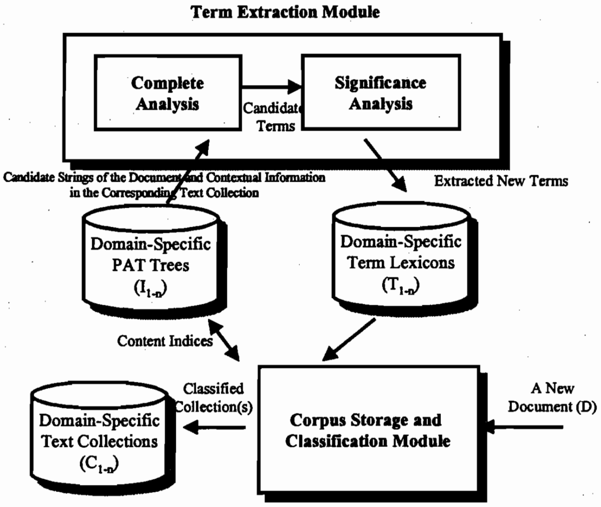

Term Extraction Module
Complete Analysis
Candidate Terms
Significance Analysis
Candidate Strings of the Document and Contextual Information in the Corresponding Text Collection
Extracted New Terms
Domain-Specific PAT Trees $(I_{1\text{-}n})$
Domain-Specific Term Lexicons $(T_{1\text{-}n})$
Content Indices
Classified Collection(s)
Domain-Specific Text Collections $(C_{1\text{-}n})$
Corpus Storage and Classification Module
A New Document (D)

Fig. 2 An abstract diagram showing the proposed method for incremental term extraction.

4-2 Experimental Results

An experiment was performed to realize the effectiveness of the proposed approach for incremental term extraction. The experiment used a Chinese online-news database from Central News Agency (CNA) in Taiwan as the testing platform. At first, a total of 1,872 political news abstracts published in July 1997 were tested. The testing database contained 5-months manually-classified documents and one-month automatically-classified documents at that stage. In this experiment, the 1,872 news documents were added in sequence for both corpus classification and term extraction. Only the new terms extracted from the politics collection were counted. Tables 1 and 2 show the obtained results. It has to point out that before the processing of term extraction the political text collection has contained 6-month of news documents, in which the sixth-month documents were automatically classified and have only 45.1% precision and 99.4% top2 recall. Meanwhile, the corresponding term lexicon is empty in the initial stage.

Table 1 shows the obtained recall and precision rates with different threshold values in the significance analysis. The correct domain-specific terms of the testing 1,872 documents were extracted manually in advance. The terms extracted with different threshold values were compared with the correct set. It can be found the best performance in terms of both recall and precision rates was that using the threshold value 2. In that case, 1,135 correct terms can be extracted and the obtained precision and recall rates were 0.78 and 0.44 respectively. Such a performance is satisfied in many applications. It is worthy to note that 258 of the extracted terms were not included in the KUH dictionary, the largest Chinese dictionary we can find, which contains more than 160,000 word entries, and is believed covers many of terminological vocabulary used in news papers.

Except the above effectiveness issues, there are other issues such as the average number of document inputs to find a new term, the average frequency as the new terms to be extracted, and how often the extracted terms can be used, etc. For this reason, Table 2 shows the detailed results with the threshold value larger than 2. It is noted that in the table term length" is the number of characters of extracted terms. Since terms with different lengths behave differently (for example three-character terms are very often personal names, and four-character or longer terms are very often compound words), the results are shown with the term length as a special parameter. From this table, it can be observed that on average every 2.41 document inputs can find new terms. Also, each extracted new terms occur 28.95 times in the one-month testing documents and was extracted at the 9.25 time on average. This indicates most of the extracted terms are not late to be extracted and many real-time reactions can be performed.

S(Y)	Total Extracted Terms(A)	No. of Correct Terms Extracted(B)	No. of Correct Terms Outside Dictionary(C)	Precision (B/A)	Recall
>1.5	2,291	1,374	297	0.60	0.53
>2	1,455	1,135	258	0.78	0.44
>2.5	723	593	172	0.82	0.23
>3	214	184	66	0.86	0.07

Table 1. The testing results for incremental term extraction with different threshold values in the significance analysis which were obtained from a total of 1,872 political news abstracts published in July, 1997.

The proposed approach has been tested extensively and found very efficient in extracting terms from online text collections. For example, as shown in Table 3 there were more than ten thousand political terms can be extracted from a total of 13,849 political news abstracts published from Aug. to Dec. in 1997. The obtained results were found similar to that extracted from one-month news abstracts. With the increase of the news documents, the frequency values of the extracted terms are obviously increased but the frequency the terms can be extracted are similar.

Term length (character N-gram)	Number of extracted new terms	Number of documents with new terms extracted	Average number of document inputs can find new terms (A)	Average frequency of the extracted new terms	Average frequency as the term can be extracted
2	776	515	3.93	34.22	9.37
3	416	325	6.04	24.60	9.09
4	171	157	12.16	19.22	8.97
5	51	49	37.28	20.35	9.18
6	17	17	109.81	27.00	8.65
7	15	15	123.60	27.40	11.20
8	6	6	274.67	13.00	9.83
9	3	3	205.67	18.00	11.33
Total N-grams	1,455	814	2.41	28.95	9.25

Table 2. The detailed results for incremental term extraction with the threshold value larger than 2 in the significance analysis, which were obtained from a total of 1,872 political news abstracts published in July 1997.

Term length (character N-gram)	Number of extracted new terms	Number of documents with new terms extracted	Average number of document inputs can find new terms (A)	Average frequency of the extracted new terms	Average frequency as the term can be extracted
2	3,376	2,502	5.75	72.12	11.41
3	4,274	3,056	4.69	31.15	9.51
4	2,408	2,021	7.10	22.23	9.17
5	694	642	21.89	25.02	9.40
6	303	295	47.20	26.25	10.29
7	145	145	95.17	33.23	14.59
8	87	87	156.90	25.08	11.86
9	52	51	265.65	24.33	13.54
Total N-grams	11,339	6,242	2.28	40.90	10.12

Table 3: The detailed results for incremental term extraction with the threshold value larger than 2 in the significance analysis, which were obtained from a total of 13,849 political news abstracts published from Aug. to Dec. in 1997.

However, there exist some difficulties to be discussed. Taking all the terms with different lengths into account, it can be found that the precision rate for three-character terms was relatively low because many frequently used single-character words and two-character words are easily combined to produce three-character terms which are not necessarily key elements for most IR and NLP applications. Close examination of the extracted terms indicates that most of them are domain-specific such as proper nouns and topic terms, which are often very important in IR applications. This phenomenon is especially significant for terms with three or more characters. While it is important to indicate that the proposed approach is weak in extracting low-frequency terms, because the extracted terms should at least occur 9.25 times and 10.12 times as in Tables 2 and 3 respectively. To deal with the extraction of low-frequency but domain-specific terms, we are considering the combination of linguistic analysis methods as that developed in the named entity extraction module.

5. Low-frequency Named Entity Extraction

For those low-frequency terms, it is hard to judge if these terms own complete word boundaries based on statistical information, because the possible patterns in their context are very limited to be investigated. To deal with the extraction of low-frequency but significant keyterms , we present another method to perform semantic completeness analysis. As found in our experiments, the presented method can be used to handle the extraction of low-frequency named entities.

Names are some symbols that represent some characters or organizations and are conventionally used to identify the named entity. Named entity extraction (NE) is to identify all named locations, named persons, named organizations, dates, monetary amounts, and percentages in text. There are several features with the named entity. First, each type of named entity owns separate rule sets. In Chinese NE, family names predict personal names quite well. Former researches dealt with NE problem with rule-based heuristic approaches. Second, each named entity plays some specific roles. This situation can be revealed by neighboring contextual conventions of the named entity. Taking Chinese news articles for example, most of the personal names appear with a title in the context to indicate the identification or profession of the person. Such context not only gives information about the character that the named entity represents, but also helps to identify more named entities.

Since named entities usually have templates in the context and can be modeled by other named entities in the same category, a preliminary method and initial experiment were therefore developed. The basic idea of the method is described as follows:

Context Dependency Estimation:

Assume a named entity x belong to class K. Context of all named entities in K can be therefore used, if L(x) similar to L(K) or R(x) similar to R(K), where L(.) is the left context and R(.) is the associated right context respectively. A three-step process of context semantic learning is designed as below.

Step 1: Given an initial named entity set K and corpus D, the first step is to generate L(K) and R(K) based on K and D.
Step 2: For each "possible" new named entity x, add x into K if P(x) > Th or (Tl < P(x) <Th and L(x) c L(K) or R(x) c R(K)), where P(x) is a trained Markovian probability function, Tl, Th are two predefined threshold values, L(x) c L(K) means that L(x) belongs to L(x), and R(x) c R(K) that belongs to R(x).
Step 3: Extend L(K), R(K) by L(x), R(x) and repeat Step 2 until the K set cannot be increased obviously.

We have done a small scale of experiments on Chinese personal named entity extraction based on the above method. The testing data size is 1.65MB of news documents, and the initial Markovian probability of personal names is based on order-one Markov model and Sinica corpus. The obtained recall and precision rates with the change of threshold values have been obtained and shown in Table 4. It is can be easily to see that based on context information the extraction accuracy can be improved.

	Probability based (baseline)	With Context Estimation (weight 0.06)	With Context Estimation (weight 0.04)	With Context Estimation (weight 0.02)	With Context Estimation (weight 0.005)	With Context Estimation (weight 0.003)	With Context Estimation (weight 0.001)
Corrected names extracted	7,768	8,608	8,608	8,632	8,778	9,067	9,073
Error names Extracted	1,123	1,188	1,193	1,524	2,423	3,351	5,074
Recall rate	0.848	0.917	**0.940**	0.942	0.958	0.990	0.990
Precision rate	0.873	0.876	**0.878**	0.849	0.783	0.730	0.641
9157 names to be extracted from 1,876 news abstracts							

Table 4: The obtained results for personal named entity extraction.

6. Similar Term Extraction

Automatic construction of a thesaurus from online text resources is important but a challenging research topic. A thesaurus is a set of items (phrases or words) plus a set of relations between these items [Jing'94] . Some researchers have used head-modifier relationships or descriptions of entities to determine similar words[Strzalkowski'95][Radev'98][Lin'98]. Others make use of lexical occurrence information to build related words[Jing’94][Crouch'92][Schutze'97]. Our research towards this topic is just in the beginning. The first step we would like to try is to extract similar terms from the set of domain-specific terms extracted based on the above term extraction methods.

Since it can extract a number of domain-specific erms which were excluded in general dictionary right now, it seems to be possible to deal with the similarity and association among these extracted terms. According to the demand of different computer processing, we simply divide the similar terms into three categories:

(1) Abbreviation : (中央研究院, 中研院)

(2) Named entity with associated title or description : (李登輝總統, 總統李登輝, 李登輝), or (網球名將張德培, 張德培)

(3) Terms different in content but similar in concept : 資訊, 電腦, 計算機

The first two types of similar terms are that similar in content, i.e., sharing common composed character strings among similar terms. The third type of similar terms has no obvious common sub-strings. Since it is more difficult to extract the third type of similar terms, in the beginning stage we just investigate the extraction of the first two types.

6-1 The Proposed Method

(1) Similarity Measurement

The proposed method is based on an assumption that similar terms frequently co-occur in the same documents. There are several ways to measure the correlation of two terms. The Dice coefficient as defined below was found more effective and therefore adopted:

$$\textbf{Dice}\ (k1,k2) = 2f_{k1k2}/(f_{k1}+f_{k2}),$$

where f_{k1}, f_{k2} andf_{k1k2} are the numbers of document occurring ***k1,k2*** and both ***k1*** and ***k2*** together , respectively.

(2) The Extraction Algorithm

The extraction algorithm used is very simple as shown below:

```
1. Term Extraction:
   1.1 Use the above PAT-Tree-Based and named entity extraction methods to extract keyterms.
2. Estimation of Similar Terms:
   2.1 choose any two keyterms k1, k2
   2.2 if k1 is substring of k2, then compute Dice(k1,k2)
   2.3 if Dice(k1,k2) > t1, where t1 is the threshold,
           then k1 and k2 is a pair of similar terms
```

6-2 Experimental Results

The first experiment is to test the accuracy of the similarity estimation. A total of 466KB CNA news articles related to the judiciary and transport subject domains were tested. Some of the experimental results are shown in Fig.1, where the horizontal axis indicates variation of different t1 values, and vertical axis indicates the corresponding ratios of recall and precision with the change of different t1 values. The results show that the precision can be high, if t1 is set at 0.5.

The second experiment is to test the accuracy using different sets and sizes of news articles. The test news were grouped manually into four sets, namely CNA11:congress/politics 1996, CNA12:congress/politics 1998, CNA21:judiciary/transport 1996, CNA22: judiciary/transport 1998. The results are shown in Table 5. It can be found that the average precision rate of 73.75% (112.5/153.25) can be achieved. Appendix 1 and 2 show some samples of extracted similar terms.

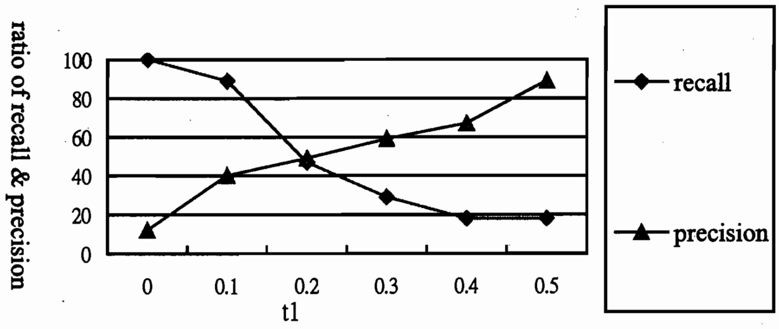


Fig.3 The ratios of recall and precision with different t1

	Text Size(MB)	Total Number of Similar Term Pairs	Number of Correct Extracted Similar Term Pairs	Obtained Precision
CNA11	11.34	329	238	72%
CNA12	3.84	153	115	75%
CNA21	5.21	66	55	83%
CNA22	2.13	65	42	65%
average	5.63	153.25	112.5	73.75%

Table 5. Obtained results of the Similar Term Extraction Experiment.

7. Conclusion

In this paper an ongoing research called Live Dictionary Construction has been introduced. Such research effort has been integrated with a number of techniques. This paper focuses on the introduction of incremental term extraction and similar term extraction. Preliminary experimental results show that th it is very promising build a dynamic dictionary with IR systems.

References

1. [Chang'99] Yuan-Lu Chang , Chun-Liang Chen and Lee-Feng Chien (1999). An Integrative Approach for Chinese Named Entity Extraction, Paper in preparation.
2. [Chen'98] Chen, Chun-Liang (1998). PAT-tree-based Natural Language Processing and Applications under Internet Environment. Master Thesis, Dept. of CS&IE, National Taiwan University,.
3. [Chien'97] Chien, Lee-Feng (1997) *PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval*, Proceedings of ACM SIGIR'97, Philadelphia, USA, pp. 50-58.
4. [Chien'99] Chien, Lee-Feng (1999) *PAT-Tree-Based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval,*, to appear on Information Processing and Management , Elsevier Press.
5. [Crouch'92] Crouch C. J. and Yang, B. (1992). Experiments in automatic statistical thesaurus construction, Proceedings of SIGIR'92.
6. [Gonnet'92] Gonnet, G. H., Baeza-yates, R. et al. (1992) *New Indices for Text: Pat Trees and Pat Arrays*. Information Retrieval Data Structures & Algorithms, pp. 66-82, Prentice Hall.
7. [Jing'94] Yufeng Jing and W. Bruce Croft (1994). An Association Thesaurus for Information Retrieval", UMass Technical Report 94-17. 1994
8. [Lewis'96] Lewis, David D. and Sparck Jones, Karen (1996) *Natural Language Processing for Information Retrieval*, Communications of the ACM, Vol. 39, No. 1, Jan. 1996, pp. 92-101.
9. [Lin'98] Dekang Lin (1998) Automatic Retrieval and Clustering of Similar Words, COLING'98.
10. [Radev'98] Dragomir R. Radev (1998) Learning Correlation between Linguistic Indicators and Semantic Constraints: Reuse of context-Dependent Descriptions of Entities, COLING'98.
11. [Schutze'97] Hinrich Schutze and Jan O. Pedersen, (1997) A Coocurrence-based Thesaurus and Two Applications to Information Retrieval", Information Processing & Management, Vol. 33, No. 3, pp. 307-318,1997.
12. [Schutze'98] Schutze, Hinrich (1998) *The Hypertext Concordance: A Better Back-of-the-Book Index,* Proceedings of the First Workshop on Computational Terminology (Computerm'98), pp. 101-104.
13. [Smadja'93] Smadja, F., (1993) *Retrieving Collocations from Text: Xtract*, Computational Linguistics, 19 (1), pp. 143-177.
14. [Wan'97] Wan, T. L., Evens, M. et al. (1997) *Experiments with Automatic Indexing and a Relational Thesaurus in a Chinese Information Retrieval System*, Journal of the American Society for Information Science,48(12), pp. 1068-1096.
15. [Wu'95] Wu, Z., Tseng, G. (1995) *ACTS: An Automatic Chinese Text Segmentation System for Full Text Retrieval*. Journal of the American Society for Information Science, 46 (2), pp. 83-96.
16. [Zernik'91] Zernik, Uri (1991) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, Publishers.

Appendix 1. Some Samples with Similar Terms from CNA news

@19980113 12:07:30 **蕭萬長**說明選擇訪問菲律賓的原因

（中央社記者梁君棣台北十三日電）**行政院長蕭萬長**今 天說明他這次訪問菲律賓的理由，主要是拜訪總部設在菲律賓的**亞洲開發銀行**，**亞銀**有五十六個會員國，中華民國也是**亞銀**的會員國之一，到**亞銀**訪問，可以進一步了解整個局勢。

==

@19980113 11:53:44 **蕭萬長**結束訪菲返國舉行記者說明

（中央社記者梁君棣台北十三日電）**行政院長蕭萬長**今天早上從菲律.........

第一、**亞銀**對當前亞洲金融風暴有見解，與我國的了解相當接近，差不多...。

第二、**亞銀**對我經濟發展表示肯定，並對這次我國處理金融風暴的政策...。

第三、**亞銀**對由行政院經濟建設委員會主任委員江丙坤所率訪問團的行動...。

蕭萬長說，在與相關人士討論時曾提到中華民國如何參加區域金融合作機制，亞銀表示將支持我國參與。

==

@19980103 12:41:06 多數**陸委會**諮詢委員指人民幣短期內會撐下去

（中央社記者許雅靜台北三日電）**行政院大陸委員會**今天召開諮詢委員會議，**陸委會**企劃處處長詹志宏表示，多數諮詢委員對大陸今年經濟不樂觀，但認爲短期內人民幣仍會撐下去，整體而言，人民幣如果貶值，對我經濟影響不大。

==

@19980116 15:34:26 **海基會**願協助河北震災救援我希望**海協會**回應
（中央社記者曾淳良台北十六日電）大陸河北張家口一月十日發生強烈地震，至少造成五十人死亡，一萬多人受傷，整個救災工作仍在極惡劣環境下進行，我方基於人道考量，主動表示願配合災區協助工作，陸委會希望 大陸海協會作出回應。

==

@19980119 18:34:45 **陸委會**指張京育未安排與汪道涵在東京會面
（中央社記者曾淳良台北十九日電）**行政院大陸委員會**主任委員張京育正在東京訪問，**大陸海峽兩岸關係協會**會長汪道涵目前也在東京，**陸委會**副主委兼發言人許柯生今天表示，張京育此行主要是參加學術研討會，並未安排與汪道涵會面。

Appendix 2. Samples of Extracted Similar Terms

- Correct abbreviations:

世界自由民主聯盟=>世盟 [0.695652]
北大西洋公約組織=>北約 [0.516129]
金門防衛司令部=>金防部 [0.642857]
中央選舉委員會=>中選會 [0.647059]
中山科學研究院=>中科院 [0.600000]
公共工程委員會=>工程會 [0.600000]
亞洲開發銀行=>亞銀 [0.666667]
中央研究院=>中研院 [0.555556]
國家安全局=>國安局 [0.631579]
李登輝總統=>李總統 [0.692105]
民主進步黨=>民進黨 (273)[0.546000]
違章建築=>違建 [0.545455]
台灣銀行=>台銀 [0.583333]
空軍總部=>空總 [0.577778]
社會福利=>社福 [0.547368]
歐洲聯盟=>歐盟 [0.784615]

- Correct description of entities

李元簇夫人徐曼雲=>徐曼雲 [0.666667]
台北市議員林瑞圖=>林瑞圖 [0.560000]
教宗若望保祿二世=>教宗 [0.523077]
俄羅斯總統葉爾勤=>葉爾勤 [0.533333]
古巴總統卡斯楚=>卡斯楚 [0.500000]
參謀總長羅本立=>羅本立 [0.695652]
立法委員王志雄=>王志雄 [0.590909]
交通部長蔡兆陽=>蔡兆陽 [0.660870]
高雄市長吳敦義=>吳敦義 [0.595745]
法務部長廖正豪=>廖正豪 [0.732394]
台灣省長宋楚瑜=>宋楚瑜 [0.617143]
台北市長陳水扁=>陳水扁 [0.566038]
印尼總統蘇哈托=>蘇哈托 [0.537313]
內政部長黃主文=>黃主文 [0.676471]
行政院長蕭萬長=>蕭萬長 [0.691814]
參謀總長唐飛=>唐飛 [0.500000]
先總統蔣公=>蔣公 [0.526316]
大使謝棟樑=>謝棟樑 [0.777778]
團長何明德=>何明德 [0.842105]
中非共和國=>中非 [0.705882]
二二八事件=>二二八 [0.666667]
副總統連戰=>連戰 [0.674121]
政變傳聞=>政變 [0.526316]
排雷工程=>排雷 [0.500000]
拜耳公司=>拜耳 [0.740741]
樞機主教=>主教 [0.510638]
赴法國=>赴法 [0.600000]
警義消=>義消 [0.666667]
智障者=>智障 [0.500000]
縣農會=>農會 [0.600000]
花蓮縣=>花蓮 [0.533333]

- Incorrect pairs:

性侵害防治委員會=>性侵害 [0.600000]
監察院糾正=>糾正 [0.625000]
海洋政策=>海洋 [0.533333]
拜耳案=>拜耳 [0.640000]

國語電話語音辨認之強健性特徵參數及其調整方法

黃儀芬 王小川
國立清華大學電機工程學系

摘要

本文對經電話線之國語語音辨認，就兩種不同的強健性解決策略進行探討。(1) 在強健型特徵參數萃取上，採用倒頻譜平均值正規化法（CMN）及自相關係數頻域正規化法（DFT-MN-AUTO）。 (2) 在特徵參數調整法方面，採用特徵參數扣減法（SBR）、階層式特徵參數扣減法（HSBR）及統計式對應法（SM）。並提出叢集模型概念及其改良架構。。以國語語音資料庫 MAT-160 進行實驗，結果顯示在有通道效應情況下，CMN 效果最佳，若有背景雜訊情況，則是 DFT-MN-AUTO 效果最好。在特徵參數調整方法方面，以訓練端及測試端皆進行階層式特徵參數扣減法運算，並採用加權式叢集模型，可得最佳辨識率。

一、 緒論

近年來，隨著電腦運算速度和儲存設備的大幅進展，使語音辨識技術得有突破性的發展，也因此，讓電腦、儀器聽懂人說的話，成爲一個非常具有實現可能性的課題。目前，在安靜環境下針對特定語者模式的技術已有相當不錯的研究成果，市面上相關的產品也紛紛出現。但是，若要更廣泛的使用語音辨識的技術，便必須再進一步發展針對不特定語者的辨認技術以及克服訓練環境語料與測試環境語料不匹配的問題。就經電話線之國語語音辨認而言，已有若干研究與探討[1,2,3,4]。

語音辨認流程可分成訓練端及測試端兩個部分。在訓練端，透過語音與詞庫資料，經由各種演算法的訓練，去預估與統計出聲學模型及語言模型。在測試端，則應用所得之模型，求出辨認結果。在聲學模型訓練的部份，著重在語料的狀態估測與切割，普遍被使用的演算法有 k 均值切割演算、維特比最佳路徑搜尋、EM 演算法…等。在語言模型訓練部份，則以樹狀結構演算最爲普遍。

電話線語音和在安靜環境下錄製的麥克風語音相較，具有較複雜的失真問題。其失真來源包含：

- 發話端之環境雜訊（ambient noise）
- 電話線通道效應（channel effect）
- 頻寬限制（bandwidth limitation）

加上各種語音環境下皆存在的語者發音變異（speaker variation）問題，構成電話線語音辨認必得處理強健性的課題。

解決電話線語音辨識強健性的問題，可從不同領域著手[5,6,7,8]。假設 x 爲訓練環境之語料波形訊號，X 爲訓練環境之語料特徵參數，Λ_x爲訓練環境下之聲學模型；y 爲訓練環境之語料波形訊號，Y 爲訓練環境之語料特徵參數，Λ_y爲訓練環境下之聲學模型。增加強健性的方法大致可分類成三種：

- 具強健能力的特徵參數求取法 $R_\tau(\cdot)$
- 在特徵參數域中估測失真偏移量 $F_\nu(\cdot)$，將 Y 扣除此一偏移量，拉近與 X 距離，以期匹配Λ_x。
- 估測模型轉換函式 $G_\eta(\cdot)$，使Λ_x能接近Λ_y。

目前最普遍的語音模型訓練方式爲求取梅爾倒頻譜參數及使用隱藏式馬可夫模型。根據生理實驗結果，人耳聽覺在頻譜效應上並非線性，梅爾倒頻譜參數的求取，即是根據相對的轉換公式，以一系列非等距非對稱的「三角帶通濾波器」模擬耳蝸在接收訊號時的情形，求出頻譜振幅，再經由對數化及餘弦轉換，可得多維的梅爾倒頻譜參數。而差分化參數可以描述發音過程中的變動特性，因此也被廣泛的採用。

語音訊號具有暫態穩定的性質，故對於同一音節或同一句話而言，雖然每次發音的長短不同，但聲道變化的過程是相似的，語音波形訊號爲此聲道共鳴後的表現。因此，可將波形訊號視爲一雙重隨機程序的結果，以隱藏式馬可夫模型來模擬。

二、 國語語音之模型設計

國語是單音節的語言，每一音節由聲母、韻母、和聲調所組成。聲母有 22 類（包含一個空聲母），韻母有 40 類（包含二個空聲母分別對應於捲舌音與齦後音），聲調

有 5 類。若不考慮聲調化，可將之歸納成 412 個音節類型。同一音節內，聲母和韻母會相互影響，產生耦合的現象；此情形在跨音節語音的韻尾與聲母、複韻的兩個韻母之間亦會出現。

只考慮音節內耦合的現象，並採用右音相關聲韻母模型。其中 22 類聲母和 6 類首韻韻母共可歸納出 94 個右音相關聲母，同理空聲母也依據首韻韻母特性分成 6 類。每個聲韻母模型分別以 3 個和 4 個狀態表示，加上 1 個靜音狀態模型。共可得 464 個狀態模型。另鑑於男女音質的差異，實驗中皆採用男女生分類模型。

三、 語音資料庫

本文中用來訓練模型的語料庫共有兩套。第一套 MAT-160[2,13]，是電話線環境下的錄音語料，係國科會補助之語音收集計畫(MAT 計畫)所錄製，其內容經過安排設計，使能涵蓋國語語音中可能出現的音節，爲主要實驗使用之語料庫。第二套爲 MIC-101，係中華電信研究所錄製，是麥克風環境下所錄製的語料庫，內容爲 2 至 4 字的短字詞彙。表 1 及表 2 分別說明其內容。

MAT-160	
語料庫來源	中華民國計算語言學學會
錄音環境	透過電話網路經由個人電腦錄製
取樣頻率	8kHz
取樣位元數	16bits
語者	女 79 人、男 81 人
每語者語料內容	12 句單音節、30 句短字詞彙、10 句平衡長句
總句數 / 總時數	8320 句 / 5.01 小時

表 1：MAT-160 語料庫

MIC-101	
語料庫來源	中華電信研究所
錄音環境	透過麥克風在安靜環境下經由個人電腦錄製
取樣頻率	8kHz
取樣位元數	16bits

語者	女 51 人、男 50 人
每語者語料內容	50 句短字詞彙
總句數 / 總時數	5050 句 / 1.44 小時

表 2：MIC-101 語料庫

訓練語料為 TEST-500，是 MAT 計畫中抽取之語音資料，在 1998 年語音辨認評比時採用之自行測試語料[14]，內容如下：

TEST-500	
語料庫來源	中華民國計算語言學學會
錄音環境	透過電話網路經由個人電腦錄製
取樣頻率	8kHz
取樣位元數	16bits
語者來源	女 15 人、男 15 人，與 MAT-160 無重複
語料內容	50 句單音節、150 句短句詞彙（text dependent） 300 句平衡長句（text independent）
總句數 / 總時數	500 句（音節總數 4736） / 0.45 小時

表 3：TEST-500 語料庫

四、 基本特徵參數及實驗

本研究中所設計之語音辨認基本系統架構中，特徵參數使用梅爾倒頻譜係數，求取法則如下表所示，此法所得的特徵參數標示為 MEL：

音框長度	256 points
音框位移	128 points
MEL 特徵參數	12-order MFCC + 12-order-delta-MFCC + 1-order-delta-log-energy + 1-order-delta-delta-log-energy

表 4：MEL 特徵參數取法

為瞭解在測試環境匹配或不匹配時的辨認結果，先就上述之訓練語料庫與測試語料庫，作基礎實驗。在實驗中，辨識率皆依照下列方式計算：

辨認率＝(正確音節總數－錯誤音節總數)／正確音節總數

錯誤音節總數＝取代音節總數＋刪除音節總數＋插入音節總數

將 MAT-160 與 MIC-101 訓練語料庫，以 MEL 爲特徵參數，分別訓練出聲學模型，以 TEST-500 爲測試語料，辨認結果如表 5。

	以 MAT-160 爲訓練語料			以 MIC-101 爲訓練語料		
	混合數			混合數		
	4	8	16	4	8	16
插入率(%)	3.53	3.82	3.61	4.22	3.86	3.93
刪除率(%)	2.22	2.11	1.94	6.65	7.45	7.47
取代率(%)	49.30	47.61	45.68	78.15	77.45	76.44
辨認率(%)	44.95	46.45	48.75	10.98	11.23	12.16

表 5：以 MAT-160 及 MIC-101 爲訓練語料之實驗結果

由此實驗結果可看出，當測試環境與訓練環境接近時，可得較佳之辨認率。而當測試環境不匹配時，辨認率即急驟下降，在混合數等於 4 和等於 16 的情況下，分別降低了 33.97%和 36.59%，降幅達 75%以上。混合數的多寡亦是決定因素，混合數的增加會提升辨認正確率，其改善主因爲降低取代型錯誤。在以 MAT-160 爲訓練語料的情況下，混合數=16 可較混合數=4 提升 3.8%的辨識率，幅度約爲 8%。

爲改進辨認正確率，本文針對(1) 強健型特徵參數萃取，以及 (2) 特徵參數調整法，作更進一步之探討。

五、 強健型特徵參數之萃取

5-1. 倒頻譜平均值正規化法（Cepstral Mean Normalization, CMN）

倒頻譜平均值正規化法是一種減低通道效應影響的方法，其原理如下：假設 x(m,n) 爲語者實際在第 m 個音框所發的第 n 點訊號值，y(m,n)爲 x(m,n)經過通道後的語音訊號值，則 x(m,n)與 y(m,n)的關係可表示如下：

$$y(m,n) = x(m,n) \otimes h(m,n) \qquad (1)$$

若假設在語者所念的一句語料內通道效應爲穩態，可移去 h(m,n)的音框引數，

成爲 h(n)。不同的通道，使 h(n)值產生差異，因此，以 y(m,n)訓練出來的模型將受通道效應的所產生的影響，若能移除 h(n)的影響，語音模型將更爲精確。以 X(m,k)、Y(m,k)及 H(k)分別代表 y(m,n)、x(m,n)及 h(n)在頻譜上的表現。

$$Y(m,k) = X(m,k)\cdot H(k) \tag{2}$$

取對數值之後，通道效應變成加法性。

$$\log Y(m,k) = \log X(m,k) + \log H(k) \tag{3}$$

等號兩邊各求平均值，可得

$$\frac{1}{M}\sum_{m=0}^{M-1}\log Y(m,k) = \frac{1}{M}\sum_{m=0}^{M-1}\log X(m,k) - \frac{1}{M}\sum_{m=0}^{M-1}\log H(k) \tag{4}$$

$$\frac{1}{M}\sum_{m=0}^{M-1}\log Y(m,k) = \frac{1}{M}\sum_{m=0}^{M-1}\log X(m,k) - \log H(k) \tag{5}$$

原對數值減其平均值，通道效應即被消除。

$$\log Y(m,k) - \frac{1}{M}\sum_{m=0}^{M-1}\log Y(m,k) = \log X(m,k) - \frac{1}{M}\sum_{m=0}^{M-1}\log X(m,k) \tag{6}$$

轉換成倒頻譜(Cepstrum)，即等於倒頻譜值減其平均值，此法稱爲倒頻譜平均值正規化(Cepstrum Mean Normalization, CMN)。

在基礎實驗所使用的 MEL 特徵參數，其中 12 階 MFCC 可視爲 log *Y(m,k)* 的線性組合，因此適用倒頻譜平均值正規化法。依據(6)式得出新的 12 階 MFCC 係數，再與其他的 12 維度未變動的特徵向量構成新的 26 階特徵參數，在接下來的敘述中，此法得到之語音參數標示爲 MEL-CMN。

5-2. 自相關係數頻域正規化法（DFT-MN-AUTO）

自相關係數頻域正規化法爲自相關係數微分法（RAS）[5,6]的變形，兩者的差異在：RAS 法爲針對消除背景雜訊而設計，DFT-MN-AUTO 法爲針對消除通道效應而設計。

若是考慮通道效應，假設 x(m,n)爲語者實際在第 m 個音框所發的第 n 點訊號值，y(m,n)爲 x(m,n)經過通道後的語音訊號值，則 x(m,n)與 y(m,n)的關係可表示如下：

$$y(m,n) = x(m,n) \otimes h(m,n) \tag{7}$$

若假設在語者所念的一句語料內通道效應為穩態，可移去 h(m,n)的音框引數成為 h(n)。

$$y(m,n)=x(m,n)\otimes h(n) \qquad (8)$$

在自相關域上，可表示成：

$$r_{yy}(m,k)=r_{xx}(m,k)\otimes h(-k)\otimes h(k) \quad ,0\le m\le M-1,0\le k\le 2N-1 \qquad (9)$$

$$r_{yy}(m,k)=\begin{cases}\sum_{j=0}^{N-1-k} y(m,j)y(m,j+k) & ,0\le k\le N-1\\ 0 & ,k=N\\ r_{yy}(m,2N-k) & ,N+1\le k\le 2N-1\end{cases} \qquad (10)$$

將第 m 個音框之自相關係數 $r_{yy}(m,k)$取 2N 點作 DFT 演算，即變成功率頻譜：

$$S_{r_{yy}}(m,f)=S_{r_{xx}}(m,f)\cdot\left|H(f)\right|^2 \qquad (11)$$

取對數值之後，

$$\log S_{r_{yy}}(m,f)=\log S_{r_{xx}}(m,f)+2\log\left|H(f)\right| \qquad (12)$$

等號兩邊各求平均值，可得

$$\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)=\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{xx}}(m,f)+\frac{2}{M}\sum_{m=0}^{M-1}\log\left|H(f)\right|$$

$$\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)=\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{xx}}(m,f)+2\log\left|H(f)\right| \qquad (13)$$

原對數值減其平均值，通道效應即被消除。

$$\log S_{r_{yy}}(m,f)-\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)=\log S_{r_{xx}}(m,f)-\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{xx}}(m,f) \qquad (14)$$

據此可得調整後的自相關係數，其求法如下：

$$\overline{r_{yy}}(m,k)=InverseDFT\left\{\exp\left(\log S_{r_{yy}}(m,f)-\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)\right)\right\}$$

$$,0\le m\le M-1,\ 0\le k\le 2N-1,\ 0\le f\le 2N-1 \qquad (15)$$

由(15)式得出經由頻域正規化調整的自相關係數，因自相關係數串之對稱特性，取前 N 點結果作爲最終訊號串輸出 $r_{yy}^{+}(m,k)$。

以 $\overline{r_{yy}^{+}}(m,k)$取代原始的語音波形訊號，求取 MFCC 特徵參數並訓練相關模型；值得注意的是，當以前述 MEL 法求取 MFCC 時，參數 c_0通常是捨棄不用的，因其代表的是訊號的強弱，對於語音特徵來說並不具鑑別度。但以自相關係數所求出的 c_0包含其它的語音訊息，故在接下來的實驗裡，嘗試兩種作法，一爲保留 c_0（取係數 c_0~c_{11}），標示爲 DFT-MN-AUTO-C0；另一作法如同 MEL，捨棄 c_0，取 c_1~c_{12}爲 MFCC，標示爲 DFT-MN-AUTO-C1。如圖 1 所示，語音波形訊號自相關係數之求取，我們採用一快速演算法：

- 將波形訊號經由 DFT 轉換至頻域。
- 計算頻域訊號的能量參數。
- 將能量參數經由 IDFT 轉回時域，即可得自相關係數，自相關係數的長度爲 2N 點。

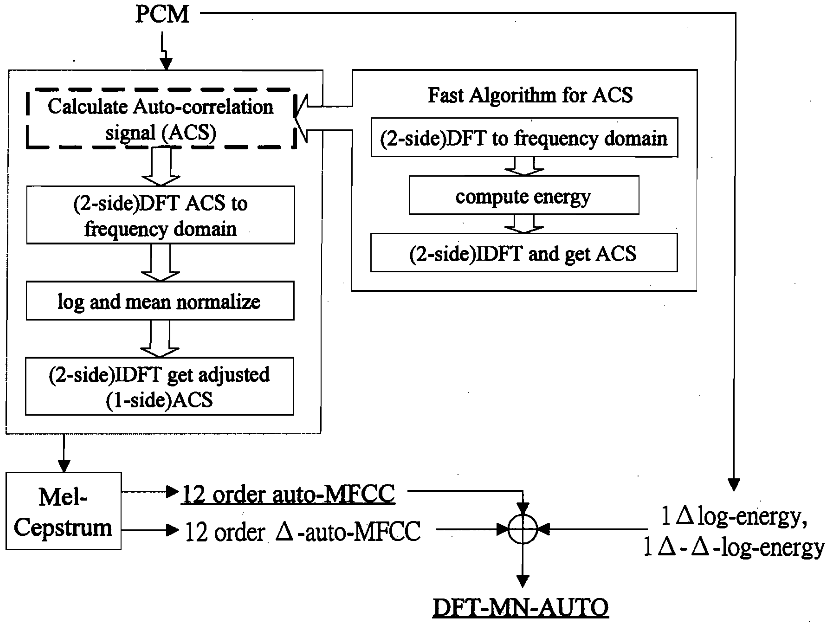


圖 1：DFT-MN-AUTO 特徵參數求取流程

5-3. 強健型特徵參數實驗

(1) 測試語料與訓練語料環境匹配實驗

以 MAT-160 爲訓練語料，並分別以 MEL-CMN、DFT-MN-AUTO-C0 及 DFT-MN-AUTO-C1 爲特徵參數，測試結果如表 6 所示。

特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
插入率(%)	2.05	2.36	1.84	1.71	2.17	2.13	3.25	3.29	2.93
刪除率(%)	2.05	1.92	1.71	1.96	1.79	1.56	2.07	1.73	1.48
取代率(%)	43.96	41.89	41.28	46.26	44.45	44.13	46.18	44.89	43.77
辨認率(%)	51.94	53.82	55.17	50.06	51.58	52.17	48.50	50.08	51.82

表 6：不同特徵參數之測試結果

由實驗可看出 MEL-CMN 特徵參數求取法的表現最佳，DFT-MN-AUTO-C0 其次，DFT-MN-AUTO-C1 最末。但與表 5 相對照，皆可獲一定程度之提升。三種強健型特徵參數中，DFT-MN-AUTO-C1 的插入型錯誤比率較 MEL-CMN 與 DFT-MN-AUTO-C0 高，因此，若搭配較精準的切音方法，應可得更佳的表現。

(2) 去除波形訊號偏移量之延伸實驗

觀察了訓練語料庫與測試語料庫波形訊號的情形，發現錄製方式不同的關係，波形訊號的平均值會產生不同的偏移特性，其統計結果如圖 2 及圖 3 所示。

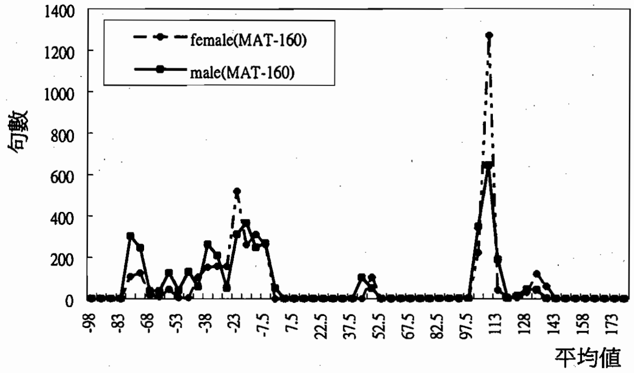


圖 2：MAT-160 波形訊號偏移特性

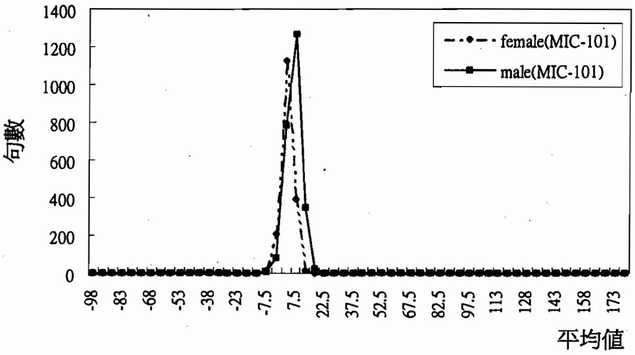


圖 3：MIC-101 波形訊號偏移特性

根據聲波的物理特性，一句聲音的波形訊號平均值應該在零附近，一如圖 3 所示透過麥克風錄製乾淨語音的統計結果。但我們發現 MAT-160 語料的波形訊號偏移特性並非如此，其可能原因爲：MAT（臺灣地區國語語音資料庫）語料是經由電話線連接到國內各大學語音實驗室進行錄音工作。各大學所使用的錄音界面是透過個人電腦上的 Dialogic Card 取樣並儲存的，若此卡的 DC 值略有偏移，則該大學所錄製的語音皆受此偏移量影響。MAT-160 語料庫便是從 MAT 計劃所錄製的聲音中整理與挑選出較佳之 160 人而成，因此，包含了數所學校錄音的結果。對照圖 2，可發現 MAT-160 的波形訊號偏移有數個鋒值，恰可符合上述所做的假設。所以，一個多處地點收音的語料庫應就系統間校準的問題多加注意。

爲了消除錄製系統間的差異對語料所造成的影響，在進行模型訓練及辨認之前，以句爲單位，對每句語料扣減其波形訊號的平均值，得出新的語料，此法標示爲 PCM-MN。以 PCM-MN 法處理語音訊號之後，再求特徵參數，其實驗結果如下，

特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
插入率(%)	2.68	2.96	2.58	1.94	2.11	2.03	3.29	3.59	3.42
刪除率(%)	2.09	1.79	1.52	2.01	1.75	1.56	2.13	1.73	1.58
取代率(%)	44.53	43.07	41.81	45.16	43.52	42.93	45.02	43.86	42.48
辨認率(%)	50.70	52.17	54.10	50.89	52.62	53.48	49.56	50.82	52.51

表 7：不同特徵參數，採用 PCM-MN 法之測試結果

由實驗可看出， PCM –MN 法對於使用自相關係數計算特徵參數的方法（DFT-MN-AUTO-C0 和 DFT-MN-AUTO-C1）具正面的效益，在混合數等於 4 與 16 的情形下，DFT-MN-AUTO-C0 有 0.83%與 1.17%的提升，DFT-MN-AUTO-C1 有 1.06%與 0.49%的提升。其原因是具 DC 偏移量的波形訊號對於自相關係數的影響是全面的，若假設受到錄音設備偏移的影響爲 $y'(m,n)=y(m,n)+bias$，在自相關域則爲 $r_{yy}'(m,k)=r_{yy}(m,k)+ bias^2+bias*\Sigma[y(m,n)+y(m,n+k)]$，最末一項造成的雜訊不僅不能從之後的方法消除，而且因與訊號相乘關係，影響值不小。當進行 PCM-MN 時，偏移值造成的影響便大幅被壓抑，故可提升辨識率結果。

然而，對 MEL-CMN 特徵參數而言，則具負面影響，在混合數等於 4 與 16 的情

形下，辨識率下降了 1.24%與 1.07%。波形訊號偏移量對頻域的影響僅在頻率爲零處，其餘頻域皆不受干擾，故對 MFCC 影響本就不大。當額外加上 PCM-MN 時，事實上強迫所有波形訊號的平均值皆爲零，對照圖 3 麥克風錄音的情況，可知波形訊號的平均值實際並非皆爲零，而是在零附近呈現一窄分佈情況，加上扣減平均值的過程中忽略了偏移值時變的可能性，故 PCM-MN 亦造成外加的雜訊來源。對 MEL-CMN 而言，因其必須收集所有音框之 MFCC 求其平均，再扣減回每一音框，故某些音框若遭較強干擾，會透過 CMN 的過程連帶影響其他音框，進而降低了整句語料之特徵參數可靠性。

(3) 訓練語料與測試語料環境不匹配實驗

將訓練語料換成 MIC-101，測試訓練語料與測試語料在環境不匹配情形下，各種強健型特徵參數的表現如下。

特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
插入率(%)	3.40	3.25	3.04	3.38	3.25	3.23	5.47	5.41	5.30
刪除率(%)	2.98	3.00	2.81	3.12	2.96	2.68	2.72	2.66	2.70
取代率(%)	64.29	63.66	63.05	66.79	66.91	66.49	66.89	65.60	64.76
辨認率(%)	29.33	30.09	31.10	26.71	26.88	27.51	24.92	26.33	27.24

表 8：不同特徵參數之測試結果

實驗中，MEL 特徵參數在環境不匹配的情形下，混合數 4 與 16 的辨識率爲 10.98%與 12.16%，以此對照，使用強健型特徵參數 MEL-CMN 與 DFT-MN-AUTO 皆可獲得改善，其中又以 MEL-CMN 的效果最佳。在混合數爲 4 與 16 的條件下，MEL-CMN 可提升辨識率 18.35%與 18.94%，DFT-MN-AUTO-C0 可提升 15.73%與 15.35%，DFT-MN-AUTO-C1 可提升 13.94%與 15.08%。

(4) 外加雜訊實驗

除了通道效應外，背影雜訊亦是造成辨識率下降的主因。因所使用的測試語料 TEST-500 可視爲較乾淨的電話線語音，爲模擬雜訊情形，對語音訊號加上不同強度的白雜訊（WHT），其實驗結果如下：

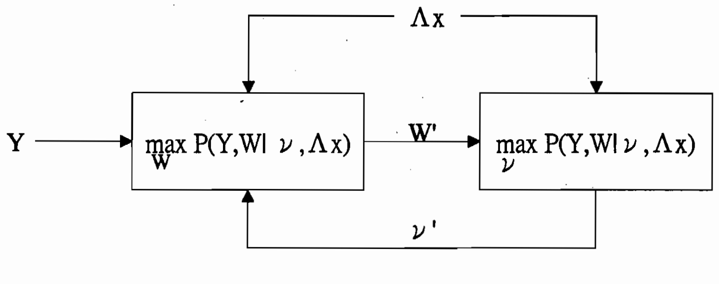
特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
無外加雜訊	51.94	53.82	55.17	50.89	52.65	53.48	49.56	50.82	52.51
SNR= 20 dB	37.36	40.87	42.14	38.86	40.98	42.24	39.07	41.74	42.27
SNR= 10 dB	17.33	19.10	21.37	22.70	25.02	27.47	25.21	27.13	28.28
SNR= 0 dB	1.44	2.26	4.08	4.88	6.07	6.57	8.24	7.40	8.79

表 9：不同特徵參數，有外加雜訊時之辨認結果

由實驗的結果可發現，DFT-MN-AUTO 法對抗雜訊的能力較強。當信號雜訊比大於 20dB 時，採用自相關係數便可獲得好處。原因是在求取自相關係數的過程中，訊號加乘可增大其訊號雜訊比值。其中 DFT-MN-AUTO-C1 的效果略好於 DFT-MN-AUTO-C0，應與所加入的訊號爲白雜訊有關，自相關域上白雜訊會影響低頻部份，而 DFT-MN-AUTO-C1 的求取正好可避開這部份的影響。。

六、 特徵參數調整法

特徵參數調整法的主要概念爲：在特徵參數域中估測與參考模型之間的失真偏移量，據之調整觀測序列特徵參數，以期符合訓練環境情狀。圖 4 爲示意圖[9]，其中 Y 代表測試語料特徵參數序列，W’代表辨識後的音節字串（包含錯誤可能），Λ_x 代表參考模型，ν 爲特徵參數調整參數。


Λx
Y
$\max_{W} P(Y,W| \nu,\Lambda x)$
W'
$\max_{\nu} P(Y,W| \nu,\Lambda x)$
ν'

$$(\nu', W') = \operatorname*{argmax}_{(\nu, W)} p(Y, W| \nu, \Lambda x)$$

圖 4：特徵參數調整法示意圖

6-1. 訊號偏移消去法（Signal Bias Removal, SBR）

訊號偏移消去法[7]爲一種估測測試環境與模型間頻道偏移量，再對原特徵向量扣除偏移量的方法，以消除測試與訓練環境間不同頻道效應的影響，簡介作法如下：

假設$Y=\{y_t\}$爲測試語音觀測序列，$X=\{x_t\}$爲符合訓練環境的語音觀測序列，兩者關係如下：

$$y_t = x_t + \overline{b} \tag{16}$$

偏差量$\bar{b}$以下列式子估出：

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu}_{s(t)}) \tag{17}$$

其中$\hat{\mu}_{s(t)}$代表第 t 個音框中，相對於訓練碼本有最大觀測機率之狀態平均值。

經由多次遞迴，SBR 可估出更準確的偏差量，在第 n 次遞迴時匹配訓練環境的觀測序列爲：

$$\widetilde{X} = \{y_t - (\overline{b}^{\,n} + \overline{b}^{\,n-1} + \ldots + \overline{b}^{\,0})\} \tag{18}$$

6-2. 層次式訊號偏移消去法（Hierarchical Signal Bias Removal, HSBR）

層次式訊號偏移消去法[8]亦是一種在特徵參數域進行調整的強健性補償方法，與 SBR 不同的是：不是對整句測試語料估出一個可能的頻道偏移量，而是對每一個音框求取音框相關（frame dependent）的偏移量。基本假設爲在一段語句中所受到的頻道偏差不全是穩態的，不同音段內的頻道情形應是變化不定的，因此針對不同音框採用不同的偏移量來模擬頻道失真。

爲估出「音框相關偏差」(frame-dependent bias)，假設$Y=\{y_t\}$爲測試語音觀測序列，$X=\{x_t\}$爲符合訓練環境的語音觀測序列，$B=\{\overline{b_t}\}$則爲要估測的音框相關偏差序列，三者關係如下：

$$y_t = x_t + \overline{b}_t \tag{19}$$

求取$\overline{b}_t$之前，先將測試語料與參考模型碼本相較較，找出每個音框所屬的叢集。假設音框被分散到 M 個叢集中，第 i 個叢集偏移量定義成：

$$b_i^c = \frac{1}{T_i}\sum_{l=1}^{T_i}(y_{t_i(l)} - \hat{\mu}_i), \qquad 1 \le i \le \mathrm{M} \tag{20}$$

其中$y_{t_i(l)}$表示測試語句中屬於叢集 i 的音框，T_i爲屬於叢集 i 的音框數目，而$\hat{\mu}_i$則

爲叢集 i 的平均值向量。音框相關偏移量可經由各個叢集偏差b^c的線性組合得出：

$$\overline{b_t} = \frac{\sum_{i=1}^{M} b_i^c w_{t(i)}}{\sum_{j=1}^{M} w_{t(i)}} \tag{21}$$

其中叢集權重$w_{t(i)}$是每個音框分別對叢集 i 的平均值做加權歐氏距離（Weighted Euclidean distance）的倒數。最後估出匹配訓練環境的觀測序列：

$$\widetilde{X} = \{(y_t - \overline{b}_t)\} \tag{22}$$

此法也和 SBR 一樣，可遞迴地減少環境間的差異性。

6-3. 統計式對應法（Stochastic Matching, SM）

統計式對應法[9,10,11]是一種基於最大機率估測（ML estimation），來求取測試環境與訓練環境特徵參數域失真的方法。因其亦爲音框獨立的偏移扣減方式，故基本假設如訊號偏移消除法，所不同的是，統計式對應法包含機率概念，除了用到參考模型的狀態平均值外，也考慮變異數的影響：

假設$Y = \{y_t\}$爲測試語音觀測序列，$X = \{x_t\}$爲符合訓練環境的語音觀測序列，兩者關係如下：

$$y_t = x_t + \overline{b} \tag{23}$$

偏差量$\bar{b}$以下列式子估出：

$$\bar{b} = \frac{\sum_{t=1}^{T} \frac{(y_t - \hat{\mu}_{s(t)})}{\Sigma_{s(t)}}}{\sum_{t=1}^{T} \Sigma_{s(t)}} \tag{24}$$

其中$\hat{\mu}_{s(t)}$和$\Sigma_{s(t)}$代表第 t 個音框中，相對於訓練碼本有最大觀測機率之狀態平均值及變異數。同樣的，統計式對應法亦與前兩種方法一樣，可遞迴地減少環境間的差異性。

6-4. 特徵參數調整法實驗

特徵參數調整法可分別對測試端與訓練端進行。

在測試端的做法爲：

1. 每句測試語料以一階動態規劃演算法得出最佳狀態序列；
2. 利用步驟 1 所得序列進行特徵參數調整；
3. 重複進行步驟 1 與步驟 2，可得不同遞迴數的辨識結果。

在訓練端的做法爲：

1. 對每句訓練語料以維特比演算法得出最佳狀態序列；
2. 利用步驟 1 所得序列進行特徵參數調整；
3. 對新的特徵參數進行叢集，得出調整模型；
4. 重複進行步驟 1 到步驟 3，可得不同遞迴數的調整模型。

在實驗中，分別對於 (1)測試端作特徵參數調整，以及(2) 訓練端與測試端的語料皆作特徵參數調整，進行測試。所使用的訓練語料皆爲 MAT-160，以 MEL 特徵參數作測試，使用混合數 16 的女男分類模型，其未作特徵參數調整時之辨認率爲 48.75%，作特徵參數調整時之結果在表 10。

		遞迴數				
		1	2	3	4	5
測試端作特徵參數調整	SBR	53.00	53.59	53.84	53.95	54.05
	HSBR	52.85	53.44	53.80	53.97	54.01
	SM	52.72	53.53	53.86	53.82	53.95
訓練端與測試端皆作特徵參數調整	SBR	54.27	54.65	54.52	54.43	54.46
	HSBR	53.80	54.62	54.65	54.77	55.03
	SM	53.36	54.52	54.48	54.50	54.54

表 10：特徵參數調整之測試結果

由實驗的結果可看出，只在測試端進行特徵參數調整，則不論是特徵參數扣減法、階層式特徵參數扣減法或統計式對應法，所獲致的改善成果都差不多，在遞迴數爲 5 的情形下，經由特徵參數調整可提升整體辨識率 5.20% ~ 5.30%，提升的幅度約爲 10.50%。

訓練端與測試端皆作特徵參數調整時，訓練端的模型是經由 5 次遞迴特徵參數調整後，進行叢集訓練而得出。比較實驗的結果，可發現，在訓練端亦進行模型參數的調整對於辨識率有 0.42% ~ 1.02%的提升。這是因爲在訓練端進行特徵參數調整時，訓練語料透過扣減偏移量的計算會使得模型模糊的程度下降（即變異數值下降），而使模型更加強健。這點可由階層式特徵參數扣減法提升最多可看出來，該法在求取特徵參數偏移量時與參考模型的精確度關連性最大。

結果顯示三種方法中，階層式特徵參數扣減法的辨識效果最好，這是因爲此法多加考慮了特徵參數偏移量的時變性，其偏移值計算是音框相關的，而特徵參數扣減法和統計式對應法的偏移值則是音框獨立的，辨認率結果也顯示沒有很大的差別。

6-5. 叢集模型建構同類參考碼本

在上述實驗中，每一音框的特徵參數是相比於最佳解碼序列中對應之狀態混合數，但是，相對應的狀態混合數並非皆是正確值，根據觀察辨識結果，發現造成辨識錯誤大部份的原因皆爲聲母或韻母的取代型錯誤，並且大半落在混淆音的叢集範圍內，故若對可能的混淆狀態混合數進行特徵參數調整的運算，對於辨識率應有所助益。因此，希望以叢集模型[12]的架構，替代原先的對應狀態混合順列。

叢集模型的概念是：對於參考模型內所有狀態的所有混合數進行叢集分類。叢集的結果可得：1.叢集模型，由所有落在該叢集的狀態混合數以線性方式合併得出；2.各狀態混合數的叢集映射表。

在測試端，計算特徵參數偏移量時，每一個音框相對應之狀態混合數以其映射叢集模型替代，辨認流程如圖 5：

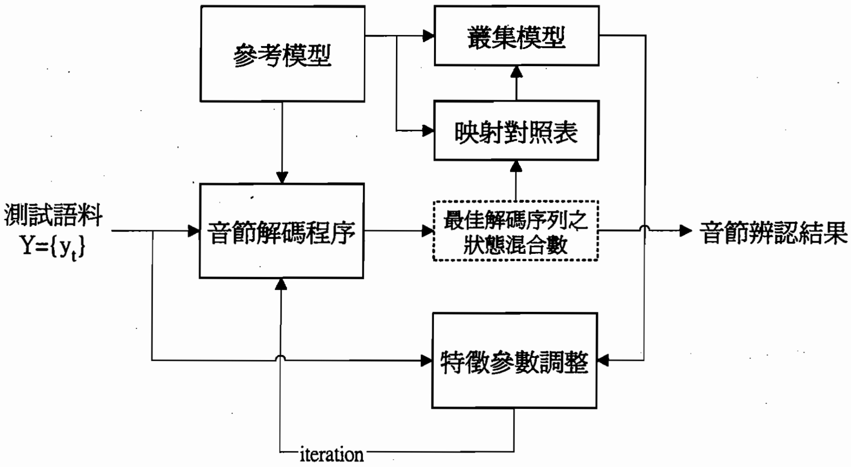


圖 5：叢集模型辨認的基本架構

在圖 5 中，匯入特徵參數調整方塊的模型參考值是透過一映射對照表，選取最相

近的叢集模型匯入特徵參數調整方塊作爲參考模型，進行特徵參數調整運算。此法標示爲 VQModel。一個混合數 16 的模型，所有狀態的所有混合數總共約有 4500 個，接下來的實驗中，將之叢集成 256 個碼本，每個碼本約可分到 10~40 個混合數。所使用的訓練語料爲 MAT-160，以 MEL 特徵參數作測試，使用混合數 16 的女男分類模型，其未作特徵參數調整時之辨認率爲 48.75%。此法先對只作測試端特徵參數調整之 SBR 方法進行測試，其結果並不理想，表 11 所示之辨認率沒有改進。

		遞迴數				
		1	2	3	4	5
只作測試端特徵參數調整(SBR)	未使用叢集模型	53.00	53.59	53.84	53.95	54.05
	使用叢集模型	53.46	53.86	53.91	53.82	53.84

表 11：使用叢集模型之實驗結果

這是因爲一個叢集模型的參數值爲落入此叢集中的所有狀態混合數平均而得，故若某測試音框對應的最佳狀態混合單位若映射到叢集的邊緣，在計算時會硬是將它拉到中心的位置，而造成大的誤差。然而，比對其音節辨認輸出時可發現，叢集模型的方式可以補償一些基本 SBR 法無法改善的部份。

爲了改善因叢集造成模型值的偏移，進而影響補償效果，因此需對叢集模型辨認的基本架構進行改良。所選取的改良作法爲：對於同一叢集的所有狀態混合數求取音框對應偏移值，再以相似機率值（likelihood probability）進行加權；因此一個音框的偏移量是經由其最佳解碼序列所落入叢集內的所有狀態混合數，計算距離並加權相加而產生的。這個方法捨棄了叢集參考模型，僅使用了映射對照表。此法標誌爲 WVQModel，實驗結果如表 12。

		遞迴數				
		1	2	3	4	5
測試端作特徵參數調整	SBR	53.12	53.82	54.14	54.16	54.22
	HSBR	53.23	53.59	54.05	54.20	54.18
	SM	52.83	53.67	53.70	53.89	54.08
訓練端與測試端皆作特徵參數調整	SBR	54.22	54.65	54.67	54.60	54.60
	HSBR	55.29	55.11	55.19	55.22	55.17
	SM	53.32	54.24	54.69	54.73	54.73

表 12：使用改良叢集模型之實驗結果

比較實驗結果，加權式模型叢集的方法皆可獲致比直接序列方法較佳的辨認率。

七、 結論

在本文中，討論了兩種不同的強健性策略：強健型特徵參數萃取及特徵參數調整法。所採用的三種的強健型特徵參數萃取方法中， MEL-CMN 對抗通道效應的能力最好，DFT-MN-AUTO-C1 對抗雜訊的能力最佳。在訓練環境與測試環境皆為電話網路的情況下，MEL-CMN 可得最佳辨識率 55.17%。在測試的三種特徵參數調整方法中，若只在測試端單方作特徵參數調整，這三種方法的效果都差不多，辨識率約可提升 5.25%。若是在訓練與測試雙方作特徵參數調整，可再提升辨識率 0.50%~1.00%。另外，提出叢集模型概念及其改良「加權式叢集模型法」，此法可再提升辨識率約 0.20%。綜上所述，以階層式特徵參數扣減法，在訓練與測試雙方作特徵參數調整，並使用加權式叢集模型運算，可得最佳辨識率 55.29%。

參考文獻

[1] 簡仁宗, “電話環境下語音辨認之研究”，國立清華大學電機工程研究所博士論文，民國八十六年六月

[2] 邱榮樑，“經電話網路之連續語音辨認及辨認技術評比方法之建立”，國立清華大學電機工程研究所碩士論文，民國八十七年六月

[3] 涂英傑，“電話環境下國語語音辨認之強健性問題”，國立臺灣大學電機工程研究所碩士論文，民國八十七年六月

[4] 謝華君，“電話網路上國語連續音節辨認的初步研究”，國立清華大學電機工程研究所碩士論文，民國八十六年六月

[5] You, Kuo-Hwei and Hsiao-Chuan Wang, “Robust Features Derived from Temporal Trajectory Filtering for Speech Recognition under the Corruption of Additive and Convolutional Noises”, Proceedings ICASSP, 1998, pp. 577-580.

[6] You, Kuo-Hwei and Hsiao-Chuan Wang, “Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences”, Speech Communication vol. 28, no. 1, pp. 13-24, 1999.

[7] Rahim, Mazin G. and Biing-Hwang Juang, “Signal Bias Removal by Maximum Likelihood

Estimation for Robust Telephone Speech Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, pp. 19-30, 1996.

[8] Rahim, Mazin G., Biing-Hwang Juang, Wu Chou and Buhrke E., "Signal conditioning techniques for robust speech recognition", IEEE Signal Processing Letters, Vol. 3, No. 4, pp. 107-109, 1996.

[9] Sankar, Ananth and Chin-Hui Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 3, pp. 190-202, 1996.

[10] Siohan, Olivier and Chin-Hui Lee, "Iterative Noise and Channel Estimation Under the Stochastic Matching Algorithm Framework", IEEE Signal Processing Letters, Vol. 4, NO. 11, pp. 304-306, 1997.

[11] Chien, Jen-Tzung, Hsiao-Chuan Wang and Lee-Min Lee, "Estimation of Channel Bias for Telephone Speech Recognition", Proceedings ICSLP 1996, pp. 1840-1843.

[12] Lawrence, Craig and Mazin Rahim, "Integrated Bias Removal Techniques for Robust Speech Recognition", Proceedings, EuroSpeech 1997, pp. 2567-2570.

[13] Wang, H. C., "MAT- A project to collect Mandarin speech data through telephone networks in Taiwan," Computational Linguistics and Chinese Language Processing, vol. 2, no. 1, pp. 73-90, 1997.

[14] Wang, H. C., "Speech research infra-structure in Taiwan – from database to performance assessment," Proceedings, 1999 Oriental Cocosda Workshop, 1999, pp. 53-56.

音框同步之雜訊補償方法在汽車語音辨識之應用

Frame Synchronous Noise Compensation for Car Speech Recognition

簡仁宗　　林敏順
國立成功大學資訊工程學系
Email : jtchien@mail.ncku.edu.tw

摘要

自動語音辨識(ASR)系統應用在雜訊環境下時，由於雜訊語音與模型參數間的不匹配，將導致辨識率明顯的下降。音框同步補償方法可從測試語音中以音框為單位作補償，每個音框先計算出語音與模型參數之線性等化因子，再根據等化因子的大小將模型參數的平均值調整量對應出來，自動找出語音隱藏式馬可夫模型之參數調整量，將模型參數調整後再作辨識。本論文提出一種強健性的方法做平均值調整函式的估測，從實驗結果得知，使用本方法可有效提升在汽車噪音環境下語音辨識的正確率。在九十公里和五十公里汽車環境下免持麥克風之詞彙辨識系統都有明顯的改善。

1. 簡介

語音是人與電腦間最自然的溝通方式，要讓電腦聽的懂人講話的聲音一直是人類努力的目標，此目標達成與否要視電腦語音辨識技術的開發成熟度而定，雖然傳統錄音間所錄得的語音已經可以達到很高的辨識率，但真正實用之語音辨識系統其應用之場所一定有程度不等的噪音存在[6][7]，若此辨識系統不做任何調整，測試語音與模型參數間的不匹配將導致辨識率明顯的下降。因此我們以汽車環境下免持聽筒之語音辨識噪音補償來進行研究。

一般而言，不同汽車上的噪音大小不同，要用來訓練(training)噪音語音模組之語料庫將會非常龐大且不易取得是不切實際的方法；另一方面汽車環境是屬於高雜訊的地方，例如、引擎輪胎轉動引起的噪音、風嘯聲、收音機或音響的噪音、汽車內說話的回音等等，所以在

測試(testing)時與語音模組不吻合的情形會非常嚴重，這些都會使語音辨識技術更加困難，我們所提的方法就是要解決上面的問題。我們以安靜房間錄下的語料庫訓練出一組語音模組，再以實際汽車裡錄得的少數語料訓練出平均值與變異數調整函式，當測試語料在測試時會根據噪音程度自動對應出平均值及變異數的調整量作補償。

我們所提出的方法是依據在汽車環境下免持麥克風語音辨識之噪音補償方法[1]做改良，這個方法在做估測模型參數調整函式時需要將噪音語料做維特比(Viterbi)切音，由於噪音語料庫經Viterbi切音之後不準確，如再依據等化因子估測調整函式效果不佳，為了改善這個缺點我們以新的方法來估測調整函式。本補償方法是可以音框同步的，主要是當聲音錄得的同時不需要等所有語音資料都收集好再進行辨識，可以以一個音框為單位，錄得一個音框後就可直接計算等化因子，再根據等化因子對應出平均值調整量，將乾淨語音模組依不同環境作不同之調整，在經過實際測試後的確可有效地改善汽車環境下的語音辨識率。

2. 噪音補償方法

這裡的噪音補償方法的理論基礎是從美國喬治亞理工學院學者Carlson 和 Clements 於1994年提出的“以投射為主之相似度量測 (projection-based likelihood measure)” [3][4]方法所延伸出來的，根據1989年AT&T 貝爾實驗室 Mansour 和 Juang [9]的觀察結果發現，任何乾淨語音的倒頻譜向量受到白色雜訊(white noise)的干擾，其向量的大小值會縮小且相位大致不變的特性，投射為主之相似度量測就是根據這個觀察結果所發展出來的抗噪音辨識演算法。但在實際汽車環境下的噪音並非白色雜訊，因此我們導入一個等化因子(equalization factor) λ_e，我們將根據等化因子訓練出平均值調整函式，再根據等化因子大小自動對應出其平均值調整量，再語音辨識時結合進去。以下為等化因子求法和相似度量測時結合平均值補償量之算法。

2-1 等化因子

一個觀測樣本$\mathbf{c}_t$與乾淨語音所訓練出來的隱藏式馬可夫模型$\Lambda_{s,m}=(\mu_{s,m},\Sigma_{s,m})$ 做相似度

量測時(其中s為狀態的引數，m為混合數的引數)，我們通常都用高斯機率密度 $P(\mathbf{c}_t|\Lambda_{s,m})=N(\mathbf{c}_t;\mu_{s,m},\Sigma_{s,m})$ 來量測，然而，在噪音的干擾下，此相似度量測的平均值向量$\mu_{s,m}$部份應該自動匹配掉雜訊語音，所以在平均值向量乘上一個線性調整因子λ形成以下的相似度量測：

$$P(\mathbf{c}_t\big|\lambda,\Lambda_{s,m})=N(\mathbf{c}_t;\lambda\mu_{s,m},\Sigma_{s,m})=(2\pi)^{-N/2}\big|\Sigma_{s,m}\big|^{-1/2}\cdot\exp\left(-\frac{1}{2}(\mathbf{c}_t-\lambda\mu_{s,m})^T\Sigma_{s,m}^{-1}(\mathbf{c}_t-\lambda\mu_{s,m})\right) \quad \textbf{(1)}$$

其中N是向量的維度$\mu_{s,m},\Sigma_{s,m}$是隱藏式馬可夫模型的平均值及變異數。

應用最佳相似度 **(maximum likelihood, ML)** 法則，可以推導出最佳等化因子λ_e如下所示:

$$\lambda_e=\arg\max_{\lambda}\log P(\mathbf{c}_t\big|\lambda,\Lambda_{s,m})=\frac{\mathbf{c}_t^T\,\Sigma_{s,m}^{-1}\,\mu_{s,m}}{\mu_{s,m}^T\,\Sigma_{s,m}^{-1}\,\mu_{s,m}} \quad \textbf{(2)}$$

此λ_e是 $\mathbf{c}_t$ 在$\mu_{s,m}$上之投射量，為隱藏式馬可夫模型引數s和m的函式，及觀察樣本$\mathbf{c}_t$的相關的函式。將式**(2)**的λ_e代回式**(1)**即為投射為主之相似度量測的計算方法。

2-2 相似度量測之參數調整

由於平均值部份用簡單的等化因子做調整會產生程度不等的偏差，我們認為平均值向量的偏差 $\mathbf{b}=\mathbf{c}_t-\lambda_e\mu_{s,m}$ 也應一併補償，平均值補償函式$\mathbf{b}(\lambda_e)$是與λ_e有關，我們對所有的$\mathbf{c}_t$與隱藏式馬可夫模組$\Lambda_{s,m}$會針對各λ_e值去統計平均值應對應的調整量$\mathbf{b}$，當語音辨識進行辨識之相似度量測時先計算出λ_e值後就可根據平均值補償函式$\mathbf{b}(\lambda_e)$自動對應出平均值調整量來作補償，以提高語音在噪音環境下的辨識率。式**(3)**為結合平均值補償函式後之相似度量測的計算方法：

$$P\big(c_t\,|\,\lambda_e,\Lambda_{s,m},b(\lambda_e)\big)=N\big(c_t;\lambda_e\mu_{s,m}+b(\lambda_e),\Sigma_{s,m}\big)=$$

$$(2\pi)^{-N/2}\big|\Sigma_{s,m}\big|^{-1/2}\exp\left(-\frac{1}{2}\big(c_t-\lambda_e\mu_{s,m}-b(\lambda_e)\big)^T\Sigma_{s,m}{}^{-1}\big(c_t-\lambda_e\mu_{s,m}-b(\lambda_e)\big)\right) \quad \textbf{(3)}$$

本方法又稱為平均值補償過相似度量測(Mean Compensated Likelihood Measure, MCLM) 。

我們將對所有的c_t與隱藏式馬可夫模組$\Lambda_{s,m}$針對各λ_e值去統計平均值的調整量，在我們的實驗中λ_e值是介於-2到4之間，每0.01為一個區間(section)共600個section，當語音辨識進行辨識之相似度量測時先計算出λ_e值後自動對應出平均值調整量作補償，以提高語音在噪音環境下的辨識率。

3. 調整函式之估測

在原始平均值和變異數調整函式的估測過程中[5]，我們需要準備一組乾淨語料庫，訓練出乾淨的語音模組，以及一組以人工方式加上不同噪音型態和噪音分貝的噪音語料庫，但要在汽車環境下做自動語音辨識系統(ASR)，這樣的作法是不太實際，因為我們無法在汽車上錄下一組與實驗室同步的語音資料，這很困難而且有諸多限制，最好的方法就是實際準備一組在噪音環境錄下的少量語料庫用來訓練調整函式。對於調整函式的估測我們提出兩種研究方法並在實驗結果中列出其辨識的結果。

3-1 噪音語料庫

汽車噪音環境下的語料庫是實際在汽車行駛中錄下的語音資料；有時速零公里怠速路況、五十公里正常路況及九十公里高速路況三組；以免持聽筒遠距離麥克風(**hands free far talking**)方式錄音，錄音時引擎發動、關閉汽車音響、關閉車窗、冷氣開到最小，使用**MDWalkman(MZ-R55)**錄音設備，麥克風為高抗雜訊麥克風型號為(**SONY ECM-717**)，麥克風置於副駕駛坐前的置物箱上。語料庫內容以人名和命令為主；其中有人名**100**個、命令**14**個。以這樣的錄音環境下我們總共錄了**1271**句，由五男五女所錄下。

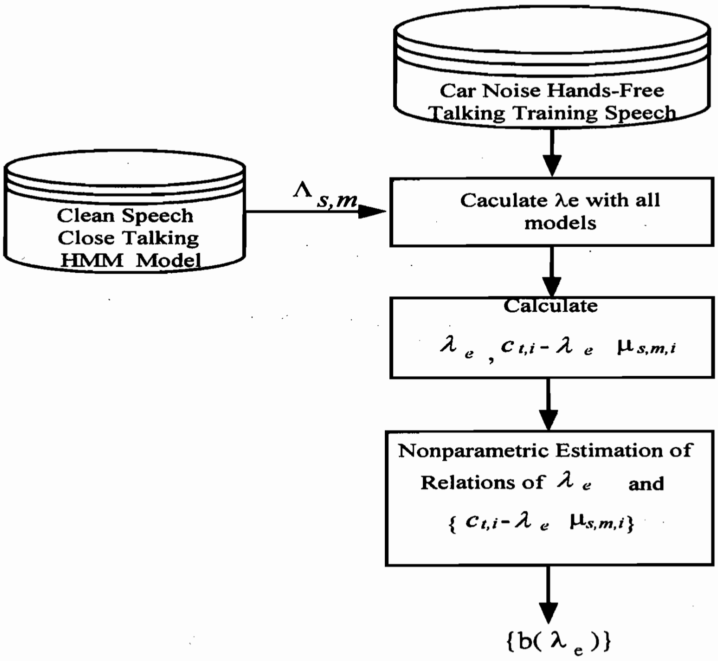
這**1271**句總共是由兩組語料庫所組成，這兩組語料分別是用來訓練補償函式用的訓練語料庫，和用來做測試用的測試語料庫。第一組是訓練補償函式用的語料庫，汽車是**TOYOTA COROLLA 1.8**分別由兩男兩女實際開車所錄下的語料總共有**480**句，零公里有**122**句、五十公里**158**句、九十公里**200**句。第二組是測試用語料庫，汽車是裕隆尖兵**1.6**分別由三男三女實際開車所錄下的語料總共有**791**句、零公里有**204**句、五十公里**263**句、九十公里**324**句。在估測補償函式方面我們又把第一組資料額外分出一些訓練語句，由原本兩男兩女再分出一男一女，分出的部份共有**222**句，零公里**54**句、五十公里**74**句、九十公里**94**句，我們將觀察訓練語句的多寡與辨識率的關係。

3-2 調整函式的估測方法

由於與傳統相類似的方法[1]需要以維特比(**Viterbi**)切音之後再根據λ_e去統計補償函式，但是噪音語料庫經維特比(**Viterbi**)切音後會有誤差，雖然仍具有很重要的統計特性，但統計特性不可靠。為了改善這個缺點我們以新的方法來統計補償函式，仍是以實際在汽車環境錄下的噪音語料庫來訓練，共有兩組語料庫；兩男兩女及一男一女語料。圖一、為我們提出的平均值調整函式**b**(λ_e)估測方法。與傳統相類似的方法[1]做比較，不同的部份在於每一個訓練音框**c**$_t$我們會與所有的語音模組都計算λ_e，之後根據λ_e收集訓練音框，我們把維特比(**Viterbi**)這個切音程式取代掉了，其進行步驟描述如下：

(1)首先，對於輸入的雜訊語音，每一個音框**c**$_t$會與所有的語音模組$\Lambda_{s,m}$都計算λ_e。等化因子的值限定在**-2**到**4**之間。

(2)等化因子λ_e計算出來後，同時也將受雜訊影響的平均值調整量$c_{t,i}-\lambda_e\mu_{s,m,i}$統計出來，$i$ 表示特徵向量的維度引數。


Car Noise Hands-Free Talking Training Speech
Clean Speech Close Talking HMM Model
$\Lambda_{s,m}$
Caculate λe with all models
Calculate
λ_e, $c_{t,i}-\lambda_e\ \mu_{s,m,i}$
Nonparametric Estimation of Relations of λ_e and $\{c_{t,i}-\lambda_e\ \mu_{s,m,i}\}$
$\{b(\lambda_e)\}$

圖一、平均值調整函式估測流程圖

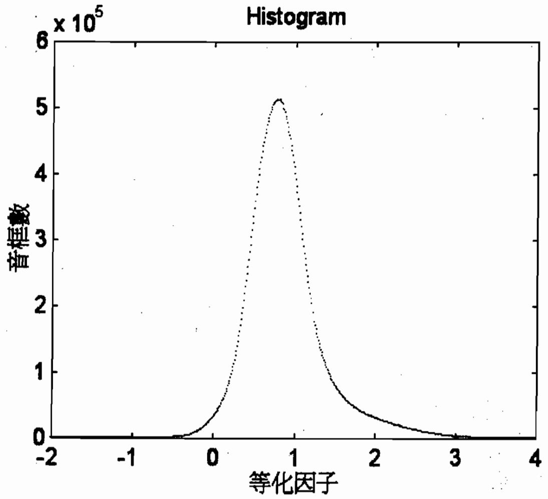
(3)從等化因子λ_e與調整量$\{c_{t,i}-\lambda_e\mu_{s,m,i}\}$的關係分佈圖(**Scatter Diagram**)中，估測出平均值補償函式**b**(λ_e)。

圖二為使用改善後估測調整函式方法的音框數與等化因子關係圖，橫軸為等化因子值介於-2 到 4 之間，縱軸為音框數出現的頻率，大部份的音框λ_e值介於 0.5 到 2 之間，也是形成一個近似常態分配的曲線。圖三為使用改善後估測調整函式的平均值調整量與等化因子在 LPC cepstrum 第一維下的關係圖，其中，橫軸為等化因子，縱軸為平均值對應的調整量。

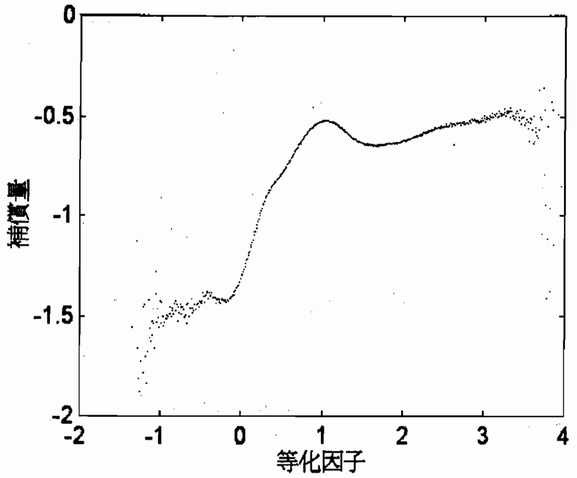
4. 實驗結果

4.1 調整參數的存取及辨識架構

語音辨識系統中，語音需先經過取樣及量化成為數位資料，實驗中所有語料的語音取樣頻率均為8kHz，以及以16bit表示每個數位點，我們可以從有效的語音取樣中抽取適當的語



圖二、等化因子之機率密度函式



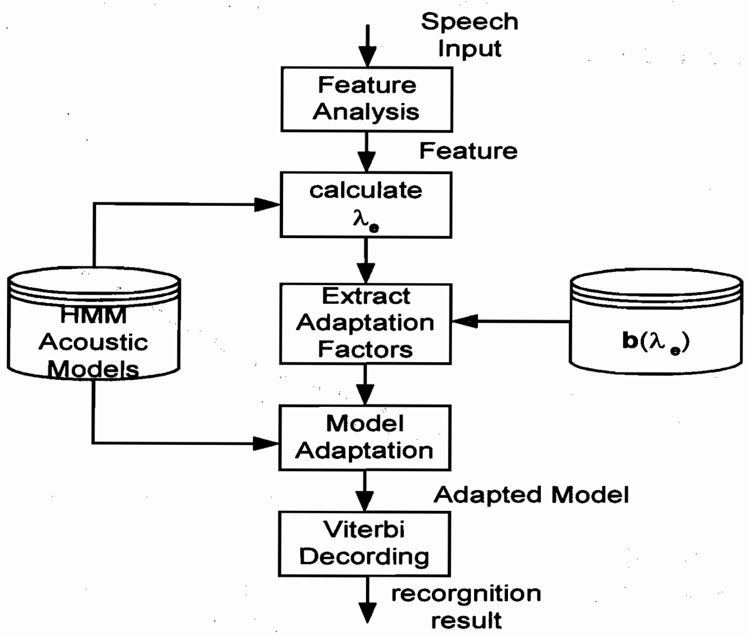
圖三、平均值調整函式在 LPC cepstrum 第 1 維下的分佈圖

音特徵值，我們的語音特徵每個音框的參數為12階 LPC cepstrum 和12階 delta LPC cepstrum 和1階 delta log energy 和1階 delta delta LOG Energy[8]。共26階。本研究方法在做補償函式分析時，會對每一階均建立一個查詢表(table)，我們的特徵參數有26維，所以平均值調整函式方面有26個查詢表，變異數調整函式方面也是26個查詢表。

訓練乾淨的隱藏式馬可夫模型參數(HMM)的語料庫(database)是以近距離麥克風之方式錄下的乾淨語料，這組語料庫共有5050句由50男及51女在安靜辦公室房間裡所錄下的，每一個人各唸二字詞、三字詞或四字詞共50句。使用的馬可夫模型參數有408個音節模組，以次音節前後相關的方式建構出467個狀態及一個背景雜訊狀態；每個狀態依實際的音框多寡分成不同數量的混合數，每個狀態至多有四個混合數。

將兩組調整函式儲存在記憶體中，於雜訊語音辨識時，就可依等化因子對應出平均值與變異數調整量，將乾淨語音模組依不同環境作不同之調整。圖四為噪音環境下的語音辨識架構；以下為整個架構的說明：

(1)當語音輸入求取參數後，由維特比解碼器(Viterbi Decoding)找出一條最佳路徑。在求取最佳路徑時，需先參考隱藏式馬可夫(HMM)語音模型的參數值依式子(2)計算出等化



圖四、系統架構流程圖

因子λ_e。

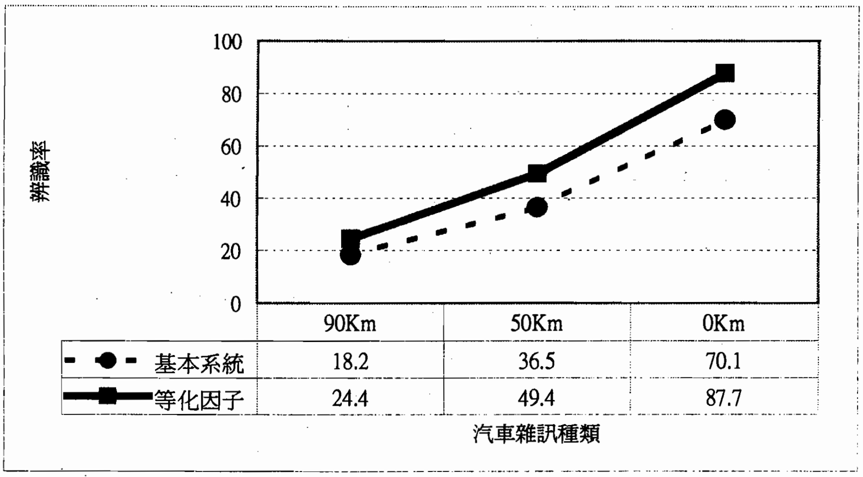
(2)根據λ_e可以由資料庫找出$b(\lambda_e)$這兩個參數；之後再進行相似度量測，此最佳路徑就是經過補償後的最佳路徑。

(3)與114個詞彙做樣型比對，找出一個最有可能的詞彙。

4-2 實驗結果

本實驗之測試語料庫包括零公里怠速路況共204個測試語句，五十公里正常路況下共263個測試語句，九十公里高速路況共有324個測試語句，它們都是由遠距離麥克風錄得的，所使用的汽車是裕隆尖兵 1.6，由三男三女所錄下的。實驗結果部份，我們列出不作補償及加入λ_e補償的辨識結果，其中 90 公里測試語料部份由原先 18.2%增加到 24.3%；50 公里測試語料部份由原先 36.5%增加到 49.4%；0 公里測試語料部份由原先 70.0%增加到 87.7%；我們可以發現辨識結果都有明顯的改善。圖五為基本系統及加入λ_e作補償辨識結果的比較圖。

實驗中用來訓練補償函式語料庫部份有兩組(兩男兩女及一男一女)訓練語料，從表一實驗結果中可以看出以兩男兩女的語料訓練調整函式會比一男一女的訓練語料有較好的辨識率，主要是因為較多的訓練語料可以訓練出較佳的調整函式。不管是90公里路況或是50公里路況，都可以看出這個現象。而且我們發現以改良後估測調整函式會比以原始的估測方法[1]的辨識率要好。我們以改良後估測調整函式的方法對這兩組訓練語料所得的比較結果如圖六所示。其中等化因子的範圍都是訂在-2到4之間。

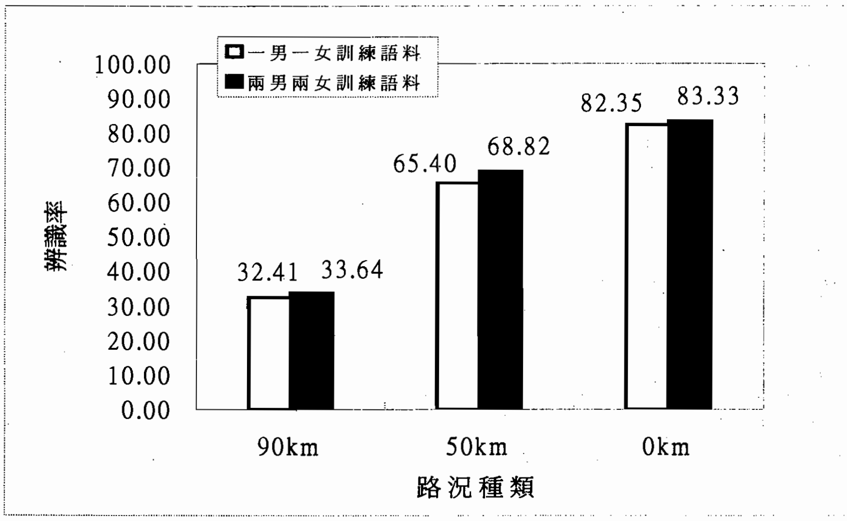


圖五、基本系統及加入λ_e作補償辨識結果的比較

				一男一女訓練語料	兩男兩女訓練語料
		基本系統	.+λ_e	.+λ_e+b	.+λ_e+b
原始的估測方法	90km語料	18.2	24.3	27.7	31.5
	50km語料	36.5	49.4	57.7	63.9
	0km 語料	70.1	87.7	82.4	86.8
改良的估測方法	90km語料			32.4	33.6
	50km語料			65.4	68.8
	0km 語料			82.4	83.3

表一、訓練語料的多寡對辨識率的影響

我們作效能評估所使用的電腦為P2-350，RAM為128MB，表二為每句語音辨識所需消耗的時間比較表。



圖六、訓練語料與辨識率比較圖

效能評估	基本系統	.+λ	.λ+b
秒/句	0.41	0.67	0.74

表 二、語音辨識的消耗時間比較表(單位：秒)

在λ_e的計算中，每句大約要耗掉0.26秒，再加上其他function call的時間，基本系統與作平均值補償大約差0.74-0.41=0.33秒，因為查表動作沒有佔CPU很多時間，所有作λ_e補償與作

λ_e及b(λ_e)補償的時間相差不多，但我們必需額外付出一些記憶體來儲存補償函式，如果等化因子的範圍為-2到4，平均值及變異數補償函式需要2(函式個數)*600(λ_e量化的點數)*26(特徵向量維度)=31200筆浮點數大小，總共要121KBytes。

如果我們以改良後的方法為主，並把等化因子的範圍限制在0~3之間，我們發現使用較小的λ_e範圍降低了些許辨識率，但所需的記憶體為1(平均值補償函式)*300(λ_e量化的點數)*26(特徵向量維度)=7800筆浮點數，約為30KBytes，可大幅減少記憶體的使用量。

實驗結果最後部份我們混合訓練語料及測試語料，以觀察不同汽車對辨識率的影響。表三是以之前的語料庫組合方式所得的最後結果，訓練語料與測試語料是由不同的車子錄得。而現在我們把含有這兩台汽車的語料都混合在一起，混合後的訓練語料有526句；其中零公里部分有118句，五十公里部分有206句，九十公里部分有202句。而混合後的測試語料共有788句；其中零公里部分有203句，五十公里部分有262句，九十公里部分有323句。表四為我們混合不同汽車語料後的辨識結果。

改良後的方式估測調整函式			λ_e範圍: 0 ~3	λ_e範圍: -2 ~ 4
	Baseline	.+λe	.+λ_e+b	.+λe+b
90km語料	18.2	24.4	32.4	33.6
50km語料	36.6	49.4	62.0	68.8
0km 語料	70.1	87.8	82.8	83.3

表三、辨識結果的比較

改良後的方式估測調整函式			λ_e範圍: 0 ~3	λ_e範圍: -2 ~ 4
	Baseline	.+λe	.+λe+b	.+λ_e+b
90km語料	26.70	33.85	40.99	41.61
50km語料	48.85	63.35	72.51	77.86
0km 語料	76.47	84.31	87.2	87.25

表四、混合不同汽車語料後的辨識結果

5. 及時展示系統

為了實際評估此演算法的效能，最好的方法就是直接線上做語音辨識，我們並不是直接開車然後在車上作展示；而是先錄製一段汽車背景雜訊，用喇叭播放出來以模擬實際的汽車噪音環境，而麥克風的位置是以免持式遠距離麥克風為主，大約與說話者的距離25到35公分之間，線上錄下一段語音做即時語音辨識。

在展示系統設計過程中我們遇到了一些問題，如背景雜訊是由喇叭播出與實際汽車環境是不相同的，我們不能以汽車上訓練出的補償函式直接用在展示系統中，所以我們採用線上錄音線上訓練補償函式的方式來解決；另一個問題是前後背景雜訊太長會嚴重地導致辨識結果不佳。為了解決這個切音問題我們使用兩階段維特比(Two Pass Viterbi Decoding)辨識方式，並且線上錄下一段背景雜訊以訓練出背景模組(Background Model)並配合Two pass Viterbi Decoding做語音辨識。

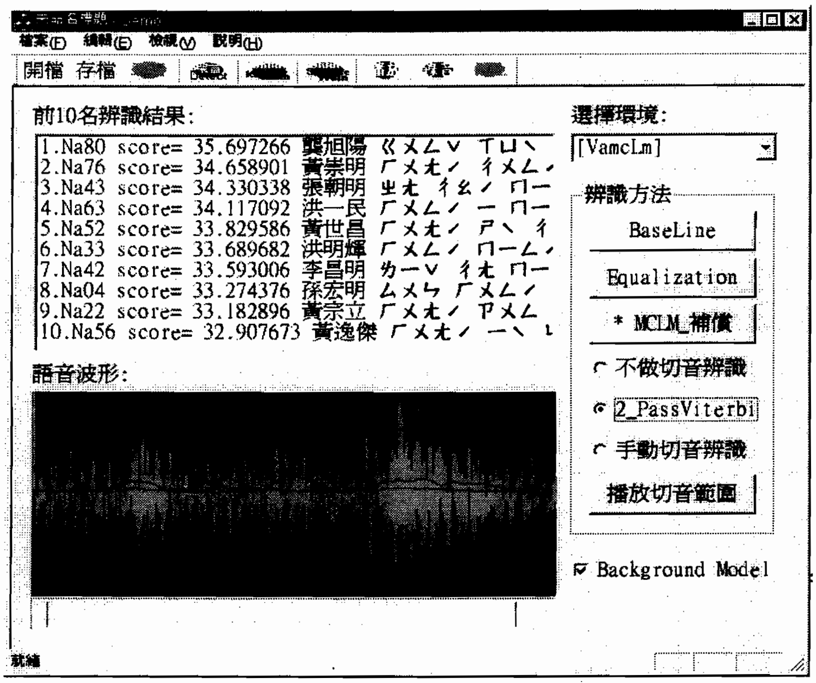
5-1 背景模組

背景模組(Background Model)的訓練方式，會依實際錄得的聲音長短，訓練出不同混合數(Mixture)的背景模組，例如64個混合數或32個混合數等等。實際的訓練方式我們是以向量量化(Vector Quantization)分析程序作群的分類，在做語音辨識時將原始乾淨語音訓練出來的背景模組用新的背景模組取代，在做模型調整時此背景模組不做調整。

5-2 兩階段維特比辨識

在噪音環境下為了可以更準確地切出語音音段，我們使用兩階段維特比的語音辨識方式。所謂兩階段維特比是針對測試語音執行兩遍維特比(Viterbi Decoding)程式，第一遍維特比先段出語音的起始點及結束點，第二遍維特比再將段出後的語音進行辨識。為了可以準確地段出語音音段我們會結合在前一節所介紹的背景模組(Background Model)，這樣可以使切音的結果更加準確。圖七是我們在Windows平台上發展出的一套系統展示介面，在這套介面上我們可線上錄製一段聲音，線上訓練出背景雜訊模組及平均值補償函式，也可線上錄音放音並及時求出辨識結果。辨識的方法有：Baseline不作任何補償的方法；Equalization以等化

因子作補償的方法；MCLM結合等化因子及平均值補償函式的辨識方法。我們並以辨識結果的前10名來觀察不同的辨識方法的辨識效果。



前10名辨識結果:

```
1.Na80 score= 35.697266 龔旭陽 ㄍㄨㄥˇ ㄒㄩˋ
2.Na76 score= 34.658901 黃崇明 ㄏㄨㄤˊ ㄔㄨㄥˊ
3.Na43 score= 34.330338 張朝明 ㄓㄤ ㄔㄠˊ ㄇㄧ
4.Na63 score= 34.117092 洪一民 ㄏㄨㄥˊ ㄧ ㄇㄧ
5.Na52 score= 33.829586 黃世昌 ㄏㄨㄤˊ ㄕˋ ㄔ
6.Na33 score= 33.689682 洪明輝 ㄏㄨㄥˊ ㄇㄧㄥˊ
7.Na42 score= 33.593006 李昌明 ㄌㄧˇ ㄔㄤ ㄇㄧ
8.Na04 score= 33.274376 孫宏明 ㄙㄨㄣ ㄏㄨㄥˊ
9.Na22 score= 33.182896 黃宗立 ㄏㄨㄤˊ ㄗㄨㄥ
10.Na56 score= 32.907673 黃逸傑 ㄏㄨㄤˊ ㄧˋ ㄐ
```

語音波形:

選擇環境: [VamcLm]

辨識方法

BaseLine

Equalization

* MCLM_補償

○ 不做切音辨識

● 2_PassViterbi

○ 手動切音辨識

播放切音範圍

☑ Background Model

圖七、Demo 系統介面

6. 結論

依我們的演算法先計算等化因子，在進行相似度量測時把這兩個調整函式結合進去，可大幅回升在噪音環境下語音辨識正確率。我們以實際的噪音環境下錄下少數語料庫訓練出調整函式，就可以有很好的語音模型調整效果。

實驗結果顯示我們以改良後的方法辨識率有提升，而且可以降低記憶體的需求量，將來如果要實際實做成晶片，把它應用在汽車噪音環境裡的對話系統或撥號系統，這是一個很實用的演算法。因為本方法需要額外的CPU時間計算等化因子λ_e，以及少量的記憶體空間儲存調整函式，相信在未來電腦硬體技術會快速發展，這些額外的需求將不成問題。

致謝:

感謝工研院電通所前瞻技術中心在本研究上的協助

參考文獻

[1] Jen-Tzung Chien and Ming-Shung Lin (1999)，"Noise Compensation approach to hands-free speech recognition in the car", Proc Workshop on Distributed System Technologies and Application , pp.80-87, Taiwan-Tainan (in Chinese)

[2] Boll, S. F. (1979), "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoustic, Speech, Signal Processing*, Vol. 27, pp. 113-120.

[3] Carlson, B. A. and Clements, M. A. (1991), "Application of a weighted projection measure for robust hidden Markov model based speech recognition", *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 921-924.

[4] Carlson, B. A. and Clements, M. A. (1994), "A projection-based likelihood measure for speech recognition in noise", *IEEE Trans. Speech and Audio Processing*, Vol. 2, pp. 97-102.

[5] Chien, J. T., Wang, H. C. and Lee L. M. (1998), "A novel projection-based likelihood measure for noisy speech recognition," *Speech Communication*, Vol. 24, no. 4, pp. 287-297, July 1998.

[6] Gong, Y. (1995), "Speech recognition in noisy environments: A survey", *Speech Communication*, Vol. 16, pp. 261-291.

[7] Lee, C. H. (1997), "On feature and model compensation approach to robust speech recognition", *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 45-54.

[8] Lee, C. H., Giachin, E., Rabiner, L. R., Pieraccini, R. and Rosenberg, A. E. (1992), "Improved acoustic modeling for large vocabulary continuous speech recognition", *Computer Speech and Language*, Vol. 6, pp. 103-127.

[9] Mansour, D. and Juang, B. H. (1989), "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoustic, Speech, Signal Processing*, Vol. 37, pp. 1659-1671.