以音韻屬性偵測擷取對話語音關鍵詞之研究

Study on Keyword Spotting using Prosodic Attribute Detection for Conversational Speech

黃昱睿　Yu-Jui Huang
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0990435@mail.ncyu.edu.tw

鐘尹蔚　Yin-Wei Chung
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0970421@mail.ncyu.edu.tw

葉瑞峰　Jui-Feng Yeh
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
ralph@mail.ncyu.edu.tw

摘要

在口語對話上，為了有效地理解使用者所要表達的資訊意義，擷取關鍵詞語是很重要的研究議題之一。本研究針對對話語音內容的關鍵詞，提出以音韻屬性來做為擷取關鍵詞的特徵。利用預先訓練的決策樹將語者語句分段成韻律詞，進一步使用支援向量機(SVM)來偵測韻律詞是否為關鍵詞。在此利用中研院語言研究所鄭秋豫所提出階層式多短語語流韻律架構與韻律詞邊界的偵測方法，而邊界偵測為利用其韻律特性建立的決策樹來偵測。最後，以音韻屬性為特徵來做為偵測的參數，藉由 SVM 分析各個韻律詞之特徵值找出焦點所在語音時間區段，藉由擷取此焦點作為關鍵詞。最後部分為針對錄製的對話語料進行實驗並分析，所得到的準確度與召回率比參照發音相似度或聲學特徵還要高，證實所提方法在關鍵詞擷取是可行的。

Abstract

It is one of most essential issues to extract the keywords from conversational speech for understanding the utterances from speakers. This thesis aims at keyword spotting from spontaneous speech for keyword detecting. We proposed prosodic features that are used for keyword detection. The prosody words are segmented from speaker's utterance according to the pre-training decision tree. The supported vector machine is further used as the classifier to judge the prosody word is keyword or not. The prosody word boundary segmentation algorithm based on decision tree is illustrated. Besides the data driven feature, the knowledge obtained from the corpus observation is integrated in the decision tree. Finally, the keyword

in the focus part are extracted using prosody features by sported vector machine (SVM). According to the experimental results, we can find the proposed method outperform the phone verification approach especially in recall and accuracy. This shows the proposed approach is operative for keyword detecting.

關鍵詞：關鍵詞語、音韻屬性、韻律詞、口述語言。

Keywords: Keyword spotting, prosodic feature, prosody word, spoken language.

一、緒論

關鍵詞辨識(Keyword spotting)在近代語音研究與應用上為一項很重要的學問，其目的是讓電腦系統能夠從語音資料裡面，自動偵測出特定的關鍵詞彙。於應用上也包含了很多層面，例如語音資料檢索(新聞報導、影片資料、電視轉播等)、自動語音轉接總機系統、查詢服務等。在人機溝通方面，以自發性語音(Spontaneous speech)為主要輸入方式的對話系統(Dialogue system)裡，因為每個人的言談風格(Speaking style)都有所差異，很難以文法(Grammar)的角度來完整分析語者所要表達的涵義。在實際應用上，考量到對話系統之即時性(Real time)，如何讓系統對於使用者的語音資訊做充分的掌握，再再影響了對話系統之實用與否。Kawahara 等人就把關鍵詞語的擷取(Keyword extraction)與確認(Verification)結合剖析器應用於對話系統，其主要步驟分為關鍵片語偵測(Key-phrase detection)、關鍵片語驗證(Key-phrase verification)、句子剖析(Sentence parsing)以及句子驗證(sentence verification)四個部分。麻省理工學院則提出漸進式的對話理解架構(Incremental understanding)，有效擷取系統所需之資訊[1]，而這樣的理念最早是由心理學者 Charpter 等人根據人類口語對話之理解程序所提出的[2]。可以想見關鍵詞與漸進式理解方式在語音對話中是很關鍵的兩個部分。因此如何將語音的關鍵詞擷取出來加以確認，便是口述語言理解(Spoken Language Understanding, SLU)上一個重要的課題。另一方面，為了增加語音辨識的正確性，很多學者提出了不同的額外語音特性來幫助辨識，喬治亞理工李錦輝教授提出以知識為背景(Knowledge based)的方式[3]，導入了語音學上的音韻屬性(Prosodic attribute)作為額外的語音辨識輔助方式。為了有效地於系統上擷取關鍵詞，本研究提出音韻屬性之關鍵詞擷取方法，藉由中研院語言學研究所鄭秋豫提出的階層式多短語語流韻律(Hierarchical Prosodic Phrase Grouping, HPG)架構概念[4][5]，偵測語者的韻律詞(Prosodic word)，以偵測為基礎的方式，參照其各種特徵而辨別是否為關鍵詞。為了增加語音辨識的正確率，國外也有很多學者也借助於其他的知識背景方式。Ali 以聲學上的音韻特徵為基礎，針對每個音素不同的特徵來區別，利用這些特徵來以音素為單位進行連續語音辨識[1]。Wieland 則針對語言模型，提出以統計方式並結合考量語意的方式，建立 Bi-gram 模型，並且使用 Beam-search Viterbi 方式來搜尋最佳語句路徑。結合這兩點並應用於口述語言的辨識[6]。Bitar 提出探討結合知識背景，以這些特徵來作為辨識方法的各種參數，除了傳統 HMM，並結合專業知識再評估，結果在語音辨識上有不錯的成效[7]。Rabiner 在 1989 年，語音辨識還尚未很成熟時，提出了利用點。其一為隱藏式馬可夫模型方式來辨識語音的概念與針對其應用方式來討論，並且說明了其兩項擁有完整的數學理念與架構，並且可以廣泛應用在各種領域。第

二點就是將其應用在語音辨識上確實有良好的效果[8]。Tatsuya Kawahara 與 Chin-Hui Lee 提到 Key-Phrase Detection 和 Verification 的結合，意即關鍵詞的擷取與組合，在對於針對富含文法結構鬆散且變化性大之特性的口述語言，系統理解其對話內容上有很重大的幫助[9]，也更加強化我們利用擷取關鍵詞來評估語言行為的重要性。

二、相關研究

在關鍵詞辨識的研究上，Rose[10]利用 HMM 建立了一個關鍵詞辨識系統，個別訓練關鍵詞與非關鍵詞部分，再利用為辨識關鍵詞的數個狀態，還有非關鍵詞的填詞器(filler)數個，架構整個辨識網絡來辨識整段語音的關鍵詞部分。Zhang[11]提出兩階段式的方法，第一階段辨識出可能性最高的音素序列，第二階段藉由混亂矩陣判斷其相似性，列出最有可能的序列，最後則擷取出信心度最高分的作為關鍵詞。Bahi[12]方法也是類似，將每個發音音節分開成字元列組合，先將可能性關鍵詞訓練好為字元列，同樣將辨識出來的各種可能字轉換成字元列，利用 HMM 方式去比對位置偵測出最有可能的關鍵詞並擷取出，此篇特點在於並沒有特意對非關鍵詞作模型。麻省理工學院的 Bazzi 研究在 HMM 辨識器下非關鍵詞的填詞器設計[13]，裡面提到關鍵詞辨識中，詞庫外字詞也是占很重要部分，分析了錯誤警報器的正確性跟整體辨識正確率相對關係與所占比例。Lee C.H.[14]在比對發聲相似性時，額外再參考相鄰的發音，利用貝氏理論中的貝氏因子計算方式，來計算出發音相似性。Kim[15]則以貝氏理論來作為評估辨識後的發音其信心度分數。另外幾種是以特徵參數作為偵測基礎訓練分類器[16][17]，利用訓練器來對要辨識的語音作分類，判定其是否為要關鍵詞。

Haizhou Li, Bin Ma, and Chin-Hui Lee 於期刊上發表的研究多語辨識[18]，提出了新的辨識單元構想，不再以音標為單元而是用實際人類發音來做為單位，並加以訓練模型。後端部分在各個語言特徵上建立了向量空間模型來儲存各種語言上每種發音的相對關係，並訓練出分類器來藉此分類辨識為何種語言，所呈現的辨識結果比用國際音標還要好上許多。

在音韻研究方面，近年以哥倫比亞大學發展出一個偵測重音部分的工具軟體 AuToBi，可以針對短語邊界偵測並且偵測發音重音部位[19]，其文章也提到這些年來各種不同的偵測重音方式，有利用聲學屬性、基頻、能量、POS 等，分析上也有 HMM、決策樹等方式。例如 Conkie 等人[20]，針對基頻與能量以及這些參數差值與差值之 delta，並結合 HMM 架構偵測重音部位。此外，也多加入了語者相關的聲學資料，在準確性上確實提升。Sridhar[21]在評估 HMM 中裡面的參數差值時，同時監督聲學屬性與句法屬性，並使用最大熵 HMM 模型來偵測重音部分所在位置。

研究對話心理學方面，我們參考了 Erteschik-shir 的著作[22]，裡面描述人類心理在對話行為上的各種情況與回應內容和情緒，語者和聽者會在對話的進行上不斷改變所掌握的不同資訊和心理所期望的內容。

長庚大學多年來一直從事於本土語音的研究，陳志宇[23]此篇論文探討同時對國台雙語的大詞彙連續語音辨識研究。早期國外已證實隱藏式馬可夫模型應用於語音辨識上的效果，楊永泰[24]將其應用於改善於中文的語音辨識，針對音素的替換改變將其用於中文系統上。余家興[25]所做之語音辨識的研究，是利用有限狀態機的方式來達到大詞彙連續語音辨識，將語音辨識上的三種主要模型，聲學模型、辭典、語言模型都建立成有限狀態機型式，這樣在擴充性上與結合上都更為容易。陳錫賢[26]研究偵測語音上的聲韻屬性偵測，包含一些鼻音、擦音、爆破音等。並結合聲韻屬性與 MFCC 來探討語

音辨識的正確率。黃冠達[27]以音長、反模型距離、聲學分數等特徵，利用 SVM 分類器來對聲母驗證分類是否屬於關鍵詞，並另外在對核心函數修改來提升最佳效能。

國內中研院語言學研究所鄭秋豫針對漢語提出的 HPG 架構對中文韻律學研究影響很大，以 Fujisaki Model 計算分析並證實每一層韻律單位的相關性且互相影響，其他研究包括自動偵測韻律邊界，各地漢語在韻律上雖然存在差異卻本質相同等[4][5][28][29][30]，此 HPG 概念對語音學上有很多重要貢獻。

三、系統介紹

本研究的系統架構依其處理程序分為訓練階段與測試階段。訓練階段中，利用語料訓練出偵測關鍵詞模型，以及用來將語音分割成韻律詞語音段的決策樹模型。測試中，以此兩模型先後偵測韻律詞邊界與各韻律詞之特徵參數，最後偵測關鍵詞並擷取出。圖 1 為系統架構圖。

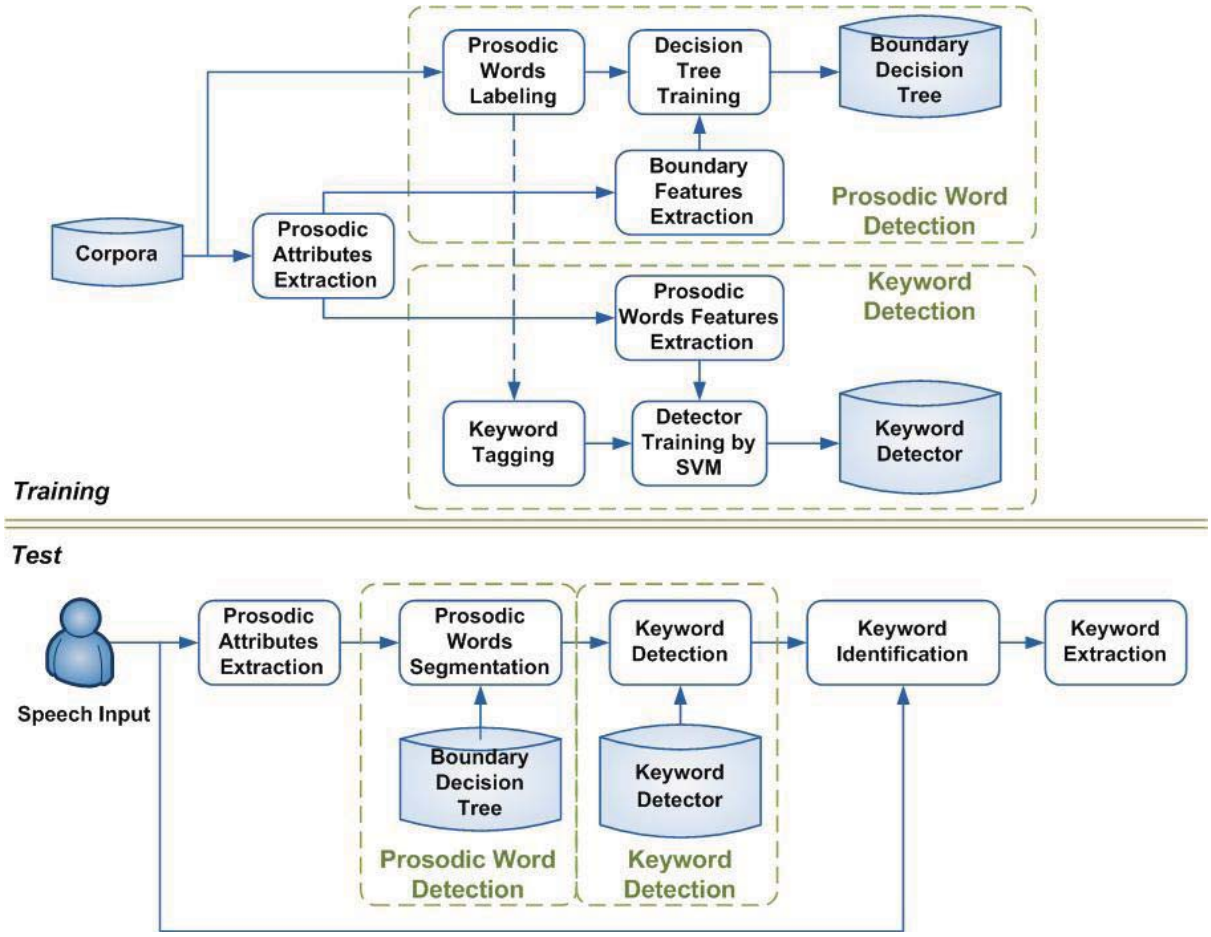


圖 1:系統架構圖

為了在實驗上偵測出韻律詞邊界用以偵測出關鍵詞，我們必須先訓練出所需要的模型。利用收集的語料庫，抽出各種音韻屬性參數(Prosodic Attributes Extraction)，包括了音高(Pitch)、音強(Intensity)、音長(Duration)。偵測韻律詞邊界模型利用階層式多短語韻律語流架構(HPG)知識，並統計分析從語料中所抽取的參數資料，最後訓練出韻律詞邊界(Prosodic Word Boundary)的決策樹(Boundary Decision Tree)，此為訓練韻律詞偵測模

型部分。而另一方面，同樣由這些音韻屬性，統計分析找出各個語音事件物理現象與相對應的參數資料，利用 SVM 訓練出超平面模型以用來分辨關鍵詞與非關鍵詞，此亦即為各個的關鍵詞偵測器(Keyword Detector)，這部分便為關鍵詞偵測部分。

而測試部分中最重要的兩個部分即為韻律詞偵測(Prosodic Word Detection)與關鍵詞偵測(Keyword Detection)，偵測這兩部分都必須仰賴訓練階段所訓練出的模型。整個實驗首先從使用者輸入的語音訊號中，抽取出各個音韻屬性。利用這些訊號參數輔以邊界決策樹模型，找出韻律詞邊界，因而將整段音訊分割成若干個韻律詞組合。根據這些韻律詞的音韻屬性，逐一以各個語音事件偵測器分析鑑定其是否屬於關鍵詞或者為非關鍵詞，偵測出關鍵詞的時間區段並擷取出來。

較早之前對於語調的研究，國內一直並無專針對漢語來研究其單位組成，所訂定的韻律單位也是沿用於國外的，主要例如音節(syllable)、韻律詞(prosodic word)、語調短語(intonation phrase)等。近幾年來，中研院語言學研究所鄭秋豫研究漢語韻律結構，重新訂製韻律單位，此結構命名為「階層式多短語語流韻律」(Hierarchical Prosodic Phrase Grouping, HPG)[4][5]。

在其相關研究中證明了中文口語語流在基頻聲學上，各層級韻律單位與邊界效應有層級關係且互相影響。該架構層級由下而上，將韻律單位定為音節(syllable, Syl)、韻律詞(prosodic word, PW)、韻律短語(prosodic phrase, PPh)、呼吸組(breath-group) 及多短語韻律句群(prosodic phrase group, PG)，即語段，共五級的韻律單位。其邊界也分為五級，由下而上依序為 B1、B2、B3、B4 與 B5。其著作中所提出的架構圖，清楚分析出語篇的各個層級概念與關係，各層級標記順序由 B5 到 B1，可逐步將整個語篇分為各層及韻律單位，最後底層可以分成到最小韻律單位音節。

本研究參考此架構，汲取韻律詞概念作為定義各個關鍵詞之單位，因此只將邊界層級收斂到 B2 層級，使音段標記後即可視為由多個韻律詞所組合而成。並且針對各個韻律詞抽取音韻屬性特徵，利用這些特徵組合來判定其是否具備某些語音特性，最後才鑑定其是否為所需之關鍵詞。圖 2 為將一音段分成韻律詞示意圖。B 標示為其邊界，可看出整段語音分割成五個韻律詞。

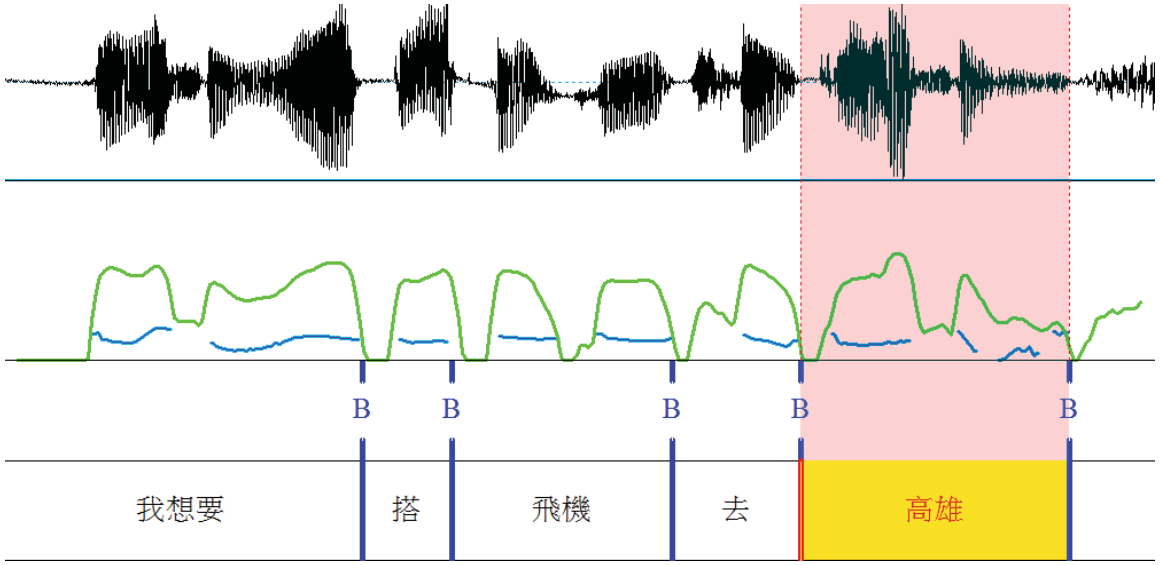


圖 2：韻律詞(Prosody word)邊界示意圖

在邊界偵測上，參考中研院語言學研究所鄭秋豫所提出的邊界特性定義[4][5]，並加入音強屬性來輔助偵測，以觀察其特性從語料庫中訓練出決策樹來作為邊界偵測法則。分析原則為在每個基頻段尾端分析其各項音韻屬性，涵蓋了此段之基頻屬性、音強屬性、以及與下一段的基頻調性。圖 3 為所訓練出決策樹原則，邊界判定上共分為 9 種類別，偵測位於哪種類別以判定是否為邊界。

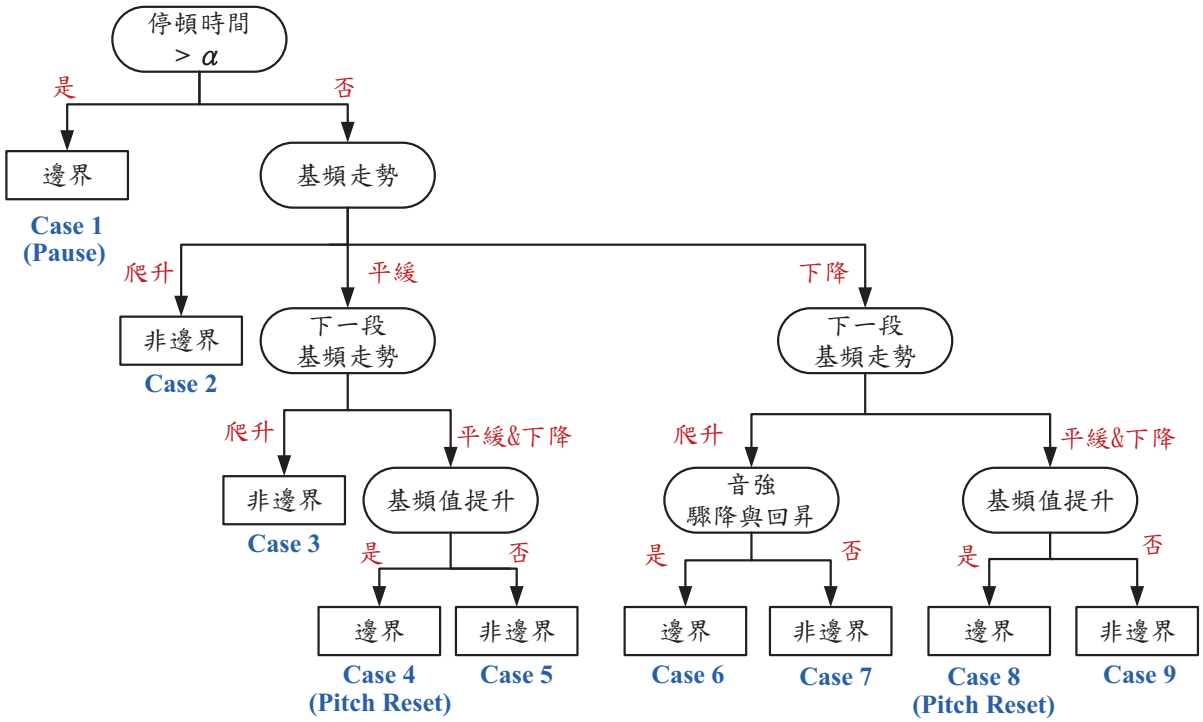


圖 3：HPG 邊界判斷決策樹

決策樹將每個基頻段的連接關係分為 9 個種類以判定是否為邊界，並且帶有不同的特徵現象，並同時表現出音韻屬性上的差異。基頻重設(Pitch reset)為邊界最重要之特徵，有此現象必為邊界處，其餘則必須比較各種基頻與音強來判定。表 1 即為各種不同類別的差異。

表 1：HPG 邊界判斷決策樹之分類

類別	特徵	音韻屬性
Case 1	韻律詞與韻律詞之間聲調轉換與停頓	停頓時間 $>\alpha$ 秒
Case 2	韻律詞的開始，並帶有上升發音必連接下個連續發音	韻律詞調性成上升走勢
Case 3	語調先平緩而後上升，必為連續語調	基頻平緩而後上升
Case 4	新的韻律詞開始之最主要特徵，即基頻重設(Pitch reset)，此為較不明顯的基頻重設	基頻值回到高點

Case 5	連貫語氣，語調大走勢呈現連貫性	基頻呈現大走勢向下或平緩
Case 6	韻律詞與韻律詞之間聲調明顯轉換	基頻下降而後上升，音強大幅度降下後提升
Case 7	韻律詞與韻律詞之間聲調轉換但調性連貫	基頻下降而後上升，音強呈現穩定走勢
Case 8	新的韻律詞開始之最主要特徵，即基頻重設(Pitch reset)，此為明顯易見的基頻重設	基頻值回到高點
Case 9	連貫語氣，語調大走勢呈現連貫性	基頻呈現大走勢向下或平緩

以下列出決策樹上運用到的音韻屬性特徵之各項參數，並逐一介紹。

(1) 停頓時間 β=0.04 秒：

在訓練語料中統計的停頓時間於 0.03 到 0.05 之間總合佔了絕大多數，但部分時間極短的停頓容易與非停頓混淆，將明顯停頓的時間判定值定在 0.04 秒。

(2)基頻走勢判斷：

要判斷其走勢需運用到基頻段的回歸方程式(式 1)，並利用線性迴歸方式出計算出其斜率(slope)，即 βi。

$$P_i(t) = \alpha_i + \beta_i t \qquad \text{(式 1)}$$

式子中 *Pi*(*t*)表第 *i* 段基頻段於時間點 *t* 之基頻值，則 βi 為 *i* 段基頻段之斜率，*bi* 為此基頻段開始時間點，*ei* 為時間結束點。而 βi 其計算方法如式 2。

$$\beta_i = \frac{\sum_{t=b_i}^{e_i}(t-\bar{t})(P_i(t)-\bar{P}_i)}{\sum_{t=b_i}^{e_i}(t-\bar{t})^2},\ t \in [b_i, e_i] \qquad \text{(式 2)}$$

$\bar{t}$ 為時間平均值，在時間軸上亦同於中間值，計算如式 3。$\bar{P}_i$ 為第 *i* 段基頻值平均值，計算如式 4。*n* 為此基頻音段取樣點數。

$$\bar{t} = \frac{1}{2}(e_i - b_i) \qquad \text{(式 3)}$$

$$\bar{P}_i = \frac{1}{n}\sum_{t=b_i}^{e_i} P_i(t) \qquad \text{(式 4)}$$

求得所需的β_i後，即可知道斜率並利用邊界條件上界 upper bound 與下界 lower bound 判定其走勢。假若β_i≧upper bound，則其走勢為向上爬升；反之，β_i≦lower bound，則為下降；而若 upper bound＞β_i＞lower bound，其走勢判定為平緩。此外，根據基頻走勢所產生的基頻重設也會有差異，case 4 的 pitch reset 值與 case 8 的 pitch reset 值也會有所差異。

本研究利用支援向量機，它為一監督式學習的二元分類器，其目的為尋找最佳化的向量分類，尤其運用在解決非線性化的問題。要訓練分類模型，必須給予標記好的語料作為訓練資料，另外需再準備一筆測試資料，SVM 利用已訓練好之模型作預測(predict)，將資料分類於相近的類別。在研究中，將測試資料分為兩類，標記為+1 的是關鍵詞，標記為-1 的是非關鍵詞，其數學式表示如式 5。

$$T_i = \begin{cases} +1, if\ \ T_i\ \ is\ semantic\ object \\ -1, \qquad\quad otherwise \end{cases} \quad \text{(式 5)}$$

作為 SVM 分析中所要用到的特徵參數，我們利用音韻屬性來構成不同的特徵，組成一個向量空間，附表為我們研究中估計之各特徵參數與計算方式。包含了音長、停頓以及位置。

以下逐一列出評估之特徵參數：

(1) 音長：

此特徵考量所有韻律詞內與音長相關的特徵，考量了基頻音段數、基頻音長、音節數、發音音長以及各音節的音長，並將所有用來評估的特徵列於附表編號 01-10。n_i 表示為第 i 個韻律詞的基頻音段數。P_{ij} 為第 i 個韻律詞中第 j 段基頻段，且 $\{P_{i1}, P_{i2}, ...P_{in}\} \subseteq PW_i$。$P_{ij}^{Dur}$ 為第 i 個韻律詞中第 j 段基頻段的基頻音長。Bi 為此韻律詞開始時間點，Ei 為韻律詞時間結束點。Syl_N_i 表示為第 i 個韻律詞的音節數。Syl_{ij}_b 為第 i 個韻律詞的第 j 個音節之起始時間，Syl_{ij}_e 為第 i 個韻律詞的第 j 個音節之結束時間。

(2) 停頓：

語言行為上可能會為了表達重要字詞，會預先出現停頓現象再以加強語氣道出關鍵詞。評估此特性，計算韻律詞起頭的停頓時間作為特徵參數。bpause 停頓起始時間點，epause 停頓結束時間點。如附表編號 11。

(3) 位置：

在語言行為與文法上，或者語者的習慣性上，重要字詞容易出現在句尾與句首。因此我們評估每一個韻律詞的起始位置作為評估參數。計算方式以該韻律詞的起始時間除於語段結束時間。而第 13 個特徵則做為輔助用，計算整句話總共多少個韻律詞。如附表編號 12-13。

以上所提出之 13 個特徵將詳細表列於附表。

本研究所提之關鍵詞語辨識(Keyword spotting)乃是指系統根據語音動作(Speech act)所欲擷取之關鍵詞之可能內容值。因為關鍵詞的擷取或語意框架(Semantic slot)之填入關係到使用者端的輸入，而使用者端的輸入則與機器端的輸出息息相關，所以 DA pair 即是指相對應之機器端與使用者端的輸出入對應關係。參考 Erteschik-shir 著作[23]中，我們知道一段對話涵蓋著兩個意念，主旨(Topic)與焦點(Focus)。人在期望求得某項訊息時，在表達出所希望得到的訊息，這項渴望得到的大範圍目標內容即為主旨。而對於回話者的內容會以語用學(Pragmatics)的角度來指觀察談話的重點，此重點即為焦點，同時其他詞語與多餘的贅詞不在注意目標。而在描述這些對話中的重要資訊時，會不自覺的做出強調動作，根據上述的主旨與焦點的行為，可以觀察韻律詞訊號面發生的音韻特徵，用以偵測出對話中關鍵字。

如圖 4 和圖 5 所示，雖然使用者答句一樣，但因語意動作不同則關鍵詞語集合與關鍵詞位置亦為不同。圖 4 主旨在於詢問交通工具訊息，所以答句焦點部分所要得到的訊息內容是計程車。圖 5 主旨在於詢問地點訊息，所以答句焦點部分所要得到的訊息內容是於安平古堡。

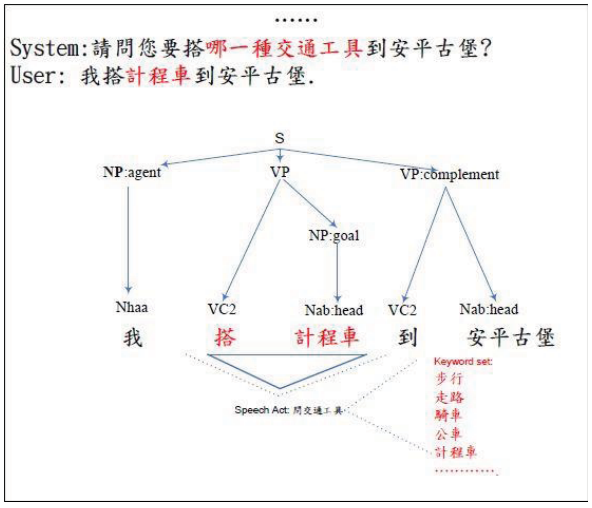


圖 4: DA pairs 動態定義關鍵詞

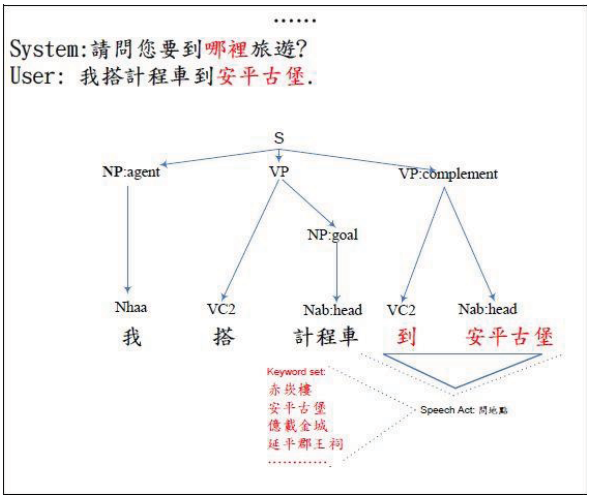


圖 5: DA pairs 動態定義關鍵詞

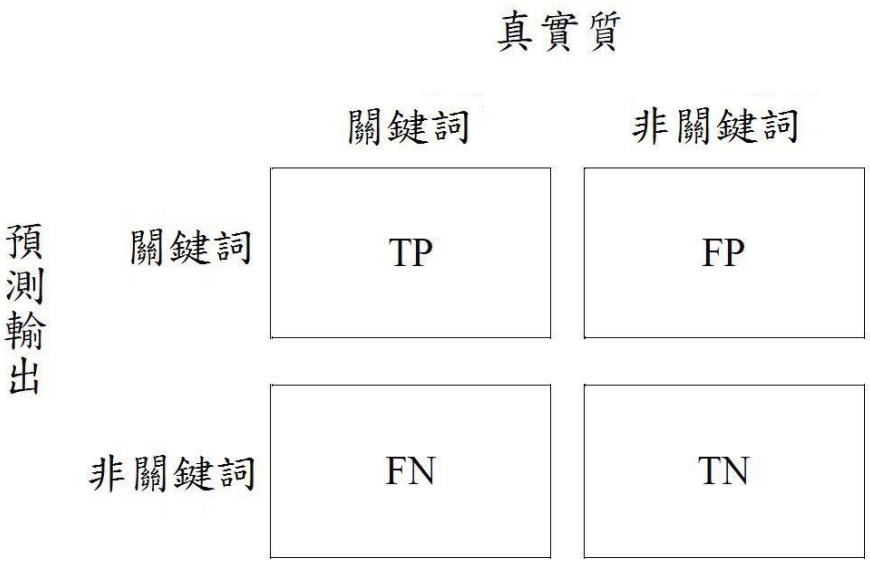
四、實驗結果與討論

實驗用語料使用國立教育廣播電臺線上廣播之錄音檔，並從中擷錄符合問答對話形式之語句，單元「好老師的 52 堂課」共截錄 247 句，單元「基測百分百」共 568 句。在「全球華文網，數位化出版品：快樂學華語」，從中的教學語音語料抽出 73 句。遠東圖書「旅遊中文開口說」共 173 句。語料總合共 1061 句。從中取 850 句作為訓練語料，剩下 211 為測試語料。所有語料檔皆標記韻律詞邊界，以及對各個韻律詞標註關鍵詞與非關鍵詞。訓練語料關鍵詞總量 850，非關鍵詞總量 2498。實驗語料關鍵詞總量 211，非關鍵詞總量 660。

本研究分析關鍵詞擷取的正確率，必須對訓練語料與測試語料皆標記上韻律詞邊界以及標註上是否為關鍵詞。如果經由系統分析出為關鍵詞，並且人工標記同為關鍵詞，則歸為正確的關鍵詞擷取，此為真陽性(True Positive, TP)。但是若標記為關鍵詞，分析出的為非關鍵詞，那麼即為錯誤，此為偽陰性(False Negative, FN)。反之，標記為非關

鍵詞時，分析出為非關鍵詞，則為真陰性(True Negative, TN)。若分析出為關鍵詞，則為偽陽性(False Positive, FP)。如圖 6 所示，每個韻律詞分析結果為四種表示並作為分析的依據。

圖 6：分析結果示意表



利用上述的結果來評估本系統的準確度(accuracy)、精確度(precision)、召回率(recall)。計算如式 6，7，8。

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad \text{(式 6)}$$

$$precision = \frac{TP}{TP + FP} \qquad \text{(式 7)}$$

$$recall = \frac{TP}{TP + FN} \qquad \text{(式 8)}$$

(1) 實驗一：人工標記邊界之 SVM 偵測關鍵詞

評估測試語料在人工標記邊界下，利用訓練好的 SVM 分類器來偵測關鍵詞，觀察其準確度，精準度與召回率，結果如表 2。實驗結果跟內部測試交叉驗證時相比，下降了約 3-5%的準確度，精準度最高的為 58%，召回率最高的是 80%。以召回率觀點評估，10 句話中約 8 句可以找到正確的關鍵詞，精準度上來說則是找到的關鍵詞，其真實值約有 58%。

表 2: 人工標記 SVM 分析結果

特徵組合	accuracy	precision	recall
4，5，12，13 (c=1，g=8)	77.16%	57.83%	68.25%

4，5，12，13 (c=10，g=16)	74.10%	52.90%	69.19%
4，5，11，12，13 (c=1，g=8)	**77.42%**	**58.17%**	69.19%
4，5，11，12，13 (c=10，g=16)	74.10%	52.94%	68.45%
3，5，6-9，12 (c=1，g=8)	75.83%	54.69%	**80.0%**
3，5，6-9，12 (c=10，g=16)	73.04%	51.25%	77.73%
4，5，6-8，12 (c=1，g=8)	74.90%	54.01%	70.14%
4，5，6-8，12 (c=10，g=16)	71.58%	49.5%	70.62%

(2) 實驗二：決策樹標記邊界之 SVM 偵測關鍵詞

評估測試語料在利用決策樹自動化偵測邊界下，利用訓練好的 SVM 分類器來偵測關鍵詞，觀察其準確度，精準度與召回率，結果如表 3。在自動化偵測邊界上，我們知道偵測邊界的精準度未滿 100%代表著會多切出更多韻律詞，也因此實驗中會將部分的關鍵詞韻律詞分為多個韻律詞，在 SVM 之判定上就會出現判斷為較多個關鍵詞，因此造成 TP 的計算量略為增加，自然伴隨著準確度、精準度、召回率的增長，所以我們在結果中可以發現數值反而比實驗一的結果提高。

表 3:決策樹標記邊界 SVM 分析結果

特徵組合	accuracy	precision	recall
4，5，12，13 (c=1，g=8)	83.38%	**70.95%**	75.33%
4，5，12，13 (c=10，g=16)	81.40%	65.41%	78.03%
4，5，11，12，13 (c=1，g=8)	**83.51%**	70.83%	75.56%
4，5，11，12，13 (c=10，g=16)	81.35%	64.91%	77.48%
3，5，6-9，12 (c=1，g=8)	82.45%	66.33%	**85.15%**
3，5，6-9，12 (c=10，g=16)	80.61%	63.00%	84.00%
4，5，6-8，12 (c=1，g=8)	80.47%	65.02%	75.33%
4，5，6-8，12 (c=10，g=16)	76.65%	58.42%	75.22%

比較對象為一個關鍵詞擷取方法[14]，其方式為利用 HTK 進行 forced alignment，針對關鍵詞進行編碼成各個相似序列，最後訓練 HMM 模型來辨識是否為關鍵詞或填充詞與(filler)。利用此方法應用於中文關鍵詞擷取，得到的結果如表 4，並與我們使用的方法做比較。參考論文的這類方式，如果出現發音類似於關鍵詞語的非關鍵詞，並無法有效的區分，因此造成正確率上比我們提出的方法低。精準度上則與我們差不多，而召回率上也少了 15%左右。

表 4：分析結果

比較對象	accuracy	precision	recall
Reference	68%	70.22%	68.45%
Label + SVM	77.42%	58.17%	80%
Decision Tree + SVM	83.51%	70.95%	85.15%

五、結論與未來研究方向

傳統的關鍵詞擷取方式，不外乎都藉著龐大關鍵詞訓練，使用聲學特徵或比對相似音來比照其相似性。本研究利用音韻屬性作為偵測特徵並結合中研院鄭秋豫所提出的 HPG 架構，以此方式偵測關鍵詞，證明其確實為一有效之方法。在 SVM 分類所應用的特徵上，對於鑑別關鍵詞與非關鍵詞有最大效果的多屬於音長類型的相關特徵，以及帶有部分文法性質意義的相對位置特徵，並實驗各種核心函數與參數以及利用這些的特徵組合測驗，訓練出最佳模型，以 SVM 去分類所獲得的結果。在人工標記邊界上的偵測，準確度約為 51%~58%上下，精準度有 68%~80%，召回率為 51%~59%。結合決策樹來偵測，會發生關鍵詞被切割為多個韻律詞之情形，被偵測出的情況而造成真陽性(True Positive, TP)上升，也造成各項評估值提高，其最後準確度約為 76%~83%上下，精準度有 58%~71%，召回率為 75%~85%。

致謝

本研究係國科會研究計劃「應用發音知識源於強健語者多樣性語音辨識之研究(NSC 99-2221-E-415-006-MY3)」之成果，承蒙. 國科會提供經費上的支援，特此申謝。

參考文獻

[1] Ali, J. Van der Spiegel, P. Mueller, G. Haentjens ,and J. Berman, “An Acoustic-Phonetic Feature-Based System for Automatic Phoneme Recognition in Continuous Speech,” ISCAS 1998.

[2] N. Chater, M. Pickering, and D. Milward. “What is incremental interpretation? ” Edinburgh Working Papers in Cognitive Science, 11:1–22, 1995.

[3] J. Li, Y. Tsao and C.H. Lee, "A Study on Knowledge Source Integration for Candidate Rescoring in Automatic Speech Recognition," ICASSP, IEEE International Conference, vol 1, pp837-840, 2005.

[4] 鄭秋豫, "語篇的基頻構組與語流韻律體現", 語言暨語言學 11(2):183-218, 2010.

[5] 鄭秋豫, "語篇韻律與上層訊息－兼論語音學研究方法與發現", 語言暨語言學 9.3:659-719, 2008.

[6] E. Wieland, F. Gallwitz, and H. Niemann. "Combining stochastic and linguistic language models for recognition of spontaneous speech." In Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, vol.1, Atlanta, May, pp 423–426, 1996.

[7] N. N. Bitar and C. Y. Espy-Wilson , "Knowledge-based Parameters for HMM Speech Recognition," ICASSP 1996.

[8] L. R. Rabiner, "A tutorial on hidden markov models and selected application in speech recognition," Proceedings of the IEEE, vol.77, no. 2, Feb. 1989.

[9] T. Kawahara, C.H. Lee, and B.H. Juang, "Flexible Speech Understanding Based on Combined Key-Phrase Detection and Verification", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, vol.6, NO. 6, pp.558-568, 1998.

[10] R. C. Rose, D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System" Acoustics, Speech, and Signal Processing, ICASSP, vol.1, Page(s): 129 - 132, 1990.

[11] P. Zhang, J. Han, J. Shao, Y. Yan, "A New Keyword Spotting Approach for Spontaneous Mandarin Speech" Signal Processing, 8th International Conference on vol.1, 2006.

[12] H. Bahi, N. Benati, "A New Keyword Spotting Approach" Multimedia Computing and Systems, ICMCS, International Conference , pp.77–80, 2009.

[13] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition," Proc. ICSLP, Beijing, 2000.

[14] H. Jiang, C.H. Lee, "A new approach to utterance verification based on neighborhood information in model space", IEEE Trans. Speech Audio Process. 11(5), pp. 425-434, 2003.

[15] T.-Y. Kim and H. Ko, "Bayesian Fusion of Confidence Measures for Speech Recognition", IEEE SIGNAL PROCESSING LETTERS, vol.12, NO. 12, Dec 2005.

[16] Y. BenAyed, D. Fohr, J. P. Haton, G. Chollet, "Improving the Performance of a Keyword Spotting System by Using Support Vector Machines", in IEEE Auto Speech Recogniton and Understanding Workshop ASRU, St, Thomas, U.S. Virgin islands, Dec 2003.

[17] R. Rose, "Confidence measures for the Switchboard database", Proc. of International Conference on Acoustics, Speech and Signal Processing, pp.511-514, 1996.

[18] H. Li, B. Ma, and C.H. Lee. "A Vector Space Modeling Approach to Spoken Language Identification", Audio, Speech, and Language Processing, IEEE Transactions on vol. 15,

NO. 1, JANUARY, pp 271-284, 2007.

[19] AuToBi. http://eniac.cs.qc.cuny.edu/andrew/autobi/index.html

[20] A. Conkie, G. Riccardi, and R. Rose. “Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events”. In Eurospeech, 1999.

[21] V. R. Sridhar, S. Bangalore, and S. Narayanan. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. IEEE Transactions on Audio, Speech & Language Processing, 16(4):797–811, 2008.

[22] N. Erteschik-shir, Information Structure: The Syntax-Discourse Interface, 2007.

[23] 陳志宇,“國台雙語大詞彙與連續語音辨認系統研究”,長庚大學碩士論文, 民國 89 年。

[24] 楊永泰, “隱藏式馬可夫模型應用於中文語音辨識之研究”, 中原大學碩士論文, 民國 89 年。

[25] 余家興,“以有限狀態機辨認大字彙中文連續語音”, 台灣大學碩士論文, 民國 93 年。

[26] 陳錫賢, “語音特定屬性之偵測與應用”, 國立清華大學碩士論文, 民國 95 年。

[27] 黃冠達, “應用支撐向量機於中文關鍵詞驗證之研究”, 台灣科技大學碩士論文, 民國 96 年。

[28] C.Y. Tseng, “Discourse Speech Tempo”. JAIST Symposium on Modeling of Speech and Audiovisual Mechanism. Ishikawa, Japan. 2011.

[29] C.Y. Tseng, and C.H. Chang, 2007. “Pause or No Pause?—Phrase Boundaries Revisited”. The 9th National Conference on Man-Machine Speech Communication（NCMMSC). 黃山, 中國, 2007.

[30] 鄭秋豫、李岳凌、鄭雲卿兩岸口語語流韻律初探—以音強及音節時程分佈為例. 海峽兩岸語言與語言生活研究 280-312. 周薦、董琨（主編）, 上海商務印書館, 2008

附表

編號	符號	特徵	計算方式
01	$P^{Num}(PW_i)$	第 *i* 個韻律詞的基頻音段數	n_i
02	$P^{Dur}(PW_i)$	第 *i* 個韻律詞的基頻總音長	$\sum_{j=1}^{n_i} P_{ij}^{Dur}$
03	$P^{Dur\_Max}(PW_i)$	第 *i* 個韻律詞的基頻最大段音長	$Max\{P_{i1}^{Dur}, P_{i2}^{Dur}, ..., P_{in}^{Dur}\}$
04	$P^{Dur\_Min}(PW_i)$	第 i 個韻律詞的基頻最小段音長	$Min\{P_{i1}^{Dur}, P_{i2}^{Dur}, ..., P_{in}^{Dur}\}$
05	$Dur(PW_i)$	第 *i* 個韻律詞的音長	$B_i - E_i - Pause(PW_i)$
06	$Syl(PW_i)$	第 *i* 個韻律詞的音節數	Syl_N_i
07	$Dur(Syl_{i1})$	第 *i* 個韻律詞的第 1 個音節長	$Syl_{i1}\_e - Syl_{i1}\_b$
08	$Dur(Syl_{i2})$	第 *i* 個韻律詞的第 2 個音節長	$Syl_{i2}\_e - Syl_{i2}\_b$
09	$Dur(Syl_{i3})$	第 *i* 個韻律詞的第 3 個音節長	$Syl_{i3}\_e - Syl_{i3}\_b$
10	$Dur(Syl_{i4})$	第 *i* 個韻律詞的第 4 個音節長	$Syl_{i4}\_e - Syl_{i4}\_b$
11	$Pause(PW_i)$	第 *i* 個韻律詞的停頓音長	$e_{pause} - b_{pause}$
12	$pos(PW_i)$	第 *i* 個韻律詞位置係數	$\frac{B_i}{E}$
13	$N(Speech)$	韻律詞總數	N